ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse24942020.pdf
https://doi.org/10.30534/ijatcse/2020/24942020

A Hybrid Approach of Lexicon-based and Corpus-based Techniques for Arabic Book Aspect and Review Polarity Detection

Raja Masadeh[1], Sa'ad Al-Azzam[2], Bassam Hammo[3]
[1]The World Islamic Sciences and Education University, Jordan, raja.masadeh@wise.edu.jo
[2]The University of Jordan, Jordan, sa3d_al3zam@hotmail.com
[3]The University of Jordan, Jordan, b.hammo@ju.edu.jo

ABSTRACT

The problem of Aspect Based Sentiment Analysis (ABSA) has been studied very well in English language, and there are many approaches have been suggested for ABSA. However, there are other languages did not find enough attention in this field. One of these languages: Arabic language that covers a large area of the world. ABSA contains several tasks. This paper addresses for both aspect polarity and review polarity detection. A hybrid approach proposed to treat polarity detection problem, which consists of corpus-based approach (review book corpus) and lexical-based approach (automatic lexicon) used for this work. The work starts by annotating dataset and cleaning it from a noisy data, after that automatically building the lexicon with sensible words with their polarity that hints to extract aspects with their polarity. Experimental results showed that the proposed system got an acceptable result.

Key words: Arabic, Corpus-based, Lexicon-based, Polarity detection.

1. INTRODUCTION

The internet has become a main part of human life and this can be noticed in many fields such as politic, sport, social media and so many topics that deal with online opinions. Sentiment analysis or opinion mining is one of most important field in both academic and commercial environments to identify the polarity of human opinions and comments in social networks, forums, blogs and reviews. The nature of human is asking other people when making decision or when buying products, so this fact attracts many companies to improve their products. The social networks, forums, personal blogs, etc., have created an important way to express and publish one's opinions. Each day hundreds of thousands of comments and reviews are added to the internet (web), which increases the need to process, store, mine and analysis these opinions to discover useful information. Mining process to a huge amount of reviews and textual sentiments and manually is costly and time-consuming. Thus, automatic mining approaches are highly useful and desirable to extract valuable information.

Opinion mining or sentiment analysis are different names in the literature for the process of analyzing and detecting the orientation of unstructured data it can be considered as a classification process which determines the opinion in a specific review as being positive or negative. The reviewers might write their comments and opinions about a certain object (a book, a product event, a topic, etc.) and the opining mining task can either consider the review is positive or negative [1].

Most of the previous studies on sentiment analysis are based on lexicon-based and corpus-based techniques [2]. In lexicon-based approach, a lexicon construction includes keywords with their polarities that helps in determining process of reviews, i.e. the initial word polarities regardless their contexts. For example, ["جيد","سعادة","رائع"] ["good", "happiness"," wonderful"] have positive polarities and ["حزن","بغيض","سئ "] ["bad", "hateful", "sadness"] have negative polarities. As for corpus-based approach, a process of sentiment analysis problem likes a text classification problem. It includes a large of corpus of data, where the review process is trained on a large polarity labeled corpus of reviews. However, the fixed lexicon approach is not effective; especially in Arabic language because there are other many words have sense (positive/negative) do not exist in database or lexicon.

Most of sentiment analysis research concerns with detecting the polarity of a whole sentence (sentence-level) regardless of aspects (aspect-level) in the sentence. ABSA (Aspect- based sentiment analysis) task [3] is concerned with identifying the aspects in each sentence and determining the sentiment polarity for each detected aspect whether it is positive or negative. The aspect task can be achieved through

determining the aspect and a whole sentence or review polarity [4].

The Arabic language is one of the most important languages in the world. It is considered one of the six official languages beside the most famous languages (i.e. English, French, Chinese, Russian, etc.) Moreover, since most of aspect sentiment analysis researches are conducted in English language and intend to apply aspect sentiment analysis task in Arabic language. The dataset that would be used in this work is obtained from book reviews (LABR) [5]. The proposed work could be summarized in three phases: annotation, preprocess and determining polarities of sentences/reviews.

The rest of this study is organized as follow: Section 2 recites the related work to opinion mining ABSA. While sections 3 includes a detailed discussion of proposed approach, which includes the dataset annotation, preprocess and polarities prediction system. Section 4 contains the results. Finally, Section 5 presents concluding of this work.

2. RELATED WORK

Most of the previous studies in ABSA field and sentiment analysis exist largely in English language and there are many approaches used in this topic. Also, supervised machine learning with set of words as features has been a standard approach to the problem of sentiment polarity classification. In this section, the state of art in this field is summarized.

Ahmad et al [6] proposed a study of analyzing Arabic sentiments about financial news. In addition to that, Y. Almas and et al extended the previous study and presented a study for sentiment extraction about financial news in three famous languages: English, Arabic and Chinese. Moreover, they developed a local grammar approach for sentiment extraction and applied it for three mentioned languages. For Arabic language, they used a financial corpus that includes 8815 text. They yielded accuracy of extraction in range between 60%-75% [7]. Samir E. Abdelrahman et al concerned of detecting polarity of words according to contextual polarities, prior polarities and related phrase. For this task, they used a Lexical Semantic Database approach and Support Vector Machine classifier. Many of features used such as word, POS of previous word, POS of next word, semantic ID, POS of semantic, prior polarity of semantic, etc. The results are 83.1, 81.6 and 83.1 for precision, recall and f-measure respectively [8].

Wagner et al. [9] suggested a system for aspect term polarity prediction, which depends on supervised machine learning approach using a combination of sentiment lexicon features and n-grams. The authors focused on restaurant domain and laptop domain for aspect term polarity prediction. The results showed that their approach is very effective and efficient. Furthermore, there are a lot of studies on sentiment polarity focused on assigning labels or scores to the whole text [10]-[12], while some studies concerned with assigning labels to sentences [13], [14] and some of researchers focused on syntactic constituents [15], [16]. Bakliwal et al. [17] presented an approach of sentiment lexicon scores as features and integrated to supervised machine learning to add standard bag of words features.

A syntax-based method is very effective in aspect polarity detection, where a simple relation is the adjectival modifier relation between a sentiment word and an aspect as in ‘fantastic food’, where fantastic affected on aspect food [18]. Zhao et al. [19] presented a study based on syntax-based method. The researchers started with dataset labeling. After that the syntactic patterns of all the annotated aspects are extracted. Then, for the unseen data, syntax trees of all sentences are obtained.

3. METHODOLOGY

This work goes through three main phases; Annotation, preprocess and aspect and review polarities determination. Where the annotation process achieved manually, and for the last two phases (Preprocess, polarities determination) achieved automatically using a created tool, which is built in C # programming language. In this section, these phases are explained in more details in the following points.

1- Annotation process: The dataset used in this project is taken from (LABR) [5], which are around 63,000 book reviews written in Arabic. The dataset used in this project consists of around 1000 book reviews. In other words, apart from a whole dataset is cut and annotates each review polarity (positive or negative), determine which the aspect term in the review and determine the polarity of the aspects. The following simple book review shows the annotation process “لم أكمل الرواية ولا أنوي ذلك التفاصيل معقدة والشخصيات مملة”. Thus, determining the main aspects in the review, which are “التفاصيل” and “الشخصيات”. After that, defining the polarity of the selected aspects, which are “negative” and negative”, respectively. Finally, determined whether the whole review polarity is positive or negative. The annotation process shows a simple example in Figure 1.

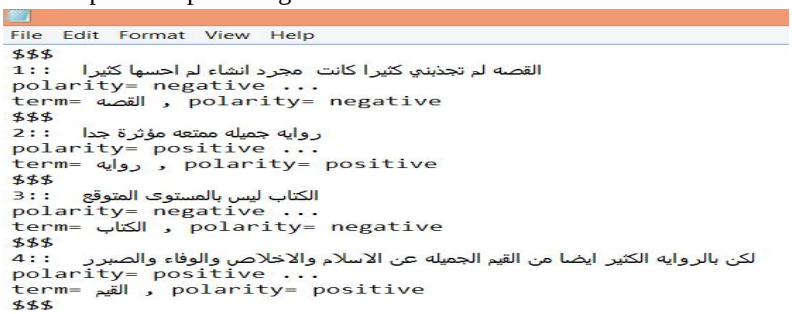

Figure 1: An example of annotation process

2- Preprocessing: Arabic text preprocessing is an important operation since most of reviews and comments are written in an unstructured format. Preprocessing phase consists of these operations: normalization, removing definition articles and removing stop words. The normalization process converts many forms of a word into a well-known form. Part of our tool is dealing with some normalization tasks such as letters diacritics removal, making " إ " ," أ " and " آ " change to " ا "(y), making both " ه " and " ة " change to " ة " at the end of word, etc. Furthermore, There is no specific rule to remove the stop words, so the most common stop words are removed in Arabic such like " في " (in), " من " (of), " على " (on), etc., and removing definition articles such as "ال" (the).

3- Determining aspect and sentence/review polarity: This phase considers the core of this work, which concerns with determining the polarity for a given aspect and a whole review. The proposed approach that treats this phase is hybrid approach of lexicon-based and corpus-based techniques. Where polarity labeled corpus of reviews manually and the tool construct an automatic lexicon, which includes the main keywords that surround the aspect with their polarities in order to utilize it in prediction polarities of test data. In detailed; firstly, getting the aspect and determining its position in the review in order to know the sense words that effect on the polarity of the aspect.

In training data, the words that are coming after and before the aspect are gained, except for aspects that exist at the end of sentence, only the word before the aspect is considered. The reason is that in Arabic language which is noticed that most of times the adjective comes after the noun. After that we get the polarity of aspects and store them in a lexicon. Whereas, in testing data the words that are coming after and before the given aspect are obtained and search for them in lexicon that is built from training data; if sense word exists in the lexicon then returns a polarity of sense word and tag it to the given aspect. However, if the word does not exist, we convert to statistics of aspects and polarities. The annotated data will be divided into train and test data. The following example shows these process, " لم أكمل الرواية ولا أنوي ذلك التفاصيل معقدة والشخصيات مملة", where the main aspects in this review are "التفاصيل" and "الشخصيات", so we find the sense word that affects and change the orientation polarities of the aspects, which are "معقدة" effects on the aspect "التفاصيل" and "مملة" effects on the aspect "الشخصيات" and change the orientation polarities, and also effects on the whole of review. After that, these sense words saved in lexicon in order to use them in test reviews.

As for tests data, the annotation will be removed from the data and test the proposed approach, which is to predict the polarity of the aspects and reviews correctly or not. Accuracy is calculated in order to evaluate the proposed approach. Figure 2 shows the process of proposed approach.

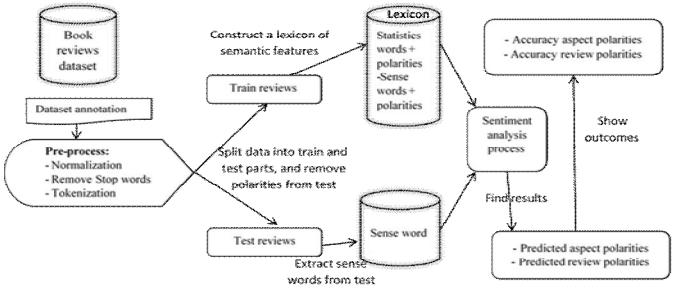

Figure 2: Aspect and review polarity determining process

4. EXPERMINTS AND RESULTS

In this section, the model is applied on collected dataset, which has 1000 annotated reviews and 1811 aspect terms. Around 800 sentences are entered to the model as a training set and 200 sentences as testing set. The model is evaluated by using accuracy measures as in (1); 20% of dataset is used for testing and 80% for training. The model concerned of aspect term polarity and review polarity detection.

$$\text{Accuracy} = \frac{\#\,Correct\ results}{\#\,All\ retrievals} \qquad (1)$$

According to the equation (1), the results showed that the accuracy if aspect term polarity detection is 80.5%, and review polarity prediction is 78%. Figure 3 shows the result.

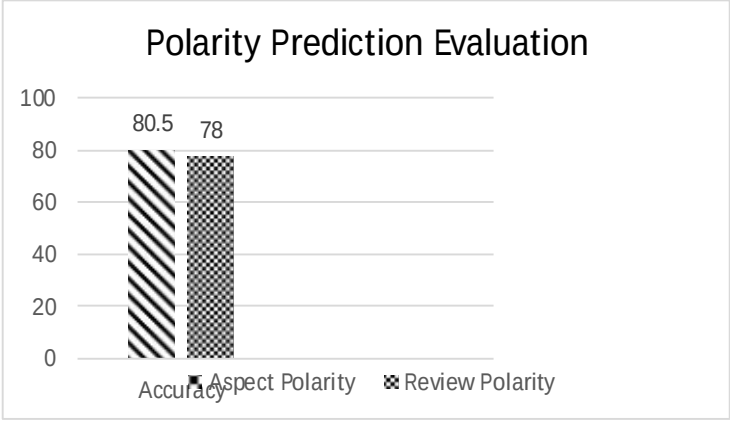

Figure 3: Aspect and review polarity results

Although the built model achieved acceptable accuracy in "aspect and review polarity determination", still there are some lacks through its behavior and functionality. In some case, the complex structure of Arabic language, sometimes the sense words doesn't exist explicitly and sometimes the sense word comes far away from the aspect. Figure 4 displays a snapshot of the result of the proposed system.

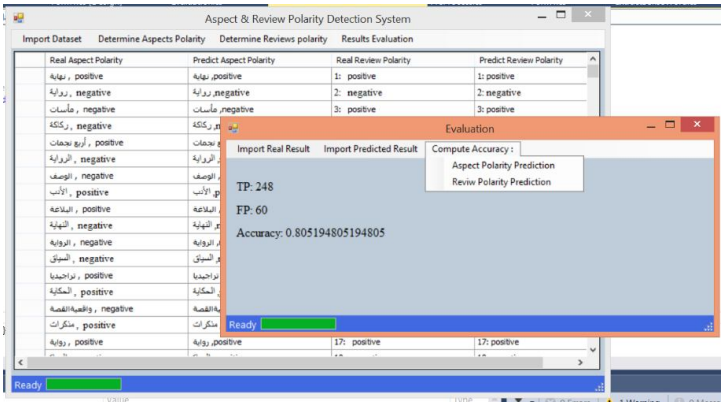

Figure 4: snapshot of the experimental results of the proposed system

5. CONCLUSION

In this work, a novel approach is suggested that tackle the problem of aspect and review polarity prediction. The proposed approach employs supervised machine learning using a combination of corpus-based and automatic lexicon-base techniques. A tool is built to achieve that using C# programing language. The work starts by annotating dataset and cleaning it from a noisy data, after that automatically building the lexicon with sensible words with their polarity that hints to extract aspects with their polarity. Experimental results showed that the proposed system got acceptable results, where the accuracy is 80.5% and 78% for aspect polarity and review polarity prediction, respectively.

REFERENCES

1. M. Rushdi Saleh, M.T. Martín Valdivia, L. A. Ureña López and J.M. Perea Ortega. **OCA: Opinion corpus for Arabic.** Journal of the American Society for Information Science and Technology, vol. 62, No. 10, pp. 2045-2054, 2011. https://doi.org/10.1002/asi.21598
2. Y. He and D. Zhou. '**Self-training from labeled features for sentiment analysis**', Information Processing, vol. 47, No. 4, pp.606–616, 2011.
3. M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi,and V. Hoste. **Semeval-2016 task 5: Aspect based sentiment analysis.** In 10th International Workshop on Semantic Evaluation (June 2016), pp. 1-13. https://doi.org/10.18653/v1/S16-1002
4. B, Liu. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, vol. 5, No. 1, pp. 1-167, 2012.
5. M. Aly and A. Atiya. **Labr: A Large scale arabic book reviews dataset**. In Meetings of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, vol. 2, pp. 494-498, 2013.
6. K. Ahmad, D. Cheng and Y. Almas. **Multi-lingual sentiment analysis of financial news streams**. In 1st International Workshop on Grid Technology for Financial Modeling and Simulation, vol. 26, p. 001. Sissa Medialab, May 2007. https://doi.org/10.22323/1.026.0001
7. Y. Almas and K. Ahmad. "**A note on extracting 'sentiments' in financial news in English, Arabic & Urdu**". The Second Workshop on Computation, al Approaches to Arabic Script-based Languages, pp. 21-22, 2007.
8. S. E. Abdelrahman, H. Mobarz, I. Farag, and M. Rashwan, "**Arabic Phrase-Level Contextual Polarity Recognition to Enhance Sentiment Arabic Lexical Semantic Database Generation**," IJACSC International Journal of Advanced Research in Artificial Intelligence, vol. 5, No. 10, 2014.
9. J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster and L. Tounsi. **Dcu: Aspect-based polarity classification for semeval task 4**, in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics and Dublin City University, pp. 223-229, August 2014. https://doi.org/10.3115/v1/S14-2036
10. B. Pang and L. Lee, "**A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts**," in Proceedings of the 42nd annual meeting on Association for Computational Linguistics, p. 271, 2004.
11. B. Pang and L. Lee. "**Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales**," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115-124, 2005.
12. B. Pang and L. Lee. "**Opinion mining and sentiment analysis**," Foundations and trends in information retrieval, vol. 2, pp. 1-135, 2008. https://doi.org/10.1561/1500000011
13. S.-M. Kim and E. Hovy. "**Determining the sentiment of opinions**," in Proceedings of the 20th international conference on Computational Linguistics, p. 1367,2004.
14. P. D. Turney, "**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, 2002.
15. R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts. "**Recursive deep models for semantic compositionality over a sentiment treebank**," in Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp. 1631-1642, 2013.
16. T. Wilson, J. Wiebe, and P. Hoffmann, "**Recognizing contextual polarity in phrase-level sentiment analysis**," in Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347-354, 2005. https://doi.org/10.3115/1220575.1220619

17. A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes. **Sentiment analysis of political tweets: Towards an accurate classifier**. Association for Computational Linguistics., pp. 49-58, June 2013.
18. K. Schouten and F. Frasincar . **Survey on aspect-level sentiment analysis**. IEEE Transactions on Knowledge and Data Engineering, vol. 28, No. 3, pp. 813-830, 2015. https://doi.org/10.1109/TKDE.2015.2485209
19. Y. Zhao, B. Qin, S. Hu, and T. Liu, "**Generalizing Syntactic Structures for Product Attribute Candidate Extraction**," In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 377-380, 2010.