Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis

Conference Paper · August 2015

DOI: 10.1109/FiCloud.2015.62

4 authors:

Mohammad AL-Smadi
Jordan University of Science and Technology
63 PUBLICATIONS **279** CITATIONS
SEE PROFILE

Omar Qawasmeh
Université Jean Monnet
6 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Bashar Bassam Talafha
Jordan University of Science and Technology
4 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Muhannad Quwaider
Jordan University of Science and Technology
33 PUBLICATIONS **536** CITATIONS
SEE PROFILE

Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis

Mohammad AL-Smadi*,
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan
*maalsmadi9@just.edu.jo

Omar Qawasmeh
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan
omar_qawasmeh@hotmail.com

Bashar Talafha
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan
talafha@live.com

Muhannad Quwaider
Computer Engineering Department
Jordan University of Science and Technology
Irbid, Jordan
mqquwaider@just.edu.jo

***Abstract*—With the prominent advances in Web interaction and the enormous growth in user-generated content, sentiment analysis has gained more interest in commercial and academic purposes. Recently, sentiment analysis of Arabic user-generated content is increasingly viewed as an important research field. However, the majority of available approaches target the overall polarity of the text. To the best of our knowledge, there is no available research on aspect-based sentiment analysis (ABSA) of Arabic text. This can be explained due to the lack of publically available datasets prepared for ABSA, and to the slow progress in sentiment analysis of Arabic text research in general. This paper fosters the domain of Arabic ABSA, and provides a benchmark human annotated Arabic dataset (HAAD). HAAD consists of books reviews in Arabic which have been annotated by humans with aspect terms and their polarities. Nevertheless, the paper reports a baseline results and a common evaluation technique to facilitate future evaluation of research and methods.**

Keywords—Sentiment Analysis; Aspect Based Sentiment Analysis; Natural Language Processing; Arabic Dataset.

I. INTRODUCTION

The rapid increase in user-generated content on the Web has demanded more advances in research oriented to sentiment and opinion mining. Sentiment analysis (SA) has become more important due to its academic and commercial potential. Although SA research has achieved prominent steps in mining the sentiment of English language [1 - 9] research in SA of Arabic text is still struggling [10 - 16]. Recently, the SA research has a shift towards having more fine-grained approaches taking into consideration the main aspects of a specific entity and how they influence the text sentiment [4]. For instance, user-generated reviews of products may not only evaluate the overall product but also express some sentiments on product specific aspects, such as price, performance, etc. Moreover, one review may contain a conflict of sentiment on different aspects (e.g. "*The food is delightful, but the prices are high*") which makes the sentiment analysis of the overall review text more challenging.

Aspect based sentiment analysis (ABSA) deals with extracting the aspects of the text main entities and identifying the sentiment the text expresses for each aspect [9]. Examples of ABSA research for English text cab be found in movie reviews [17], electronic products reviews [2], and restaurants [1, 8]. However, to the best of our knowledge there is no research in ABSA of Arabic text. A recent survey for SA in Arabic text shows that none of the research conducted in Arabic SA has considered ABSA [10].

The raising interest in SA of Arabic text has demanded having publically available datasets. The datasets OCA and AWATIF are pioneering in this field. Opinion Corpus for Arabic (OCA) consists of 500 movie reviews divided equally among the positive/negative classes [18, 19]. Whereas the AWATIF dataset was generated with a consideration of many linguistic issues [20]. For instance, from the annotators point of view how being aware of certain linguistic features of subjectivity and sentiment analysis would affect their decision. Another example is the large scale dataset of book reviews in Arabic language (LABR) [21]. LABR has around 63k book reviews with rating scale of 1-5 each. The scale starts from negative reviews of rate 1 to positive reviews of rate 5. However, none of the previous datasets was prepared to foster ABSA research. In this paper we aim to tackle this problem and bridge the gap between SA research in general and ABSA in particular.

This research aims at having a publically available dataset serves as a benchmark for the research of ABSA in Arabic texts. In order to meet the research goals, this paper presents a human annotated Arabic dataset (HAAD) prepared to support the research of ABSA with machine-readable dataset prepared carefully to support a set of research tasks such as aspect extraction and polarity detection, aspect category identification and category polarity detection (see section II). The process of data collection and annotation is discussed in section III. The dataset and the related tasks have been evaluated with a baseline approach which is discussed in section IV. Section V concludes this research and sheds the light on future plans and work.

II. ABSA RELATED TASKS

HAAD has been prepared to cover the following research tasks:

A. T1: Aspect Term Extraction

Given a review sentence, this task deals with extracting all the possible aspect terms with respect to the review domain (i.e. Book reviews in our case) reviewed by the sentence. Examples of annotated aspect terms are: (الابطال /Actors, الكتاب / Book). The extraction of aspects is done regardless to their polarity. For instance, conflict and neutral aspect terms should be extracted as well.

B. T2: Aspect Term Polarity

Depending on previous task (T1), this task focuses on assigning the extracted aspects to the polarity class (positive, negative, conflict, and neutral). The conflict case happens when both positive and negative sentiment is expressed by the same aspect term or category (e.g. *"روايه جميله ولكنها معقده بعض الشيء" / "An interesting novel but a bit complicated"*).

C. T3: Aspect Category Identification

Having a predefined aspect categories (see table II) and a collection of review sentences (without any annotations), this task investigate the ability of assigning each review sentence to one or more aspect category. The difference between this task and T1 is that the aspect terms are more fine-grained and should appear in the review sentence, whereas the aspect category is coarser category of the sentence and do not appear in the review sentence. Moreover, the aspect categories are not identified using aspect terms in the sentence, but rather inferred using sense words, adjectives, or context of the sentence meaning.

D. T4: Aspect Category Polarity

Having that the aspect categories of the review sentences are given, this task investigates the possibilities of assigning a specific polarity (positive, negative, conflict, and neutral) to each aspect category.

Tasks T1 and T2 can be targeted together and suitable for research inspecting which aspect terms influence the review polarity and how (aspect term and aspect term polarity). Whereas tasks T3 and T4 are completely separated from the other two tasks (in terms of dataset annotation) and can be a target of research focusing on which aspect categories influence the reviewer sentiment and how (aspect category and its polarity).

III. DATASET COLLCTION AND ANNOTATION

A. Data Collection

HAAD reviews have been selected out of the (LABR) a large scale dataset of book reviews in Arabic language [21]. LABR has around 63k book reviews with rating scale of 1-5 each. However, LABR lacks information related to ABSA research in general to the related tasks in particular (aspect terms (T1), aspect term polarity (T2), aspect category (T3), and aspect category polarity (T4)). Only overall review polarity is provided as a rating scale of 1-5. Positive reviews are rated 4 or 5, and negative reviews are those with ratings 1 or 2. Reviews with rating 3 are considered neutral.

The selection of candidate reviews for HAAD was group based, where 7 groups of 3 graduate students each collaborated on selecting 400 reviews for each group covering different books, different reviews of negative or positive reviews (i.e. with ratings of 1,2,4,5 respectively) out of the LABR dataset. At the end of the data collection task we had a first version of the HAAD consists of 2389 Arabic book reviews annotated with aspect terms (T1), aspect term polarity (T2), aspect category (T3), and aspect category polarity (T4). However, after linguistic review some reviews where left out to have a second version of 1513 reviews. HAAD contains of 2838 aspect term occurrences, 1296 out of them are distinct aspect terms.

B. Anotation Process

The annotated dataset contains information related to the four discussed tasks: information related to aspect terms (T1), aspect term polarity (T2), aspect category (T3), and aspect category polarity (T4). To the best of our knowledge, there is no publically available text annotation tool that supports efficiently Arabic text annotation. However, for the sake of annotation, the annotators used the BRAT web-based annotation tool [22] which was configured to meet the needs of the annotation phase. Fig. 1 depicts an example for an annotated sentence in BRAT using the annotation configuration (see next section), as done by annotators.

The annotation process has been done by 7 groups selected out of post-graduate students enrolled to the course of Natural Language Processing (CS-722) in winter term 2014 at Jordan University of Science and Technology. Each group - consists of 3 members - was asked to annotate 400 distinct reviews selected out of LABR for different Books with different polarities. Internally, group members were asked to annotate and review their peers' annotations. All participants were native Arabic speakers. Out of students annotations we had 2389 annotated sentences formatted as XML. In the second phase the whole dataset was reviewed by the course instructor who is also a native Arabic speaker and holds a Ph.D. degree in computer science. Some of the reviews were deleted because of linguistic problems and invalid selection of aspect terms which forms 37% of the whole dataset to have only 1513 reviews in the second version of HAAD.

The groups had a training session on annotation using BRAT and were given guidelines for the required annotation as follows: (a) **Aspect terms and polarities.** During this stage, the annotators were asked to annotate all single/multiple terms that refer to specific terms (e.g. الابطال /Actors, الكتاب / Book) of the target entity (i.e. Books in our case). The aspect terms were annotated as they appeared in the original review even if they were misspelled. For each annotated aspect term, the annotators were asked to provide a polarity value (positive, negative, conflict, neutral). Table I summarizes the aspect terms distribution over the sentiment class in both training and testing datasets. (b) **Aspect categories and polarities.** In this stage, annotators were asked to assign each review sentence to corresponding aspect categories (e.g. المشاعر/ Feelings, الحبكه / Plot, الاسلوب / Style) and provide the polarity (positive, negative, conflict, neutral) of each aspect category. Table II summarizes the aspect categories distribution over the sentiment class.

TABLE I. ASPECT TERMS AND THEIR POLARITIES

Dataset	Polarity				Total
	Positive	*Negative*	*Conflict*	*Neutral*	
Train	1252	855	26	126	2259
Test	124	432	1	22	579
Overall Dataset	1376	1287	27	148	2838

```
<sentence id="915">
    <text>ميزة هذا الكتاب أن الأبطال هم عرب مثلنا</text>
    <aspectTerms>
        <aspectTerm term="الأبطال" polarity="positive" from="19" to="26"/>
        <aspectTerm term="الكتاب" polarity="positive" from="9" to="15"/>
    </aspectTerms>
    <aspectCategories>
        <aspectCategory category="المزايا" polarity="positive"/>
    </aspectCategories>
</sentence>
```

Fig. 1. Example from HAAD annotated and presented using XML.

TABLE II. CATEGORIES DISTRIBUTION OVER THE SENTIMENT CLASS.

Category	Polarity								*Total*	
	Positive		Negative		Conflict		Neutral			
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
الوقت / Time	1	0	4	5	0	0	0	0	5	5
الهوامش / Margins	2	0	2	4	0	0	0	0	4	4
المؤلف / Author	23	0	34	21	0	0	0	0	57	21
المشاعر/ Feelings	87	9	48	22	0	0	0	0	135	31
المزايا / Benefits	75	6	12	0	0	0	0	0	87	6
اللغات / Languages	6	0	1	7	0	1	0	0	7	8
الطائفيه / Sectarianism	5	0	6	5	0	0	0	0	11	5
السياق / Context	95	0	62	45	1	0	3	1	161	46
السلبيات / Negatives	0	0	42	44	1	1	0	0	43	45
الخاتمه / Epilogue	18	0	17	4	0	0	0	0	35	4
الحبكه / Plot	104	11	38	22	4	0	1	0	147	33
التقييم / Rating	70	0	100	24	1	0	7	0	178	24
الاماكن / Places	4	0	4	1	0	0	0	0	8	1
الاسلوب / Style	200	5	112	64	8	0	2	0	322	69
Total	**690**	**31**	**482**	**268**	**15**	**2**	**13**	**1**	**1200**	**302**

C. *Annotation Format*

As there is no publically available other dataset for ABSA of Arabic text we aim to provide this dataset as a benchmark one for Arabic ABSA research. However, for ABSA research in English, benchmark datasets have different annotations Schemes [9]. For instance the restaurants reviews dataset [1] uses category-based annotation using six categories (e.g. Food, Price, etc.) and four polarity labels (Positive, Negative, Conflict, Neutral) on the level of the whole sentence. In the research of [2] for product reviews feature-based sentiment analysis is used to classify negative/positive reviews (e.g. Digital Camera, Feature: picture quality, 250 positive reviews, 35 negative reviews). Among possible schemes, the SemEval2014 Task4 dataset schema has been selected for this task [9]. Semantic Evaluation (SemEval) is a prestigious workshop in the domain of natural language processing. SemEval 2014 had a task (Task4) on ABSA research where the organizers provided a benchmark dataset formatted based on XML for restaurants and laptops reviews in English. The selected schema was used to configure the BRAT tool

annotation and a mapping layer where used to map the BRAT annotated file to the SemEval-Task4 compliant XML file.

As depicted in Fig. 1, each review sentence in the HAAD dataset is annotated using the following XML tags:

- **<aspectTerm term=" " polarity=" " from="" to=""/>** XML element for each occurrence of an aspect term. In addition to the aspect term polarity its location in the text is provided based on start and end index of text characters.
- **<aspectCategory category=" " polarity=" "/>** XML element for each occurrence of an aspect term category.

The XML-based dataset is available publically for non-commercial research on a repository prepared for sharing and dissemination purposes[1].

IV. BASELINE EVALAUATION

As we aim to provide a benchmark dataset for ABSA of Arabic reviews, the dataset is provided with baseline evaluation for the four tasks discussed in section II. The baseline evaluation is based on [9] and is explained as follows:

A. Approaches of Baseline Evaluation

T1: Aspect term extraction baseline: the baseline tags all the tokens in the test dataset if they are listed in the human annotations list of aspect terms from the training dataset.

T2: Aspect term polarity baseline: for each aspect term *t* in the test sentence *s*, the baseline checks if *t* has been seen in the training sentences. If yes, the baseline retrieves the *d* most similar sentences of the training to *s*, and assigns to the aspect term *t* the most frequent polarity in the *d* sentences. The Dice coefficient similarity measure is used to compute the distance between sentences *s* and *d*. If not, if *t* has not been seen in the training set, *t* is assigned the most frequent aspect term polarity label in the training set.

T3: Aspect category extraction baseline: for every test sentence *s*, the baseline retrievers the *d* most similar sentences in the training set (as in T2 baseline). The *c* most frequent aspect category label of *d* is then assigned to *s*.

T4: Aspect category polarity baseline: each aspect category *c* in the test senesce *s* is assigned the most frequent polarity label *l* that have the *d* most similar training sentences to *s*. Sentence similarity is computed as in T2 baseline. If *c* is not seen in the training sentences, then the most frequent aspect category polarity label *l* in the whole training set is assigned to aspect category *c*.

B. Evaluation Measures

In order to evaluate aspect term extraction (T1) and aspect category detection (T3), the F_1 measure is computed:

$$F_1 = \frac{2 \,.\, P \,.\, R}{P + R}$$

Precision (P) and recall (R) are computed as follows:

$$P = \frac{|S \cap G|}{|S|}$$

$$R = \frac{|S \cap G|}{|G|}$$

where S is the set of aspect terms or aspect categories annotations out of the test sentences in T1 and T3 respectively. G is the set of the gold (correct) aspect terms or aspect categories for the same test set.

In order to evaluate aspect term polarity (T2) and aspect category polarity (T4), the accuracy of approach being used is measured. The accuracy measure is defined as the number of correctly predicted polarities divided by the total number of aspect term polarity or aspect category polarity annotations.

C. Baseline Results

Evaluating the prepared dataset using the baseline evaluation has led to the results presented in Table III. In task 1 (T1) the results was (P = 0.209877), (R = 0.264249), and (F1 = 0.233945). The accuracy for aspect term polarity was (Accuracy = 0.297064) (#Correct/#All: 172/579). Whereas the results for task 3 (T3) was (P = 0.151815), (R = 0.151815), and (F1 = 0.151815) where the baseline approach retrieved all the aspect term categories in the test set. In task 4 (T4) the accuracy for the category polarity identification was (Accuracy = 0.425743) (#Correct/#All: 129/303).

The dataset is equipped with a tool for common evaluation technique. Future researchers can use this tool to compare their results with the Test-gold dataset file to compute the same measures used in the baseline approach. This will help them to use HAAD as a reference benchmark dataset for ABSA research of Arabic text.

TABLE III. BASELINE RESULTS

Task	Results	
	F_1	*Accuracy*
T1: Aspect Term Extraction	0.233945	
T2: Aspect Term Polarity		0.297064
T3: Aspect Category Identification	0.151815	
T4: Aspect Category Polarity		0.425743

V. CONCLUSIONS AND FUTURE WORK

This paper provides a dataset for aspect-based sentiment analysis of Arabic text. Moreover, this dataset has been designed to serve as a reference benchmark dataset for the ABSA of Arabic text. The human annotated Arabic dataset (HAAD) consists of 1513 books review sentences. The sentences were selected form the LABR dataset [21] and annotated based on the guidelines of SemEval 2014: Task4 for ABSA [9].

HAAD has been prepared to cover different research tasks related to the domain of ABSA. Four tasks have been considered covering aspect terms extraction (T1), aspect term

[1] . https://github.com/msmadi/HAAD.git.

polarity identification (T2), aspect category selection (T3), and aspect category polarity identification (T4). Moreover, HAAD provides baseline results to evaluate the aforementioned four tasks. HAAD is supported with a common evaluation technique to compare conducted research and approaches with the baseline results.

As future plans, we aim to extend HAAD with more reviews form the domain of book reviews and to include other domains such as electronic products, restaurants, and hotels. Nevertheless, we are working on more advanced approaches for ABSA of Arabic text which will be evaluated on HAAD.

ACKNOWLEDGMENT

The authors are grateful to all the participants who supported in collecting and annotating the dataset and to the organizers of "SemEval 2014 - Task4: Aspect Based Sentiment Analysis" who allowed us to use their annotation guidelines and evaluation framework. This research is supported by Jordan University of Science and Technology, Research Grant Number: 20150164.

REFERENCES

[1] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content". Proceedings of the 12th International Workshop on the Web and Databases, Providence, Rhode Island, 2009.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews". Proceedings of the 10th KDD, pp. 168–177, Seattle, WA, 2004.

[3] S.-M. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text". Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 1– 8, Sydney, Australia, 2006.

[4] B. Liu, "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2012.

[5] S. Moghaddam and M. Ester, "Opinion digger: an unsupervised opinion miner from unstructured product reviews". Proceedings of the 19th CIKM, pp. 1825–1828, Toronto, ON, 2010.

[6] M. Tsytsarau and T. Palpanas. "Survey on mining subjective data on the web". Data Mining and Knowledge Discovery, 24(3):478–514, 2012.

[7] Z. Zhai, B. Liu, H. Xu, and P. Jia. "Clustering product features for opinion mining". Proceedings of the 4th International Conference of WSDM, pp. 347–354, Hong Kong, 2011.

[8] S. Brody and N. Elhadad. "An unsupervised aspect-sentiment model for online reviews". Proceedings of NAACL, pages 804–812, Los Angeles, CA, 2010.

[9] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis". In Proceedings of the *8th International Workshop on Semantic Evaluation*, SemEval 2014, Dublin, Ireland.

[10] N. Al-Twairesh,, H. Al-khalifa, and A. Al-Salman. "Subjectivity and Sentiment Analysis of Arabic:Trends and Challenges." *The ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*. 2014.

[11] A. Khurshid, D. Cheng, and Y. Almas. "Multi-lingual sentiment analysis of financial news streams." In *Proc. of the 1st Intl. Conf. on Grid in Finance*. 2006.

[12] Y. Almas,, and A. Khurshid. "A note on extracting 'sentiments' in financial news in English, Arabic & Urdu." In *The Second Workshop on Computation, al Approaches to Arabic Script-based Languages*, vol. 21, pp. 21-22. 2007.

[13] M. Elhawary and M. Elfeky. "Mining Arabic business reviews." In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pp. 1108-1113. IEEE, 2010.

[14] N. Farra, C. Elie, R. Abou Assi, and H. Hajj. "Sentence-level and document-level sentiment mining for Arabic texts." In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pp. 1114-1119. IEEE, 2010.

[15] N. Abdulla, M. Mahyoub, M. Shehab, and M. Al-Ayyoub. "Arabic sentiment analysis: Corpus-based and lexicon-based." In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. 2013.

[16] M. Itani, L. Hamandi, R. N. Zantout, and I. Elkabani. "Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes." In *Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on*, pp. 192-197. IEEE, 2012.

[17] T. T. Thet, J. Na, and C. S.G. Khoo. " Aspect-based sentiment analysis of movie reviews on discussion boards". *J. Inf. Sci.* 36, 6,pp 823-848, December 2010.

[18] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega. "Bilingual experiments with an arabic-english corpus for opinion mining." In *RANLP*. 2011.

[19] 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* 62, no. 10 (2011): 2045-2054.

[20] M. Abdul-Mageed and M. T. Diab. "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis." In *LREC*, pp. 3907-3914. 2012.

[21] M. Aly and A. Atiya. "LABR: Large-scale Arabic Book Reviews Dataset". *Association of Computational Linguistics (ACL)*, Bulgaria, August 2013.

[22] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. "BRAT: a web-based tool for NLP-assisted text annotation". In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 102-107, 2012.