A Review on Corpus Annotation for Arabic Sentiment Analysis

Latifah Almuqren[2,3], Arwa Alzammam[1], Shahad Alotaibi[2], Alexandra I. Cristea[3], and Sarah Alhumoud[1]

[1] Computer Science Department, Al-Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia
[2] Computer Science Department, Princess Nourah University, Riyadh, Saudi Arabia
[3] Computer Science Department, Warwick University, UK

As L.Almuqren@wawick.ac.uk, Alzammam@sm.imamu.edu.sa, shahdTaleb92@gmail.com , sohumoud@imamu.edu.sa, A.I,Cristea@warwick.ac.uk

Abstract. Mining publicly available data for meaning and value is an important research direction within social media analysis. To automatically analyze collected textual data, a manual effort is needed for a successful machine learning algorithm to effectively classify text. This pertains to annotating the text adding labels to each data entry. Arabic is one of the languages that are growing rapidly in the research of sentiment analysis, despite limited resources and scares annotated corpora. In this paper, we review the annotation process carried out by those papers. A total of 27 papers were reviewed between the years of 2010 and 2016.

1 Introduction

Today, social media has become a key part of many people's life. Social media provides means of communication that allow people to share their sentiments, opinions, and thoughts. By mining the content of social media, targeting opinions, valuable trends and feedback about topics or products could be inferred.

One of the known techniques in mining and analyzing social media is Sentiment Analysis (SA). SA involves a number of areas, such as computational linguistics, natural language processing and text analytics. Its goal is to capture the user sentiment about many aspects, for instance to detect product preferences and guide company strategies [1].

Whilst sentiment analysis is a current hot topic, especially for the English language, SA in Arabic text remains an inadequately researched area, due to the unique nature and structure of the Arabic language [2]. For instance, whist Modern Standard Arabic (MSA) is the formal Arabic language that is used in news and media,

people in social media rarely use MSA in their daily interaction. Instead they use dialectal Arabic, which is different from a region to another even in the same country. This variety of forms of the Arabic dialectal written language, adds to the challenges of the Arabic SA [3].

For the same reason, there is a lack of a resources and corpora for Arabic SA [3], [4], and this is the area this paper targets. In order to perform sentiment analysis, a crucial and potentially expensive, in that it is mostly manual, step is needed, which is *corpus annotation*. Leech [5] defined this step as assigning interpretative information to a document collection for mining use. Leech [5] coined the interpretative information as *linguistic information* and he distinguished between the corpus and *interpretive information.* The importance of the annotation process results from making a machine-readable version of the meta-data by annotating the corpus, for training a machine learning classifier [4].

There are dissimilar levels of corpus annotation, i.e. syntactic annotation, which is the process of parsing every sentence in the corpus and labelling it with its structure grammar, semantic annotation and POS tagging that is, labelling every word in the corpus with its appropriate part of speech label. At the semantic level, several labels are used in the annotation process (positive, negative and neutral). Semantic annotation is used for sentiment mining purpose [6]. Positive and negative sentiments can occur within the same tweet, which is considered as a challenge to the annotation process [6].

The two main approaches used to annotate a corpus, which are: the manual approach that depends on human labor, and the automatic approach that uses an annotation tool and crowdsourcing [6]. Each approach has its advantages and disadvantages. To assure the high quality of the manual annotation process, it should consider the precision, speed, and consistency when assigning annotators [5]. In addition, the annotation process needs clear guidelines, to ensure the consistency between annotators.

This report provides a comprehensive review of the annotation process for Arabic corpus, highlighting the gaps and similarities in the annotation process in the considered research. The report starts by reviewing the methodology that was used in this research. Next, it provides the annotation procedures. Finally, a conclusion is drawn.

2 Methodology

In this paper, we reviewed 56 papers on the topic of Arabic corpora annotation following the methodology in [7], focusing on several angles, such as the research corpora type, annotation process, and verification methodology. The aim of this review is to highlight the gaps and similarities in the annotation process in the considered research. Starting with comprehensive searches in different electronic databases, such as Google Scholar, ERIC, Science Direct, Sage, and Springer Link. a total of 56 papers in annotation and sentiment analysis were collected, published between the years of 2010 and 2016. The search terms used in the search process were: “Annotation”, “Arabic annotation”, “sentiment analysis”, “Twitter annotation”, “Arabic sentiment analysis”. The results are then filtered to 27 papers focusing on Arabic Sentiment Analysis including conference proceedings; peer-reviewed articles; conference workshops. Several papers concentrating on annotation different languages other than Arabic were eliminated.

3 Arabic Corpora Annotation for Sentiment Analysis

This section explains the different stages in the annotation process through the considered literature. The first subsection explains the different corpus types that are annotated by the authors. The second subsection explains the annotation processes and procedures with regards to the number of annotators’ specialty and annotation type. The third subsection is on the verification process that proceeds the annotation to ensure the quality of annotation.

3.1 Corpus Type

Resources in Arabic language like corpora, lexicons and datasets is still scarce compared to other languages that have high user generated content in the web [8] [9]. Created an annotated corpus is a crucial step for sentiment analysis [26]. The quality of the annotation has a direct relation to the accuracy of the classifier. Several attempts have been made to accomplishing Arabic corpora annotation using different types of data during the period under consideration, from 2010-2016. The majority of papers focused on Twitter corpora, 15 papers as in Table 1.

Other studies, 7 of them, annotated reviews and comments. Reviews were mainly from restaurants, movies, books, and products. Reviews were mainly from restaurants, movies, books, and products [6], [9], [10] and [11]. Comments are scraped from hotel reservation or TV programs’ websites [12], [13] and [14]. The rest of the literature used different kinds of websites mainly news [15], [16], and [12]; chat [21], web forum [21] Facebook [19], and the Pen Arabic Tree Bank [31].

Table 1 Reviewed corpora types

Corpora	Paper count	Paper
Twitter	15	[4], [20], [21], [22], [23], [11], [15], [24], [25], [26], [27], [28], [2], [29], [30].
Facebook	1	[19]
Newspapers	3	[15],[16], [12]
Penn Arabic Tree bank	1	[31]
Reviews and comments	7	[6],[9], [10] ,[11], [12], [13], [14]
Chat website	1	[21]
web forum	1	[21]

3.2 Annotation Process

In the collection of papers on Arabic corpus annotation we have studied, each research adopts a different way to annotate the dataset they are processing. In this section, we summarize these methods, including defining data such as, the number of annotators needed and the labels used in each research.
Annotation can be done on different levels, such as *sentence level*, *word level* or both. Also, there are different ways that can be adopted for annotating a dataset; for example, the annotation can be done *manually* by a number of native speakers. This means they are asked to go through the data set and annotate each sentence or words (depending on the level of annotation). Then they have to assign a label for each sentence/word based on their sentiment to positive, negative, neutral, objective, mixed or other different labels, which also can be given to them beforehand. Another way to annotate a data set is to use the human resources that are available on the Internet, through what is called *crowdsourcing*. Among the papers we covered in this research, 85% of them annotated their dataset manually, and the remaining 15% used crowdsourcing tools.

Most papers under consideration used the manual annotation method, and most of those, 21 papers, annotated the corpora on a *sentence level.* That is 91% of the total number of the manually annotated corpora. While only two researches [32] and [22] conducted their annotation manually on a *word level*. The preference towards sentence level annotation is possibly due to it being simpler to do, and potentially deal with issues such as sarcasm and a mixture of differently charged words in the same sentence. Moreover, it is a speedier method, easier to process by human annotators.

The researchers on [33] and [16] used the manual annotation method with the help of tools that were used to facilitate the annotation process and to reduce time and workload for the annotators. The first research used the Amazon Mechanical Turk service using an API called Boto, to annotate the dataset manually [34]. And the second research developed their own tool which is named MANDIAC [16]. Another research [27] used the manual method involving five Arabic native speakers. Three of them were a linguistic specialist. Since in this research they built a lexicon for Arabic idioms and proverbs, therefore they needed a linguistic specialist to provide an accurate annotation for tweets that have idioms.

The annotation method using crowdsourcing tool on a sentence level, was implemented by researchers in [17], [24], [9], and [25]. In [17], [24], and [25] the annotation was carried out by volunteers on the web. Annotators were asked to register their usernames and passwords. Then they annotated the set of tweets that were shown to them one by one. An interesting different way to annotate the dataset involves using *games*. In this research [9], the players of the game were asked to annotate a set of sentences which are given to them, as they advanced with the level of the game. The next Table 2 summarizes the number of papers, and which type and level of annotation process it use.

Table 2 Annotation Types and levels

Annotation process type	**Paper Count**	**Paper**
Manually on a sentence level	21	[30], [21], [23], [10], [11], [31], [28], [29], [2], [15], [14], [20], [8], [13], [35], [26], [33], [16], [27], [18], [12].
Manually on a word Level	2	[32], [22].
Crowdsourcing Tools on a sentence level	4	[17], [24], [9], [25].

The number of annotators also differ between researchs, out of the 27 papers under consideration, nine of which depend on a total of three native Arabic speakers to annotate their dataset [30], [23], [17] [10], [33], [28], [29], [18], [25]. Authors in [27] annotated their dataset with the help of five people. Moreover, five researches asked two people in [21], [31], [2], [20], [24]. This is depicted in Figure 1 which shows the number of annotators in all 27 researches covered in this paper, and the percentage of each category.

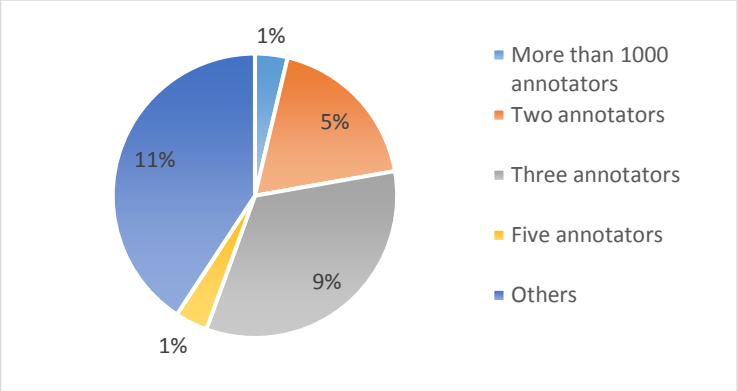

Figure 1 Number of annotators and its relation to the total number of search papers.

Authors in [11] used a site for reviewing books which is "Goodreads" [36] to help them with the annotation process where they collected 16486 users reviews and decided the polarity of the comment based on the number of stars given for a review. The rest of the 11 research papers either depended on reviews' corpora to rate movies, books or hotels or used crowdsourcing tools. Therefore, didn't mention the number of users helped with the annotation process. As of the factors that may had effects on emitting the number of annotators in those search papers, it can be due to the fact that the annotation process was done by random users on the web. Although it should have been mentioned in either case even if it was done by random users whose characteristics are not known. In [9] the authors developed a game; with the players annotating the data. The players could not move to the next level until all the team members agreed on the same result of annotation. This may seem to enforce the agreement brutally and hinder the game experience, by motivating the players to select whatever makes them agree with the other player to move into the next level.

3.3 Verification

The accuracy of the annotation results has an impact on the sentiment analysis quality. The observations' labels have to be verified manually or by calculating a coefficient. Different kinds of verification methods are used with the annotation

methods ranging from simple similarity index by counting the number of agreements to more elaborate and accurate methods that measure agreements that occur by chance only. For instance, The Kappa coefficient is used with categorical data to measure reliability between two observers assigning cases to a set of k categories [37]. Annotating observation is a tedious task that is prone to human biases and errors. In this subsection, we discuss the verification measures that adopted by the authors under consideration to this survey. Some studies implemented Kappa coefficient and others used manual ways, like deleting any data that has contradicting opinions, moderate the opinion with the help of another annotator, or promote discussion between the annotators. That is, in the case of a disagreement between two annotators, a third one is brought in, and the majority-voting principle is used to finalize the decision, as in [30], [23] and [29]. The Majority-voting was also used to choose the final label in [17], [25] and [24]. In [25] there are two groups of annotators: supervisor annotators, which are three and one of them is the paper author, and non-supervisor annotators, which are the users of their tool. The observation was considered rated correctly if there was a match between the labels assigned by the supervisor annotators and the labels assigned by non-supervisor annotator. In [21], [18], [10], [31], [28] and [20] the authors used the Kappa coefficient to measure the inter-annotators agreements. The results of Kappa coefficient in [18] is 0.454, in [10] the average is 0.45, in [31] the generic set had an agreement of 0.321, while on the topic-specific set it was 0.397. In [28] it is about 0.96, in [20] the average is 0.51 and in [21] there are four data sets that listed Sequentially with their agreement results: DARDASHA has 0.89, TAGREED has 0.85, TAHRIR has 0.85, and MONTADA has 0.88. For conflict resolution, in [2] and [27], if the annotators gave different labels, then they discuss and agree to choose one of them. If they could not choose one, then they ignore the observation which affects the size of the corpus. The authors in [15] deleted any review that had no specific orientation towards positive or negative.. In the research carried out by [17] the verification process was not mentioned, however, data has to be annotated by at least three users of their annotation tool and the user can delete any empty data, duplicated data or any data written using English letters. In [33], they used a public tool for sets of three annotators, but if there was a conflict between them in annotating an observation, this observation is ignored. All papers that annotated their corpora based on the review rating, as in number or stars or points, like [11], [8] and [35] did not mention the method of verification.

4 Conclusion

Arabic as a rich morphology language with scares resources and lack of corpora. This research provides a review on the recent annotation approaches carried out in the process of Arabic Sentiment Analysis for the creation of Arabic language corpora. The review covered studies started between the years 2010 and 2016, which resulted

in 27 papers. In the reviewed papers annotation process is carried out via two main approaches: manual annotation and crowdsourcing. In this review 85% of the literature under consideration did the annotation manually, the remaining, used crowdsourcing tools. Moreover, the most annotated corpus was the micro-blogging platform Twitter 56%, other platforms %18 where covering Facebook, Newspapers and Penn Arabic Tree bank. While the rest of the corpus was reviews and comments 26%. There is a need to build and provide Arabic annotated corpora that researchers can use directly without the burden of going through the stage of annotation. This will provide more space to focus on other important research areas like building a better classifier that encompasses dialectal processing.

References

[1] A. Kaur and V. Gupta, 'A survey on sentiment analysis and opinion mining techniques', *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 4, pp. 367–371, 2013.

[2] Mourad, A. and Darwish, K., 'Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs.', in *In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2013, pp. 55–64.

[3] N. Y. Habash, 'Introduction to Arabic Natural Language Processing', in *Synthesis Lectures on Human Language Technologies*, 3(1) vols, 2010, pp. 1–187.

[4] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, 'Subjectivity and sentiment analysis of Arabic: Trends and challenges', in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 148–155.

[5] G. Leech, 'Corpus annotation schemes', *Literary and linguistic computing*, vol. 8, no. 4, pp. 275–281, 1993.

[6] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[7] A. alOwisheq, S. alHumoud, N. alTwairesh, and T. alBuhairi, 'Arabic Sentiment Analysis Resources: A Survey', in *Social Computing and Social Media*, 2016, pp. 267–278.

[8] M. 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*, 2011.

[9] A. A. Al-Subaihin, H. S. Al-Khalifa, and A. S. Al-Salman, '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 and Services*, 2011, pp. 543–546.

[10] S. Alhazmi, W. Black, and J. McNaught, 'Arabic SentiWordNet in relation to SentiWordNet 3.0', *2180*, vol. 1266, no. 4, p. 1, 2013.

[11] M. A. Aly and A. F. Atiya, 'LABR: A Large Scale Arabic Book Reviews Dataset.', in *ACL (2)*, 2013, pp. 494–498.

[12] H. B. Asmaa MOUNTASSIR and Ilham BERRADA, 'Sentiment Classification on Arabic corpora: A preliminary cross-study', *Lavoisier*, vol. 16, no. 1279–5127, p. 121, Jan. 2013.
[13] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar, 'An Opinion Analysis Tool for Colloquial and Standard Arabic', *In The Fourth International Conference on Information and Communication Systems (ICICS)*, 2013.
[14] F. R. Maurizio Tesconi, C. A. Salvatore Minutoli, and Andrea Marchetti, 'KAFnotator: a multilingual semantic text annotation tool', in *5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*, Hong Kong, 2010.
[15] S. M. A. Aqil M. Azmi, 'Aara'– a system for mining the polarity of Saudi public opinion through e-newspaper comments', *Journal of Information Science*, vol. 40, no. 3, pp. 398–410, Jun. 2014.
[16] O. Obeid *et al.*, 'MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization', *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, 2016.
[17] R. M. Duwairi and I. Qarqaz, 'Arabic sentiment analysis using supervised classification', in *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, 2014, pp. 579–583.
[18] M. Abdul-Mageed and M. Diab, 'SANA:A Large Scale Multi-Genre,Multi-Dialect Lexicon fo rArabic Subjectivity and Sentiment Analysis', 2014.
[19] M. Maamouri and A. Bies, 'Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools', in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Stroudsburg, PA, USA, 2004, pp. 2–9.
[20] M. Abdul-Mageed, H. AlHuzliy, D. AbuElhija, and M. Diab, 'DINA: AMulti-Dialect Dataset for Arabic Emotion Analysis', *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, no. 2016.
[21] M. Abdul-Mageed, S. Kübler, and M. Diab, 'Samar: A system for subjectivity and sentiment analysis of arabic social media', in *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, 2012, pp. 19–28.
[22] S. Alhumoud, T. Albuhairi, and M. Altuwaijri, 'Arabic Sentiment Analysis using WEKA a Hybrid Learning Approach', Nov. 2015.
[23] S. Al-Osaimi and K. Badruddin., 'Role of Emotion icons in Sentiment classification of Arabic Tweets', *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems.*
[24] R. Duwairi, 'Sentiment Analysis for Dialectical Arabic', *ICICS*, 2015.
[25] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, 'Sentiment analysis in arabic tweets', in *Information and communication systems (icics), 2014 5th international conference on*, 2014, pp. 1–6.
[26] K. N. N. El-Makky, E. A. Alaa El-Ebshihy, S. M. Omneya Hafez, and Shimaa Ibrahim, 'Sentiment Analysis of Colloquial Arabic Tweets', presented at the The 3rd ASE International Conference on Social Informatics (SocialInformatics 2014), Harvard University, Cambridge, MA, USA, 2014.

[27] H. Ibrahim, S. Abdou, and M. Gheith, 'Idioms-Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis', *International Journal of Computer Applications*.
[28] H. Ibrahim, S. Abdou, and M. Gheith, 'Sentiment Analysis For Modern Standard Arabic And Colloquial', *International Journal on Natural Language Computing (IJNLC)*, vol. 4.
[29] A. Shoukry and A. Rafea, 'Preprocessing Egyptian Dialect Tweets for Sentiment Mining', *AMTA*.
[30] A. Shoukry and A. Rafea, 'Sentence-level Arabic sentiment analysis', in *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, 2012, pp. 546–550.
[31] M. Abdul-Mageed and M. T. Diab, 'Subjectivity and sentiment annotation of modern standard arabic newswire', in *Proceedings of the 5th Linguistic Annotation Workshop*, 2011, pp. 110–118.
[32] S. Alhumoud, T. Albuhairi, and W. Alohaideb, 'Hybrid Sentiment Analyser for Arabic Tweets using R', *Conference Paper*, Nov. 2015.
[33] M. Nabil, M. Aly, and A. Atiya, 'ASTD: Arabic Sentiment Tweets Dataset', *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2515–2519, 2015.
[34] 'Boto API', *GitHub*. [Online]. Available: https://github.com/boto/boto. [Accessed: 05-Feb-2017].
[35] S. R. E.-B. Hady ElSahar, 'Building Large Arabic Multi-domain Resources for Sentiment Analysis', presented at the 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 2015, vol. 9042, pp. 23–34.
[36] 'Goodreads', *Goodreads*. [Online]. Available: https://www.goodreads.com/. [Accessed: 27-Nov-2016].
[37] J. Cohen, 'A coefficient of agreement for nominal scales', *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.