A Prototype for a Standard Arabic Sentiment Analysis Corpus

Mohammed N. Al-Kabi
Computer Science Department
Zarqa University
P. O. Box 2000
13110 Zarqa -Jordan
malkabi@zu.edu.jo

Mahmoud A. Al-Ayyoub
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan
maalshbool@just.edu.jo

Izzat M. Alsmadi
Computer Science Department
University of New Haven
West Haven, CT 06516, USA
ialsmadi@newhaven.edu

Heider A. Wahsheh
Computer Science Department
College of Computer Science
King Khaled University
Abha, Saudi Arabia
heiderwahsheh@yahoo.com

Abstract: *The researchers in the field of Arabic sentiment analysis (SA) need a relatively big standard Arabic sentiment analysis corpus to conduct their studies. There are a number of existing Arabic datasets; however they suffer from certain limitations such as the small number of reviews or topics they contain, the restriction to Modern Standard Arabic (MSA), or not being publicly available. Therefore, this study aims to establish a flexible and relatively big standard Arabic sentiment analysis corpus that can be considered as a pillar and cornerstone to build larger Arabic corpora. In addition to MSA, this corpus contains reviews written in the five main Arabic dialects (Egyptian, Levantine, Arabian Peninsula, Mesopotamian, and Maghrebi group). Furthermore, this corpus has other five types of reviews (English, mixed MSA & English, French, mixed MSA & Emoticons, and mixed Egyptian & Emoticons). This corpus is released for free to be used by researchers in this field, where it is characterized by its flexibility in allowing the users to add, remove, and revise its contents. The total number of topics and reviews of this initial copy is 250 and 1,442, respectively. The collected topics are distributed equally among five domains (classes): Economy, Food-Life style, Religion, Sport, and Technology, where each domain has 50 topics. This corpus is built manually to ensure the highest quality to the researchers in this field.*

Keywords: *Sentiment analysis; opinion mining; making of Arabic corpus; Arabic reference corpus; Maktoob Yahoo!*

1. Introduction

The Arabic language is a Semitic language originated in the Arabian Peninsula. It is one of the first common Semitic languages, such as Amharic, Hebrew, Tigrinya, and Aramaic. The Modern Standard Arabic (MSA) is the language mainly used in the media (Radio, TV, news bulletin, books, journals, newspapers, ads, etc.) and it is the language used in education and official correspondence. MSA dates back to the end of the eighteenth century, and it is the official language of 27 countries worldwide. These countries are located in the Arab world spanning the regions from Southwest Asia to Northwest Africa including the horn of Africa. There is no consensus on the total number of Arabic native speakers; researchers estimate that this number ranges between 280 and 400 million Arabic native speakers. Therefore, it is the fifth most used language in the world, and one of the official languages of the United Nations (UN) since 1974. The MSA is a descendant of the Classical Arabic (CA) language (the language of the holy Qur'an) that was used in the 6th century [1-5]. MSA and CA are different mainly in style and vocabulary.

The Arabic language used in the Arab world is divided into two main versions: Modern Standard Arabic (MSA) and Colloquial (dialectal) Arabic. The MSA has no variants while Colloquial Arabic has many regional variants (dialects). MSA is used mainly in formal speeches and interviews, formal print media, education, media, official correspondence, etc. Colloquial Arabic represents the real native spoken language used to communicate at homes, markets, offices, etc. Before the Internet era, Colloquial Arabic was known mainly in the spoken form not written form [3]. Arabic has a wide number of dialects that vary greatly between different Arab countries, cities, towns, and villages. Arabic speech divides into two main types: Bedouin and sedentary, and all of these dialects are sedentary dialects. The variations (vernaculars) in Arabic dialects are positively proportional to their geographical distances.

The Web 2.0 era offers to its users around the globe the ability to generate content (user-generated content

(UGC)). Web blogs and microblogs (Facebook, Twitter, Google +, etc.) have a huge amount of UGC that represent an important source of invaluable information about the users' needs, trends, opinions, etc. Extracting and analyzing such information manually is not an option. Therefore, such huge amount of UGC needs efficient and effective algorithms to be implemented. Different languages and dialects are used to generate this huge amount of UGC. Online social networking services like Facebook, Instagram, Twitter, YouTube, Google+, Vine, LinkedIn, Yahoo!, Pinterest, and Tumblr are used to collect the necessary datasets to conduct sentiment analysis and opinion mining.

This paper aims to lay a cornerstone for the creation of standard Arabic corpus for sentiment analysis and opinion mining. Furthermore, this corpus is suitable to be used for Arabic text classification studies. Such corpus can be enlarged or contracted according to the requirements of different researchers. The Maktoob Yahoo! website is used to collect 250 topics distributed equally among five selected topics (Economy, Food-Life style, Religion, Sport, and Technology); however, the numbers of collected reviews for each topic are not equal. The total number of collected reviews is 1,442 written by 865 unique IDs (including 63 reviews written by users with no IDs). The numbers of collected reviews for each domain are (arranged from largest to smallest): Sport (465 reviews), Religion (378 reviews), Economy (222 reviews), Food-Life style (222 reviews), and Technology (155 reviews). The average number of collected reviews per topic is 5.76. Our initial investigation on the collected Arabic reviews from the Maktoob Yahoo! website shows that around 64%, 19%, 6%, and 3% of these reviews are written in MSA, Egyptian dialect, Levantine dialect, and English respectively. Furthermore, some topics are related to more than one domain. Therefore, there are 7 subclasses: Arts, Economy, Education, Food, Politics, Religion, and Sport. This corpus can be used for studies that include supervised learning, as well as those that include creating sentiment lexicons for sentiment analysis studies. Corpora are created either automatically or manually. We prefer to use the manual approach to ensure that we get the highest possible quality. The creation of this corpus consumes a lot of time to collect and annotate different topics and reviews. The main contribution of this study is creating a multi-domain and multi-dialect Arabic dataset. The annotation is manually conducted to guarantee the best results. This corpus can be downloaded from https://drive.google.com/file/d/0B1847AXYiV_geFVneVRVdG9fQVk/view?usp=sharing.

The remainder of this paper is organized as follow. The next section (Section 2) exhibits related works to the creation of corpora. Section 3 presents our proposed standard Arabic sentiment analysis corpus with a highlight on the merits and deficiencies of this corpus. Section 4 shows a preliminary discussion of the collected corpus and the results of the analysis performed on it. Section 5 presents concluding remarks about this paper and discussion of future plans to enlarge this corpus.

2. Related Work

As mentioned in the previous section, sentiment corpora are created either automatically or manually. Our proposed corpus is created manually to guarantee the highest possible quality. This section presents some of the studies that are closely related to this one.

Sarmento et al. designed a rule-based system supported by a sentiment lexicon to automatically build a corpus for sentiment analysis. They focused on comments posted on an online newspaper about political entities. The experiments they conducted revealed that negative comments are relatively easier to recognize than positive ones due to irony and polarity inversion and shifting [6]. Such challenges motivated researchers to build a number of specialized corpora. For example, Bosco et al. [7] constructed a corpus to deal with irony in Italian. Zhang et al. [8] constructed a corpus to deal with polarity shifting in English.

Pak and Paroubek showed in their study how to collect a corpus for emotion analysis from Twitter. They collected a corpus of 300,000 tweets in English evenly distributed among three classes: the class of positive emotions such as happiness and joy, the class of negative emotions such as sadness and anger and the class of objective text expressing no opinion. The authors performed linguistic analysis of the constructed corpus and made some interesting remarks such as the one about the strong emotional indication of some Part-of-Speech (POS) tags [9].

In a very impressive work, Ptaszynski et al. took the largest corpus of Japanese blogs consisting of five billion words and designed a system to automatically annotate it for affect analysis. In addition to dealing with a massively sized corpus, the proposed system worked on both word-level and sentence-level, dealt with subjectivity and considered an extended set of emotion classes (not just positive/negative) for the purpose of affect analysis [10]. Another work on the Japanese language is that of Shiramatsu et al. [11]. The authors developed Social Opinions and Concerns for Ideal Argumentation (SOCIA) for the purpose of concern assessment in Japanese regional communities. To facilitate public involvement in such a task, the authors developed O2 that based on Linked Open Data

(LOD) to facilitate Japanese public involvement in regional communities. O2 is an e-Participation web platform.

In addition to building and annotating corpora, the field of sentiment analysis benefits from building and annotating lexicons as lexicon based (unsupervised) and semi-supervised approaches to sentiment analysis are among the most studied approaches. Baccianella et al. [12] presented the third version of the SentiWordNet lexicon, which was constructed by annotating the synsets of the famous WordNet lexicon according to the sentiment they convey.

Rushdi-Saleh et al. [13, 14] presented in their study an Opinion Corpus for Arabic (OCA) composed of only 500 movie reviews written in Arabic. This work is considered one of the earliest works on Arabic sentiment analysis corpora. The polarities of these 500 Arabic reviews are distributed equally among the positive and negative classes. Therefore, their corpus consists of 250 positive Arabic reviews and 250 negative Arabic reviews. The corpus was constructed manually and then translated to English for comparison purposes. While the effort to construct the OCA corpus is considered pioneering due to the limited resources for Arabic sentiment analysis, it does suffer from certain drawbacks. It is relatively small in size, lacks any neutral or objective reviews, and is restricted in a domain, which motivated other researchers to build their own corpora.

AWATIF is a sentence level multi-genre corpus of (MSA) labelled for subjectivity and sentiment analysis (SSA) is presented by Abdul-Mageed and Diab [15]. They used two types of annotation guidelines: simple (SIMP) and linguistically-motivated and genre-nuanced (LG), to find out how annotators would accomplish the labelling of MSA sentences under three different conditions. The total number of MSA sentences 10,729 collected from three sources: 2,855 MSA sentences collected from multi-domain collection of news wire stories called ATB1V3, 5,342 MSA sentences collected from 30 Wikipedia Talk Pages (WTP), and 2,532 threaded conversations collected from seven Web forums (WFs). Our corpus is characterized by the inclusion of MSA and various Arabic dialects while AWATIF is limited to MSA.

One of the earliest works approaching Arabic sentiment analysis using an automatically constructed sentiment lexicon was that of Al-Ayyoub, BaniEssa and Alsmadi [16]. To test their approach, the authors collected a small dataset of 900 tweets evenly distributed among the positive, negative and neutral classes. Their results contradict that of [6] in the sense that their tool was more efficient in detecting positive tweets than it was with negative and neutral ones.

In his Master's thesis, Abdulla [17], manually collected two datasets: a large one consisting of more than 10,000 reviews and relatively smaller one consisting of 2,000 reviews. Earlier versions of these datasets were used in [18-22]. The smaller dataset consisted of tweets written mainly in MSA and Jordanian dialect. These tweets are distributed equally among the positive and negative classes. On the other hand, the larger dataset consisted of other dialects and languages as well as other classes such as negative and spam. This dataset was later used to automatically construct a sentiment lexicon for the Arabic language.

Different Arabic Sentiment Analysis studies use datasets and corpora of different sizes, and in most cases these are constructed by the researchers who conduct these studies. Khasawneh et al. [23] constructed a small dataset that consists of only 1,000 Arabic reviews collected from two social media websites (Facebook and Twitter). Two studies conducted by Al-Kabi et al. [24, 25] constructed a dataset that includes 1,080 Arabic reviews, and these reviews use MSA and colloquial Arabic. In a relevant work, Al-Kabi et al. [26] constructed a dataset consists of 4,050 Arabic, Emoticons, and English reviews. So it is larger and more diverse than the previous dataset.

Aly and Atiya presented, in their study [27], a dataset called Large-scale Arabic Book Review (LABR). The authors claimed that it is the largest Arabic sentiment analysis dataset, since it consists of over 63,257 Arabic reviews about 2,131 Arabic books. The reviews were collected from a website that uses a 5-star rating system. Each review provides such a rating which makes the automatic collection of the dataset a feasible task. Similar to the other works on Arabic corpora, this dataset lacks diversity in terms of the Arabic dialects included. Being a publicly available tool, it was easy for other researcher such as [28] to perform a more extensive set of experiments on the LABR dataset in order to improve the best known accuracy for it.

Another important work based on the LABR dataset was that of AL-Smadi et al. [29] in which the authors selected 1,513 reviews out of the LABR dataset and annotated them for aspect-based sentiment analysis (ABSA). The annotation of this dataset (called HAAD) was performed according to the SemEval2014 Task4 guidelines.[1] The authors provided baseline experiments on the resulting dataset. In a follow-up work, Obaidat et al. [30] showed how to get a higher accuracy than the baseline experiments of [29] using lexicon-based approaches. Finally, a recent paper [31] exploited ABSA in order to evaluate Arabic news affect on readers taking the Israel-Gaza

[1] http://alt.qcri.org/semeval2014/task4/

conflict of 2014 as a case study. The authors collected a large number of Facebook posts and comments, but ended up annotating only 2,265 posts due to the nuance difficulties in the ABSA annotation of such a tricky case study.

3. Standard Arabic Sentiment Analysis Corpus

This section is dedicated to present the merits and deficiencies of the collected corpus besides showing its contents and structure. The collected corpus is stored inside MS Access database that has five tables, where each table is dedicated to one of the five domains (classes): Economy, Food-Life style, Religion, Sport, and Technology. The researchers identified 62 spam/irrelevant reviews, for which they used the code IRR within the polarity field. Most of the spam reviews are ads trying to attract people to some businesses, and there are some spam reviews that try to provoke people to revolute against their rulers, for example, in the sports domain. We notice a good portion of users accuse the Maktoob Yahoo! website of helping to provoke conflicts in the Arab World.

The structure of the database tables used is shown in Figure 1. Figure 1 shows the design view of each of the five tables used to store this corpus. Each table has 16 columns (fields) as shown in Figure 1.

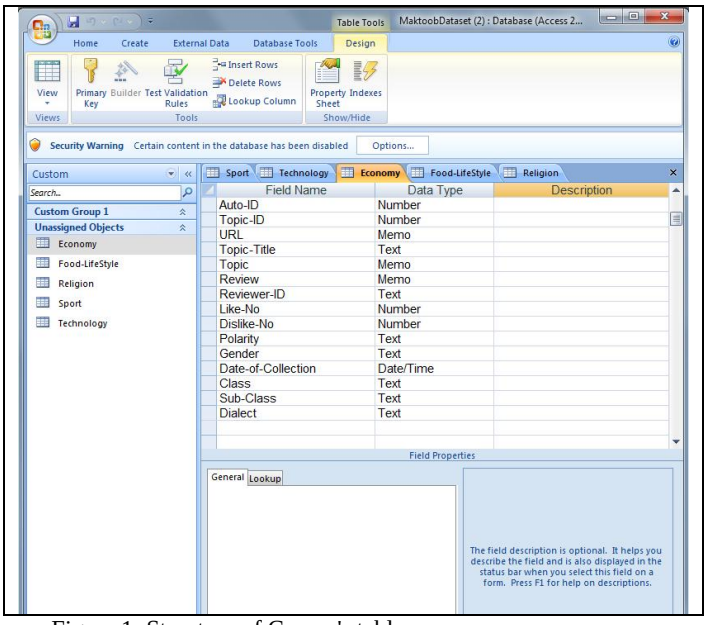

Figure 1. Structure of Corpus's tables

The ID of each review is identified within (Auto-ID) as shown in Figure 1, and each collected topic has an id (Topic-ID). The URL of each topic and reviews are stored within the (URL) field. Furthermore, there are columns for topic titles, topic content, reviews and reviewer IDs. The other columns (Like-No, Dislike-No, Polarity, Gender, Date-of-Collection, Class, Sub-Class, and Dialect) are used to store the number of likes, number of dislikes, polarity values, the expected or published gender of the reviewer, date of the collection of each review, domain of each topic, Sub domain of each topic, and the dialect of each review, respectively.

Table 1. Summary of the Size and Content of the Corpus

Domain Name	Number of Arabic Topics	Number of Reviews	Avg. No. of Reviews per Topic
Economy	50	222	4.44
Food-Life Style	50	222	4.44
Religion	50	378	7.56
Sport	50	465	9.3
Technology	50	155	3.1
Total	**250**	**1442**	**5.76**

The size and the content of the constructed corpus are shown in Table 1. Table 1 shows clearly that the 250 collected topics are distributed equally on the five domains (Economy, Food-Life style, Religion, Sport, and Technology). The authors have no control on the number of reviews about each of these topics. The average number of opinions per topic reflects the interest of people in that domain, so we deduce the sport domain is the most interested domain among the five selected domains, followed by religious topics, followed by two domains that have same average Economy and Food-Life style. Last and not least interested domain is the Technology domain.

4. Experiments and Results

In this section, a number of preliminary analysis and the results are shown. Table 2 shows the percentages of dialects used to express the reviews and comments about a different aspect of life in the Arab world.

Table 2 and Table 3 show preliminary analysis to the Arabic opinion mining corpus under consideration. These analyses are related to the topics published by the Maktoob Yahoo! website and use MSA to be understandable by all Arabs. Therefore, these two tables exhibit results about different authors within the 5 domains under consideration.

Table 2. Summary of the Size and Content of the Corpus

Domain Name	No. of Characters (no spaces)	No. of Characters (with spaces)	No. of Words	Avg. Words per Title
Economy	2172	2609	438	8.76
Food-Life Style	1507	1807	300	6.00
Religion	33375	3986	611	12.22
Sport	2217	2660	443	8.86
Technology	2104	2538	434	8.68
Total	**41375**	**13600**	**2226**	**8.9**

Table 2 shows the sizes of the titles of different published articles measured in characters and words. The shortest titles are used within Food-Life Style domain and longest titles are used within religion domain. The full list of the five domains sorted by ascending size is: Food-Life Style, Technology, Economy, Sport, and Religion. The average number of words per title (last column) in table 2 is computed by dividing each number in the number of words column by 50 since we have 50 topics in each domain.

Table 3. Summary of the Size of the Topic contents

Domain Name	No. of Characters (no spaces)	No. of Characters (with spaces)	No. of Words	Avg. Words per Topic
Economy	100390	120611	20218	404.36
Food-Life Style	31666	38170	6502	130.04
Religion	96769	116815	20045	400.9
Sport	35389	42435	7045	140.9
Technology	46948	56254	9305	186.1
Total	**311162**	**374285**	**63115**	**252.46**

Table 3 shows the size of the topic contents of different published articles measured in characters and words. The shortest articles are used within Food-Life Style domain and longest articles are used within an economy domain. The full list of the five domains sorted by ascending size of article's contents is: Food-Life Style, Sport, Technology, Religion, and Economy. The two sorted lists of domains by the sizes of article's titles and article's contents are different. Hence, there is no relation between the sizes of the titles and the sizes of article's contents. The average number of words per topic (last column) in table 3 is computed by dividing each number in the number of words column by 50 since we have 50 topics in each domain.

Table 4. Summary of the Size of the Review Contents

Domain Name	No. of Characters (no spaces)	No. of Characters (with spaces)	No. of Words	Avg. Words per Opinion
Economy	43158	52695	9526	42.91
Food-Life Style	20600	24994	4393	19.79
Religion	71031	86719	15687	41.50
Sport	53573	65144	11574	24.89
Technology	10362	12596	2234	14.41
Total	**198724**	**242148**	**43414**	**28.7**

Table 4 shows the size of the opinions about different published articles measured in characters and words. The shortest opinions are used within technology domain and longest articles are used within an economy domain. The full list of the five domains sorted by ascending size of opinion's contents is: Technology, Food-Life Style, Sport, Religion, and Economy. Hence we can deduce that social media users in the Arab world who interested in economy likes to write more than others who are interested in other domains, followed by users who are interested to write their opinions about religious articles. Table 1 and Table 4 show that Technology domain has the lowest interest by social media users in the Arab world. The average number of review's words (last column) in table 4 is computed by dividing each number in the number of words column by the corresponding number of opinions presented in table 1. Moreover, table 4 exhibits that average length of the whole collection which is equal to 28.7. The computations of the review lengths include spam reviews.

Table 5 is based on the 1442 reviews collected and shown in Table 1. The dialect of only 146 reviews can not be identified. Therefore, Table 5 is based on 1296 reviews only.

Table 5. Summary of Corpus's Languages

Language Name	Percentage
MSA	848 / 1296 = 65.43%
Egyptian	241 / 1296 = 18.59%
Levantine	74 / 1296=5.70%
English	40 / 1296 = 3.08%
Arabian Peninsula	32 / 1296 = 2.46%
Mesopotamian group	32 / 1296 = 2.46%
French	10 / 1296 = 0.771%
MSA + English	7 / 1296 = 0.54%
Arabizi	5 / 1296 = 0.385%
Maghrebi group	4 / 1296 = 0.308%
MSA + Emoticons	2 / 1296 = 0.154%
Egyptian + Emoticons	1 / 1296 = 0.077%

Table 5 shows clearly the distribution of the languages used by the users in the Arab world to express their comments and reviews. It is clear that the vast majority (64.081%) still use MSA since it is supposed to be the most comprehendible version of Arabic across the 27 countries in which it represents the official language. Egypt, the most populated country in the Arab world, reached a population of 94 million as announced by the country's official state information service [32]. So, it is no wonder that the second language used in this part of the world is the Egyptian dialect.

In this study, by Levantine, we mean Levantine Arabic, the dialect spoken in Jordan, Syria, Lebanon and Palestine. The third common dialect used is Levantine (6.155%), since it is used mainly in Levant region (Eastern Mediterranean) that includes four countries (Syria (20 million), Jordan (8 million), Lebanon (5 million), and Palestine (4.5 million) with 40 million residents and substantially large presence in the media in the Arab world [33]. The three top commonly used dialects in this region are followed by

English, Arabian Peninsula group, Mesopotamian group, French, MSA + English, Arabizi, Maghrebi group, MSA + Emoticons, and Egyptian + Emoticons, respectively as shown in Table 5.

Table 6 shows the distribution of the polarities among the five classes under consideration. The process of the manual annotation of this modest corpus consumes a lot of time, and in many cases we found that no two human professionals can assign the same polarities to a number of confusing reviews. Therefore, we decide to conduct a study about the manual annotation of Arabic reviews and comments. Not all reviews in this corpus can be annotated by human as positive, negative or neutral easily. Table 6 shows that majority of reviews in the economy and religion domains are negative, while the majority of reviews in Food-Life Style, Sport, and Technology domains are positive. The last row in table 6 shows the majority of collected reviews in this corpus are negative and the overall percentage of irregular and spam reviews is 4.3%.

Table 6. Summary of Polarity Distribution among the 5 Domains

Domain Name	No. of Positive Polarities	No. of Negative Polarities	No. of Neutral Polarities	No. of Irregular/Spam Polarities	Un Known
Economy	33 (14.86%)	130 (58.55%)	27 (12.16%)	32 (14.41%)	0 (0%)
Food-Life Style	84 (37.83%)	68 (30.63%)	31 (13.96%)	39 (17.56%)	0 (0%)
Religion	87 (23.01%)	247 (65.34%)	38 (10.05%)	6 (1.58%)	0 (0%)
Sport	216 (46.45%)	149 (32.04%)	23 (4.94%)	25 (5.37%)	52 (11.18%)
Technology	58 (37.41%)	36 (23.22%)	26 (16.77%)	25 (16.12%)	10 (6.45%)
Total	**478**	**630**	**145**	**127**	**62**
Average	33.15%	43.68%	10.05%	8.80%	4.3%

The distribution of Internet users according to their gender among the five domains is presented in table 7. The authors of this study could not identify the gender of 21% of all Maktoob Yahoo! who wrote the reviews we collected in this study. Therefore, two percentages are presented in table 7. The top percentages include the unknown gender values while the bottom percentages exclude the unknown gender values. Table 7 shows clearly that the males constitute the majority within all domains, the highest percentage of females was within Food-Life Style domain. The overall percentages show that female Maktoob Yahoo! users constitute one-fourth of male Maktoob Yahoo! users.

Table 7. Summary of Gender Distribution among the 5 Domains

Domain Name	Male	Female	Unknown Gender	Total
Economy	130 (59%) (86%)	21 (0.9%) (14%)	71 (32%)	222
Food-Life Style	117 (53%) (67%)	57 (0.25%) (33%)	48 (22%)	222
Religion	276 (73%) (84%)	52 (14%) (16%)	50 (13%)	378
Sport	292 (63%) (80%)	73 (16%) (20%)	100 (21%)	465
Technology	99 (64%) (82%)	22 (14%) (18%)	34 (22%)	155
Total	**914**	**225**	**303**	**1442**
Average	**63% 80%**	**16% 20%**	**21%**	

5. Conclusion and Future Work

This paper shows the creation of a flexible preliminary Arabic corpus for sentiment analysis and opinion mining for Arabic reviews and comments. This corpus characterizes by its flexibility, where each of its users can add, delete or revise it. Arabic comments and reviews within this corpus constitute mainly of Arabic comments (MSA and Colloquial (dialectal) Arabic). Our analysis shows that most of the users of Maktoob Yahoo! prefer to use MSA (65.43%), to enable different Arab visitors to comprehend their comments. Additionally this corpus has few comments and reviews that used English, French, Emoticons, etc. This corpus consists of 250 topics equally divided among the five classes (domains) we choose to include in this corpus. We plan to expand this corpus by adding more classes, and adding a new domain for political topics and reviews, and another one for Arts and stars. Moreover, we plan on adding more topics and reviews to each of the five domains within this corpus. Our future plan includes more detailed analysis of the collected topics and reviews, besides releasing the new corpus to be used freely by different researchers in the field of sentiment analysis, text mining, and data mining.

References

[1] A Guide to Arabic - 10 facts about the Arabic language. Available at: http://www.bbc.co.uk/languages/other/arabic/guide/facts.shtml [Online; accessed September-2015].

[2] Arabic language. Available at: http://www.arabicegypt.com/news/facts-about-the-arabic-language [Online; accessed September-2015].
[3] Arabic Language. Available at: http://en.wikipedia.org/wiki/Arabic_language [Online; accessed September-2015].
[4] The Arabic Language. Available at: http://www.vistawide.com/arabic/arabic.htm [Online; accessed September-2015].
[5] Semitic languages. Available at: http://en.wikipedia.org/wiki/Semitic_languages [Online; accessed September-2015].
[6] Sarmento L. , Carvalho P. , Silva M. J., and Oliveira E. d., "Automatic creation of a reference corpus for political opinion mining in user-generated content," in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09). ACM, New York, NY, USA, 2009, pp. 29-36.
[7] Bosco C., Patti V., Bolioli A., "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT," IEEE Intelligent Systems, vol. 2, no. 2, 2013, pp. 55 - 63.
[8] Zhang X., Li S., Zhou G., Zhao H., "Polarity Shifting: Corpus Construction and Analysis," in Proceedings of the 2011 International Conference on Asian Language Processing (IALP '11), 2011, pp. 272-275.
[9] Pak A., Paroubek P., "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010, pp. 1320-1326.
[10] Ptaszynski M., Rzepka R., Araki K., and Momouchi Y., "Automatically annotating a five-billion-word corpus of Japanese blogs for affect and sentiment analysis," in Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 89-98.
[11] Shiramatsu S., Hirata N., Swezey R. M. E., Sano H., Ozono T., and Shintani T., "Gathering Public Concerns from Web towards Building Corpus of Japanese Regional Concerns," in Proceedings of the IEEE 2012 IIAI International Conference on Advanced Applied Informatics, 2012, pp. 248 - 253.
[12] Baccianella S., Esuli A., and Sebastiani F., "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp. 2200-2204.
[13] Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-López L. A., and Perea-Ortega J. M. , "Bilingual Experiments with an Arabic-English Corpus for Opinion Mining," in Proceedings of Recent Advances in Natural Language Processing, 2011, pp. 740–745.
[14] Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-López L. A. , and Perea-Ortega J. M., "OCA: Opinion Corpus for Arabic," Journal of the Association for Information Science and Technology, vol. 62, no. 10, 2011, pp. 2045–2054.
[15] Abdul-Mageed M., and Diab M., "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, 2012.
[16] Al-Ayyoub M., Bani Essa S., Alsmadi I. "Lexicon-Based Sentiment Analysis of Arabic Tweets." International Journal of Social Network Mining (IJSNM), 2(2), 2015, pp.101–114.
[17] Abdulla N. A., 2014. Towards Building a Sentiment Analysis Tool for Colloquial and Modern Standard Arabic Reviews. Master's thesis. Computer Science Department, Jordan University of Science and Technology, Irbid, Jordan.
[18] Abdulla N.A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., "Arabic Sentiment Analysis: Lexicon-based and Corpus-based." In Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT 2013), 2013, 3-5 Dec. 2013, pp.1-6.
[19] Al-Kabi M.N., Abdulla N.A., Al-Ayyoub M., "An analytical study of Arabic sentiments: Maktoob case study." In Proceedings of 8th International Conference for Internet Technology and Secured Transactions (ICITST), 2013, 9-12 Dec. 2013, pp.89-94.
[20] Abdulla N.A., Al-Ayyoub M. and Al-Kabi M.N. "An extended analytical study of Arabic sentiments." International Journal of Big Data Intelligence (IJDBI), 1(2), 2014, pp.103–113.
[21] Abdulla N. A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., Al-Kabi, M. N., and Al-rifai S. "Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis." International Journal of Information Technology and Web Engineering (IJITWE), 9(3), 2014, pp. 55-71.
[22] Abdulla N.A., Majdalawi R., Mohammed S., Al-Ayyoub M., Al-Kabi M.N., "Automatic Lexicon Construction for Arabic Sentiment Analysis." In Proceedings of 2nd International Conference on

Future Internet of Things and Cloud (FiCloud 2014), 2014, 27-29 Aug. 2014, pp.547-552.
[23] Khasawneh R. T., Wahsheh H. A., Al-Kabi, M. N., Alsmadi I. M. "Sentiment Analysis of Arabic Social Media Content: A Comparative Study." The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), December 9-12, 2013, London, UK, 6 pages.
[24] Al-Kabi M., Gigieh A., Alsmadi I., Wahsheh H., and Haidar M. "An opinion analysis tool for colloquial and standard Arabic." In The Fourth International Conference on Information and Communication Systems (ICICS 2013), 6 pages. Irbid, Jordan, (April 23-25, 2013).
[25] Al-Kabi M. N., Gigieh A. H., Alsmadi I. M., Wahsheh H. A., Haidar M. M., "Opinion Mining and Analysis for Arabic Language." International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher, 5(5), 2014, pp. 181-195.
[26] Al-Kabi M., Al-Qudah N. M., Alsmadi I., Dabour M., Wahsheh H., "Arabic / English Sentiment Analysis: An Empirical Study." In The Fourth International Conference on Information and Communication Systems (ICICS 2013), 6 pages. Irbid, Jordan, (April 23-25, 2013).
[27] Aly M., Atiya A., "LABR: A Large Scale Arabic Book Reviews Dataset," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, (Volume 2: Short Papers), 2013, pp. 494-498.
[28] Al Shboul B., Al-Ayyoub M., Jararweh Y. "Multi-Way Sentiment Classification of Arabic Reviews." In the Sixth International Conference on Information and Communication Systems (ICICS 2015). Amman, Jordan, (April, 2015).
[29] AL-Smadi M., Qawasmeh O., Talafha B., Quwaider M. "Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis." In Proceedings of 3rd International Conference on Future Internet of Things and Cloud (FiCloud 2015), 2015, Aug. 2015.
[30] Obaidat I., Mohawesh R., Al-Ayyoub M., Al-Smadi M., Jararweh Y. "Enhancing the Determination of Aspect Categories and Their Polarities in Arabic Reviews Using Lexicon-Based Approaches." In Proceedings of the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT 2015), 2015, Dec. 2015.
[31] Al-Smadi M., Al-Sarhan H., Al-Ayyoub M., Jararweh Y., Benkhelifa E. "Using Aspect-Based Sentiment Analysis to Evaluate Arabic News Affect on Readers." In Proceedings of the 8th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2015), 2015, Dec. 2015.
[32] Population. Available at: http://www.sis.gov.eg/En/Templates/Articles/tmpArticles.aspx?CatID=19#.VBRHYPmSwmA [Online; accessed September-2015].
[33] Levant, http://en.wikipedia.org/wiki/Levant [Online; accessed September-2015].