Corpora for sentiment analysis of Arabic text in social media

ITANI, Maher, ROAST, Chris <http://orcid.org/0000-0002-6931-6252> and AL-KHAYATT, Samir

Citation:

ITANI, Maher, ROAST, Chris and AL-KHAYATT, Samir (2017). Corpora for sentiment analysis of Arabic text in social media. In: 8th International Conference on Information and Communication Systems (ICICS). IEEE, 64-69. [Book Section]

Corpora For Sentiment Analysis Of Arabic Text In Social Media

Maher Itani, Chris Roast, and Samir Al-Khayatt
Department of Computing
Communication and Computing Research Centre
Sheffield Hallam University
Sheffield, England

***Abstract*— Different Natural Language Processing (NLP) applications such as text categorization, machine translation, etc., need annotated corpora to check quality and performance. Similarly, sentiment analysis requires annotated corpora to test the performance of classifiers. Manual annotation performed by native speakers is used as a benchmark test to measure how accurate a classifier is. In this paper we summarise currently available Arabic corpora and describe work in progress to build, annotate, and use Arabic corpora consisting of Facebook (FB) posts. The distinctive nature of thesecorpora is that it is based on posts written in Dialectal Arabic (DA) not following specific grammatical or spelling standards. The corpora are annotated with five labels (positive, negative, dual, neutral, and spam). In addition to building the corpus, the paper illustrates how manual tagging can be used to extract opinionated words and phrases to be used in a lexicon-based classifier.**

Keywords—sentiment analysis, corpora, Arabic language, social media.

I. Introduction

NLP applications operate mainly on textual data. Different applications such as text categorization, machine translation and sentiment analysis require a corpus for training, testing and validation. A corpus is a large set of text built using different sources, and is often has metadata (such as: labels andPart of Speech (POS) tags) associated with it or to any of its components words, phrases, sentences, documents, etc.. Such corpora are used by various kinds of classifiers [1] such as Naïve Bayes (NB), decision tree (DT), Support Vector Machines (SVM), k-nearest neighbors (kNN), etc.. Classifiers using annotated corpora are designed and used for a variety of purposes, such as predicting movie sales, question answering, and other applications [2-10]. However, to test such classifiers, and in a supervised learning context, the classes of the text used to train and test the classifier is required. The class of each record is specified by 'human classifiers' - native speakers of the language who read the text and label or mark it according to a predetermined set of rules[11]. Such rules require that the human annotators draw upon their own, often tacit, native linguistic knowledge. Reliability can be assessed based upon agreement among human annotators, and measured in terms of Inter Annotator Agreement (IAA).

In the sentiment analysis context, classifiers aim to identify whether given posts are: positive, negative, neutral, etc. Hence the training of such classifiers is dependent upon reliable examples of this data, as corpora. Textual posts such as movie or product reviews available online are good examples of such opinion rich sources. However, the huge number of online posts makes the manual extraction and classification of the opinions embedded in these posts an infeasible task. In addition, public sentiment posts are generally "noisy" – they use informal language and dialects with less regard for correct spelling and grammatical rules. In the remainder of this paper we refer to such informal uses of Arabic as Dialectal Arabic (DA).

Research related to building corpora is limited for the Arabic language when compared with the English language. Authors in [9, 10, 11] attempt to partially fill this gap. Arabic resources become scarcer when we consider the sentiment classification of DA text such as that found in social media. In this work we focus on this problem by describing an approach to building such corpora by developing annotated Facebook corpora.

Below, section 2 provides a brief overview about the Arabic language and its characteristics, section 3 summarises currently available corpora. Section 4 describes our data collection process. Section 5 summarises the preprocessing approach applied on collected data. We describe different actions done during manual tagging in section 6, section 7 summarises characteristics of the built corpora and we conclude in section 8.

II. The Arabic language

In this section we summarise the main features of the Arabic language and its characteristics to clarify the focus of this research and associated data. The Arabic language is considered amidst top six major languages of the world [9, 10, 12]. The number of native speakers exceeds 200 million and it is the formal language used ? in over twenty countries.

There are three different forms of Arabic language?[10]: Classical Arabic; Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Classical Arabic is the language of Qur'an, the holy book of Islam, one of the world's major religions. Modern Standard Arabic (MSA) is the dialect used in education, books, television, newspapers, and in conversation among educated Arabs who have different local dialect. Local dialects (also known as colloquial Arabic) exist based on geographical location and country - even within the same country the dialect may vary in different areas.

The local dialects can be roughly divided as follows [13]:

- Dialects of Iraq and Kuwait
- Dialects of Sudan and Egypt.
- Dialects of Lebanon, Syria, Jordan, and Palestine.
- Dialects of the Gulf (Iraq, KSA, UAE, Kuwait, Qatar, Bahrain, and Yemen.).
- Dialects of Libya, Tunisia, Algeria, and Morocco.

Our work focuses on local dialects as these are commonly used by Arab users on social media. The lack of differentiation between specific dialects also reflects the often de-localised nature of posts to social media.

The Arabic alphabet consists of 28 letters containing both long vowels and consonants. Arabic text has right to left alignment. Unlike English, the shape of Arabic letters are not fixed, they change according to their location in the word. Moreover, short vowels, or diacritics, exist for Arabic language. These diacritics change the meaning of the word and its pronunciation, for example: كَتَبَ (katabah) means “wrote” whereas كُتُبْ (kutub) means “books”. Although the same letters are used in both words, the diacritics significantly change the meaning, grammar and pronunciation.

In addition to the complexity of diacritics and the different ways a letter can be written in Arabic, the Arabic language includes is morphologically complex.

Morphology: In addition to having different morphological features for genders such as شرب (Shariba, which means “he drank”) and شربت (Sharibat, which means “she drank”), the Arabic language also has specific morphology different than regular singular and plural and is used specifically when the word (whether it was verb, subject, adjective, etc.) is referring to two, example ولد (“walad”, which means “a boy”), ولدان (“waladan”, which means “two boys”), and أولاد (“awlad”, which means “more than two boys”).

Pronunciation: Some pronunciations do not exist in English and some other European languages such as "Gh" as in "Gharb" غرب for 'West' or "Ghareeb" غريب for 'strange') also "Kh" (as in Sheikh شيخ) and finally "Dh" (as in "Dhuhur" ظهر for afternoon or "Dhil" ظل for shadow) which can't be easily pronounced in other languages.

The characteristics mentioned above make developing resources for Arabic language such as corpora, and tools such as classifiers a more complex task when compared to the English language.

III. Current Corpora

Corpora are normally built from a variety of representative sources, governed by their intended use. For instance, an application that will be used to classify sentiment of social media posts will not benefit from a corpus of poems or a corpus of scientific articles, but would benefit more from a corpus consisting of social media posts. Existing research shows a number of different sources and approaches to developing corpora:

1. Corpora built by mining data from databases [14-19].
2. Corpora manually built by treating written text [20, 21].
3. Corpora built by downloading and processing webpages [22, 23].
4. Corpora built by downloading social media posts [24, 25].
5. Corpora built by downloading subsets of different corpora [26-28].
6. Corpora build from combinations of the above [29, 30].
7. Corpora derived from spoken language [31].

There are many Arabic corpora available for text categorisation. These include: the publicly available Quranic corpus [32] that consists of one text file that includes syntactic and morphological annotation of the Quran; and,a set of Arabic corpora [33] collected from online Arabic websites (mainly newswire) and categorised per topic (science, sports, etc.). Unfortunately, these are of little help when we consider sentiment analysis because they are not opinionated and are not sentiment tagged. Hence, despite providing the morphological and grammatical features of the words in the corpora, there is no sentiment information present.

One way to compensate for the lack of sentiment tagged corpora is to translate existing tagged corpora found in other languages. For instance, authors in [34] used movie reviews written in English after translating them and modifying the sentence structure to fit the classifier used.

One frequently used English corpus is the movie review corpus constructed in [35] that contains 1000 positive reviews and 1000 negative reviews, same corpus was used in [36, 37]. Reference [38] illustrates how a crawler can be used to download web documents for languages with no NLP resources. Also available are web resources [39, 40] that provide illustrations of how to build a corpus. Concerning Arabic language, authors in [41] describe the compilation stage related to building an Arabic corpus with different levels of analysis: syntactic, semantic, morphological, and lexical. First, authors in [42] explain the need for such a corpus in grammar, semantics, lexicography and NLP and raise important questions that can guide researches when analysing a corpus. CLARA [43] studies MSA texts collected from periodicals, books and miscellaneous texts. Al-Hayat Online Newspaper [44] and An-Nahar Online Newspaper [45] are two other Arabic corpora that cover different topics where texts are collected from online versions of the two newspapers. Authors also highlight the areas where progress is still limited concerning tools development such as translators, tokenisers?, POS taggers, stemmer, vocalisers?, and word disambiguation resolvers. Table 1 lists other corpora mentioned in the literature of sentiment analysis of Arabic text. Although some corpora exist for text domain classification (arts, news, etc.), we are not aware of any annotated to support sentiment analysis. Our work aims to help address this.

TABLE I. ARABIC CORPORA

Paper	Source
[1]	Forums of different domains
[14, 24, 46 - 49]	Social media
[34]	Movie Reviews
[50]	Written Documents
[51-53]	Web Pages
[54]	Treebanks

Although our source of data is similar to those mentioned in [14, 24, 46 - 49], our annotation is similar to that of [11] but adds a 'spam' class. This gives the five classes: negative, positive, spam, dual, and neutral.

IV. DATA COLLECTION

Facebook was chosen for data collection since it is the social media with the biggest number of users, more than 1.4 billion users [55], it is the one preferred by Arabs. In addition, it allows posts of larger sizes when compared with other social media such as Twitter whose post size is limited to 140 characters [56]. Corpora are either built using crawlers or collected manually. Although crawlers have the advantage of collecting large numbers of posts, they do require preprocessing to remove unwanted data [9]. Any data collection from online sources is also subject to the ethical and legal constraints governing re-use of the data for research. In the case of Facebook not only might crawling not generate clean data, it is also excluded by Facebook's terms and conditions.

Hence, we developed our corpus by manually copying posts in DA from Facebook groups. The posts consist of textual data posted by users as comments on posts written by the pages' administrators. The size of posts ranged from one word to a paragraph containing many [give a specific number] sentences.

Two corpora were built for two different domains: news, and arts. The news corpus (NC) consists of 1000 posts collected from “Al Arabiyya” News Facebook page [57] and the arts corpus (AC) consists of 1000 posts collected from “The Voice” Facebook page [58]. For example, authors of posts were anonymised and the copyright conditions upon Facebook groups’ posts were verified. The limit of 1000 posts in each corpora was kept to for practical reasons. However the posts were also filtered during manual tagging to limit ambiguous cases.

In keeping with ethical requirements to limit the identification of individuals, all posts were copied, excluding individual’s Facebook identities.

Although DA includes different dialects, reflecting the non-localised nature of social media and its contributors, no attempt was made to differentiate dialects. Hence, the Facebook posts were treated as a reflection of the aggregate DA evident in social media. Out of interest, a later assessment of the posts indicated only 5% could be associated with a specific dialect with the remaining 95% being common to all dialects.

V. PREPROCESSING

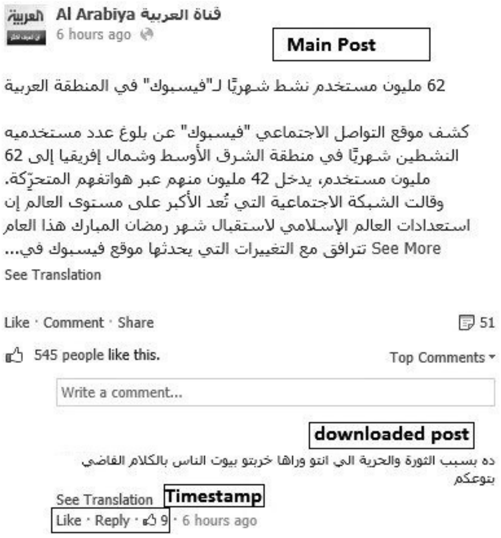
After data collection, posts were preprocessed in three different stages:

Removing redundancies: online users tend to post the same text more than one time in the same thread, either to show passion towards an object (like cheering for an artist), to express hatred or other negative emotions (using curse words and offensive language), or to spread a spam, i.e., to post a hyperlink referring to another website or Facebook page. So since the corpus will be used for text categorisation or sentiment analysis, posts in a corpus should be unique since the classifier will not gain extra knowledge from redundant posts.

Removing time stamps: each post has a time stamp that mentions when the post was written. For instance, in Facebook, this stamp is relative to the time the post is being seen. A time stamp may say the post has been posted two minutes ago. Timestamps are of no significance and therefore we removed them from collected posts.

Removing Likes: A "like" in Facebook terminology is a link that can be clicked by users to show that users agree with what has been posted regardless of its sentiment. In other words, a post with many likes is not necessarily a positive one.

Figure 1 shows a sample of the main post posted by the FB page administrators and the downloaded comment posted by an online user.



Al Arabiya قناة العربية

6 hours ago

62 مليون مستخدم نشط شهريًا لـ"فيسبوك" في المنطقة العربية

كشف موقع التواصل الاجتماعي "فيسبوك" عن بلوغ عدد مستخدميه
النشطين شهريًا في منطقة الشرق الأوسط وشمال إفريقيا إلى 62
مليون مستخدم، يدخل 42 مليون منهم عبر هواتفهم المتحرّكة.
وقالت الشبكة الاجتماعية التي تُعد الأكبر على مستوى العالم إن
استعدادات العالم الإسلامي لاستقبال شهر رمضان المبارك هذا العام
تترافق مع التغييرات التي يحدثها موقع فيسبوك في... See More

See Translation

Like · Comment · Share 51

545 people like this. Top Comments

Write a comment...

ده بسبب الثورة والحرية الي انتو وراها خربتو بيوت الناس بالكلام الفاضي
بتوعكم

See Translation

Like · Reply · 9 · 6 hours ago

Fig. 1. Sample of a downloaded post in context.

VI. Manual Tagging

Following preprocessing, four expert native Arabic speakers tagged the collected posts. Each expert human taggers could read, write, and speak MSA in addition to having a good understanding of other Arabic dialects. Any posts where the dialect details were considered to be unfamiliar the expert consulted native speakers of the relevant dialect, such as: Egyptian, Iraqi, and Tunisian. However, the occurrence of such posts was limited (less than 5%). Manual tagging employed the following rules:

Negative: if the post expresses a negative sentiment or feeling such as sadness, pessimism, hostility, etc.. For example:

للاسف كان ذلك على حساب يسرى

(Unfortunately that was on Yusra's expense)

Positive: if the posts express a positive sentiment or feeling such as enthusiasm, happiness, optimism, etc.. For example:

مبروك مراد

(Congratulations Murad)

Dual: if the post contains negative and positive sentiments regardless of the frequency of positive and negative patterns. For example:

مراد أخذ اللقب عن جدارة واستحقاق وموتوا بغيظكن ياحساد

(Murad deserves the title, die haters)

Spam: if the post is inviting users to join or “Like” a Facebook page. For example:

: السلام عليكم ممكن تنشرون هذا البيج

https://www.facebook.com/pages/%D9%85%D8...

(Greetings, can you spread this page)

Neutral: if the post is informative or expressing no sentiment, example:

مراد شو شعورك ان ربحت احلى صوت وشو شعورك ان خسرت ؟

(Murad how would you feel if you win or lose the competition?)

In addition to giving a tag for each post, the human tagger extracted the words and phrases from each post that were behind giving the post its class. These extract were used to form a lexicon for the relevant domain. In sentiment analysis, a lexicon is a dictionary of words and phrases an assigned polarity based on sentiment. For instance, “good” is positive, “bad” is negative”, and “car” is neutral. For example, in the post below:

مبروك مراد

(Congratulations Murad)

The word مبروك (which means congratulations) was added to the lexicon with a positive polarity.

In order to strengthen validity of the manual annotation, only posts on which all four annotators agreed were added to the corpora, others were discarded. Beforehand the initial Inter Annotator Agreement (IAA) was 97%. Downloading and processing posts continued until a target of 2000 annotated posts was achieved (with 100% IAA following the discards). The same validation rules were applied to the extracted lexicons. In addition to removing redundant entries.

VII. Corpora Charactersitcs

The output of the manual tagging was a set of posts each one belonging to one of the five classes and three sets of lexicons (negative, positive, and spam) forming a lexicon. Table 2 below shows the number of posts of each class in each corpus. Table 3 shows the number of lexicons of each set extracted from each corpus. In addition to their use in sentiment classification and despite the small size of the set, spam lexicon was introduced to fill what may be a gap in publicly available resources such as the ones in [59]. Our analysis shows that opinionated posts of different classes exist in different domains (like news and arts) and in close ratios.

TABLE II. Frequency of Posts of Each Class

	AC	NC	Total
Negative	224	230	454
Positive	233	236	469
Dual	151	161	312
Spam	197	193	390
Neutral	195	180	375
Total	1000	1000	2000

TABLE III. Frequency of Extracted Lexicon Entries

	AC	NC	Total
Negative	743	678	1421
Positive	684	573	1257
Spam	96	43	139
Total	1523	1294	2817

The numbers of extracted lexicon entries show that social media posts constitute a good source to build an opinionated lexicon: out of the 2817 extracted lexicons, 2509 were unique yielding in ~11% redundancy rate. Although the corpus is relatively small, the numbers show that on average; at least one new lexicon can be extracted from each post. As for the upper threshold to this, a much bigger corpus needs to be annotated to see show? at which number of posts, or corpus size, will no new lexicons appear. Moreover, the numbers of extracted lexicons per class indicate that the frequency of lexicons in a corpus is not strongly domain dependent. The lexicon commonality between domains is ~8% for Negative, ~14% for Positive and ~10% for Spam. Further analysis using bigger corpora is needed to see if this independence and commonality stands.

It is worth mentioning that extraction of the opinionated lexicons (bad, good, etc.) does not necessarily match the class of the post, i.e., negative lexicons may be extracted from a positive post and vice versa depending on the interpretation of the ? overall sentiment of the post. For instance a post would say: “The minister reported that the economic situation will get better”, although this post contains the positive lexicon “better”, the post itself is considered neutral since it is only reporting what someone else has said without expressing the user’s own opinion. Complete details concerning constructing the lexicon will appear in future work.

AC contains 12053 words with an average of 12 words per post whereas NC contains 8423 words with an average of 8 words per post. Results analysed in [11] show that ~25% of the corpus size is enough to classify the corpus if a lexicon-based classifier is used which means that the remaining 75% is insignificant to a lexicon based classifier as they are non-opinionated words and would not affect the classification.

VIII. CONCLUSION

The Arabic language is one of the top major languages in the world and is used in 22 countries. Different forms exist for Arabic language (classical, modern and dialectal). DA is used by internet user to post textual data on social media such as Facebook. On the other hand, sentiment analysis of Arabic text posted in social media needs tagged corpora. In this paper, we have provided an account of work in progress focusing upon how such corpora were built from Facebook posts. The ultimate aim behind the corpus is to be used by a lexicon to classify the sentiment of social media posts. Within the same process we provided an account of how to extract lexicons for a corpus. Although our work focused on Facebook, same process can be adopted when dealing with tweets (textual data posted on Twitter) and comments on other social media such as Instagram, MySpace, Linkedin, etc. However, any such processing would be subject to platform specific terms and conditions.

Designed (or constructed?) corpora were, subsequently, used in a lexicon based classifier. The performance of the classifier was determined upon comparing its results against the annotation done by the human taggers. Details of implementation are beyond the scope of this paper and will appear in future publications. Moreover, future work will include additional annotations besides working on increasing the size of the corpora. Finally, using the corpora, the classifier validated with performances ranging between 73% and 96%. The corpora are available upon contacting the authors.

REFERENCES

[1] El-Halees, A., 2011. Arabic opinion mining using combined classification approach.

[2] Jin, X., Li, Y., Mah, T. and Tong, J., 2007, August. Sensitive webpage classification for content advertising. In Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising (28-33). ACM.

[3] Mishne, G. and Glance, N.S., 2006, March. Predicting Movie Sales from Blogger Sentiment. In AAAI spring symposium: computational approaches to analyzing weblogs (155-158).

[4] Shikalgar, N.R. and Badgujar, D., 2013. Online Review Mining for forecasting sales. International Journal for research in Engineering & Technologies (IJRET) December.

[5] Tatemura, J., 2000, January. Virtual reviewers for collaborative exploration of movie reviews. In Proceedings of the 5th international conference on Intelligent user interfaces (272-275). ACM.

[6] Somasundaran, S., Wilson, T., Wiebe, J. and Stoyanov, V., 2007, March. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In ICWSM.

[7] Stoyanov, V., Cardie, C. and Wiebe, J., 2005, October. Multi-perspective question answering using the OpQA corpus. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (923-930). Association for Computational Linguistics.

[8] Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[9] Izwaini, S., 2003, March. Building specialised corpora for translation studies. In Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics.

[10] The Arabic Language. 2013. [Online] Available at www.al-bab.com [Accessed 17 July 2016]

[11] Itani, M.M., Zantout, R.N., Hamandi, L. and Elkabani, I., 2012, December. 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 (192-197). IEEE.

[12] Official Languages, Un.Org, United Nations, 2016. [Online] Available at: http://www.un.org/en/sections/about-un/official-languages/ [Accessed 17 July 2016]

[13] What is Spoken Arabic / the Arabic Dialects?, 2015, [Online] Available at: http://www.myeasyarabic.com/site/what_is_spoken_arabic.htm [Accessed 17 July 2016]

[14] Houngbo, H. and Mercer, R.E., 2014, June. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In Proceedings of the First Workshop on Argumentation Mining (19-23).

[15] Lita, L.V., Schlaikjer, A.H., Hong, W. and Nyberg, E., 2005, July. Qualitative dimensions in question answering: Extending the definitional QA task. In PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (Vol. 20, No. 4, 1616). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[16] Carlson, L., Marcu, D. and Okurowski, M.E., 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue (85-112). Springer Netherlands.

[17] Samy, D., Sandoval, A.M., Guirao, J.M. and Alfonseca, E., 2006. Building a Parallel Multilingual Corpus (Arabic-Spanish-English). In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC.

[18] Dukes, K. and Habash, N., 2010, May. Morphological Annotation of Quranic Arabic. In LREC.

[19] Rytting, C.A., Rodrigues, P., Buckwalter, T., Novak, V., Bills, A., Silbert, N.H. and Madgavkar, M., 2014. ArCADE: An Arabic Corpus of Auditory Dictation Errors. ACL 2014, 109.

[20] Megyesi, B.B., Hein, A.S. and Johanson, E.C., 2006, May. Building a swedish-turkish parallel corpus. In Proceedings of the Fifth International Conference on Language Resources and Evaluation.

[21] El-Haj, M. and Koulali, R., 2013. KALIMAT a multipurpose Arabic Corpus. In Second Workshop on Arabic Corpus Linguistics (WACL-2) (22-25).

[22] Arabic Linguistic Blog. 2014. [Online]. Available at: http://archive.is/Ep1a [Accessed 27 Dec 2015]

[23] King Saud University Corpus of Classical Arabic. 2012. [Online] Available at: http://ksucorpus.ksu.edu.sa/?p=43. [Accessed 27 Dec 2015]

[24] Hamouda, A.E.D.A. and El-taher, F.E.Z., 2013. Sentiment analyser for arabic comments system. Int. J. Adv. Comput. Sci. Appl, 4(3).

[25] Hamouda, S.B. and Akaichi, J., 2013. Social networks’ text mining for sentiment classification: The case of Facebook’statuses updates in the ‘Arabic Spring’era. International Journal Application or Innovation in Engineering and Management, 2(5), 470-478.

[26] Roberts, K., 2009, August. Building an annotated textual inference corpus for motion and space. In Proceedings of the 2009 Workshop on Applied Textual Inference (48-51). Association for Computational Linguistics.

[27] Bahloul, R.B., Elkarwi, M., Haddar, K. and Blache, P., 2014, September. Building an Arabic Linguistic Resource from a Treebank: The Case of Property Grammar. In International Conference on Text, Speech, and Dialogue (240-246). Springer International Publishing.

[28] Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W., 2004, September. The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, 466-467).

[29] Al-Sabbagh, R. and Girju, R., 2012, May. YADAC: Yet another Dialectal Arabic Corpus. In LREC (2882-2889).

[30] AbdelRaouf A, Higgins CA, Pridmore T, and Khalil M. 2010, Building a multi-modal Arabic corpus (MMAC). International Journal on Document Analysis and Recognition (IJDAR). 13(4):285-302.

[31] Oostdijk, N., 1999. Building a corpus of spoken Dutch. In CLIN.

[32] Quranic Arabic Corpus, 2011. [Online] Available at: http://corpus.quran.com/download/default.jsp. [Accessed 7 Dec 2015]

[33] Abdelali A, Cowie J, and Soliman H. 25th to 28th of July 2005, Building a modern standard Arabic corpus. In workshop on computational modeling of lexical acquisition. The split meeting. Croatia,

[34] Farra, N., Challita, E., Assi, R.A. and Hajj, H., 2010, December. Sentence-level and document-level sentiment mining for arabic texts. In 2010 IEEE International Conference on Data Mining Workshops (1114-1119). IEEE.

[35] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (79-86). Association for Computational Linguistics.

[36] Pang, B. and Lee, L., 2004, July. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (271). Association for Computational Linguistics.

[37] Pang, B. and Lee, L., 2005, June. 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 (115-124). Association for Computational Linguistics.

[38] Scannell, K.P., 2007, September. The Crúbadán Project: Corpus building for under-resourced languages. In Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop (Vol. 4, 5-15).

[39] Sinclair, J., 2005, [Online] Available at: Developing Linguistic Corpora: a Guide to Good Practice, http://icar.univ-lyon2.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf. [Accessed 17 July 2016]

[40] Burnard, L. (1998) ‘Using SGML for Linguistic Analysis: the case of the BNC‘ [online] Available from http://users.ox.ac.uk/~lou/wip/Boston/howto.htm [Accessed 27 Dec 2015]

[41] Alansary S, Nagi M, Adly N., 2007, Building an International Corpus of Arabic (ICA): progress of compilation stage. In7th international conference on language engineering, Cairo, Egypt, 5-6.

[42] Alansary S, Nagi M, and Adly N., 2008. Towards analyzing the international corpus of Arabic (ICA): Progress of morphological stage. In8th International Conference on Language Engineering, Egypt 2008 Dec.

[43] Zemánek, P., 2001, July. CLARA (Corpus Linguae Arabicae): An Overview. In Proceedings of ACL/EACL Workshop on Arabic Language

[44] Al-Hayat Online Newspaper, 2011. [Online Available at: http://www.alhayat.com/. [Accessed 27 Dec 2015]

[45] Annahar Online Newspaper. 2015. [Online]. Available at: http://www.annahar.com/ [Accessed 27 Dec 2015]

[46] Refaee, E. and Rieser, V., 2014, May. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In LREC (2268-2273).

[47] Hajjem, M., Trabelsi, M. and Latiri, C., 2013. Building comparable corpora from social networks. In BUCC, 7th Workshop on Building and Using Comparable Corpora, LREC, Reykjavik, Iceland.

[48] Akra D. and Jarrar M., 2014, Towards Building a Corpus for Palestinian Dialect.

[49] Al-Sulaiti L and Atwell ES., 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, 11(2), 135-171.

[50] Abdul-Mageed M, Diab MT, and Korayem M., 2011, Subjectivity and sentiment analysis of modern standard arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 587-591.

[51] Mustafa, M. and Suleman, H., 2011. Building a Multilingual and Mixed Arabic-English Corpus. In Proceedings Arabic Language Technology International Conference (ALTIC).

[52] Saad, M.K. and Ashour, W., 2010, November. Osac: Open source arabic corpora. In 6th ArchEng Int. Symposiums, EEECS (Vol. 10).

[53] Abdul-Mageed, M., Kübler, S., and Diab, M., 2012, ‘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, 19–28.

[54] Statistic Brain Research Institute, 2016. [Online]. Available at http://www.statisticbrain.com/facebook-statistics/ [Accessed 17 July 2016]

[55] Arab Social Media Report, 2013, [Online]. Available at http://www.arabsocialmediareport.com/Facebook/LineChart.aspx?&PriMenuID=18&CatID=24&mnu=Cat [Accessed 27 Dec 2015]

[56] Twitter Developer Documentation Overview, 2016, [Online] Available at: https://dev.twitter.com/overview/api/counting-characters. [Accessed 27 Dec 2015]

[57] Al-Arabiya Facebook Page, 2011, [Online] Available at: http://www.facebook.com/AlArabiya. [Accessed 27 Dec 2015]

[58] MBCTheVoice Facebook Page, 2011. [Online] Available at: http://www.facebook.com/MBCTheVoice [Accessed 27 Dec 2015]

[59] Salameh, M., Mohammad, S.M., Kiritchenko, S., 2016, ‘Arabic Sentiment Analysis and Cross-lingual Sentiment Resources ‘ [online] Available from http://saifmohammad.com/WebPages/ArabicSA.html [Accessed 1 Feb 2017]