I.J. Modern Education and Computer Science, **2019, 1, 33-38**
Published Online January 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.01.04

Twitter Benchmark Dataset for Arabic Sentiment Analysis

Donia Gamal, Marco Alfonse, El-Sayed M.El-Horbaty and Abdel-Badeeh M.Salem
Computer Science Department, Faculty of computer and information sciences, Ain Shams University, Cairo, Egypt
Email: donia.gamaleldin@cis.asu.edu.eg, marco@fcis.asu.edu.eg, {shorbaty, absalem}@cis.asu.edu.eg

Received: 19 September 2018; Accepted: 17 October 2018; Published: 08 January 2019

Abstract—Sentiment classification is the most rising research areas of sentiment analysis and text mining, especially with the massive amount of opinions available on social media. Recent results and efforts have demonstrated that there is no single strategy can mutually accomplish the best prediction performance on various datasets. There is a lack of existing researches to Arabic sentiment analysis compared to English sentiment analysis, because of the unique nature and difficulty of the Arabic language which leads to shortage in Arabic dataset used in sentiment analysis. An Arabic benchmark dataset is proposed in this paper for sentiment analysis showing the gathering methodology of the most recent tweets in different Arabic dialects. This dataset includes more than 151,000 different opinions in variant Arabic dialects which labeled into two balanced classes, namely, positive and negative. Different machine learning algorithms are applied on this dataset including the ridge regression which gives the highest accuracy of 99.90%.

Index Terms—Arabic Dialects, Arabic Sentiment Analysis, Arabic Opinion Mining, Twitter, Arabic Benchmark Dataset, Machine Learning.

I. Introduction

In the last few years, there has been a tremendous rise in the utilization of microblogging online platforms such as Twitter. The majority of individuals start to express their opinions and feelings on numerous things such as movies and many topics of events on the web [1]. Impelled by that growth, organizations and media companies are aggressively looking for ways to mine the data in Twitter for information about what individuals think and feel about their items, services, and products.

Sentiment Analysis (SA) - also called Opinion Mining (OM) - contains various subtasks, such as detection of subjectivity, classification of polarity sentiments, summarization of reviews, and classification of emotions [2]. OM can be viewed as a classification methodology that intends to decide whether a specific text is written to express an opinion in a positive or a negative way regarding an object (e.g., a topic, movie, product, person, or candidate). SA is typically performed utilizing one of two main approaches; a) lexicon-based approach, in which rules extracted from the linguistic study of a language are applied to the SA and b) Machine Learning (ML) approach which relies on the famous ML algorithms to solve SA as a classification task [3]. Current research focuses on building frameworks for English SA [4]; however, minor work has been done on different languages. In general, OM is carried out on textual documents in three different basic levels [5] which are document, sentence, and aspect levels. This research is focused on the document level [6].

There are two main principles that are shared between numerous social networking services; the short length of their text messages and language varieties. For instance, Facebook has a cutoff of 420 characters for posting a status [7] and Twitter has a limitation of up to 140 characters [8]. Likewise, language variations give an expansive variety of short forms, structures, and irregular words, particularly for youth generations. These two characteristics incite noteworthy data sparseness condition and hence affect the performance and efficiency of sentiment classifiers gained from such noisy and uproarious data.

In this paper, the construction of the benchmark dataset of Arabic Dialect Tweets along with the tools and algorithms utilized in this research are explained in detail.

The rest of the paper is organized as follows: Section II describes the steps of collecting and preprocessing of Arabic tweets. Section III shows the list of the machine learning algorithms applied to the collected dataset. The last section concludes the paper and illustrates the future work.

II. The Constructing of the Dataset

The construction of the dataset has many fundamental stages as shown in Fig 1. These stages are gathering the Arabic Dialect tweets dataset, preprocessing tweets and annotations which includes removing non-Arabic letters, tokenizing, removing stop words, removing repeating characters, removing URLs and user mentions, removing hashtags and retweets, removing diacritics, handling emoticons, normalizing Arabic analogous letters, labeling tweets and removing and handling skewness of data.

A. *Collecting Arabic Dialects Tweets Dataset*

The Arabic language is viewed as one of the top 10 main languages that are used on the web, however it is acknowledged as a poor content language, unlike English with very few numbers of web pages containing Arabic reviews and feedbacks [9]. Tweepy API - a free Application Program Interface (API) - is used to access Twitter data and statistics [10]. By setting the language to Arabic (Lang= AR), and the geolocation (the latitude and longitude) to the Middle East, Arabic tweets are now ready to be fetched. Different Arabic dialect phrases are used as keywords for querying twitter while collecting positive and negative tweets. More than 151,000 tweets from various topics are gathered. These tweets are written in Modern Standard Arabic (MSA) and Colloquial Arabic [11]. Table I depicts the details of the dataset.

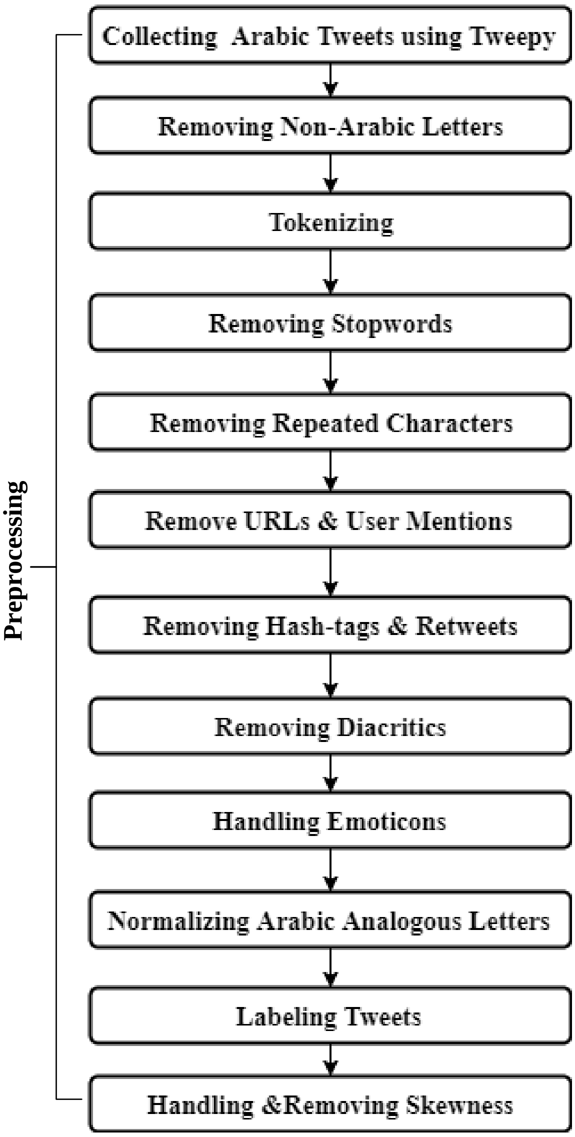

Preprocessing
Collecting Arabic Tweets using Tweepy
Removing Non-Arabic Letters
Tokenizing
Removing Stopwords
Removing Repeated Characters
Remove URLs & User Mentions
Removing Hash-tags & Retweets
Removing Diacritics
Handling Emoticons
Normalizing Arabic Analogous Letters
Labeling Tweets
Handling &Removing Skewness

Fig.1. The Twitter Benchmark Dataset Construction

Table I. Data Description

Label / Term	Positive	Negative	Total
Number of Tweets	75,774	75,774	151,548
Number of Words	82,576	127,760	210,336

This research is concerned with the Arabic language with different dialects, including the Egyptian colloquial. Large portions of internet users who compose and write in these languages are expressing their opinions, sentiment, and emotions with a sarcastic and mocking way. The sarcastic text is excluded because it gives a misleading meaning. Sarcasm is not so regularly used in client reviews, feedback, and criticism about items, products, and services, however, it is used very commonly in any political discussions, which make political sentiments incorrectly classified [12].

B. Preprocessing Tweets and Annotations

The tweets have been chosen manually to ensure that they hold only one opinion, not sarcastic or subjective. The chosen tweets are then passed to the process of removing all user-names, pictures, retweets, user mentions, hashtags, emoticons, URLs and all non-Arabic letters to be easily manipulated and dealt with.

Step 1: Removing Non-Arabic Letters

To remove the noise from the data, non-Arabic letters such as (!, -, *, &) are removed by iterating on all the tweet words.

Step 2: Tokenizing

Tokenization is only a segmentation process of the sentences. In this progression, Tweets will be tokenized or fragmented with the assistance of splitting text by spaces and punctuation. Then the tweet is checked for uniqueness, which ensures that sentence does not contain any repeating words.

Step 3: Removing Arabic Stop Words

A set of undesirable Arabic stop words are excluded to facilitate the data processing. 243 words are used for evacuating unneeded words (i.e. كما, من, الى, بعد, لهم, قد). Removing undesirable words is an essential task when mining unstructured data.

Step 4: Removing Repeated Characters

As the social network users write statuses and tweets in free text, ungrammatical and informal Arabic language, tweets may contain numerous syntax errors. These errors highly influence the mining process which makes it very difficult and troublesome. For example, one of the common mistakes, web users may repeat a character in a word like "جمييييييل" instead of "جميل", which means "Beautiful". This step solves these issues to repack words to its correct and right syntax.

Step 5: Removing Hashtags and Retweets

In this step, URL links (e.g. http://twitter.com), special words (e.g. RT which means retweet), and hashtags (e.g. "#WorldCup2018") are removed.

Step 6: Removing Diacritics

Diacritics are utilized as a part of the dialect, above the word's letters to change the pronunciation and in some cases the meaning of the words. However, in the utilized dialects in social media websites, diacritics are infrequently used, and in most of the cases, they are utilized only for decorations reasons. For example: “حَياة” which means life is replaced with “حياة” after removing diacritics.

Step 7: Handling Emoticons

Additionally, Twitter users utilize symbols, for example,": D" and ";)" to express their sentiment and emotions. These emoticons, also called emoji, express valuable information to the SA. Consequently, in order to detect the sentiment out of the emotions, they are labeled as well. For instance an emotion is tagged as happy if the utilized symbols are ":)", ": D" and" :')" and tagged as sad if the used symbols are ":(", ":'(" [13]. Emotions tags will influence the classification procedure as they hold a sentiment.

Step 8: Normalizing Arabic Analogous Letters

Users generally have a tendency to write similar Arabic letters based on their choice, particularly the Hamza (ء) when exists with the Alef (ا) letter can be maintained differently based on its position. Generally, we are not worried about its right spelling since we realize that the text context demonstrates its proposed meaning (e.g. we write انا instead of أنا). Moreover, the Taa' marbuota (ة), appears only at the end of wordsـة——)) and has kind of a similarity in shape to the Haa' marbuota letter (ه) at the end of the word as wellـه——)), web users regularly write the Haa' marbuota and the Taa' marbuota alternatively. In this way, every shape of the Alef with the Hamza is exchanged to a normal Alef, similarly, all Taa' marbuota to Haa' marbuota letter.

Step 9: Labeling Tweets

There are 4404 phrases [14] that are commonly used in expressing the sentiment, these phrases are selected and annotated manually as positive (+) and negative (-) sentiments. These phrases are used as keywords in the query for tweets, so the retrieved tweets were labeled based on these keywords and the frequency of positive/negative words appearing in the text as shown in Fig.2.

Table II shows an example for labeling a tweet.

Table II. Example of Labeling Tweet

Tweet	التفاؤل وقت الفشل ذكاء الثقة بالنفس وقت الياس قوة الاصرار رغم المعوقات نجاح بحد ذاته
Positive (6 words)	{'التفاءل', 'ذكاء ', 'الثقة ', 'قوة ', 'الاصرار', 'نجاح'}
Negative (3 words)	{'الفشل', 'الياس', 'المعوقات'}
Label	Positive Tweet

```
Input:   Preprocessed Tweets
Output:  Labeled Tweets
For each tweet in Retrieved Tweets:
    Initialize positive count to zero
    Initialize negative count to zero
    For each word in tweet:
        If word is in Positive-Phrases:
            Increment positive count
        Else If word is in Negative-Phrases:
            Increment negative count
    If positive count > negative count And (negative count/positive count) <= 0.5:
        Label tweet as positive
    Else :
        Label tweet as negative
```

Fig.2. Tweets Labeling Algorithm

Step 10: Removing and Handling Skewness

After performing all these normalization steps, all tweets are processed by removing duplicate tweets to guarantee the uniqueness of the dataset in the next phases.

At the point when the training dataset is imbalanced, then building helpful and efficient classification models can be a particularly challenging and difficult endeavor due to the dataset bias. Class imbalance presents an issue in the utilization of traditional classification algorithms as they attempt to build models with the objective of maximizing the accuracy of the classification [15]. To avoid these problems associated with class imbalance, 151,500 tweets are selected from the whole retrieved tweets which exceed 400,000 tweets. Then the selected tweets are annotated consisting of 75,774 positive tweets and 75,774 negative tweets.

III. Experiment and Evaluation

Machine Learning Algorithms (MLA) are applied in SC; it is tied with learning structures and includes different classification algorithms for text. Different supervised learning algorithms are applied on the proposed dataset in which the data is divided into training and testing data.

The Term Frequency -Inverse Document frequency (TF-IDF) is applied on the dataset as a feature extraction method [16]. Term frequency calculates the frequency of every token in the tweet. TF-IDF value demonstrates the significance of a token to a document within the tweets.

This work utilizes the Naive Bayes (NB), Adaptive Boosting (AdaBoost), Support Vector Machines (SVM), idge Regression (RR), and Maximum Entropy (ME) algorithms [17, 18] for classifying the dataset and comparing the results obtained using these algorithms to come up with an efficient algorithm that results in an accurate classification.

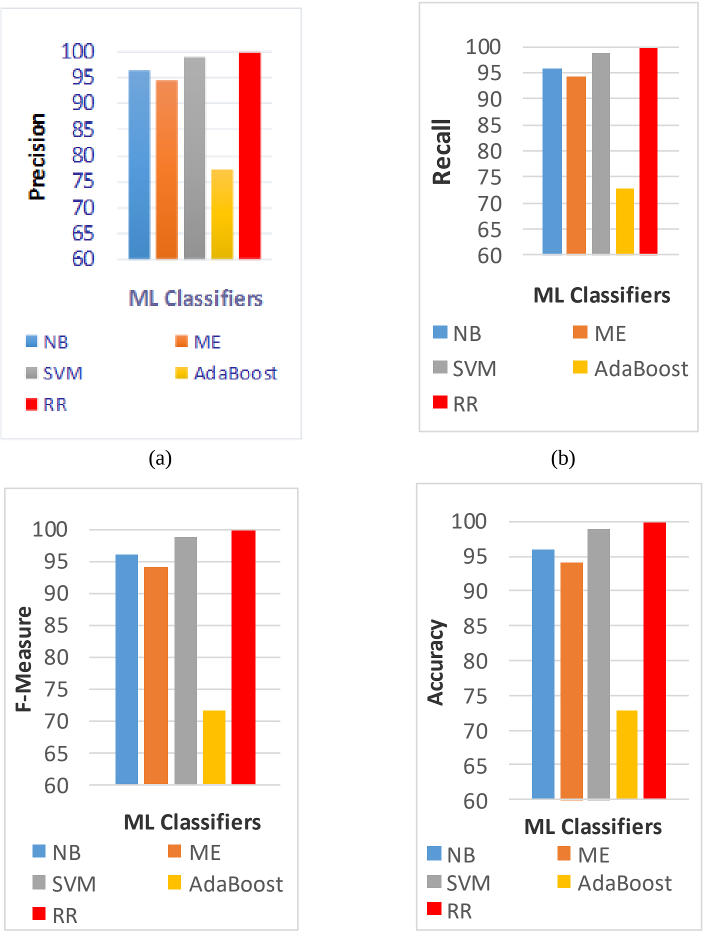
The experiments were conducted using Natural Language Tool Kit (NLTK) [19], Scikit-learn [20] in python 3.6 with a memory of 16GB RAM. Evaluation of these models is done using the cross validation process. One of the standard strategies is utilized, as a part of numerous MLA is k-fold cross validation [21] with k = 10. The measurements utilized for assessing the classification performance are accuracy, precision, recall and f-measure [22].

Table III indicates the results acquired from subsequent analyzing of various supervised learning algorithms using 10-Fold cross validation.

Table III. Performance of Machine Learning Algorithms

Rating Classifier	Precision	Recall	F-measure	Accuracy
NB	96.22%	95.98%	95.97%	95.98%
AdaBoost	77.27%	72.82%	71.66	72.83%
SVM	98.95%	98.94%	98.94%	98.94%
ME	94.48%	94.22%	94.21%	94.22%
RR	99.90%	99.90%	99.90%	99.90%

The RR gives a better accuracy than other MLA for SC on the twitter dataset. Fig. 3 visualizes the performance of NB, AdaBoost, ME, RR and SVM in terms of precision, recall, F-measure and accuracy.



(a) (b)

(c) (d)

Fig.3. Sentiment Classification Dataset Results. a) Evaluating precision with different MLA. b) Evaluating recall with different MLA. c) Evaluating f-measure with different MLA. d) Evaluating accuracy with different MLA.

From fig. 3, it's noticed that the RR classifier outperforms NB, SVM, Ada-Boost, and ME on the dataset. For all metrics values, Ada-Boost gives the lowest value. The SVM and NB give also high values for all metrics in range from 95% to 99%.

IV. Conclusions and Future Work

In this paper, a sentiment dataset of modern standard Arabic and Egyptian dialects on twitter is offered for SA classification tasks. The steps that are performed to construct the first release of the dataset using many idioms/proverbs for querying twitter are reported. Moreover, the algorithms applied for SC on the dataset achieved 99.90% accuracy using RR and TF-IDF.

For future work, this line of research and studies will be kept up for enhancing and upgrading the dataset with further enlargement by adding more annotated opinion idioms, extra proverbs and old wisdoms. This research focused on Arabic; however, the proposed method of constructing dataset can be utilized with any other languages.

References

[1] Uysal, Alper Kursat, and Yi Lu Murphey. "Sentiment classification: Feature selection based approaches versus deep learning.", *Proceedings of IEEE International Conference Computer and Information Technology (CIT)*, pp. 23-30. IEEE, 2017.

[2] Abdul-Mageed Muhammad, and Mona T. Diab., AWATIF: A Multi-Genre corpus for modern standard Arabic subjectivity and sentiment analysis, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), pp. 3907-3914, 2012.

[3] Medhat Walaa, Ahmed Hassan, and Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5, no. 4, pp. 1093-1113, 2014.

[4] Korayem Mohammed, David Crandall, and Muhammad Abdul-Mageed, Subjectivity and sentiment analysis of arabic: A survey, Proceedings of International conference on advanced machine learning technologies and applications, pp. 128-139. Springer, Berlin, Heidelberg, 2012.

[5] Dehkharghani, Rahim, Berrin Yanikoglu, Yucel Saygin, and Kemal Oflazer. "Sentiment analysis in Turkish at different granularity levels.", *International Journal of Natural Language Engineering*, vol. 23, no. 4, pp. 535-559, 2017.

[6] Märkle-Huß, Joscha, Stefan Feuerriegel, and Helmut Prendinger. "Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures." *In Proceedings of the 50th Hawaii International Conference on System Sciences*, pp. 1142-1151. HICSS, 2017.

[7] Pudaruth, Sameerchand, Sharmila Moheeputh, Narmeen Permessur, and Adeelah Chamroo. "Sentiment Analysis from Facebook Comments using Automatic Coding in NVivo 11.", *International Journal of Advances in Distributed Computing and Artificial Intelligence Journal (ADCAIJ)*, vol. 7, no. 1,pp. 41-48, 2018.

[8] Trupthi, M., Suresh Pabboju, and G. Narasimha. "Sentiment analysis on twitter using streaming API." Proceedings of IEEE 7th International In Advance Computing Conference (IACC), pp. 915-919. IEEE, 2017.

[9] Elhawary Mohamed, and Mohamed Elfeky, Mining Arabic business reviews, Proceedings of the IEEE International Conference on Data Mining Workshops, IEEE Computer Society, pp. 1108-1113, 2010.

[10] Roesslein, Joshua. "tweepy Documentation", http://docs.tweepy.org/en/v3.5.0/, 2009 [last accessed July 2018]

[11] https://www.arabacademy.com[last accessed July 2018]

[12] Liu Bing. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5, no. 1 pp. 1-167, 2012.

[13] Ahmed Soha, Michel Pasquier, and Ghassan Qadah, Key issues in conducting sentiment analysis on Arabic social media text, Proceedings of 9th International Conference on Innovations in Information Technology (IIT), pp. 72-77. IEEE, 2013.

[14] Aly Mohamed, and Amir Atiya. Labr: A large scale arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics , vol. 2, pp. 494-498. 2013.

[15] Seiffert Chris, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano, Building Useful Models from Imbalanced Data with Sampling and Boosting, In Proceedings of Florida Artificial Intelligence Research Society (FLAIRS) conference, pp. 306-311. 2008.

[16] Wawre Suchita V., and Sachin N. Deshmukh, Sentiment classification using machine learning techniques, International Journal of Science and Research (IJSR) 5, no. 4, pp. 819-821, 2016

[17] Zhao Jun, Kang Liu, and Liheng Xu, Sentiment analysis: mining opinions, sentiments, and emotions, International Journal of Computational Linguistics, Vol. 42, No. 3, pp. 595-598, 2016.

[18] Pozzi Federico Alberto, Elisabetta Fersini, Enza Messina, and Bing Liu. Sentiment analysis in social networks. Morgan Kaufmann, 2016.

[19] https://www.nltk.org/ [last access July 2018]

[20] http://scikit-learn.org/ [last access July 2018]

[21] Pang Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the Association for Computational Linguistics (ACL-02) conference on Empirical methods in natural language processing, Vol. 10, pp. 79-86. 2002.

[22] Kouloumpis Efthymios, Theresa Wilson, and Johanna D. Moore, Twitter sentiment analysis: The good the bad and the omg!, *In Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM 11)*, no. 538-541, pp. 538-541, 2011.

Authors' Profiles

Donia Gamal is a teaching and research assistant in Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt and received the B.Sc. degree with very good with honors in 2013 and M.SC degree in 2019. Her research interests: Sentiment Analysis, Machine Learning, and Artificial

Intelligence. She has 3 publications in refereed international journals and conferences.

Marco Alfonse is a Lecturer at the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt. He got Ph.D. of Computer Science since August 2014, University of Ain Shams. His research interests: Semantic Web, Ontological Engineering, Machine Learning, Medical Informatics, and Artificial Intelligence. He has 34 publications in refereed international journals and conferences.

El-Sayed M. El-Horbaty, He received his Ph.D. in Computer science from London University, U.K (1985)., his M.Sc. (1978) and B.Sc (1974) in Mathematics From Ain Shams University, Egypt. His work experience includes 44 years as an in Egypt (Ain Shams University), Qatar(Qatar University) and Emirates (Emirates University, Ajman University and ADU University). He worked as Deputy Dean of the faculty of IT, Ajman University (2002-2008). He is working as a Vice Dean of the faculty of Computer & Information Sciences, Ain Shams University (2010-2017). Prof. El-Horbaty is current areas of research are parallel and distributed computing, combinatorial optimization, image processing, cloud computing, e-health and mobile cloud computing. His work appeared in journals such as Parallel Computing, International Journal of Mobile Network Design and Innovation, International Journal of bio-Medical Informatics and e-health, and International journal of Computers and Applications (IJCA), Applied Mathematics and Computation, and International Review on Computers and software. Also he has been involved in more than 26 conferences.

Abdel-Badeeh M Salem is a Professor of Computer Science since 1989 at Ain Shams University, Egypt. His research includes intelligent computing, knowledge-based systems, biomedical informatics, and intelligent e-learning. He has published around 250 papers in refereed journals and conferences. He has been involved in more than 400 Conferences and workshops as a Keynote Speaker, Scientific Program Committee, Organizer and Session Chair. He is a member of many national and international informatics associations.

How to cite this paper: Donia Gamal, Marco Alfonse, El-Sayed M.El-Horbaty, Abdel-Badeeh M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.1, pp. 33-38, 2019.DOI: 10.5815/ijmecs.2019.01.04