



Building Online Social Network Dataset for Arabic Text Classification

Ahmed Omar¹ , Tarek M. Mahmoud^{1,2},
and Tarek Abd-El-Hafeez¹ 

¹ Computer Science Department, Faculty of Science, Minia University,
EL-Minia, Egypt

{Ahmed.omar, d.tarek, tarek}@mu.edu.eg

² Canadian International College (CIC), Cairo, Egypt

Abstract. Social networking sites have spread widely in recent years, and through them, a large amount of data is shared in all its forms: text, photo, voice, and video. It also allows communication with users through different forms such as chat, comments, and Posts, and the most exchanged content is in the form of text data. These results a large volume of data displayed to each user. This encouraged and attracted the attention of researchers to make an effort to analyze and work on this large amount of data available for free on the online social networks, most efforts focus on Twitter and English data. Building dataset is the most time-consuming and the most important part of the text classification process. Despite the increase in the number of Arabic users and the increase in Arabic content on online social Networks (OSN), there is a scarcity in Arabic datasets collected from social networks for text classification purpose. So In this paper, Arabic social dataset was built to be used in text classification purpose. our dataset was gathered from Facebook, it consists of 25,000 posts were collected from different Facebook pages and were classified into ten categories, politics, economics, sport, religion, technology, TV, ads, foods, health, and porno. The dataset was assessed to ten Arabic local speakers and Facebook users to evaluate the validity of the dataset made. We used a RapidMiner tool to evaluate and compute the performance of our dataset. We obtained a classification accuracy of 95.12%.

Keywords: Text mining · Text classification · Online social network
Arabic text · Arabic dataset

1 Introduction

Research in social media has become a point of interest from many researchers because of the increasing field of online social networks in most platforms. Social Networks are nowadays the most familiar interactive media to communicate, share, and publish an unlimited amount of human life information. Communications mean the exchange of particular types of content, including text, photo, audio, and video data. Online Social Networks supply very little support to prevent unwanted data on user timeline. Sometimes the shared information may be vulgar or not wanted and it is inevitable to see it. Facebook, for example, gives users the ability to declare who is allowed to add

data to their walls. (i.e., friends, friends of friends, or defined groups of friends). In Facebook, no data checking for the contents happen and hence it is highly likely that offensive content gets posted without unchecking or filter no matter of the users [1].

Most of the exchanged data over the Social Networks are in the text format. Text mining is a technique made together with data mining, machine learning, and information retrieval. Text mining may also point out as text data analysis or data mining in which significant information can be retrieved from the text. To make text preprocessing or prerequisites, the phases are parsing, tokenization, normalization, etc. [2]. Text classification techniques will be used for automatically labeling a set of categories based on contents of each text data. The classification will be one of the following categories (politics, economics, sport, religion, technology, TV, ads, foods, health, and porno).

Online Social Networks (OSN) are used by different languages' speakers. It is not only used by English speakers. There are many users on OSN use other languages than English e.g. Arabic. Arabic is fourth one of the top ten languages on the internet in June 2017 [3]. The need and attention in classifying Arabic texts have increased recently, due to a lot of reasons: The Arabic language is very rich with contents, there are about 184 million Arab Internet users and a large percentage of them cannot read English [3]. In addition to, the online Arabic contents have grown quickly in the last decade, exceeding 3% of the entire online contents and is ranked the eighth in the whole internet content [4]. However, there is lack of language resources and text processing techniques for the Arabic language [5].

This paper presents a dataset collected from Arabic Facebook pages, the dataset contains 25,000 posts, collected automatic and manually labeled to 10 categories, and then we apply some text processing techniques which include, removing non-Arabic letters, removing word suffix and prefix, normalization, and transformation. The labeling phase includes two stages, in the first, we labeled each post to a specific class, then we ask 10 Facebook users who are Arabic native speakers to labeled the posts, and according to users' feedbacks, some posts classification are changed.

The next sections are organized as follows: Sect. 2 contains related works review. In Sect. 3 the data collection methodology is viewed. Section 4 contains results and evaluation followed by Sect. 5 which contains the conclusion.

2 Related Work

The dataset building is different in terms of the research purpose, such as Natural Language Processing (NLP) and Text Mining. The dataset differs also in the collect sources i.e. websites, social networks, news pages, and blogs. Also, datasets vary in size and language. Some datasets are available for free and some of them are available commercially. In recent years interest is begun to build datasets from online social networks, because of the large amounts of data available in it. There are no free standard datasets available for the Arabic text classification research, unlike English text classification, so researchers rely on collecting their dataset for each research point [4]. Few research efforts were done for Arabic datasets building.

Al-Kabi et al. [6] collect 4050 comments from social media such as Facebook, YouTube, Twitter, Digg, and Yahoo. These comments were used to build a dataset in

Arabic and English languages, SocialMention and Twenz tools were used to gather comments together with reviews in Arabic and English language. The dataset was classified to three classes only, political news, commercial, and academic. Three classification models were used to evaluate the dataset ((Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor algorithm (K-NN)), and the conducted results showed that the Naïve Bayes algorithm gave the best results for both SocialMention and Twenz tools with an accuracy of 66.2% and 45.3%, respectively.

Abdul-Mageed et al. [7] created annotated data comprising a four of datasets contain different dialects: the first dataset contains 2798 chat message collected randomly from of an Egyptian room chat session in Maktoob chat, the second dataset contains 3015 Arabic tweets collected from Twitter. The third dataset consists of 3008 sentences, was collected from 30 Talk Pages on Wikipedia. The fourth one comprises 3097 sentences Web forum collected from a larger pool of threaded conversations pertaining to different varieties of Arabic, the topics covered in this forum is religion or politics. The proposed system by Abdul-Mageed et al. gives the best accuracy with the first dataset, which achieves 84.65% in subjectivity classification.

Yin et al. [8] built a dataset for short text classification, the data was collected from two micro blogs and contains five classes, Politics, Economy, Education, Entertainment, and S&T. Semi-supervised learning and SVM were used to improve the accuracy of the classification, and the highest performance of this algorithm in terms of precision and recall was 80.49% and 81.77%, respectively.

Al-Tahrawi and Al-Khatib [4] used Al-Jazeera News dataset to evaluate the performance of polynomial networks classifier. Alj-News dataset was collected from Al-jazeera News Arabic Website. The dataset contains 1500 Arabic documents divided evenly into five classes: Politics, Science Economic, Art, and Sport. And the performance in terms of precision and recall was 90% and 89%, respectively.

Al Mukhaiti et al. [9] built a dataset for Arabic Sentiment Analysis. The data was collected from Facebook, YouTube, Twitter, Keek, and Instagram, and contains 2009 tweets/review. A system developed by Siddiqui et al. [10] was used to evaluate the dataset, the evaluation metrics were precision, recall, and accuracy and the results were 75.9%, 79.8, and 77.7%, respectively.

Alayba et al. [11] built a dataset for Arabic Sentiment Analysis on health services, the dataset contains 2026 tweets collected from twitter using twitter API, three machine learning algorithms were used: Naïve Bayes (NB), Logistic Regression (LR), and SVM, with a change on the size of training set and test set in three phases, The accuracy results were between 85% and 91% and the best classifiers was SVM using linear support vector.

3 Data Collection

The most time-exhaustion and the most important phase of text mining is Data collection [9]. In this paper, we will explain the overview of the dataset development process. This process divided into three phases, data Acquisition phase, data filtering phase, and data labeling phase. Figure 1 depicts the dataset development phases.

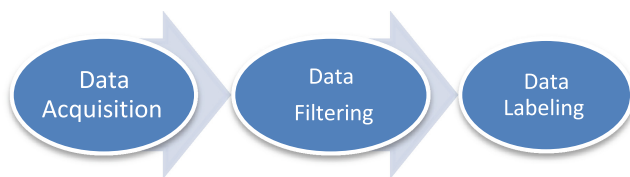


Fig. 1. Dataset development phases

The following steps depict the dataset development phases:

1. Collect data by crawling the Arabic Facebook pages.
2. Filter the data collected in the previous step
 - a. Removing URLs
 - b. Removing non-Arabic
 - c. Removing repeated posts.
3. Manually labeled the filtered posts to one of the ten categories chosen.

We chose ten categories/classes for the dataset, these ten classes cover most of the social network topics, and the classes are politics, economics, sport, religion, technology, TV, ads, foods, health, and porno. Figure 2 shows the process of building the proposed dataset. The algorithm used for building the dataset can be summarized as:

3.1 Data Acquisition Phase

We have collected about 40,000 Arabic Facebook posts. To collect the posts we developed a web browser to collect the data automatically, it has the ability to collect posts, comments, and replies, by automatically scrolling down the Facebook page to show all posts from the page date of creation, then gathering all the posts and save them in a text file.

3.2 Data Filtering Phase

This phase included of removal of the following types of posts, URLs only posts, non-Arabic posts, and repeated posts. Some posts contain only URL(s), this URL is in English letters and does not useful in Arabic dataset. Arabic Facebook pages sometimes share non-Arabic posts, English or Franco Arabic (Arabic spellings with English letters and digits). The last type is repeated posts, different pages may publish the same post, or a page may re-post an old post, the post was added only once. In case that the post contains an Arabic text plus to URLs, digits, non-Arabic letters, or emotions symbols, this post will be filtered, the Arabic text only will be saved, and the remaining parts will be removed. Table 1 depicts some examples.

3.3 Data Labeling Phase

The filtered posts were used in the labeling phase, wherein the filtered posts were labeled as politics, economics, sport, religion, technology, TV, ads, foods, health, or

```

Input: D: array of strings
Input: P: array of strings
Input: j:=0, k:=0
Input: C:{"politics",
"economics", "sport", "religion", "technology", "TV", "ads", "foods", "health", "porno"}
I. Scroll Through the Facebook page and save posts(D)
II.  Foreach  $D_i$  in D do
    a.  If not(( $D_i := \text{URL}$ ) or ( $D_i := \text{non-Arabic}$ ) or ( $P$  contains  $D_i$ )) then
        i.   $P[j] := D_i$ 
        ii.  $j := j+1$ 
    b.  End if
    c.  If not(( $D_i$  contains URL) or ( $D_i$  contains non-Arabic)) then
        i.  Remove URL or non-Arabic
III.  End for
IV.  Foreach  $P_i \in P$  do
    a.  Load the post  $P_i$  in the Browser and read the class selected by the user (C)
    b.   $C_i[k] := P_i$ 
    c.  Append  $P_i$  to  $C_i$  Text File
    d.   $K := k+1$ 
V.    End for
VI.  End.
    
```

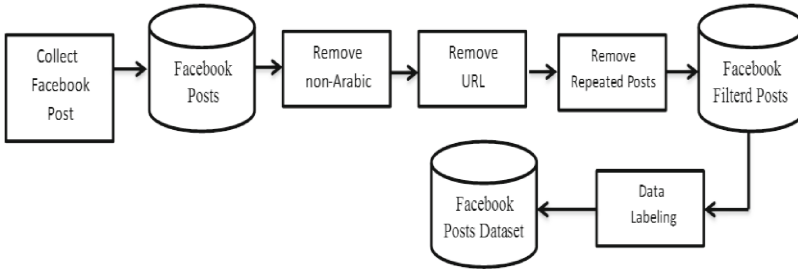


Fig. 2. Arabic dataset building process

Table 1. Filtering examples

Post	Action-Reason
هدف بوليفيا في شبك التانغو Bolivia's goal	Filtered-delete non-Arabic
http://onbe.in/1K6M2Up	Removed-Only URL
matsh alahly walzmark Isa bkrh alsa3h 10	Removed-non-Arabic
جھزي مطبخك بحلول مبتكرة bit.ly/CC_kitchen_utensils	Filtered-delete URL
☺☺ اختبارات الحساسية لمعرفة سبب المرض	Filtered-delete emotions
لمدة نصف ساعة EGS02051C018 تم إيقاف الورقة المالية	Filtered-delete non-Arabic

porno. No system accessible for assessing posts for Arabic text classification. So, we divide the labeling phase to two parts, first As a speaker of the Arabic language, I categorized the data collected, by reading and characterize each post to one of the previous ten classes. Table 2 views example of the classified posts.

To assess our labeling made during the first part, the dataset was given to ten Arabic native speakers who additionally confirm the validity of the dataset created. We

Table 2. Labeling phase

Post	Labeled as
مطلوب سيارات حديثة بسائق أو بدون لفترات طويلة بشيراتون المطار	Ads
النيل للتأجير التمويلي تستهدف طرح أسهمها بالبورصة لتمويل الأنشطة الجديدة	Economics
البيكنج بودر هو سر نجاح التوست أعرفي المقادير بالتفصيل من هنا	Food
علاج السمنة الموضوعية بالميزو ثيرابي بعيادة الشفاء هو تقنية آمنة طبية	Health
مشاورات نائب وزير الخارجية المصرى على هامش جنيف	Politics
يارب مع نهاية اليوم نور لي قلبي، ويسر لي حالي، وفرج عني كربى وهمي	Religion
تعرف علي رسالة حسام حسن للاعبه قبل مواجهة الزمالك في كأس مصر	Sport
تسريب مواصفات هاتف سامسونج	Technology
محمد هنيدي يبدأ تسجيل حلقات جديدة من سوبر هنيدي	TV
ممكن إنسانة جادة لعلاقة ممتعة دلح وحب	Porno

**Fig. 3.** Webpage screenshot from a PC**Fig. 4.** Webpage screenshot from a smartphone

built an online web page to help the users to assess the dataset remotely, at any time, from any location and from any device, PC or smartphone. In Figs. 3 and 4 screenshots of the web page from PC and smartphone, are shown respectively.

4 Results and Evaluation

Data Classification is a two steps process: (1) the training (or learning) phase and (2) the test (or evaluation) phase where the actual class of the instance is compared with the predicted class. If the hit rate is acceptable to the analyst, the classifier is accepted as being capable of classifying future instances with unknown class [12].

Our dataset building process contains three phases: 1. Data Acquisition, 2. Data Filtering, and 3. Data Labeling. Table 3 views the total number of posts in each class after phase 3.

Table 3. Labeling dataset phase

Class	No. of posts
Sport	2500
Politics	2500
TV	2500
Technology	2500
Religion	2500
Economic	2500
Food	2500
Porno	2500
Ads	2500
Health	2500
Total	25000

To evaluate our dataset, RapidMiner Studio Professional 7.6 was used to analyze data, RapidMiner is an open-source platform independently used for data mining [13]. RapidMiner is code-free software for designing advanced analysis processes with machine learning, data and text mining and business analytics, and predictive analytics, through its graphical user interface, data mining processes can be easily designed and executed. There are operators for tokenization, stemming, and stop words filtering. RapidMiner provides extensions of data loading, data transformation, data modeling, and data visualization methods. One of the most useful extensions in RapidMiner is the Text Processing package, which includes operators that support text mining. RapidMiner has an important feature, it can process a lot of languages including the Arabic language.

We applying RapidMiner operators: Naive Bayes, k-Nearest Neighbors (k-NN), support vector machine SVM, Performance (for classification) and Apply model (for testing). Evaluation metrics includes weighted mean recall, Weighted mean Precision, Kappa statistic, and Accuracy. The weighted mean recall is the average of recall calculated per class. Weighted Mean Precision is the average of precision obtained per class. Kappa Statistic (the accuracy varies from 0 to 1) measures the approval, of

prediction with the true class and it means that the classifier is in total agreement with a random classifier. The accuracy is defined as the ratio of numbers of correctly classified posts to the total number of posts. Recall and Precision are defined as:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{1}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{3}$$

Where TP, TN, FP, and FN refer to: Truly Positive, Truly Negative, Falsely Positive, and Falsely Negative claims of the classifier respectively.

Table 4 shows the result of testing our dataset on RapidMiner tool. It gives the accuracy of 95.12% when using SVM model, which gives the highest accuracy among used models. Figure 5 depicts the output result from RapidMiner. Comparing our results with other available dataset mentioned in Sect. 2 related work, show that the accuracy of our dataset is high compared to mentioned results as shown in Table 5.

Table 4. Evaluation results

Evaluation metric	Value
Accuracy	95.12%
Weighted mean recall	95.12%
Weighted mean precision	95.32%
Kappa	0.946

accuracy: 95.12%

	true ads	true politi...	true econ...	true food	true health	true porno	true relig...	true sports	true tech...	true tv	class pr...
pred. ads	732	1	4	1	1	0	1	0	0	2	98.65%
pred. poli...	0	655	25	0	3	0	3	7	11	13	91.35%
pred. eco...	8	60	699	0	14	6	11	5	35	19	81.56%
pred. food	1	2	0	745	8	1	0	4	3	0	97.51%
pred. he...	1	3	2	2	714	1	1	0	3	2	97.94%
pred. por...	0	2	1	1	1	741	1	0	2	4	98.41%
pred. reli...	0	0	1	1	4	0	727	0	0	0	99.18%
pred. sp...	2	1	2	0	0	0	1	730	2	6	98.12%
pred. tec...	3	16	15	0	4	0	3	4	690	3	93.50%
pred. tv	3	10	1	0	1	1	2	0	4	701	96.96%
class rec...	97.60%	87.33%	93.20%	99.33%	95.20%	98.80%	96.93%	97.33%	92.00%	93.47%	

Fig. 5. RapidMiner results

Table 5. Comparison of some existing works

Evaluation metric	Proposed dataset	Al-Kabi et al. [6]	Abdul-Mageed et al. [7]	Yin et al. [8]	Al-Tahrawi and Al-Khatib [4]	Al Mukhaiti et al. [9]	Alayba et al. [11]
Accuracy	95.12%	66.2%	84.65%	–	–	77.7%	91%
Recall	95.12%	–	–	81.7%	90%	79.8	–
Precision	95.32%	–	–	80.4%	89%	75.9%	–

5 Conclusions

In recent years, online social networking sites have spread widely, and data is published and shared in large quantities every moment. Social networking sites do not give users ability to filter or categorize content on their walls. Therefore, we have created in this paper a dataset of online Arabic text collected from the Facebook to be used in the text classification process. We chose Arabic because of the paucity of the Arabic dataset available while online Arabic content is increasing and online Arab users are growing. The dataset building process divided into three phases, data Acquisition, data filtering, and data labeling phase. The dataset was collected from Arabic Facebook pages, then in data filtering phase the URLs, non-Arabic, and repeated posts were removed from the dataset. Finally, in the labeling phase, each post was given a label from the ten chosen categories and ten Facebook Arabic users were involved in the labeling process. To evaluate our dataset RapidMiner tool was used and the performance achieved in terms of accuracy was 95.12% with the SVM model. Our dataset will help researchers in the field of short Arabic text processing.

References

1. Bodkhe, R., Ghorpade, T., Jethani, V.: A novel methodology to filter out unwanted messages from OSN user's wall using trust value calculation. In: Proceedings of the Second International Conference on Computer and Communication Technologies, pp. 755–764. Springer (2016)
2. Ghosh, S., Roy, S., Bandyopadhyay, S.: A tutorial review on text mining algorithms. *Int. J. Adv. Res. Comput. Commun. Eng.* **1**(4), 7 (2012)
3. Internet World Stats: Internet World Users by Language (2017). <http://www.internetworldstats.com/stats7.htm>. Accessed 13 Sep 2017
4. Al-Tahrawi, M.M., Al-Khatib, S.N.: Arabic text classification using polynomial networks. *J. King Saud Univ. Comput. Inf. Sci.* **27**(4), 437–449 (2015)
5. Al-Sallab, A., Baly, R., Hajj, H., Shaban, K.B., El-Hajj, W., Badaro, G.: AROMA: a recursive deep learning model for opinion mining in Arabic as a low resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **16**(4), 25 (2017)
6. 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), pp. 23–25 (2013)
7. Abdul-Mageed, M., Diab, M., Kübler, S.: SAMAR: subjectivity and sentiment analysis for Arabic social media. *Comput. Speech Lang.* **28**(1), 20–37 (2014)

8. Yin, C., Xiang, J., Zhang, H., Wang, J., Yin, Z., Kim, J.-U.: A new SVM method for short text classification based on semi-supervised learning. In: 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), pp. 100–103. IEEE (2015)
9. Al Mukhaiti, A.J.S., Siddiqui, S., Shaalan, K.: Dataset built for Arabic sentiment analysis. In: International Conference on Advanced Intelligent Systems and Informatics, pp. 406–416. Springer (2017)
10. Siddiqui, S., Monem, A.A., Shaalan, K.: Sentiment analysis in Arabic. In: International Conference on Applications of Natural Language to Information Systems, pp. 409–414. Springer (2016)
11. Alayba, A.M., Palade, V., England, M., Iqbal, R.: Arabic language sentiment analysis on health services. arXiv preprint [arXiv:1702.03197](https://arxiv.org/abs/1702.03197) (2017)
12. Borges, L.C., Marques, V.M., Bernardino, J.: Comparison of data mining techniques and tools for data classification. In: Proceedings of the International C* Conference on Computer Science and Software Engineering, pp. 113–116. ACM (2013)
13. RapidMiner Documentation. <https://docs.rapidminer.com/>. Accessed 10 Sep 2017