Building Large Arabic Multi-domain Resources for Sentiment Analysis

Hady ElSahar and Samhaa R. El-Beltagy

Center of Informatics Sciences, Nile University, Cairo, Egypt

hadyelsahar@gmail.com, samhaa@computer.org

Abstract. While there has been a recent progress in the area of Arabic Sentiment Analysis, most of the resources in this area are either of limited size, domain specific or not publicly available. In this paper, we address this problem by generating large multi-domain datasets for sentiment analysis in Arabic. The datasets were scrapped from different reviewing websites and consist of a total of 33K annotated reviews for movies, hotels, restaurants and products. Moreover we build multi-domain lexicons from the generated datasets. Different experiments have been carried out to validate the usefulness of the datasets and the generated lexicons for the task of sentiment classification. From the experimental results, we highlight some useful insights addressing: the best performing classifiers and features, the effect of introducing lexicon based features and factors affecting the accuracy of sentiment classification in general. All the datasets, experiments code and results have been made publicly available for scientific purposes.

1 Introduction

In the past few years, Sentiment analysis has been the focus of many research studies due to the wide variety of its potential applications. Many of these studies have relied heavily on available resources mostly in the form of polarity annotated datasets [15–17, 20] or sentiment lexicons such as SentiWordNet [5].

At the same time, the Arabic language has shown rapid growth in terms of its users on the internet, moving up to the 4th place in the world ranking of languages by users according to internetworldstats[1]. This, along with the major happenings in the Middle East, show a large potential for sentiment analysis and consequently an urgent need for more reliable processes and resources for addressing it.

Because of that, there has been an increasing interest and research in the area of Arabic sentiment analysis. However, The Arabic Language remains under resourced with respect to available data. This can be attributed to the fact that most resources developed within studies addressing Arabic sentiment analysis, are either limited in size, not publicly available or developed for a very specific domain.

[1] http://www.internetworldstats.com/stats7.htm

adfa, p. 1, 2011.

Having said that, a handful of recently published work addresses the issue of availing large Arabic resources for sentiment analysis [4, 6, 12]. In this presented work, we follow in the footsteps of these, by creating a large multi-domain datasets of annotated reviews which we publicly avail to the scientific community. The datasets cover the following domains: movies, hotels, restaurants and products and are made up of approximately 33K reviews. Furthermore we make use of each of the generated datasets to build domain specific sentiment lexicons.

We make use of the multi-domain generated lexicons to perform extensive experiments benchmarking a wide range of classifiers and feature building methods for the task of sentiment classification. Experimental results provide useful insights with respect to the performance of various classifiers, the effect of different content representations, and the usefulness of the generated lexicons when used solely and when combined with other features. Furthermore, we study the effect of document length and richness with subjective terms on the performance of the sentiment classification task, with the aim to find the document criteria which affects the performance of the sentiment classification the most.

2 Related work

Building sentiment analysis resources for the Arabic language, has been addressed by a number of researchers. For sentiment annotated corpora Rushdi-Saleh et al. [18] presented OCA; a dataset of 500 annotated movie reviews collected from different web pages and blogs in Arabic. Although the dataset is publicly available, it is limited in size and only covers the movie reviews domain.

Abdul-Mageed & Diab [1] presented the AWATIF multi-genre corpus of Modern Standard Arabic labeled for subjectivity and sentiment analysis. The corpus was built from different resources including the Penn Arabic Treebank, Wikipedia Talk Pages and Web forums. It was manually annotated by trained annotators and through crowd sourcing. The dataset targets only Modern Standard Arabic (MSA) which is not commonly when writing reviews on most websites and social media. Moreover the dataset is not available for public use.

LABR [4, 12] a large dataset of 63K, polarity annotated, Arabic Book reviews scrapped from www.goodreads.com. On this site, each review is rated on a scale of 1 to 5 stars which the authors have mapped to a sentiment polarity. The dataset was then used for the tasks of sentiment polarity classification and rating classification. The large scale dataset is publicly available for use; however it only covers the domain of book reviews.

For sentiment lexica, as a part of a case study exploring the challenges in conducting sentiment analysis on Arabic social media, El-Beltagy et al. [7] developed a sentiment lexicon including more than 4K terms. The lexicon was semi-automatically constructed through expanding a seed list of positive and negative terms by mining conjugated patterns and then filtering them manually. El-Sahar & El-Beltagy [8] presented a fully automated approach to extract dialect sentiment lexicons from twitter streams using lexico-syntactic patterns and point wise mutual information.

More recently, SANA, a large scale multi-genre sentiment lexicon was presented [2, 3]. SANA is made up of 224,564 entries covering Modern Standard Arabic, Egyptian Dialectal Arabic and Levantine Dialectal Arabic. SANA is built from different resources including The Penn Arabic Treebank [10], Egyptian chat logs, YouTube comments, twitter and English SentiWordNet. Some of the lexicon components were built manually, others were obtained using automatic methods such as machine translation. Various techniques were used to evaluate the generated lexicon. The lexicon is not publicly available.

3 Building the Datasets

Finding and extracting of Arabic reviewing content from the internet is considered to be a hard task relatively to English [18]. This is due to the smaller number and activity of Arabic based e-commerce & reviewing websites over the internet, also that many Arabic speakers use the English language or Arabic transliterated in Roman characters to write their reviews. All this has had a big impact on reducing the amount of pure Arabic reviews on the internet. Fortunately, recently the Arabic reviewing content over the internet has shown a significant growth, moreover new reviewing websites has been built. In this study we make use of the available reviewing Arabic content over the internet to create multi-domain datasets reliable for the task of sentiment analysis.

3.1 Dataset Generation

For the automatic generation of annotated datasets, we utilize the open-source Scrapy [2]framework, which is a framework for building custom web crawlers. The datasets cover four domains as follows:

1. **Hotel Reviews (HTL)**: For the hotels domain 15K Arabic reviews were scrapped from TripAdvisor[3]. Those were written for 8100 Hotels by 13K users.
2. **Restaurant Reviews (RES):** For the restaurants domain two sources were scrapped for reviews: the first is Qaym[4] from which 8.6K Arabic reviews were obtained, and the second is TripAdvisor from which 2.6K reviews were collected. Both datasets cover 4.5K restaurants and have reviews written by over 3K users

[2] www.scrapy.org

[3] www.tripadvisor.com

[4] www.Qaym.com

3. **Movie Reviews (MOV):** The movies domain dataset was built out of scrapping 1.5K reviews from Elcinemas.com covering around 1K movies.
4. **Product Reviews (PROD):** For the Products domain, a dataset of 15K reviews was scraped from the Souq[5] website. The dataset includes reviews from Egypt, Saudi Arabia, and the United Arab Emirates and covers 6.5K products for which reviews were written by 7.5K users.

Each of websites above provides for each review, the text of the review as well as a rating entered by the reviewer. The rating reflects the overall sentiment of the reviewer towards the entity s/he reviewed. So, for each review, the rating was extracted and normalized into one of three categories: positive, negative, or mixed using the same approach adopted by Nabil et al. and Pang et al. [12, 14]. To eliminate any irrelevant and re-occurring spam reviews, we eliminate all redundant reviews.

3.2 Datasets Statistics

In order to better understand the nature of the various collected datasets, a set of statistics reflecting the number of reviews each contains, the number of users who contributed to the reviews, the number of items reviewed, and the polarity distribution of the reviews, was generated for each dataset. These are presented in Table 1 and Figure 1. The total number of collected unique reviews is approximately 33K. The dataset that had the most redundancy was the PROD dataset where out of the total 14K scrapped reviews, only 5K were unique. This can be attributed to the fact that lots of reviews in this dataset consist of only one word, which increases the probability of redundancy.

As shown in figure 1, the total the number of positive reviews is far larger than that of negative reviews in all of the datasets. The same phenomenon was also observed by the authors of LABR [4], the dataset collected for book reviews.

Figure 2 shows a box plot representing the average number of tokens per review for each dataset. As can be seen, the movie reviews in the MOV dataset are by far the longest. By examining the dataset, it was observed that this is due to the fact that people tend to write long descriptions of the movie they are reviewing, within their review. On the other hand, the PROD dataset tends to have the shortest reviews. Later in this paper we investigate the effect of the length of document on the process of sentiment classification.

Table 1.Summary of Dataset Statistics

	HTL	RES#1	RES#2	MOV	PROD	ALL
#Reviews	15579	8664	2646	1524	14279	42692
#Unique Reviews	15562	8300	2640	1522	5092	**33116**
#Users	13407	1639	1726	416	7465	24653
#Items	8100	3178	1476	933	5906	19593

[5] www.souq.com

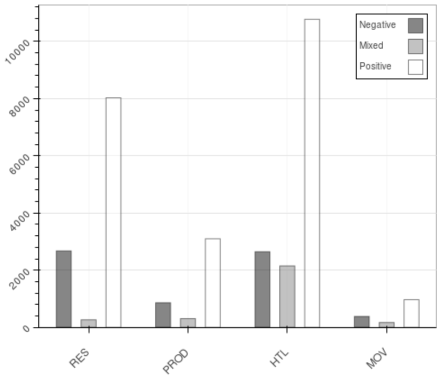

Fig. 2. Number of reviews for each class

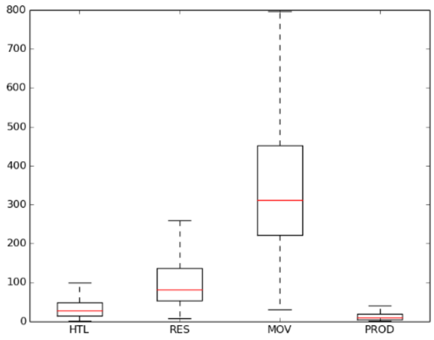

Fig. 1. Box plot showing number of tokens for each of the datasets

4 Building Lexicons

In this section we introduce a method to generate multi-domain lexicons out of the collected reviews datasets. The approach followed is a semi-supervised one, that makes use of the feature selection capabilities of Support Vector Machines [21] to select the most significant phrases contributing to accuracy of sentiment classification and is very similar to that presented by Nabil, Aly and Atiya [12]

To build a lexicon, we follow an approach that generates unigrams and bi-grams from the collected documents. For selecting the set of most significant features we utilize 1-norm Support Vector Machines [21] displayed in (1). 1-norm support vector machines use the L1 penalty $\|\beta\|_1$ calculated as shown in (2).
L1 regularization has proven to be superior to L2(3) regularization when the number of features is larger than the number of samples, or in other words when there are many irrelevant features [13], which is our case as we use all the extracted n-grams as features.

$$\arg min_{\beta,\beta_0} \sum_{i=1}^{n} [1 - y_i(x_i^T\beta + \beta_0)]_+ + \lambda\|\beta\|_1 \qquad (1)$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad (2)$$

$$\|x\| = \sqrt{\sum_{i=1}^{n} |x_i|^2} \qquad (3)$$

In addition to the previously generated multi-domain datasets, we make use of the LABR dataset for book reviews [4, 12] in order to generate multi-domain lexica covering the book reviews domain as well. We use each of the datasets individually and split each into two parts: 80% for training & validation (we use this as well to generate our lexicons), and 20% for testing. The aim of testing is to access the usability of learned lexicons on classifying unseen data.

Out of the training examples we start by building a bag of words model for each dataset where features are the set of unigrams and bigrams and values are simple word counts. Since we are interested in generating a sentiment lexicon of positive and negative terms only, we use only reviews tagged with a positive or negative class.

We use cross validation to tune the soft margin parameter C ($\lambda = 1/2C$). Higher values of C add a higher penalty for the misclassified points rather than maximizing the separation margin. So the optimization problem will lead to a larger number of selected features to reduce the misclassified errors. Lower values of C, result in smaller vectors which are more sparse, leading to a lower number of selected features which might lead to underfitting when the selected features are not enough for the classification process. The best performing classifier is the classifier with the highest accuracy with the least amount of selected features.

After this step, we rank the non-zero coefficients of the best model parameters and map them to the corresponding unigram and bigram features. Features with the highest positive value coefficients are considered to be the highest discriminative features of the documents with +ve sentiment. On the other hand, n-grams which corresponds to the highest negative value coefficients are considered to indicate a –ve sentiment. Based on this, we automatically label the n-grams with the corresponding class label.

This process was repeated for each of datasets. The resulting unigrams and bigrams from each, was then reviewed by two Arabic native speaker graduate students. The reviewers were asked to manually filter incorrect or irrelevant terms and to keep only those which match with their assigned label thus indicating positive or negative sentiment.

The result of this process was a set of domain specific lexicons extracted from each dataset. In addition, we combined all lexicons into one domain general lexicon; sizes of the different lexicons are shown in table 2.

Table 2. Summary of lexicon sizes

	HTL	RES	MOV	PROD	LABR	ALL
# Selected features	556	1413	526	661	3552	6708
# Manually filtered	218	734	87	369	874	1913

5 Experiments:

In this section we design a set of experiments, aiming to: a) validate the usefulness of the generated datasets and lexicons, b) provide extensive benchmarks for different machine learning classifiers and feature building methods over the generated datasets to aid future research work. The experiments consisted of three variations which are described in the following subsections.

5.1 Dataset Setups:

Experiments on each of the generated datasets were done independently. We also ran the experiments on the LABR book reviews dataset [12]. We explore the problem of sentiment classification as a 2 class classification problem (positive or negative) and a 3 class classification problem (positive, negative and mixed). We ran the experiments using 5-fold cross validation. Moreover, we re-ran our trained classifiers on the 20% unseen testing dataset to make sure that our models are not over fitted. All experiments were carried out using both balanced and unbalanced datasets, but due to paper length limitations the experiments carried out on unbalanced datasets, are documented in a separate report.

5.2 Training features:

For building feature vectors we applied several methods that have been widely utilized before in sentiment classification such as word existence, word count [17, 20] and TFIDF [18].

We also used Delta TFIDF [11]. This method is a derivative of TFIDF in which each n-gram is assigned a weight equal to the difference of that n-gram's TFIDF scores in the positive and negative training corpora as represented in (4). In this equation, $V_{t,d}$ is the Delta TFIDF value for term t in document d, $C_{t,d}$ is the number of times term *t* occurs in document *d*, P_t and N_t are the number of positive and negative labeled documents in the training set with the term *t* while $|P|$ and $|N|$ are the sizes of the positive and negative labeled documents in the training sets.

$$V_{t,d} = C_{t,d} * \log_2\left(\frac{|P|}{P_t}\right) - C_{t,d} * \log_2\left(\frac{|N|}{N_t}\right) \quad (4)$$

This method promises to be more efficient than traditional TFIDF, especially in the reviews domain as common subjective words like "Good", "Bad", "Excellent" are likely to appear in a large number of documents leading to small IDF values, even though these terms are highly indicative. At the same time, these terms don't re-occur frequently within the same document, as users tend to use synonyms to convey the same meaning, which overall results in smaller values of TF*IDF.

Another type of feature representation was examined. In this representation schemed, a feature vector is comprised entirely of entries from previously generated lexicons. A document is then represented by the intersection of its terms with the lexicon terms or simply by the matches in the document from the lexicon, and their count.
We apply this feature representation method once by using domain specific lexicons on each of their respective datasets and another, using the combined lexicon. We refer to those feature representation methods as Lex-domain and Lex-all respectively.

The experiments examine the effect of combining feature vectors generated from Lex-domain and Lex-all, with those generated from TF-IDF, Delta-TFIDF and Count. The effect of this step is discussed in details in the next section.

5.3 Classifiers:

For the training and classification tasks, experiments were done using Linear SVM, Logistic regression, Bernoulli Naive Bayes, K nearest neighbor and stochastic gradient descent. The linear SVM parameters were set using cross validation.

Combining different features, classifiers and dataset setups resulted in 615 experiments for each of the datasets. The detailed experiments results and the source code of the experiments have been made publically available for research purposes[6], but a summary of what the authors think are the most important experiments, is presented in the next sub-section

Table 3. Ranking of calssifiers by average accuracy

Classifier	Accuracy	
	2 Classes	3 Classes
Linear SVM	**0.824**	**0.599**
Bernoulli NB	0.791	0.564
LREG	0.771	0.545
SGD	0.752	0.544
KNN	**0.668**	**0.469**

[6] http://bit.ly/1wXue3C

Table 4. Average accuracy associated with of each of the feature representations with and without combining lexicon based features

	Features	Lexicon	LABR	MOV	RES	PROD	HTL	Average
2 Class	Lex-domain	N/A	0.727	0.703	0.811	0.740	0.859	0.768
	Lex-all	N/A	0.746	0.739	0.826	0.732	0.868	0.782
	Count	None	0.806	0.710	0.810	0.725	0.866	0.783
		Lex-domain	0.810	0.703	0.816	0.745	0.874	0.790
		Lex-all	**0.812**	0.733	0.819	0.745	0.873	0.796
	TFIDF	None	0.739	0.552	0.761	0.723	0.730	0.701
		Lex-domain	0.786	0.723	0.819	0.751	0.876	0.791
		Lex-all	0.783	**0.743**	0.836	0.758	0.876	**0.799**
	Delta-TFIDF	None	0.739	0.535	0.745	0.694	0.746	0.692
		Lex-domain	0.771	0.704	0.831	0.752	0.884	0.789
		Lex-all	0.779	0.721	**0.846**	**0.759**	**0.887**	0.798
3 Class	Lex-domain	None	0.510	0.503	0.578	0.524	0.630	0.549
	Lex-all	None	0.529	0.491	0.607	0.494	0.649	0.554
	Count	None	0.603	0.497	0.563	0.520	0.669	0.570
		Lex-domain	0.605	0.484	0.579	0.532	0.669	0.574
		Lex-all	**0.606**	**0.526**	0.589	**0.537**	**0.671**	**0.586**
	TFIDF	None	0.546	0.348	0.513	0.473	0.575	0.491
		Lex-domain	0.578	0.520	0.581	0.536	0.653	0.574
		Lex-all	0.577	0.510	0.599	0.510	0.661	0.572
	Delta-TFIDF	None	0.527	0.340	0.471	0.442	0.549	0.466
		Lex-domain	0.555	0.503	0.588	0.531	0.656	0.566
		Lex-all	0.567	0.476	**0.606**	0.505	0.669	0.565

6 Results and Discussion:

This section highlights some of the experiments performed seeking answers for the proposed research questions. We present below the results recorded by experimenting on the balanced datasets. In the detailed experiments report we also present the results for the unbalanced datasets.

6.1 Best performing Classifiers and features

Comparing the performance of different classifiers, we average the accuracy of each classifier over all datasets using all feature building methods, the results are shown in Table 3.

It can be observed, that both the 2 class and 3 class classification problems yielded the same ranking for best and worst classifiers. Linear SVM proved to be the best preforming classifier over all datasets scoring a significant difference than the rest of the classifiers. While the worst preforming classifier was the K Nearest Neighbor.

To compare the effect of employing different feature representation methods, we calculate the accuracy of each one of them averaged over all classifiers;results are shown in Table 4. For the 2 class classification problem, the top three feature representation methods were Delta-TFIDF, TFIDF and Count, feature vectors when combined with Lex-all, the feature vectors of the combined lexicon. The same three feature representation methods also ranked on top, for the 3 class classification problem.

The least performing were the feature vectors of TF-IDF and Delta-TFIDF when used solely without combining with any lexicon based features vectors, with a 10% drop in the accuracy than the top performing feature representations.

6.2 Accuracy of Lexicon based Features solely and combined with other features

Lexicon based feature vectors of Lex-domain and Lex-all solely achieved a fair performance in comparison with the best performing features with less than a 2% drop in the average accuracy. These results were obtained using unseen test data different from the one used to build lexicons.

Given that the maximum length of the Lexicon based features Lex-Domain and Lex-all is 2K, while other feature vectors can grow up to several millions. This proves that Lexicon based features generated from a sample of the datasets can lead to much simpler classifiers.

Combining lexicon based features with other features provided large improvements on the total accuracy, with 10% in cases of TFIDF and Delta-TFIDF and 2% in case of Counts.

Using domain general features Lex-all rather than Lex-domain, doesn’t show a significant difference in the total accuracy in our case, as the length of the generated lexicons are relatively small and although they are multi-domain,all of them are generated from the reviews domain where very similar language seems to be used across domains

6.3 Effect of document length and richness with subjective terms on sentiment classification

In order to show the effect of document length and subjectivity richness on the performance of sentiment classification, we label each of the misclassified documents out of the 2 class classification problem with its number of terms and subjectivity score. The subjectivity score is the sum of all polarities of the subjective terms in the document. To calculate this score we rely on the generated domain general lexicon to detect subjective words in a document and their polarity. Additionally, a set of negation detection rules were crafted and used to flip the term polarity if it happened to be negated.

Then, documents of similar size and subjectivity score are grouped together and the average error rate is calculated for each. The resulting error rates were used to plot the heat map shown in Fig3. From the resulting heat map, the following can be observed:

- The error rate increases as the subjectivity score tends to zero and decreases as the subjectivity grows in any of the positive or negative directions. This applies for small, mid-range and relatively long documents.
- For extremely long documents, a document's subjectivity seizes to correlate with the error rate. We find that documents longer than 1K words with very high subjectivity, achieve very high error rates. This is probably because longer documents allow more topic drifts, criticizing other subjects or having comparisons with other entities which are not handled explicitly by any of our classifiers.
- Extremely short documents by definition have a very limited number of terms and hence cannot have a high subjectivity score, which often results from matching with multiple entries in subjectivity lexicon. As a result, the majority of extremely short documents end up with high error rates.
- Finally, we find that the error rate for mid-range documents is slightly shifted to the positive side. At the same time, the maximum subjectivity scores on the positive side are higher than on the negative side which is consistent with the observation that negative terms are less frequently used than positive terms [9, 19].

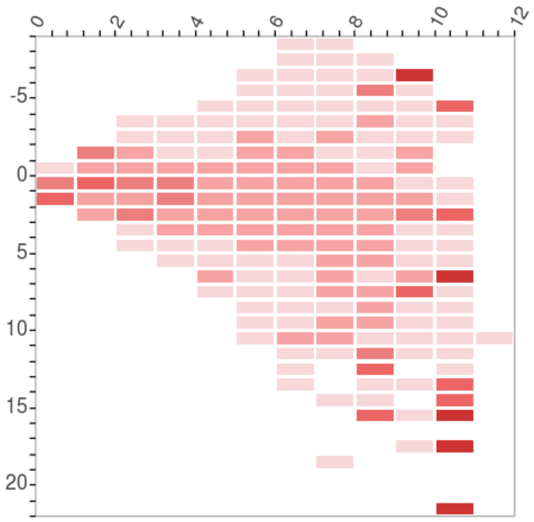

Fig. 3. Heat map showing the error rate for various document lengths and subjectivity score groups. The horizontal axis shows the log of the document lengths, while the vertical axis represents the subjectivity scores and the color gradient is the error rate (the darker the worse).

7 Conclusion and Future work:

In this study, we introduced large multi-domain datasets for sentiment analysis, scrapped from multiple websites that support reviews in the domains of movies, hotels, restaurants and products. Moreover we presented a multi-domain lexicon of 2K entries extracted from the datasets.

Although the generated lexicon isn't very large, experimental results have shown that abstracting reviews by lexicon based features only, achieved a relatively fair performance for the task of sentiment classification.

An extensive set of experiments was performed for the sake of benchmarking the datasets and testing their viability for both two class and three class sentiment classification problems. Out of the experimental results, we highlighted that the top performing classifier was SVM and the worst was KNN, and that the best performing feature representations were the combination of the lexicon based features with the other features.

Finally according to the error analysis on the task of sentiment classification, we find that the document length and richness with subjectivity both affect the accuracy of sentiment classification, in which; sentiment classifiers tend to work better when the documents are rich with polar terms of one class, i.e., high values of subjectivity score. However, this often doesn't hold when the document length is extremely short or long.

Although the generated datasets cover multiple domains, they are all generated only from reviews. Thus, their usefulness for social media sentiment analysis, is yet to be studied. This might include generation of additional datasets to cover cases that doesn't show up in the reviews domain but common in social media like advertisements and news. This is a motivation for future research work.

8 References

1. Abdul-Mageed, M., Diab, M.: AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. Lrec. 3907–3914 (2012).
2. Abdul-mageed, M., Diab, M.: SANA : A Large Scale Multi-Genre , Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1162–1169 (2014).
3. Abdul-Mageed, M., Diab, M.: Toward Building a Large-Scale Arabic Sentiment Lexicon. Proceedings of the 6th International Global WordNet Conference. pp. 18–22 (2012).
4. Aly, M., Atiya, A.: LABR: A Large Scale Arabic Book Reviews Dataset. Aclweb.Org. 494–498 (2013).

5. Baccianella, S. et al.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). pp. 2200–2204 (2010).

6. Badaro, G. et al.: A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining. ANLP 2014. pp. 176–184 (2014).

7. El-Beltagy, S., Ali, A.: Open Issues in the Sentiment Analysis of Arabic Social Media : A Case Study. in proceedings of 9th International Conference on Innovations in Information Technology (IIT). pp. 215 – 220 (2013).

8. Elsahar, H., El-Beltagy, S.: A Fully Automated Approach for Arabic Slang Lexicon Extraction from Microblogs. Computational Linguistics and Intelligent Text Processing. pp. 79–91 Springer Berlin Heidelberg (2014).

9. Jerry, B., Osgood, C.: The pollyanna hypothesis. J. Verbal Learning Verbal Behav. 8, 1, 1–8 (1969).

10. Maamouri, M. et al.: The penn arabic treebank: Building a large-scale annotated arabic corpus. NEMLAR Conference on Arabic Language Resources and Tools. pp. 102–109 (2004).

11. Martineau, J. et al.: Delta TFIDF: An Improved Feature Space for Sentiment Analysis. Proc. Second Int. Conf. Weblogs Soc. Media (ICWSM. 29, 490–497 (2008).

12. Nabil, M. et al.: LABR: A Large Scale Arabic Book Reviews Dataset. arXiv Prepr. arXiv1411.6718. (2014).

13. Ng, A.: Feature selection, L 1 vs. L 2 regularization, and rotational invariance. ICML. (2004).

14. Pang, B. et al.: Thumbs up? Sentiment Classification using Machine Learning Techniques. Conf. Empir. Methods Nat. Lang. Process. (EMNLP 2002). 79–86 (2002).

15. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. the 42nd annual meeting on Association for Computational Linguistics. pp. 271–278 (2004).

16. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. 1, (2005).

17. Pang, B., Lee, L.: Thumbs up? Sentiment classification using machine learning techniques. Proc. Conf. Empir. Methods Nat. Lang. Process. July 6-7, 2002, Philadephia, Pennsylvania, USA. 79–86 (2002).

18. Rushdi-Saleh, M., Martin-Valdivia, T.: OCA: Opinion corpus for Arabic. J. Am. Soc. Inf. Sci. Technol. 62.10. 2045–2054 (2011).

19. Taboada, M. et al.: Lexicon-Based Methods for Sentiment Analysis. , (2011).

20. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. pp. 417–424 (2002).

21. Zhu, J. et al.: 1 -norm Support Vector Machines. Advances in neural information processing systems 16.1. pp. 49–56 (2004).