LABR: A Large Scale Arabic Book Reviews Dataset

Mohamed Aly
Computer Engineering Department
Cairo University
Giza, Egypt
mohamed@mohamedaly.info

Amir Atiya
Computer Engineering Department
Cairo University
Giza, Egypt
amir@alumni.caltech.edu

Abstract

We introduce LABR, the largest sentiment analysis dataset to-date for the Arabic language. It consists of over 63,000 book reviews, each rated on a scale of 1 to 5 stars. We investigate the properties of the the dataset, and present its statistics. We explore using the dataset for two tasks: *sentiment polarity classification* and *rating classification*. We provide standard splits of the dataset into training and testing, for both polarity and rating classification, in both balanced and unbalanced settings. We run baseline experiments on the dataset to establish a benchmark.

1 Introduction

The internet is full of platforms where users can express their opinions about different subjects, from movies and commercial products to books and restaurants. With the explosion of social media, this has become easier and more prevalent than ever. Mining these troves of unstructured text has become a very active area of research with lots of applications. **Sentiment Classification** is among the most studied tasks for processing opinions (Pang and Lee, 2008; Liu, 2010). In its basic form, it involves classifying a piece of opinion, e.g. a movie or book review, into either having a *positive* or *negative* sentiment. Another form involves predicting the actual rating of a review, e.g. predicting the number of stars on a scale from 1 to 5 stars.

Most of the current research has focused on building sentiment analysis applications for the English language (Pang and Lee, 2008; Liu, 2010; Korayem et al., 2012), with much less work on other languages. In particular, there has been little work on sentiment analysis in Arabic (Abbasi et al., 2008; Abdul-Mageed et al., 2011; Abdul-Mageed et al., 2012; Abdul-Mageed and Diab, 2012b; Korayem et al., 2012), and very few, considerably small-sized, datasets to work with (Rushdi-Saleh et al., 2011b; Rushdi-Saleh et al., 2011a; Abdul-Mageed and Diab, 2012a; Elarnaoty et al., 2012). In this work, we try to address the lack of large-scale Arabic sentiment analysis datasets in this field, in the hope of sparking more interest in research in Arabic sentiment analysis and related tasks. Towards this end, we introduce **LABR**, the **L**arge-scale **A**rabic **B**ook **R**eview dataset. It is a set of over 63K book reviews, each with a rating of 1 to 5 stars.

We make the following contributions: (1) We present the largest Arabic sentiment analysis dataset to-date (up to our knowledge); (2) We provide standard splits for the dataset into training and testing sets. This will make comparing different results much easier. The dataset and the splits are publicly available at www.mohamedaly.info/datasets; (3) We explore the structure and properties of the dataset, and perform baseline experiments for two tasks: *sentiment polarity classification* and *rating classification*.

2 Related Work

A few Arabic sentiment analysis datasets have been collected in the past couple of years, we mention the relevant two sets:

OCA Opinion Corpus for Arabic (Rushdi-Saleh et al., 2011b) contains 500 movie reviews in Arabic, collected from forums and websites. It is divided into 250 positive and 250 negative reviews, although the division is not standard in that there is no rating for *neutral* reviews i.e. for 10-star rating systems, ratings above and including 5 are considered positive and those below 5 are considered negative.

AWATIF is a multi-genre corpus for Modern Standard Arabic sentiment analysis (Abdul-

Number of reviews	63,257
Number of users	16,486
Avg. reviews per user	3.84
Median reviews per user	2
Number of books	2,131
Avg. reviews per book	29.68
Median reviews per book	6
Median tokens per review	33
Max tokens per review	3,736
Avg. tokens per review	65
Number of tokens	4,134,853
Number of sentences	342,199

Table 1: **Important Dataset Statistics**.

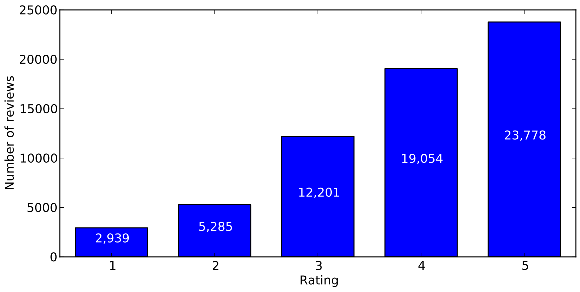

Figure 1: **Reviews Histogram**. The plot shows the number of reviews for each rating.

Mageed and Diab, 2012a). It consists of about 2855 sentences of news wire stories, 5342 sentences from Wikipedia talk pages, and 2532 threaded conversations from web forums.

3 Dataset Collection

We downloaded over 220,000 reviews from the book readers social network www.goodreads.com during the month of March 2013. These reviews were from the first 2143 books in the list of *Best Arabic Books*. After harvesting the reviews, we found out that over 70% of them were not in Arabic, either because some non-Arabic books exist in the list, or because of existing translations of some of the books in other languages. After filtering out the non-Arabic reviews, and performing several pre-processing steps to clean up HTML tags and other unwanted content, we ended up with 63,257 Arabic reviews.

4 Dataset Properties

The dataset contains 63,257 reviews that were submitted by 16,486 users for 2,131 different books.

Task		Training Set	Test Set
1. Polarity Classification	B	13,160	3,288
	U	40,845	10,211
2. Rating Classification	B	11,760	2,935
	U	50,606	12,651

Table 2: **Training and Test sets**. **B** stands for balanced, and **U** stands for Unbalanced.

Table 1 contains some important facts about the dataset and Fig. 1 shows the number of reviews for each rating. We consider as *positive* reviews those with ratings 4 or 5, and *negative* reviews those with ratings 1 or 2. Reviews with rating 3 are considered neutral and not included in the polarity classification. The number of positive reviews is much larger than that of negative reviews. We believe this is because the books we got reviews for were the most popular books, and the top rated ones had many more reviews than the the least popular books.

The average user provided 3.84 reviews with the median being 2. The average book got almost 30 reviews with the median being 6. Fig. 2 shows the number of reviews per user and book. As shown in the Fig. 2*c*, most books and users have few reviews, and vice versa. Figures 2*a-b* show a box plot of the number of reviews per user and book. We notice that books (and users) tend to have (give) positive reviews than negative reviews, where the median number of positive reviews per book is 5 while that for negative reviews is only 2 (and similarly for reviews per user).

Fig. 3 shows the statistics of tokens and sentences. The reviews were tokenized and "*rough*" sentence counts were computed (by looking for punctuation characters). The average number of tokens per review is 65.4, the average number of sentences per review is 5.4, and the average number of tokens per sentence is 12. Figures 3*a-b* show that the distribution is similar for positive and negative reviews. Fig. 3*c* shows a plot of the frequency of the tokens in the vocabulary in a log-log scale, which conforms to Zipf's law (Manning and Schütze, 2000).

5 Experiments

We explored using the dataset for two tasks: (a) **Sentiment polarity classification**: where the goal is to predict if the review is *positive* i.e. with rating 4 or 5, or is *negative* i.e. with rating 1 or 2; and (b)

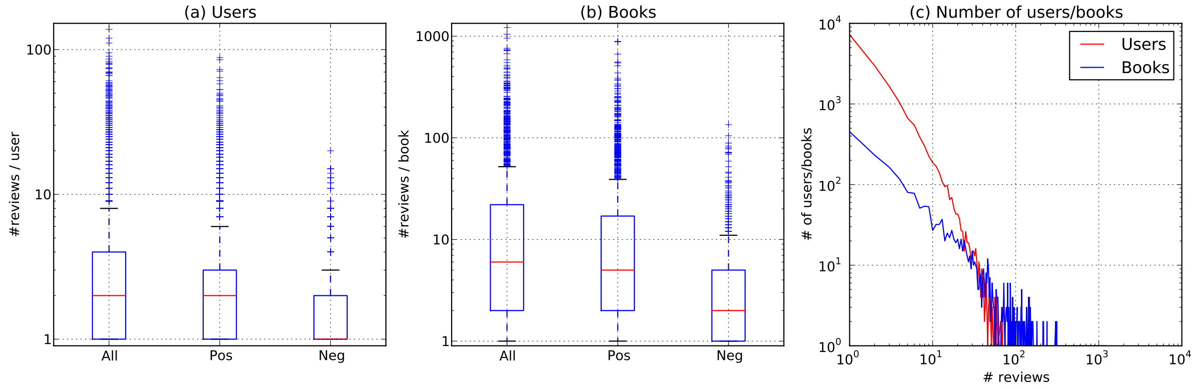

Figure 2: **Users and Books Statistics**. *(a)* Box plot of the number of reviews per user for all, positive, and negative reviews. The *red* line denotes the median, and the edges of the box the *quartiles*. *(b)* the number of reviews per book for all, positive, and negative reviews. *(c)* the number of books/users with a given number of reviews.

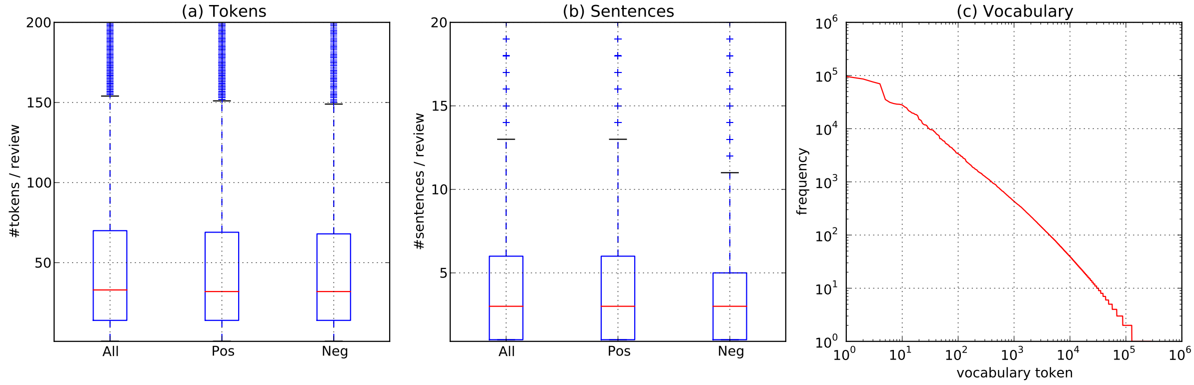

Figure 3: **Tokens and Sentences Statistics**. *(a)* the number of tokens per review for all, positive, and negative reviews. *(b)* the number of sentences per review. *(c)* the frequency distribution of the vocabulary tokens.

Rating classification: where the goal is to predict the rating of the review on a scale of 1 to 5.

To this end, we divided the dataset into separate training and test sets, with a ratio of 8:2. We do this because we already have enough training data, so there is no need to resort to cross-validation (Pang et al., 2002). To avoid the bias of having more positive than negative reviews, we explored two settings: (a) a ***balanced*** split where the number of reviews from *every* class is the same, and is taken to be the size of the smallest class (where larger classes are down-sampled); (b) an ***unbalanced*** split where the number of reviews from every class is unrestricted, and follows the distribution shown in Fig. 1. Table 2 shows the number of reviews in the training and test sets for each of the two tasks for the balanced and unbalanced splits, while Fig. 4 shows the breakdown of these numbers per class.

Tables 3-4 show results of the experiments for both tasks in both balanced/unbalanced settings. We tried different features: unigrams, bigrams, and trigrams with/without tf-idf weighting. For classifiers, we used Multinomial Naive Bayes, Bernoulli Naive Bayes (for binary counts), and Support Vector Machines. We report two measures: the *total* classification accuracy (percentage of correctly classified test examples) and *weighted* F1 measure (Manning and Schütze, 2000). All experiments were implemented in Python using scikit-learn (Pedregosa et al., 2011) and Qalsadi (available at pypi.python.org/pypi/qalsadi).

We notice that: (a) The total accuracy and weighted F1 are quite correlated and go hand-in-hand. (b) Task 1 is much easier than task 2, which is expected. (c) The unbalanced setting seems eas-

Features	Tf-Idf	Balanced			Unbalanced		
		MNB	BNB	SVM	MNB	BNB	SVM
1g	No	0.801 / 0.801	0.807 / 0.807	0.766 / 0.766	0.887 / 0.879	0.889 / 0.876	0.880 / 0.877
	Yes	0.809 / 0.808	0.529 / 0.417	0.801 / 0.801	0.838 / 0.765	0.838 / 0.766	0.903 / 0.895
1g+2g	No	0.821 / 0.821	0.821 / 0.821	0.789 / 0.789	0.893 / 0.877	0.891 / 0.873	0.892 / 0.888
	Yes	0.822 / 0.822	0.513 / 0.368	0.818 / 0.818	0.838 / 0.765	0.837 / 0.763	**0.910 / 0.901**
1g+2g+3g	No	0.821 / 0.821	0.823 / 0.823	0.786 / 0.786	0.889 / 0.869	0.886 / 0.863	0.893 / 0.888
	Yes	**0.827 / 0.827**	0.511 / 0.363	0.821 / 0.820	0.838 / 0.765	0.837 / 0.763	**0.910 / 0.901**

Table 3: **Task 1: Polarity Classification Experimental Results**. *1g* means using the unigram model, *1g+2g* is using unigrams + bigrams, and *1g+2g+3g* is using trigrams. *Tf-Idf* indicates whether tf-idf weighting was used or not. *MNB* is Multinomial Naive Bayes, *BNB* is Bernoulli Naive Bayes, and *SVM* is the Support Vector Machine. The numbers represent *total* accuracy / *weighted* F1 measure. See Sec. 5.

Features	Tf-Idf	Balanced			Unbalanced		
		MNB	BNB	SVM	MNB	BNB	SVM
1g	No	0.393 / 0.392	0.395 / 0.396	0.367 / 0.365	0.465 / 0.445	0.464 / 0.438	0.460 / 0.454
	Yes	0.402 / 0.405	0.222 / 0.128	0.387 / 0.384	0.430 / 0.330	0.379 / 0.229	0.482 / 0.472
1g+2g	No	0.407 / 0.408	0.418 / 0.421	0.383 / 0.379	0.487 / 0.460	0.487 / 0.458	0.472 / 0.466
	Yes	0.419 / 0.423	0.212 / 0.098	0.411 / 0.407	0.432 / 0.325	0.379 / 0.217	0.501 / 0.490
1g+2g+3g	No	0.405 / 0.408	0.417 / 0.420	0.384 / 0.381	0.487 / 0.457	0.484 / 0.452	0.474 / 0.467
	Yes	**0.426 / 0.431**	0.211 / 0.093	0.410 / 0.407	0.431 / 0.322	0.379 / 0.216	**0.503 / 0.491**

Table 4: **Task 2: Rating Classification Experimental Results**. See Table 3 and Sec. 5.

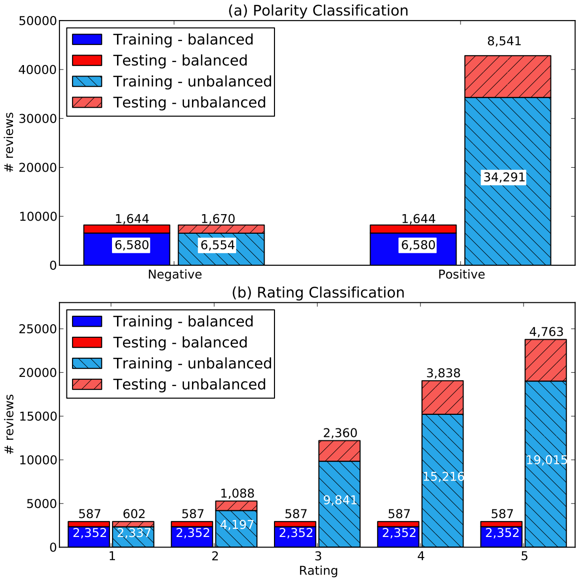

Figure 4: **Training-Test Splits**. *(a)* Histogram of the number of training and test reviews for the polarity classification task for *balanced* (solid) and *unbalanced* (hatched) cases. *(b)* The same for the rating classification task. In the balanced set, all classes have the same number of reviews as the smallest class, which is done by down-sampling the larger classes.

ier than the balanced one. This might be because the unbalanced sets contain more training examples to make use of. (d) SVM does much better in the unbalanced setting, while MNB is slightly better than SVM in the balanced setting. (e) Using more ngrams helps, and especially combined with tf-idf weighting, as all the best scores are with tf-idf.

6 Conclusion and Future Work

In this work we presented the largest Arabic sentiment analysis dataset to-date. We explored its properties and statistics, provided standard splits, and performed several baseline experiments to establish a benchmark. Although we used very simple features and classifiers, task 1 achieved quite good results (~90% accuracy) but there is much room for improvement in task 2 (~50% accuracy). We plan next to work more on the dataset to get sentence-level polarity labels, and to extract Arabic sentiment lexicon and explore its potential. Furthermore, we also plan to explore using Arabic-specific and more powerful features.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*.

Muhammad Abdul-Mageed and Mona Diab. 2012a. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Muhammad Abdul-Mageed and Mona Diab. 2012b. Toward building a large-scale arabic sentiment lexicon. In *Proceedings of the 6th International Global Word-Net Conference*.

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Muhammad Abdul-Mageed, Sandra Kübler, and Mona Diab. 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*.

Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. 2012. A machine learning approach for opinion holder extraction in arabic language. *arXiv preprint arXiv:1206.1011*.

Mohammed Korayem, David Crandall, and Muhammad Abdul-Mageed. 2012. Subjectivity and sentiment analysis of arabic: A survey. In *Advanced Machine Learning Technologies and Applications*.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*.

Christopher D. Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *EMNLP*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

M. Rushdi-Saleh, M. Martín-Valdivia, L. Ureña-López, and J. Perea-Ortega. 2011a. Bilingual experiments with an arabic-english corpus for opinion mining. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

M. Rushdi-Saleh, M. Martín-Valdivia, L. Ureña-López, and J. Perea-Ortega. 2011b. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*.