LABR: A Large Scale Arabic Book Reviews Dataset

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

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

Abstract

We introduce 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, downloaded from goodreads.com. We investigate the properties of the the dataset, and present its statistics.

Statistics

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

Sample Reviews

قرأت الرواية بعد ما عرضها علينا شخص اثق في رأيه عن الكتابة الابداعية والي اطور اكثر من اسلوبي في الكتابة . في اثناء القراءة اعجبت كثيرا بإبداع دكتور احمد خالد توفيق وعندما بدأت اتابع واقرأ كتاباته القديمة نوعا شعرت بفقر ذهني الابداعي وقررت اني اقرأ المزيد والمزيد قبل خروج اول كتاب لي على الساحة حتى لا احكم على نفسي بالاعدام

qr>t AlrwAyp bEd mA ErDhA ElyA $xS Avq fy r>yh En AlktAbp AlAbdAEyp wAny ATwr Aktr mn Aslwby fy AlktAbp . fy AvnA' AlqrA'p AEjbt kvyrA b<bdAE dktwr AHmd xAld twfyq wEndmA bd>t AtAbE wAqr> ktAbth Alqdymp nwEAF $Ert bfqr *hny AlAbdAEy wqrrt Any Aqr> Almzyd wAlmzyd qbl xrwj Awl ktAb ly ElY AlsAHp HtY lA AHkm ElY nfsy bAlAEdAm

I read the novel after it was introduced to me by somebody whose opinion about creative writing I trust, and so that I would develop my style in writing. During reading, I liked very much the creativity of Doctor Ahmed Khalid Tawfik, and when I started to follow and read his old writings, I felt how poor my creative mind is, and I decided that I would read more and more before I have my first book out in the field so that I would not doom myself

---

حتى الان ,استغرب اسلوب الكاتب -ليس سيئا -و لكن يبدو و كأنه يتعمد التعقيد ليضفي بعض البلاغة في وصف اكثر الطبقات " الشعبية في مصر . بعض الجملاغات الشديدة في وصف الدانى و المشاهد ولكن ايضا في الكتاب ما يجعلك تبتسم حينما تقرأ مشهدا قد شهدته في الميكروباص من قبل او اكثر من ذلك حينما ترى نفسك و قد وصفك الكاتب وصفا دقيقا كأحد ركاب الميكروباص

HtY AlAn ,Astgrb Aslwb AlkAtb -lys sy}A -w lkn ybdw w k>nh ytEmd AltEqyd lyDfy bED AlblAgp fy wSf Akvr AlTbqAt " AlSEbyp fy mSr. bED AlmblAgAt AlSdydp fy wSf AlnAs w AlmSAhd wlkn AyDA fy AlktAb mA yjElk tbtsm HynmA tqr> mShdA qd $hdth fy AlmykrwbAS mn qbl Aw Akvr mn *lk HynmA trY nfsk w qd wSfk AlkAtb wSfA dqyqA k>Hd rkAb AlmykrwbAS

Till now I find the writer's style strange, though it is not bad, but it seems as if he intentionally makes it more complex to add some eloquence in his depiction of the lower ranks of the soceity in Egypt. Some extreme hyperboae in the description of people and scenes, but also the book contains what makes you smile when you read about a scene that you personally witnessed in the microbus before, or even more when you see yourself described accurately by the writer as a microbus passenger

---

كتاب أكثر من رائع وهو فعلا اسم على مسمى أشعر مع كل فقرة منه أن حياتي تتجدد بالفعل

ktAb >kvr mn rA}E whw fElA Asm ElY msmY >$Er mE kl fqrp mnh >n HyAty ttjdd bAlfEl

A more than fantastic book, and it really lives up to its name I feel with every paragraph that my life is actually renewing

---

عن خادمة تسافر من الفلبين إلى الكويت لا شيء جديد، والأحداث بالنسبة لي لم تكن منطقية لا أنصح بها

En xAdmp tsAfr mn Alflbyn <lY Alkwyt lA Sy' jdyd، wAl>HdAv bAlnsbp ly lm tkn mnTqyp lA >nSH bhA

About a maid that travels from the Philippines to Kuwait, nothing new, and I didn't find the plot logical, I don't recommend it

---

لاتعليق من جمالها

lAtElyq mn jmAlhA

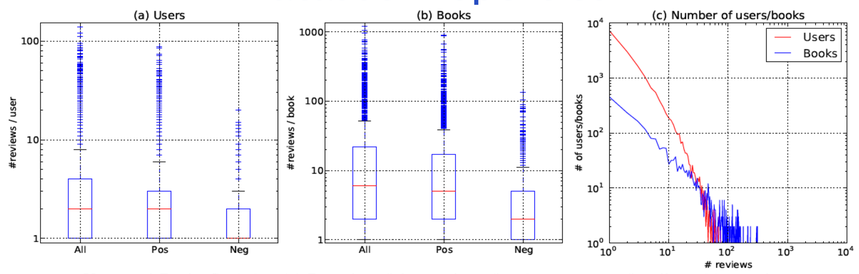
I can't comment because of its beauty

ممل جدا

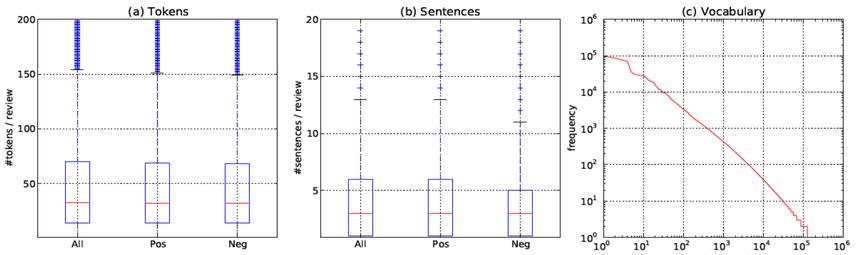
mml jdA

very boring

Dataset Properties



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.



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.

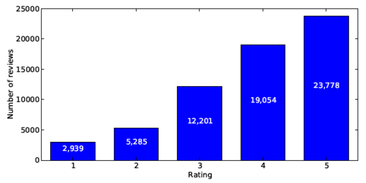
Experiments

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**

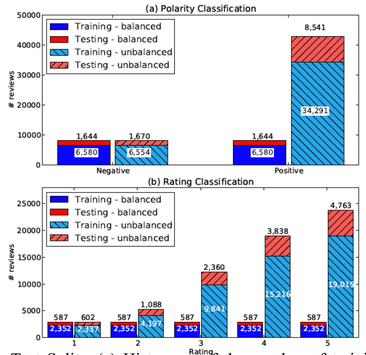
Task 1: Polarity Classification Experimental Results. 1g means using the unigram model, 1g+2g is using unigrams + bigrams, and 1g+2g+3g is using trigrams. Tf-Idf indicates whether tf-idf weighting was used or not. MNB is Multinomial Naive Bayes, BNB is Bernoulli Naive Bayes, and SVM is the Support Vector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

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**

Task 2: Rating 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.



Reviews Histogram. The plot shows the number of reviews for each rating.



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 therating classification task. In the balanced set, allclasses have the same number of reviews as thesmallest class, which is done by down-samplingthe larger classes.

Conclusion

In this work we presented the largest Arabic sentimen 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 simplefeatures and classifiers, task 1 achieved quitegood results (~90% accuracy) but there is muchroom for improvement in task 2 (~50% accuracy).

Dataset available at:
www.mohamedaly.info/datasets

Mohamed Aly and Amir Atiya. LABR: A Large Scale Arabic Book Reviews Dataset. Association of Computational Linguistics (ACL), August 2013.