ASTD: Arabic Sentiment Tweets Dataset

Mahmoud Nabil
Computer Engineering
Cairo University
Giza, Egypt
mah.nabil@cu.edu.eg

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

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

Abstract

This paper introduces **ASTD**, an Arabic social sentiment analysis dataset gathered from Twitter. It consists of about 10,000 tweets which are classified as objective, subjective positive, subjective negative, and subjective mixed. We present the properties and the statistics of the dataset, and run experiments using standard partitioning of the dataset. Our experiments provide benchmark results for 4 way sentiment classification on the dataset.

1 Introduction

Arabic sentiment analysis work is gaining large attention nowadays. This is mainly due to the need of a product that can utilize natural language processing technology to track and analyze the public mood through processing social data streams. This calls for using standard social sentiment analysis datasets. In this work we present **ASTD** (**A**rabic **S**entiment **T**weets **D**ataset) an Arabic social sentiment analysis dataset gathered from Twitter. We discuss our method for gathering and annotating the dataset, and present its properties and statistics through the following tasks: (1) 4 way sentiment classification (2) Two stage class classification; and (3) sentiment lexicon generation. The contributions in this work can be summarized as:

1. We present an Arabic social dataset of about 10k tweets for subjectivity and sentiment analysis gathered from.
2. We investigate the properties and the statistics of the dataset and provide standard splits for balanced and unbalanced settings of the dataset.
3. We present a set of benchmark experiments to the dataset to establish a baseline for future comparisons.
4. We make the dataset and the used experiments publicly available[1].

2 Related Work

The detection of user sentiment in texts is a recent task in natural language processing. This task is gaining a large attention nowadays due to the explosion in the number of social media platforms and the number of people using them. Some Arabic sentiment datasets have been collected (see Table 1). (Abdul-Mageed et al., 2014) proposed the SAMAR system that perform subjectivity and sentiment analysis for Arabic social media where they used different multi-domain datasets collected from Wikipedia TalkPages, Twitter, and Arabic forums. (Aly and Atiya, 2013) proposed LABR, a book reviews dataset collected from GoodReads. (Rushdi-Saleh et al., 2011) presented an Arabic corpus of 500 movie reviews collected from different web pages. (Refaee and Rieser, 2014) presented a manually annotated Arabic social corpus of 8,868 Tweets and they discussed the method of collecting and annotating the corpus. (Abdul-Mageed and Diab, 2014) proposed SANA, a large-scale, multi-domain, and multi-genre Arabic sentiment lexicon. The lexicon automatically extends two manually collected lexicons HUDA (4,905 entries) and SIFFAT (3,325 entries). (Ibrahim et al., 2015) built a manual corpus of 1,000 tweets and 1000 microblogs and used it for sentiment analysis task. (ElSahar and El-Beltagy, 2015) introduced four datasets in their work to build a multi-domain Arabic resource (sentiment lexicon). (Nabil et al., 2014) and (El-Sahar and El-Beltagy, 2015) proposed a semi-supervised method for building a sentiment lexicon that can be used efficiently in sentiment analysis.

[1] https://github.com/mahmoudnabil/ASTD

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519,
Lisbon, Portugal, 17-21 September 2015.

Data Set Name	Size	Source	Type	Cite
TAGREED (TGRD)	3,015	Tweets	MSA/Dialectal	(Abdul-Mageed et al., 2014)
TAHRIR (THR)	3,008	Wikipedia TalkPages	MSA	(Abdul-Mageed et al., 2014)
MONTADA (MONT)	3,097	Forums	MSA/Dialectal	(Abdul-Mageed et al., 2014)
OCA(Opinion Corpus for Arabic)	500	Movie reviews	Dialectal	(Rushdi-Saleh et al., 2011)
AWATIF	2,855	Wikipedia TalkPages/Forums	MSA/Dialectal	(Abdul-Mageed and Diab, 2012)
LABR(Large Scale Arabic Book Reviews)	63,257	GoodReads.com	MSA/Dialectal	(Aly and Atiya, 2013)
Hotel Reviews (HTL)	15,572	TripAdvisor.com	MSA/Dialectal	(ElSahar and El-Beltagy, 2015)
Restaurant Reviews (RES)	10,970	Qaym.com	MSA/Dialectal	(ElSahar and El-Beltagy, 2015)
Movie Reviews (MOV)	1,524	Elcinemas.com	MSA/Dialectal	(ElSahar and El-Beltagy, 2015)
Product Reviews (PROD)	4,272	Souq.com	MSA/Dialectal	(ElSahar and El-Beltagy, 2015)
Arabic Twitter Corpus	8,868	Tweets	MSA/Dialectal	(Refaee and Rieser, 2014)

Table 1: Arabic sentiment data sets

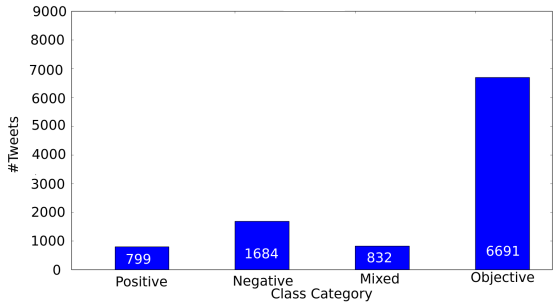

Figure 1: **Tweets Histogram**: The number of tweets for each class category. Notice the unbalance in the dataset, with much more objective tweets than positive, negative, or mixed.

Total Number of conflict free tweets	10,006
Subjective positive tweets	799
Subjective negative tweets	1,684
Subjective mixed tweets	832
Objective tweets	6,691

Table 2: Twitter dataset statistics

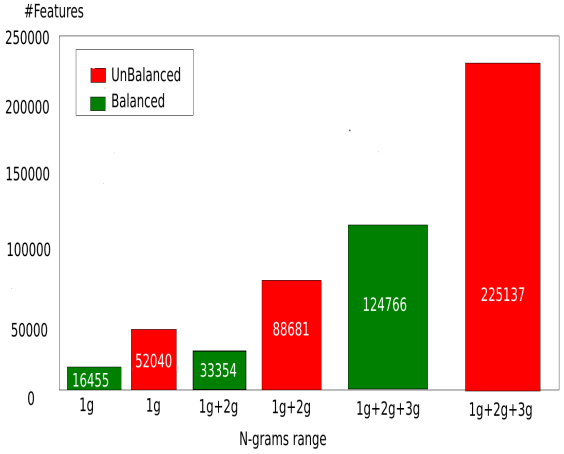

Figure 3: **Feature Counts.** Number of unigram, bigram, and trigram features per each class category.

3 Twitter Dataset

3.1 Dataset Collection

We have collected over 84,000 Arabic tweets. We downloaded the tweets over two stages: In the first stage we used SocialBakers [2] to determine the most active Egyptian Twitter accounts. This gave us a list of 30 names. We got the recent tweets of these accounts till November 2013, and this amounted to about 36,000. In the second stage we crawled EgyptTrends [3], a Twitter page for the top trending hash tags in Egypt. We got about 2500 distinct hash tags which are used again to download the tweets. We ended up obtaining about 48,000 tweets. After filtering out the non-Arabic tweets, and performing some pre-processing steps to clean up unwanted content like HTML, we ended up with 54,716 Arabic tweets.

3.2 Dataset Annotation

We used Amazon Mechanical Turk (**AMT**) service to manually annotate the data set through an API called Boto[4]. We used four tags: objective, subjective positive, subjective negative, and subjective mixed. The tweets that are assigned the same rating from at least two raters were considered as conflict free and are accepted for further processing. Other tweets that have conflict from all the three raters were ignored. We were able to label around 10k tweets. Table 2 summarizes the statistics for the conflict free ratings tweets.

3.3 Dataset Properties

The dataset has 10,006 tweets. Table 2 contains some statistics gathered from the dataset. The histogram of the class categories is shown in Fig. 1,

[2] http://www.socialbakers.com/twitter/country/egypt/

[3] https://twitter.com/EgyptTrends

[4] https://github.com/boto/boto

	Tweet	Translation	Rate
1	اكثر شعور يوجع ! ^ #لما تجوع في بيت مو بيتكم ☺**	Feeling that hurts ^ ! #To starve ina house not yours	Negative
2	محبين البرنامج يزيدوا :(	Fans of El-Bernameg are increasing :)	Positive
3	#كفاية اسفاف	#stop smallness	Negative
4	الطاقة البشرية اذا ما احسن استغلالها هى رصيدا و ليست عبئا قوتنا فى عددنا	Human energy if properly exploited is an asset and not a burden our strength in our numbers	Positive
5	احيي الشيخ حسن عبد البصير امام مسجد سيدى جابر الذي رفض تعليمات الأوقاف بنفاق مرسي في خطبة الجمعة تعلموا الاستقامة أيها #الاخوان الكاذبون	I greet Sheikh Hassan AbdelBassir Imam Sidi Gaber mosque,who refused the instructions of the endowments to hypocrite Morsi in his Friday sermon learn the integrity liars brotherhood	Mixed
6	هل تُتوج ايتكو مدريد بلقب الليجا الأحد القادم؟ #برشلونة	Is Atletico Madrid going to be crowned La Liga next Sunday? # Barcelona	Objective

Figure 2: **ASTD tweets examples**. The English translation is in the second column, the original Arabic review on the middle column, and the rating shown in right.

Number of tweets	10,006
Median tokens per tweet	16
Max tokens per tweet	45
Avg. tokens per tweet	16
Number of tokens	160,206
Number of vocabularies	38,743

Table 3: **Twitter Dataset Statistics**..

where we notice the unbalance in the dataset, with much more objective tweets than positive, negative, or mixed. Fig. 2 shows some examples from the data set, including positive, negative, mixed ,and objective tweets.

4 Dataset Experiments

In this work, we performed a standard partitioning to the dataset then we used it for the sentiment polarity classification problem using a wide range of standard classifiers to perform 4 way sentiment classification.

4.1 Data Preparation

We partitioned the data into training, validation and test sets. The validation set is used as a mini-test for evaluating and comparing models for possible inclusion into the final model. The ratio of the data among these three sets is 6:2:2 respectively.

Fig. 4 and Table 4 show the number of tweets for each class category in the training, test, and validation sets for both the balanced and unbalanced settings. Fig. 3 also shows the number of n-gram counts for both the balanced and unbalanced settings.

4.2 4 Way Sentiment Classification

We explore using the dataset for the same set of experiments presented in (Nabil et al., 2014) by applying a wide range of standard classifiers on the balanced and unbalanced settings of the dataset. The experiment is applied on both the token counts and the Tf-Idf (token frequency inverse document frequency) of the n-grams. Also we used the same accuracy measures for evaluating our results which are the weighted accuracy and the weighted F1 measure.

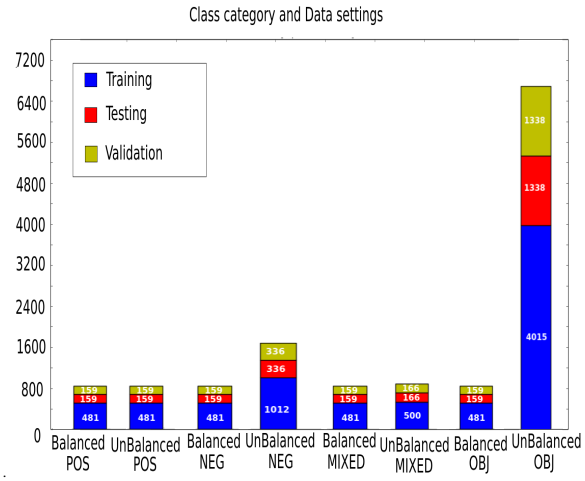

Figure 4: **Dataset Splits.** Number of tweets for each class category for training, validation, and test sets for both balanced and unbalanced settings.

Table 5 shows the result for each classifier after training on both the training and the validation set and evaluating the result on the test set (i.e. the train:test ratio is 8:2). Each cell has numbers that represent **weighted accuracy / F1 measure** where the evaluation is performed on the test set. All the experiments were implemented in Python using Scikit Learn[5]. Also the experiments were performed on a machine with Intel® Core™ i5-4440

[5] http://scikit-learn.org/

		Balanced				Unbalanced			
		Positive	Negative	Mixed	Objective	Positive	Negative	Mixed	Objective
Tweets Count	Train Set	481	481	481	481	481	1012	500	4015
	Test Set	159	159	159	159	159	336	166	1338
	Validation Set	159	159	159	159	159	336	166	1338
Features Count	unigrams	16,455				52,040			
	unigrams+bigrams	33,354				88,681			
	unigrams+bigrams+trigrams	124,766				225,137			

Table 4: **Dataset Preparation Statistics**. The top part shows the number of reviews for the training, validation, and test sets for each class category in both the balanced and unbalanced settings. The bottom part shows the number of features.

Features	Tf-Idf	Balanced			Unbalanced		
		1g	1g+2g	1g+2g+3g	1g	1g+2g	1g+2g+3g
MNB	No	0.467/0.470	0.487/0.491	**0.491/0.493**	0.686/0.604	0.684/0.590	0.682/0.584
	Yes	**0.481/0.484**	**0.491/0.492**	0.484/0.485	0.669/0.537	0.670/0.539	0.669/0.538
BNB	No	0.465/0.446	0.431/0.391	0.392/0.334	0.670/0.540	0.669/0.537	0.669/0.537
	Yes	0.289/0.184	0.255/0.110	0.253/0.107	0.669/0.537	0.669/0.537	0.669/0.537
SVM	No	0.425/0.421	0.443/0.440	0.431/0.425	0.644/0.611	0.679/0.625	0.679/0.616
	Yes	0.451/0.450	0.469/0.467	0.461/0.460	**0.687/0.620**	**0.689/0.624**	**0.691/0.626**
Passive Aggressive	No	0.421/0.422	0.447/0.443	0.439/0.435	0.639/0.609	0.664/0.621	0.671/0.616
	Yes	0.448/0.449	0.469/0.469	0.459/0.458	0.641/0.616	0.671/0.633	0.677/0.632
SGD	No	0.282/0.321	0.324/0.276	0.311/0.261	0.318/0.276	0.360/0.398	0.386/0.423
	Yes	0.340/0.295	0.409/0.382	0.415/0.388	0.664/0.557	0.671/0.557	0.669/0.551
Logistic Regression	No	0.451/0.447	0.448/0.444	0.440/0.435	0.682/0.621	0.694/0.620	0.693/0.614
	Yes	0.456/0.456	0.454/0.454	0.451/0.449	0.680/0.576	0.676/0.562	0.675/0.557
Linear Perceptron	No	0.395/0.399	0.428/0.426	0.429/0.425	0.480/0.517	0.656/0.622	0.649/0.618
	Yes	0.437/0.436	0.456/0.455	0.440/0.439	0.617/0.602	0.650/0.625	0.648/0.629
KNN	No	0.288/0.260	0.283/0.251	0.285/0.244	0.653/0.549	0.654/0.547	0.651/0.540
	Yes	0.371/0.370	0.406/0.406	0.409/0.409	0.665/0.606	0.663/0.611	0.666/0.615

Table 5: **Experiment 1: 4 way Classification Experimental Results**. *Tf-Idf* indicates whether tf-idf weighting was used or not. *MNB* is Multinomial Naive Bayes, *BNB* is Bernoulli Naive Bayes, *SVM* is the Support Vector Machine, SGD is the stochastic gradient descent and KNN is the K-nearest neighbor. The numbers represent weighted accuracy / F1 measure where the evaluation is performed on the test set. For example, 0.558/0.560 means a weighted accuracy of 0.558 and an F1 score of 0.560.

CPU @ 3.10GHz (4 cores) and 16GB of RAM.

From table 5 we can make the following observations:

1. The 4 way sentiment classification task is more challenging than the 3 way sentiment classification task. This is to be expected, since we are dealing with four classes in the former, as opposed to only three in the latter.

2. The balanced set is more challenging than the unbalanced set for the classification task. We believe that this because the the balanced set contains much fewer tweets compared to the unbalanced set. Since having fewer training examples create data sparsity for many n-grams and may therefore leads to less reliable classification.

3. SVM is the best classifier and this is consistent with previous results in (Aly and Atiya, 2013) suggesting that the SVM is reliable choice.

5 Conclusion and Future Work

In this paper we presented **ASTD** an Arabic social sentiment analysis dataset gathered from twitter. We presented our method of collecting and annotating the dataset. We investigated the properties and the statistics of the dataset and performed two set of benchmark experiments: (1) 4 way sentiment classification; (2) Two stage classification. Also we constructed a seed sentiment lexicon from the dataset. Our planned next steps include:

1. Increase the size of the dataset.

2. Discuss the issue of unbalanced dataset and text classification.

3. Extend the generated method either automated or manually.

Acknowledgments

This work has been funded by ITIDA's ITAC project number CFP-65.

References

Muhammad Abdul-Mageed and Mona T Diab. 2012. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914.

Muhammad Abdul-Mageed and Mona Diab. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Mohammed Aly and Amir Atiya. 2013. Labr: Large scale arabic book reviews dataset. In *Meetings of the Association for Computational Linguistics (ACL), Sofia, Bulgaria*.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 23–34. Springer.

Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. 2015. Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*.

Mahmoud Nabil, Mohamed A. Aly, and Amir F. Atiya. 2014. LABR: A large scale arabic book reviews dataset. *CoRR*, abs/1411.6718.

Eshrag Refaee and Verena Rieser. 2014. An arabic twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14), Reykjavik, Iceland, may. European Language Resources Association (ELRA)*.

Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. 2011. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054, October.