AraSenTi-Lexicon: A Different Approach

Hadeel AlNegheimish[1(✉)], Jowharah Alshobaili[2], Nora AlMansour[1], Rawan Bin Shiha[1], Nora AlTwairesh[1], and Sarah Alhumoud[3]

[1] College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
{halnegheimish, twairesh}@ksu.edu.sa, nora.al-mansour@hotmail.com, rawan.binshiha@gmail.com
[2] College of Computer, Qassim University, Buraydah, Saudi Arabia
j.alshobaili@qu.edu.sa
[3] Computer Science Department, Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia
sohumoud@imamu.edu.sa

Abstract. With the spread of social media, the demand for automated systems that analyze these massive amounts of data on the Web is increasing. One domain for these systems is sentiment analysis(SA). SA is designed to extract sentiment from text; this is often accomplished by using lexicons that indicate the sentiment polarity of words. While there are many English lexicons that are available, there is a lack of Arabic lexicons. In previous work, an attempt was made to generate an Arabic sentiment lexicon extracted from Twitter using the Pointwise Mutual Information (PMI) statistical method. In this paper, we extend the work by using two different statistical approaches: Chi-Square and Entropy to generate the lexicons. Intrinsic and extrinsic evaluation was conducted to compare the three lexicons. The results showed the superiority of PMI.

Keywords: Sentiment analysis · Arabic sentiment lexicon · Lexicon generation · Dialectal arabic · Twitter

1 Introduction

Since the creation of Web 2.0 technology, information exchange through the internet has increased rapidly. This new technology gave the power of sharing information not only to the data manager as its predecessor did, but also to the normal user of the web, which in turn led to the social media revolution. Social media gives people the opportunity to interact with each other directly and freely;allowing them to share news or information, express their feelings or opinions, make comments on events or articles, or even make new relationships both personal and professional. This flood of data in social media requires time and effort to read, evaluate, and analyze manually, pressing the need to have an automated system that could extract valuable insight efficiently. Accordingly, this has led to the emergence of the new research field of Sentiment Analysis (SA).

SA is concerned with classifying text into the sentiment polarity that it holds i.e. (positive, negative, neutral). SA has many beneficial aspects. For example, companies

G. Meiselwitz (Ed.): SCSM 2017, Part II, LNCS 10283, pp. 226–235, 2017.
DOI: 10.1007/978-3-319-58562-8_18

can use it to analyze customer comments and evaluate their satisfaction with the company's products. This feedback provides valuable information that could help them when making their marketing strategies [1]. SA can also be used to determine the user's desires and thus determine the appropriate advertisements based on the type of product the user has commented upon.

One approach to SA is based on using sentiment lexicons. Sentiment lexicons are compiled lists of words with their polarity (positive, negative) [2]. Sentiment intensity could also be provided; it indicates the strength in which the sentiment is being conveyed. In previous work, AlTwairesh et al. [3] generated tweet-specific Arabic sentiment lexicons using two approaches. One of these approaches utilizes the statistical measure Pointwise Mutual Information (PMI). In this paper, we use the same datasets used in [3], but propose two new statistical approaches that exploit the Entropy and Chi-Square measures. We then test and evaluate these lexiconsand compare their results with the results of PMI lexicons published in [3].

This paper is organized as follows: Sect. 2 reviews the related work on sentiment lexicon generation. Section 3 presents the details of the datasets used to generate the lexicons. Section 4 describes the new approaches used to generate the new lexicons. Section 5 details the conducted intrinsic and extrinsic evaluationof the new lexicons while Sect. 6 presents and discusses the results. Finally, we conclude the paper in Sect. 7.

2 Related Work

A sentiment lexicon contains words that are classified as positive, negative and sometimes neutral. The lexicon could contain in addition to the polarity of the word, a score that indicates the sentiment intensity. There are three approaches to generating sentiment lexicons [2]: manual approach, dictionary-based approach, and corpus-based approach. The manual approach as the name implies is done manually, but is usually done in conjunction with automated approaches as a correction step. The dictionary-based approach exploits relations found in a dictionary such as synonyms and antonyms to derive the polarity of words. Most of the works under this approach utilize WordNet e.g. [4–7]. Arabic sentiment lexicons generated using this approach e.g. [8, 9].

The corpus-based approach utilizes a corpus and a set of sentiment bearing words. Words are extracted from the corpus and compared to the set of sentiment words using different statistical methods that measure semantic similarity. Statistical approaches that are commonly used include PMI, and Chi-Square [2]. The PMI is a measure for the strength of association between two words in a corpus, i.e. the probability of the two words to co-occur in the corpus [10]. It has been adapted in sentiment analysis as a measure of the frequency of a word occurring in positive text to the frequency of the same word occurring in negative text. Turney [11]; Turney and Littman [12], was the first work that proposed to use this measurement in sentiment analysis. Other works that used this statistical measure are [13] for English and [14] for Arabic. As for the Chi-Square measure [15] used it for building a sentiment lexicon and their work was adopted in this paper also.

3 Dataset

Since we continue on the work of [3] we will use the same dataset and present here an overview of the dataset and how it was collected. Using the Twitter API, a large dataset of Arabic tweets was collected. The dataset collection was done in two phases. In the first phase, tweets that contained the emoticons ":)" (to be considered positive) and ":(" (to be considered negative) and their "lang" field was set to Arabic were collected during two months. In the second phase, a seed list of 10 Arabic positive words and 10 Arabic negative words were used as search keywords to collect tweets. Accordingly, tweets that contained the positive emoticon or positive keywords were grouped into a set that designated positive tweets and tweets that contained the negative emoticon or negative keywords were grouped into a set that designated negative tweets.

The number of collected tweets was around 6.3 million. However, due to the informal nature of Twitter data; preprocessing and cleaning was conducted on the tweets and the result after filtering and cleaning was 2.2 million Arabic tweets. Statistics of the dataset are shown in Table 1.

Table 1. Dataset statistics

	Positive tweets	Negative tweets	Total
Number of tweets collected	4068571	2272564	6341135
After cleaning and filtering	1480563	745363	2225926
Number of Tokens	21797720	12217401	34015121

4 Lexicon Generation

In this paper, we build on the previous work [3], to explore other approaches in scoring Arabic sentiment lexicons, utilizing entropy and chi-square methods.

These approaches are used to determine the intensity of the polarity of each word in the lexicon, using the frequencies of each word in positive and negative datasets, and are further detailed in the following subsections. However, they do not tell us whether the word is positive or negative. The sign of each, or direction of polarity, is determined in a uniform way, by comparing the conditional probability of the lexicon given its polarity. Concretely:

$$Sign = \begin{cases} 1 & if\ P(c|neg) < P(c|pos) \\ -1 & otherwise \end{cases} \tag{1}$$

where;

$$P(c|i) = \frac{freq(c,i)}{freq(c)} \tag{2}$$

where

c: is the word,
i: is the polarity (positive or negative).
freq(c,i) is the frequency of word *c* in dataset i:the (positive or negative).
freq(c) is the frequency of word *c* in the whole dataset.

Next, the sign is multiplied by the word score found by each of the following formula (3, 5), to determine the word's intensity.

4.1 AraSenTi-Entropy

Entropy [16] is often used in Information Theory to measure expected information content; in the case of two labels, entropy is highest when the data is evenly distributed, and lowest when all of the data is under one label. In our context, a word can either be positive or negative, so entropy can be used to measure the intensity of a word's polarity. If the entropy is high, it means that the word occurs in comparable frequency in both positive and negative text, which means that the word has weak polarity. On the other hand, if the entropy is low, it means that the word has a strong polarity, as it occurs in some sentiment significantly more than the other.

Knowing that entropy has an inverse relationship with a word's polarity, given the frequencies of words in positive and negative datasets, we find AraSenTi-Entropy lexicon scores based on the following equation:

$$Score(c) = sign * \frac{1}{-\sum_{i \in \{pos,neg\}^{p_i}} log_2 p_i} \tag{3}$$

where:

$$p_i = \frac{freq(c,i)}{freq(c)} \tag{4}$$

In the case where the word appears in one polarity only, the score is set to sign $\times$ 1, as Eq. 3 ill be undefined with the denominator being zero.

4.2 AraSenTi-ChiSq

A chi-square test is used to check the validity of some null-hypothesis by evaluating the statistical significance of the difference between observed and expected values.

In the context of sentiment analysis, the intensity of polarity of the word is determined by evaluating the null-hypothesis: "The frequency of the occurrences of a word is the same in positive and negative text". As in AraSenTi-Entropy, frequencies of words in positive and negative text are the sole determinants of scores.

The exact formula for AraSenTi-ChiSq lexicon, was based on the work of [15], and is detailed below:

$$Score(c) = X^2(c) = sign * \sum_{y \in (pos,neg)} \frac{\{freq(c,y) - \overline{freq}(c,y)\}^2}{\overline{freq}(c,y)} \quad (5)$$

where:

$$\overline{freq}(c,y) is\ the\ expected\ freq\ and\ X^2(c) \geq 0$$

Basically, the score will be the sum of square differences of frequencies normalized by the frequency under each polarity. If the null hypothesis holds, the expected value of frequency (or the frequency under the other polarity), will be equivalent to the original one, and the score will be zero (the intensity of polarity is low). In the case where a word appears under one polarity only, the denominator is set to 1 instead of 0, and the score would be most extreme.

5 Evaluation

To evaluate the performance of the generated lexicons, two evaluation methods were performed; intrinsic and extrinsic. In the intrinsic evaluation, AraSenTi-Entropy, AraSenTi-ChiSq and AraSenTi-PMI [3] lexicons were compared with each other. However, in the extrinsic evaluation, the lexicons were evaluated for their utility in classifying sentiment of three different datasets of Arabic tweets.

5.1 Intrinsic Evaluation

In this evaluation method, the three lexicons were compared to each other to determine the percentage of agreement, i.e. how many words did the lexicons agree on their polarity. Table 2 shows the number of positive and negative words used in this evaluation for each lexicon with a total of 93,295 words.

Table 2. The number of positive and negative words in the lexicons

Lexicon	Positive	Negative
AraSenTi-PMI	56434	36861
AraSenTi-ChiSq	58697	34598
AraSenTi-Entropy	56304	36991

In Table 3, the result of this evaluation is illustrated, and from it, you can notice that the highest agreement percentage was between AraSenTi-PMI [3] and AraSenTi-Entropy. In general, the agreements between the lexicons were very high.

Table 3. The percentage of agreement for the lexicons

Lexicons	Agreement
AraSenTi-PMI & AraSenTi-ChiSq	97.30%
AraSenTi-PMI & AraSenTi-Entropy	99.86%
AraSenTi-Entropy & AraSenTi-ChiSq	97.44%

5.2 Extrinsic Evaluation

We conducted an extrinsic evaluation for the three lexicons to observe the performance of the lexicons on different datasets. We evaluated the lexicons using the same datasets from the previous work which are AraSenTi-Tweet dataset [3] and two external datasets ASTD [17] and RR [18]. Information of these datasets is illustrated in Table 4.

Table 4. Datasets used in the extrinsic evaluation.

Dataset	Positive	Negative	Total
AraSenTi-Tweet	4329	5804	10133
ASTD	797	1682	2479
RR	876	1941	2817

In addition, we computed the balanced F-score (Favg), precision (P) and recall (R) to measure the performance of the lexicons for the positive and negative categories by the following formulas:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{8}$$

Where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. Then we calculated the F-score as follow:

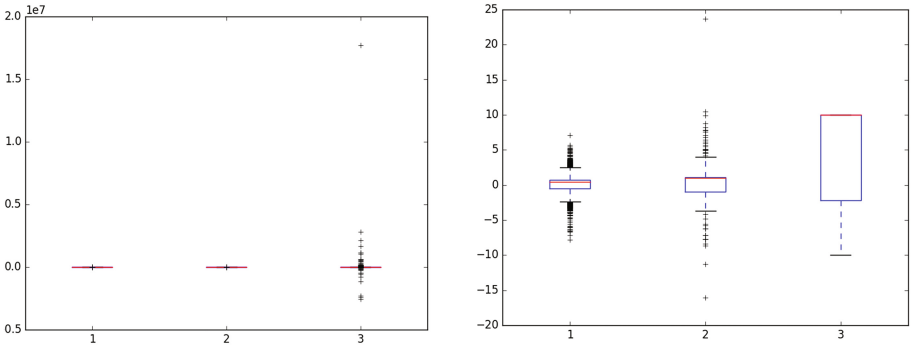
$$F_{\text{avg}} = \frac{F_{\text{pos}} + F_{\text{neg}}}{2} \tag{9}$$

For AraSenTi-Entropy and AraSenTi-ChiSq lexicons we followed the same approach used with AraSenTi-PMI [3] lexicon in the previous work. We classified the tweets into positive or negative according to the sum of the sentiment score of the words in each tweet. The threshold we used to classify the data into positive or negative

was initially zero. As such, if the sum of the sentiment score of the words in a tweet is greater than zero then the tweet is considered to be a positive tweet. Otherwise the tweet is considered to be a negative tweet. Additionally, we experimented with other values of the threshold to get the best results, we used 0, 0.5 and 1.

6 Results and Discussion

First, it is worth mentioning that the scores for AraSenTi-ChiSq lexicons were clipped to remain between −10 and 10, as there were a few outliers too great in magnitude, affecting its performance. Figure 1 shows the distribution of scores for the different lexicons before and after clipping. In Fig. 1(a), we observe that outliers in ChiSq are great in magnitude, reaching a max of around 1.8 × e7. In Fig. 1(b), after clipping the AraSenTi-ChiSq to a min −10 and max 10. Note the similarities between the plots for PMI and Entropy.



(a) (b)

Fig. 1. (a) Boxplot of the distribution of the raw scores, as per the formulas defined previously. (b) Shows the distribution of scores after clipping ChiSq, where 1,2,3 is PMI, Entropy, and ChiSq lexicons, respectively.

The results of classifying the datasets using this simple approach with varying levels of threshold, θ = [0, 0.5, 1], are displayed in Tables 5, 6 and 7 respectively. It is evident that AraSenTi-PMI lexicon performs best regardless of the chosen threshold (with max F_{avg} = 85.22%) for the AraSenTi dataset, and AraSenTi-Entropy close behind it in all experiments. AraSenTi-ChiSq has little variation across experiments, indicating that the differences in thresholds chosen are negligible to the sum of chi-square scores which determines the class. AraSenTi-ChiSq hasworse performance overall, but it is most drastic in the AraSenTi dataset (with max F_{avg} = 71.82%).

Table 5. Results with theta 0.

Lexicon	Dataset	Positive			Negative			F_{avg}
		P	R	F	P	R	F	
AraSenTi-PMI	AraSenTi	86.45	80.07	83.14	84.85	89.89	87.3	**85.22**
	ASTD	38.64	73.4	50.63	78.03	44.77	56.9	**53.77**
	RR	50.26	66.89	57.39	82.43	70.12	75.78	**66.59**
AraSenTi-Entropy	AraSenTi	83.17	78.86	80.96	83.66	87.15	85.37	83.17
	ASTD	37.76	71.64	49.45	76.63	44.05	55.94	52.7
	RR	46.86	63.93	54.08	80.52	67.28	73.31	63.7
AraSenTi-ChiSq	AraSenTi	74.81	59.19	66.09	71.86	83.95	77.44	71.77
	ASTD	36.2	67.63	47.16	73.94	43.52	54.79	50.98
	RR	45.11	56.85	50.3	77.93	68.78	73.07	61.69

Table 6. Results with theta 0.5.

Lexicon	Dataset	Positive			Negative			F_{avg}
		P	R	F	P	R	F	
AraSenTi-PMI	AraSenTi	88.92	75.29	81.54	82.28	92.45	87.07	84.31
	ASTD	39.21	64.99	48.91	75.91	52.26	61.9	55.41
	RR	53.7	57.19	55.39	80.1	77.74	78.9	67.15
AraSenTi-Entropy	AraSenTi	86.08	74.7	79.99	81.59	90.27	85.71	82.85
	ASTD	37.7	63.61	47.34	74.43	50.18	59.95	53.65
	RR	49.38	54.57	51.85	78.47	74.76	76.57	64.21
AraSenTi-ChiSq	AraSenTi	74.98	59.13	66.12	71.87	84.11	77.51	71.82
	ASTD	36.14	67.25	47.01	73.8	43.7	54.89	50.95
	RR	45.15	56.85	50.33	77.95	68.83	73.11	61.72

Table 7. Results with theta 1.

Lexicon	Dataset	Positive			Negative			F_{avg}
		P	R	F	P	R	F	
AraSenTi-PMI	AraSenTi	70.61	79.66	79.99	94.62	86.69	83.18	70.61
	ASTD	57.72	47.32	74.7	59.16	66.03	56.68	57.72
	RR	49.43	53.13	78.53	83.46	80.92	67.03	49.43
AraSenTi-Entropy	AraSenTi	87.03	73.13	79.48	80.83	91.23	85.72	82.6
	ASTD	39.09	61.61	47.83	74.98	54.52	63.13	55.48
	RR	49.89	50.8	50.34	77.61	76.97	77.29	63.82
AraSenTi-ChiSq	AraSenTi	75.13	58.95	66.06	71.83	84.29	77.56	71.81
	ASTD	36.22	67.25	47.08	73.87	43.88	55.06	51.07
	RR	45.13	56.62	50.23	77.88	68.93	73.13	61.68

Table 8. Performance of ChiSq before clipping, it was invariant across experiments

	AraSenTi-ChiSquare (Before Clipping)						
	Positive			Negative			F_{avg}
	P	R	F	P	R	F	
AraSenTi	70.67	79.95	75.02	81.94	73.27	77.36	76.19
ASTD	32.88	57.21	41.76	68.77	44.65	54.15	47.96
RR	38.09	49.09	42.9	73.58	63.99	68.45	55.68

Table 8 shows the performance of ChiSq before clipping, which was static across experiments. We can see that aside from AraSenTiFavg, clipping the scores improved its performance. The degradation in AraSenTi dataset can be attributed to the loss of relative polarity for words with scores greater than the limits.

All lexicons perform best on AraSenTi dataset, with a difference of 20 points or more in Favg. AraSenTi lexicons capture the idiosyncrasies of Twitter data, which apparently does not map well to other benchmark datasets, which may contain modern standard Arabic or other dialects.

For AraSenTi-PMI and AraSenTi-Entropy, the effect of varying threshold decreases the F_{avg} of AraSenTi dataset, but improves it for the other datasets: ASTD and RR. This is expected since the lexicons, which had been extracted from the AraSenTi dataset, have zero-median scores (as can be seen from box plots above). Furthermore, raising the threshold decreases the number of false positives, which increases the true negatives. The amount of negative data in both ASTD and RR far exceeds the amount of positive data, so such an effect is desirable

7 Conclusion

In this paper, we attempted to address a gap of the lack of Arabic sentiment lexicons that are generated from Twitter data. A previous attempt was achieved by exploiting the PMI statistical measure in [3]. New statistical approaches were investigated, these are: ChiSquare and Entropy. Intrinsic and extrinsic evaluations were conducted on the three lexicons. The results show that the performance of the lexicon that was generated using PMI outperforms other lexicons. However, the accuracy achieved from the other lexicons on the experimental datasets was very satisfying.

Acknowledgments. This work was partially funded by Deanship of Scientific Research at Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia on 2015 with grant number 360911.

References

1. Heerschop, B., Hogenboom, A., Frasincar, F.: Sentiment lexicon creation from lexical resources. In: Abramowicz, W. (ed.) BIS 2011. LNBIP, vol. 87, pp. 185–196. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21863-7_16
2. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, Cambridge (2015)
3. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A.: AraSenTi: large-scale twitter-specific arabic sentiment lexicons. In: Proceedings of the 54th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Berlin, Germany (2016)
4. Kamps, J.: Using wordnet to measure semantic orientations of adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (2004)
5. Williams, G.K., Anand, S.S.: Predicting the polarity strength of adjectives using WordNet. In: Third International AAAI Conference on Weblogs and Social Media (2009)
6. Rao, D., Ravichandran, D.: Semi-supervised polarity lexicon induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 675–682. Association for Computational Linguistics (2009)
7. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, pp. 2200–2204 (2010)
8. Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W.: A large scale Arabic sentiment lexicon for Arabic opinion mining. ANLP 2014. 165 (2014)
9. Eskander, R., Rambow, O.: SLSA: a sentiment lexicon for standard Arabic. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2545–2550. ACL, Lisbon, Purtogal (2015)
10. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguistics. **16**, 22–29 (1990)
11. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
12. Turney, P., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. National Research Council Canada, NRC Institute for Information Technology; National Research Council Canada (2002)
13. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. J. Artif. Intell. Res. **50**, 723–762 (2014)
14. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How translation alters sentiment. J. Artif. Intell. Res. **54**, 1–20 (2015)
15. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In: EMNLP-CoNLL, pp. 1075–1083 (2007)
16. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier, New York (2011)
17. Nabil, M., Aly, M., Atiya, A.F.: ASTD: Arabic sentiment tweets dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2515–2519 (2015)
18. Refaee, E., Rieser, V.: An Arabic twitter corpus for subjectivity and sentiment analysis. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, European Language Resources Association (ELRA), Reykjavik, Iceland (2014)