AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons

Nora Al-Twairesh[1,2], Hend Al-Khalifa[2], AbdulMalik Al-Salman[1]

Computer Science Department[1], Information Technology Department[2]
College of Computer and Information Sciences
King Saud University
{twairesh,hendk,salman@ksu.edu.sa}

Abstract

Sentiment Analysis (SA) is an active research area nowadays due to the tremendous interest in aggregating and evaluating opinions being disseminated by users on the Web. SA of English has been thoroughly researched; however research on SA of Arabic has just flourished. Twitter is considered a powerful tool for disseminating information and a rich resource for opinionated text containing views on many different topics. In this paper we attempt to bridge a gap in Arabic SA of Twitter which is the lack of sentiment lexicons that are tailored for the informal language of Twitter. We generate two lexicons extracted from a large dataset of tweets using two approaches and evaluate their use in a simple lexicon based method. The evaluation is performed on internal and external datasets. The performance of these automatically generated lexicons was very promising, albeit the simple method used for classification. The best F-score obtained was 89.58% on the internal dataset and 63.1-64.7% on the external datasets.

1 Introduction

The past decade has witnessed the proliferation of social media websites which has led to the production of vast amounts of unstructured text on the Web. This text can be characterized as objective, i.e. containing facts, or subjective i.e. containing opinions and sentiments about entities. Sentiment Analysis (SA) is the research field that is concerned with identifying opinions in text and classifying them as positive, negative or neutral. SA of English has been thoroughly researched; however research on SA of Arabic has just flourished.

Arabic is ranked fourth among languages on the web although it is the fastest growing language on the web among other languages (Internet World Stats, 2015). Arabic is a morphologically rich language where one lemma can have hundreds of surface forms; this complicates the tasks of SA. Moreover, the Arabic language has many variants. The formal language is called Modern Standard Arabic (MSA) and the spoken language differs in different Arabic countries producing numerous Arabic dialects sometimes called informal Arabic or colloquial Arabic. The language used in social media is known to be highly dialectal (Darwish and Magdy, 2014). Dialects differ from MSA phonologically, morphologically and syntactically and they do not have standard orthographies (Habash, 2010). Consequently, resources built for MSA cannot be adapted to dialects very well.

The informal language used in social media and in Twitter in particular makes the SA of tweets a challenging task. The language on social media is known to contain slang, nonstandard spellings and evolves by time. As such sentiment lexicons that are built from standard dictionaries cannot adequately capture the informal language in social media text. Therefore, in this paper we propose to generate Arabic sentiment lexicons that are tweet-specific i.e. generated from tweets. We present two approaches to generating Arabic sentiment lexicons from a large dataset of 2.2 million tweets. The lexicons are evaluated on three datasets, one internal dataset extracted from the larger dataset of tweets and two external datasets from the literature on Arabic SA. Moreover, the lexicons are compared to an external Arabic lexicon generated also from tweets. A simple lexicon-based method is used to evaluate the lexicons.

This paper is organized as follows: Section 2 reviews the related work on sentiment lexicon generation. Section 3 describes the details of the datasets used to generate the lexicons and how they were collected. Section 4 presents the approaches used to generate the lexicons. Section 5 details the experimental setup while Section 6 presents and analyzes the results. Finally, we

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 697–705,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

conclude the paper and present potential future work in Section 7.

2 Related Work

Words that convey positive or negative sentiment are fundamental for sentiment analysis. Compiling a list of these words is what is referred to as **sentiment lexicon generation**. There are three approaches to generate a sentiment lexicon (Liu, 2012): *manual approach, dictionary-based approach, and corpus-based approach*. The **manual approach** is usually not done alone since it is time consuming and labor intensive. It is used however, in conjunction with automated approaches to check the correction of the resulting lexicons from these approaches. In this section we review popular English and Arabic sentiment lexicons in the literature.

2.1 English Sentiment Lexicons

In the **dictionary based approach** as the name implies a dictionary is used by utilizing the synonym and antonym lists that are associated with dictionary words. The technique starts with a small set of sentiment words as seeds with known positive or negative orientations. The seed words are looked up in the dictionary then their synonyms and antonyms are added to the seed set and a new iteration starts. The process ends when no new words are found. A manual inspection is usually done after the process ends to correct errors. A majority of studies under this approach used the WordNet with different approaches for expanding the list such as distance-based measures (Kamps, 2004; Williams and Anand, 2009) and graph-based methods (Blair-Goldensohn et al., 2008; Rao and Ravichandran, 2009). Pioneering work in this approach is the construction of SentiWordNet by (Esuli and Sebastiani, 2005). Initially, they started with a set of positive seeds and a set of negative seeds then expanded the sets using the synonym and antonym relations in WordNet. This formed a training set which they used in a supervised learning classifier and applied it to all the glosses in WordNet, the process is run iteratively. Then in a following attempt (Esuli and Sebastiani, 2006), a committee of classifiers based on the previous method were used to build SentiWordNet which contains terms that are associated with three scores for objectivity, positivity and negativity, where the sum of the scores is 1. The latest version is SentiWordNet 3.0 (Baccianella et al., 2010).

As for **corpus-based approaches**, the words of the lexicon are extracted from the corpus using a seed list of known sentiment words and different approaches to find words of similar or opposite polarity. One of the earliest work in this approach was that of (Hatzivassiloglou and McKeown, 1997), where they utilized connectives e.g. and, but, etc. between adjectives in a corpus to learn new sentiment words not in the seed list. Turney, (2002); Turney and Littman, (2002) used the once popular AltaVista search engine to find the sentiment of a certain word through calculating the association strength between the word and a set of positive words minus the association strength between the word and a set of negative words. The association strength was measured using Pointwise-Mutual Information (PMI). The result is the sentiment score of the word, if it is positive this means the word is strongly associated with positive polarity and as such its polarity will be positive and if it is negative the word's polarity will be negative. The magnitude indicates the sentiment intensity of the word. We used PMI to generate one of the lexicons in this paper.

After the emergence of sentiment analysis as an evolving research field, several lexicons were constructed according to the approaches mentioned above. In the Bing Liu's lexicon (Hu and Liu, 2004), which falls under the dictionary-based method, the WordNet was exploited to infer the semantic orientation of adjectives extracted from customer reviews. The lexicon only provides the prior polarity of words: positive or negative, the sentiment intensity of the words was not calculated. Another popular sentiment lexicon is the MPQA subjectivity lexicon (Wilson et al., 2005) which was constructed by manually annotating the subjective expressions in the MPQA corpus. The words were annotated with four tags: positive, negative, both and neutral then further classified as strong or weak to denote intensity. We use these two lexicons in the generation of the other lexicon in this paper.

With the proliferation of social media websites, the need for lexicons that can capture the peculiarities of social medial language emerges. As such, many solutions for sentiment analysis of social media and Twitter in particular initiate by developing sentiment lexicons that are extracted from Twitter (Tang et al., 2014; Kiritchenko et al., 2014).

2.2 Arabic Sentiment Lexicons

Generating sentiment lexicons for Arabic has gained the interest of the research community lately. Consequently, we found several efforts for generating these lexicons. A recent effort to build a large scale multi-genre multi dialect Arabic sentiment lexicon was proposed by (Abdul-Mageed and Diab, 2014). However, it covers only two dialects: Egyptian and Levantine and is not yet fully applied to SSA tasks. Badaro et al., (2014) constructed ArSenL, a large scale Arabic sentiment lexicon. They relied on four resources to create ArSenL: English WordNet (EWN), Arabic WordNet (AWN), English SentiWordNet (ESWN), and SAMA (Standard Arabic Morphological Analyzer). Two approaches were followed producing two different lexicons: the first approach used AWN, by mapping AWN entries into ESWN using existing offsets thus producing ArSenL-AWN. The second approach utilizes SAMA's English glosses by finding the highest overlapping synsets between these glosses and ESWN thus producing ArSenL-Eng. Hence ArSenL is the union of these two lexicons. Although this lexicon can be considered as the largest Arabic sentiment lexicon developed to date, it is unfortunate that it only has MSA entries and no dialect words and is not developed from a social media context which could affect the accuracy when applied on social media text.

Following the example of ArSenL, the lexicon SLSA (Sentiment Lexicon for Standard Arabic) (Eskander and Rambow, 2015) was constructed by linking the lexicon of an Arabic morphological analyzer Aramorph with SentiWordNet. Although the approach is very similar to ArSenL, since both use SentiWordNet to obtain the scores of words, the linking algorithm used to link the glosses in Aramorph with those in SentiWordNet is different. SLSA starts by linking every entry in Aramorph with SentiWordNet if the one-gloss word and POS match. Intrinsic and extrinsic evaluations were performed by comparing SLSA and ArSenL which demonstrated the superiority of SLSA. Nevertheless, SLSA like ArSenL does not include dialect words and cannot accurately analyze social media text.

Mohammad et al., (2015), generated three Arabic lexicons from Twitter. Three datasets were collected from Twitter: the first was tweets that contained the emoticons:":)" and ":(", the second was tweets that contained a seed list of positive and negative Arabic words as hashtags and the third was also from tweets that contained Arabic positive and negative words as hashtags but these were dialectal words. Then using PMI three lexicons were generated from these datasets: Arabic Emoticon Lexicon, Arabic Hashtag Lexicon and Dialectal Arabic Hashtag Lexicon. Our approach in generating one of the lexicons is very similar and thus we use one of their lexicons in the experiments to compare with our lexicons. The best performing lexicon was the Dialectal Arabic Hashtag Lexicon therefore we use it in this paper to compare and evaluate our lexicons.

3 Dataset Collection

We followed the approaches in previous work on SA of English Twitter to collect the datasets. As in (Go et al., 2009; Pak and Paroubek, 2010) we utilized emoticons as noisy labels to construct the first dataset EMO-TWEET. Tweets containing the emoticons: ":)" and ":(" and the rule "lang:ar" (to retrieve Arabic tweets only) were collected during November and December 2015. The total number of Tweets collected is shown in Table 1.

Davidov et al., (2010) and Kiritchenko et al., (2014) used hashtags of sentiment words such as #good and #bad to create corpora of positive and negative tweets, we adopted a similar method to theirs. Initially, we tried collecting tweets that contain Arabic sentiment words with hashtags but the search results were too low. We designated this result to a cultural difference in using hashtags between the western and eastern societies. Arabs do not use hashtags in this way. Accordingly we opted to use the sentiment words as keywords without the hashtag sign and the number of search results was substantial. Tweets containing 10 Arabic words having positive polarity and 10 Arabic words having negative polarity were collected during January 2016. The keywords are in Table 2 and the number of tweets collected in Table1. These results constitute our second dataset KEY-TWEET.

Retweets, tweets containing URLs or media and tweets containing non-Arabic words were all excluded from the dataset. The reason for excluding tweets with URLs and media is that we found that most of the tweets that contain URLS and media were spam. We also noticed that although we had specified in the search query that the fetched tweets should be in Arabic "lang:ar" some of the tweets were in English and other languages. So we had to add a filter to eliminate tweets with non-Arabic characters.

In total, the number of collected tweets was around 6.3 million Arabic tweets in a time span of three months. After filtration and cleaning of the tweets, the remaining were 2.2 million tweets.

	EMO-TWEET		*KEY-TWEET*	
	Positive Emoticon :)	*Negative Emoticon :(*	*Positive keywords*	*Negative keywords*
Total number of tweets collected	2,245,054	1,272,352	1,823,517	1,000,212
After cleaning and filtering	1,033,393	407,828	447,170	337,535
Number of Tokens	12,739,308	5,082,070	9,058,412	7,135,331

Table 1: Number of collected tweets, number of tweets in datasets after cleaning and filtering and number of tokens in each dataset.

Positive Keywords	*English Translation*	*Negative Keywords*	*English Translation*
سعادة sEAdp	Happiness	محزن mHzn	Sad
خير xyr	Good	مؤسف m&sf	Regrettable
تفاؤل tfA&l	Optimism	للأسف ll>sf	Unfortunately
أعجبني >Ejbny	I like it	فاشل fA$l	Failing, unsuccessful
نجاح njAH	Success	تشاؤم t$A&m	Pessimism
فرح frH	Joy	سيء sy'	Bad
إيجابي <yjAby	Positive	سلبي slby	Negative
جيد jyd	Good	إهمال <hmAl	Negligence
ممتاز mmtAz	Excellent	خطأ xT>	Wrong
رائع rA}E	Fabulous	مؤلم m&lm	Painful

Table 2: Positive and negative keywords used to collect tweets.

4 Lexicon Generation

Two sentiment lexicons were extracted from the datasets of tweets using two different approaches. We call the first **AraSenTi-Trans** and the second **AraSenTi-PMI**. The approaches are presented in the following subsections.

4.1 AraSenTi-Trans

The datasets of tweets were processed using the MADAMIRA tool (Pasha et al., 2014). MADAMIRA is a recent effort by Pasha et al. (2014) that combines some of the best aspects of two previous systems used for Arabic NLP: MADA-Morphological Analysis and Disambiguation of Arabic (Habash and Rambow, 2005; Roth et al., 2008; Habash et al., 2009; Habash et al., 2013) and AMIRA (Diab et al., 2007). MADAMIRA, on the other hand, improves on these two systems with a solution that is more robust, portable, extensible, and faster.

The MADAMIRA tool identifies words into three types: ARABIC, NO_ANALYSIS and NON_ARABIC. This feature was used to eliminate tweets containing non-Arabic words and to distinguish MSA words from dialect words as NO_ANALYSIS words can be identified as dialect words or misspelled words or new words made up by tweepers (twitter users). According to the POS tags provided by MADAMIRA, we extracted only nouns, adjectives, adverbs, verbs and negation particles in an effort to eliminate unwanted stop words.

Then we utilized two popular English sentiment lexicons that were used in previous work on English and Arabic sentiment analysis: the Liu lexicon (Hu and Liu, 2004) and the MPQA lexicon (Wilson et al., 2005).

Most previous papers on Arabic SA that used these lexicons just translated them into Arabic, yet we tried a different approach. MADAMIRA provides an English gloss for each word identified as ARABIC, the gloss could be one, two or three words. We used this gloss to compare with the Liu lexicon and MPQA lexicon using the following heuristics:

- If all the word’s glosses are positive in both lexicons or found in one lexicon as positive and do not exist in the other lexicon: classify as positive.
- If all the word’s glosses are negative in both lexicons or found in one lexicon as negative and do not exist in the other: classify as negative.
- If the word’s glosses have different polarities in the lexicons or are (both) in MPQA: add to both list.
- Else: all remaining words are classified as neutral.

Although this approach could contain some errors, a manual check can be performed to clean up. The manual cleanup is time consuming but it is a one-time effort that requires only a few days (Liu, 2012). Accordingly we gave the automatically generated lists of positive, negative, both, and neutral words to two Arabic native speakers to review and correct the errors. We found that 5% of the neutral words were incorrectly misclassified as neutral while they were sentiment bearing words. Also 10% of the positive words were misclassified as negative, and 15% of the negative words were misclassified as positive. The lists were corrected accordingly. We can conclude that using translated English lexicons does not always give us accurate classification of polarity. This result could be due to mistranslations or cultural differences in classifying sentiment as demonstrated by (Mohammad et al., 2015; Mobarz et al., 2014; Duwairi, 2015). Accordingly, we propose a different approach to generating another lexicon in the following section.

4.2 AraSenti-PMI

The second lexicon was also generated from the dataset of tweets but through calculating the pointwise mutual information (PMI) measure for all words in the positive and negative datasets of tweets. The PMI is a measure of the strength of association between two words in a corpus, i.e. the probability of the two words to co-occur in the corpus (Church and Hanks, 1990). 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, (2002); Turney and Littman, (2002) was the first work that proposed to use this measure in sentiment analysis. They used the once popular AltaVista search engine to find the sentiment of a certain word through calculating the PMI between the word and a set of positive words minus the PMI between the word and a set of negative words. Other works that used PMI to generate sentiment lexicons can be found in (Kiritchenko et al., 2014; Mohammad et al., 2015).

The frequencies of the words in the positive and negative datasets of tweets were calculated respectively then the PMI was calculated for each as follows:

$$PMI(w,pos) = \log_2 \frac{freq(w,pos)*N}{freq(w)*freq(pos)} \quad (1)$$

where *freq(w,pos)* is the frequency of the word *w* in the positive tweets, *freq(w)* is the frequency of the word *w* in the dataset, *freq(pos)* is the total number of tokens in the positive tweets and *N* is the total number of tokens in the dataset. The PMI of the word associated with negative tweets is calculated in the same way *PMI(w,neg).* The sentiment score for word *w* will be:

$$Sentiment\ Score(w)=PMI(w,pos)-PMI(w,neg) \quad (2)$$

This was calculated for all words that occurred in the dataset five times or more, the reason for this is that the PMI is a poor estimator of low-frequency words (Kiritchenko et al., 2014), so words occurring less than 5 times were excluded. Also for words that are found in the set of positive tweets but not in the set of negative tweets or vice versa, Equation 2 would give us a sentiment score of ∞, which would highly affect the calculation of the sentiment of the whole tweet. Since the absence of a word from the negative dataset does not require that the word's sentiment is positive or vice versa; as such we calculated the sentiment score of such words as in Equation 1, *PMI(w,pos)* for words occurring only in the positive tweets and *PMI(w,neg)* for words occurring only in the negative tweets.

4.3 Lexicons Coverage

The number of positive and negative entries in each of the lexicons is shown in Table 3. The details of the lexicon of (Mohammad et al., 2015) are also shown since this lexicon will be used in the experiments in the following section for evaluation and comparison purposes. Mohammad et al., (2015) generated three lexicons, however they demonstrated that the Dialectal Arabic Hashtag Lexicon (DAHL) gave the best results and accordingly we use this lexicon in the experiments in this paper. From Table 3, we can see the high coverage of the generated lexicons AraSenti-Trans and AraSenti-PMI when compared to DAHL. In addition we manually examined the three lexicons of (Mohammad et al., 2015) and found that they were not cleaned. They contained non-Arabic words and hashtags that do not convey sentiment. This put a question mark on the validity of the lexicons and the number of entries reported. Our datasets were cleaned from non-Arabic words and punctuation, so the generated lexicons all contain valid Arabic words.

Lexicon	*Positive*	*Negative*	*Total*
AraSenti-Trans	59,525	71,817	131,342
AraSenti-PMI	56,938	37,023	93,961
DAHL	11,947	8,179	20,126

Table 3: Details of the generated lexicons and the lexicon they will be compared to.

5 Evaluation

To evaluate the performance of the tweet-specific lexicons, we performed a set of experiments using a simple lexicon-based approach, hence no training and/or tuning is required. We performed a two-way classification on the datasets (positive or negative). We leave the problem of three and four way classification (positive, negative, neutral, mixed) for future work. We evaluated the generated lexicons on a dataset of 10,133 tweets extracted from the larger datasets of tweets EMO-TWEET and KEY-TWEET. The tweets were manually annotated by three annotators that are Arabic native speakers. The conflict between annotators was resolved by majority voting. We will call this dataset AraSenTi-Tweet. We also evaluated the generated lexicons on two external datasets of tweets: ASTD by (Nabil et al., 2015) and RR by (Refaee and Rieser, 2014). We extracted only the tweets that were labeled as positive or negative from these datasets. The details of all the datasets used in the experiments are illustrated in Table 4. We plan to release the dataset and the generated lexicons for the public.

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

Table 4: Datasets used in the evaluation of the generated lexicons.

Negation significantly affects the sentiment of its scope and consequently affects the evaluation of the lexicons. Accordingly, we propose to evaluate the generated lexicons in two settings: with and without negation handling. We also compare the performance of the generated lexicons with a lexicon that was generated in a very similar approach to one of the lexicons.

Since the datasets are unbalanced, we will report the performance measures of the macro-averaged F-score (F_{avg}), precision (P) and recall (R) of the positive and negative classes as follows:

$$P= TP/(TP+FP) \quad (3)$$

$$R=TP/(TP+FN) \quad (4)$$

$$F=2*PR/P+R \quad (5)$$

where in the case of the positive class: TP is the number of positive tweets classified correctly as positive (true positive), FP is the number of negative tweets falsely classified as positive (false positive), and FN is the number of positive tweets falsely classified as negative (false negatives). The same holds for the negative class. Then the F-score is calculated as:

$$F_{avg} = \frac{F_{pos}+F_{neg}}{2} \quad (6)$$

5.1 Setup A: No Negation Handling

For the **AraSenTi-Trans** lexicon, we use the simple method of counting the number of positive and negative words in the tweet and whichever is the greatest denotes the sentiment of the tweet. The results of applying this method on the different datasets are illustrated in Table 5.

As for the **AraSenTi-PMI** lexicon, the sentiment score of all words in the tweet were summed up. The natural threshold to classify the data into positive or negative would be zero, since positive scores denote positive sentiment and negative scores denote negative sentiment. However, according to (Kiritchenko et al., 2014) other thresholds could give better results. Consequently, we experimented with the value of this threshold. We set it to 0, 0.5,and 1 and found that the best results were obtained when setting the threshold to 1. As such if the sum of the sentiment scores of the words in a tweet is greater than one, then the tweet is classified as positive, otherwise the tweet is classified as negative.

5.2 Setup B:Negation Handling

We also experimented with handling negation in the tweet, by compiling a list of negation particles found in the tweets and checking if the tweet contains a negation particle or not.

For the **AraSenTi-Trans** lexicon, if the tweet contains a negation particle and a positive word, we do not increment the positive word counter. However, for tweets containing negative words and negation particles we found that not incrementing the negative word counter degraded the accuracy, so we opted to increment the negative word counter even if a negation particle is found in the tweet.

Moreover, we experimented with adjusting the score of negation particles in the **AraSenTi-PMI** lexicon. After several experiments, we found that adjusting the score of the negation particles to -1 was the setting that gave the best performance.

6 Discussion and Results

The results of the first experimental setup for the two generated lexicons AraSenti-Trans and AraSenti-PMI are presented in Table 5. For the RR dataset and AraSenti-Tweet dataset, the superiority of the AraSenti-PMI lexicon is evident. The F_{avg} of applying the AraSenti-PMI lexicon on the RR dataset is 63.6% while the F_{avg} of applying the AraSenti-PMI lexicon on the AraSenti-Tweet dataset is 88.92%. As for the ASTD dataset, applying the AraSenti-Trans lexicon gave better results with an F_{avg} of 59.8%.

In Table 6, the results of the lexicon-based method with negation handling are presented. The results of using the DAHL lexicon on the same datasets are also reported for comparison.

First of all, the effect of negation handling on performance is significant, with increases of (1-4%) on all datasets. Although the two lexicons AraSenti-Trans and AraSenti-PMI handled negation differently but the increase for every dataset was almost the same: the ASTD dataset +4%, the RR dataset +1% and the AraSenti-Tweet dataset +2% and +1% respectively.

When comparing the performance of the generated lexicons AraSenti-Trans and AraSenti-PMI with the DAHL lexicon, we find that our lexicons presented better classification results on all datasets.

Finally, although the two lexicons were extracted from the same dataset, we find that their performance varied on the different datasets. The best performance for the ASTD dataset was when the AraSenti-Trans lexicon was used. However, the best performance for the RR and AraSenti-Tweet datasets was when the AraSenti-PMI lexicon was used. Moreover, albeit the simple lexicon-based method used in the evaluation, we find that the performance is encouraging. Several enhancements could be made such as incorporating Arabic valence shifters and certain linguistic rules to handle them.

	Lexicon									
DataSet	AraSenti-Trans					AraSenti-PMI				
	Positve		Negative		F_{avg}	Positve		Negative		F_{avg}
	P	R	P	R		P	R	P	R	
ASTD	43.92	90.21	90.74	45.42	**59.80**	37.24	77.79	78.26	37.87	50.70
RR	40.66	89.95	89.99	40.75	56.05	46.01	73.74	83.72	60.95	**63.60**
AraSenti-Tweet	63.14	95.43	94.48	58.44	74.11	85.73	89.37	91.81	88.9	**88.92**

Table 5: Results of the first experimental setup without negation handling on the generated lexicons AraSenti-Trans and AraSenti-PMI.

	Lexicon														
DataSet	AraSenti-Trans					AraSenti-PMI					DAHL				
	Positve		Negative		F_{avg}	Positve		Negative		F_{avg}	Positve		Negative		F_{avg}
	P	R	P	R		P	R	P	R		P	R	P	R	
ASTD	46.24	86.32	89	52.44	**63.10**	38.06	56.59	73.26	56.36	54.61	36.4	43.16	70.47	64.27	53.36
RR	41.31	86.3	87.84	44.67	57.55	52.03	49.77	77.77	79.29	**64.70**	38.06	38.58	72.11	71.66	55.10
AraSenti-Tweet	66.27	90.76	90.49	65.54	76.31	91.16	84.57	89.08	93.88	**89.58**	76.35	62.88	75.53	85.48	74.58

Table 6: Results of the second experimental setup with negation handling on the generated lexicons AraSenti-Trans and AraSenti-PMI and on the external lexicon DAHL

7 Conclusion

In this paper, two large-scale Arabic sentiment lexicons were generated from a large dataset of Arabic tweets. The significance of these lexicons lies in their ability to capture the idiosyncratic nature of social media text. Moreover, their high coverage suggests the possibility of using them in different genres such as product reviews. This is a possible future research direction.

The performance of the lexicons on external datasets also suggests their ability to be used in classifying new datasets. However, there is much room for improvement given the simple method used in evaluation. This simple lexicon-based method could be further enhanced by incorporating Arabic valence shifters and certain linguistic rules to handle them. We also plan to make the classification multi-way: positive, negative, neutral and mixed.

Acknowledgments

This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (GSP-36-332).

References

Muhammad Abdul-Mageed and Mona Diab. 2014. SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. In *In Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. *ANLP 2014*:165.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, volume 14, pages 339–348.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Kareem Darwish and Walid Magdy. 2014. Arabic Information Retrieval. *Foundations and Trends in Information Retrieval*, 7(4):239–342.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2007. Automated methods for processing arabic text: from tokenization to base phrase chunking. *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer*.

Rehab M Duwairi. 2015. Sentiment analysis for dialectical Arabic. In *6th International Conference on Information and Communication Systems (ICICS), 2015*, pages 166–170. IEEE.

Ramy Eskander and Owen Rambow. 2015. SLSA: A Sentiment Lexicon for Standard Arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2550, Lisbon,Purtogal, September. ACL.

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, volume 6, pages 417–422.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*:1–12.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580. Association for Computational Linguistics.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*, pages 102–109.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *HLT-NAACL*, pages 426–432. Citeseer.

Nizar Y Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the as-*

sociation for computational linguistics, pages 174–181. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Internet World Stats. 2015. Internet World Stats. November.

Jaap Kamps. 2004. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Hanaa Mobarz, Mohsen Rashown, and Ibrahim Farag. 2014. Using Automated Lexical Resources in Arabic Sentence Subjectivity. *International Journal of Artificial Intelligence & Applications*, 5(6):1.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2015. How Translation Alters Sentiment. *Journal of Artificial Intelligence Research*, 54:1–20.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. ASTD: Arabic Sentiment Tweets Dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2010)*, Valleta,Malta. European Language Resources Association (ELRA).

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 117–120. Association for Computational Linguistics.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In *COLING*, pages 172–182.

Peter D Turney. 2002. 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*, pages 417–424. Association for Computational Linguistics.

Peter Turney and Michael L Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report, National Research Council Canada, NRC Institute for Information Technology; National Research Council Canada.

Gbolahan K Williams and Sarabjot Singh Anand. 2009. Predicting the Polarity Strength of Adjectives Using WordNet. In *Third International AAAI Conference on Weblogs and Social Media*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.