SentiArabic: A Sentiment Analyzer for Standard Arabic

Ramy Eskander
Ramitechs
New York, USA
eskander@ramitechs.com

Abstract

Sentiment analysis has been receiving increasing interest as it conveys valuable information in regard to people's preferences and opinions. In this work, we present a sentiment analyzer that identifies the overall contextual polarity for Standard Arabic text. The contribution of this work is threefold. First, we modify and extend SLSA; a large-scale Sentiment Lexicon for Standard Arabic. Second, we build a sentiment corpus of Standard Arabic text tagged for its contextual polarity. This corpus represents the training, development and test sets for the proposed system. Third, we build a lightweight lexicon-based sentiment analyzer for Standard Arabic (SentiArabic). The analyzer does not require running heavy computations, where the link to the lexicon is carried out through a morphological lookup as opposed to conducting a rich morphological analysis, while the assignment of the sentiment is based on a simple decision tree that uses polarity scores as opposed to a more complex machine learning approach that relies on lexical information, while negation receives special handling. The analyzer is highly efficient as it achieves an F-score of 76.5% when evaluated on a blind test set, which is the highest results reported for that set, and an absolute 3.0% increase over a state-of-the-art system that uses deep-learning models.

Keywords: Sentiment Analysis, Sentiment Lexicon, Sentiment Corpus, Standard Arabic

1. Introduction

Sentiment analysis is the process of applying computational approaches to identify attitudes, emotions and opinions in text, speech and visual data. While there is an enormous number of sentiment analysis tools for English and other common languages, Arabic has received less focus due to lack of resources, corpora and lexicons in particular, in addition to the complexity of its morphological and syntactical systems, which incurs ambiguity and requires extensive processing.

In this work, we present SentiArabic, a lightweight lexicon-based sentiment analyzer for Standard Arabic. For a given text, the system identifies the contextual polarity (*Positive* or *Negative*) of the underlying sentiment. Our main contribution in this paper can be summarized as follow:

1. We modify and extend our previous sentiment resource SLSA, a large-scale Sentiment Lexicon for Standard Arabic (Eskander and Rambow, 2015). SLSA includes 34,281 entries, where an entry consists of a lemma, a part-of-speech (POS) tag, the corresponding English gloss, and three sentiment scores; positive, negative and objective, where the objective score is calculated as *1 - (positive score + negative score)*. The new extension follows the same structure but with a higher quality and better normalization of the polarity scores.

2. We create a new sentiment corpus of Standard Arabic text where each sentence is tagged for its polarity. The corpus is divided into three sets for the training, development and test of the sentiment analyzer.

3. We build SentiArabic, a lightweight sentiment analyzer for Standard Arabic. SentiArabic is based on the lexicon in (1), and is trained and evaluated based on the corpus in (2). SentiArabic is "lightweight" in the sense that it does not perform heavy computations on the underlying text. Instead, for every word in the input text, the corresponding lexicon entry is retrieved by using a maximum-likelihood lookup of the lemmas and POS tags as opposed to running morphological analysis in order to retrieve the lemma and POS information. The system then applies a decision tree of polarity scores in order to determine the overall polarity of the underlying text as opposed to conducting a more complex machine learning approach that uses lexical information. In addition, negation receives special handling as it affects the polarity of the following context by flipping the positive and negative scores.

The rest of the paper is structured as follows. We first review the related work in section 2, and then we present the implementation of our system in section 3. The evaluation and results are discussed in section 4, before we conclude with a discussion of future work in section 5.

2. Related Work

One of the early attempts on Arabic sentiment analysis is the work presented by (Abdul-Mageed et al., 2011), where they classified subjective text into four classes; positive, negative, neutral and mixed. The system applies Support Vector Machines (SVMs) on manually annotated data extracted from the Penn Arabic Treebank (PATB) (Maamouri et al., 2004), The authors showed that using a polarity lexicon and stem lemmatization has a considerable impact on

the performance, A subsequent system is SAMAR (Abdul-Mageed et al., 2014). SAMAR is based on a corpus of social-media text annotated for subjectivity and sentiment polarity, and uses a combination of language independent and language-specific feature sets. The authors conducted different evaluations on different parts of the corpus, and obtained an upper-bound F-score of 73.6% on polarity classification.

Mourad and Darwish (2013) used the MPQA lexicon (Wilson et al., 2005) to perform subjectivity and sentiment classification of both Standard Arabic news articles and dialectal Arabic microblogs from Twitter. The system applies Naive Bayes classification and achieves F-scores of 75.3% and 78.3% for the classification of subjectivity and sentiment, respectively.

Another system was presented by El-Makky et al. (2014). The system does sentiment classification for colloquial Arabic tweets, and uses a combination of lexicon-based sentiment orientation algorithms and supervised learning using SVMs. The system achieves an average F-score of 79.5%.

Nabil et al. (2015) manually annotated a dataset of about 10K Arabic tweets, and experimented with a set of machine learning techniques. The best results were obtained by applying SVMs on a set of unigram, bigram and trigram features, yielding a weighted F-score of 62.6%.

El-Beltagy and Ali (2013) proposed an unsupervised lexicon-based system for sentiment classification of Egyptian Arabic. The system relies on the positive and negative weights in the assignment of the overall contextual polarity, and achieves an overall accuracy of 81.8% on a dataset of 500 tweets.

Sallab et al. (2015) developed deep learning models for sentiment analysis of Standard Arabic. They tested the use of Recursive Auto Encoder models on a PATB dataset annotated for sentiment polarity by Abdul-Mageed et al. (2011). The system outperforms the state-of-the-art systems evaluated on the same dataset, where it achieves an F-score of 73.5%. In this paper we compare to this system, and show how our lightweight system outperforms by an additional F-score increase of 3.0% on the same dataset.

Eskander and Rambow (2015) proposed a lexicon-based sentiment analyzer that is based on SLSA and uses Linear SVMs for binary sentiment classification (positive and negative). The system achieves an F-score of 68.6% on the test set developed by Abdul-Mageed et al. (2011). We compare to this system as well.

Other systems are domain specific such as the system presented by Aly and Atiya (2013) for book reviews, and the system developed by A.M. Alayba (2017) for reviews on health services. Both systems apply supervised learning approaches using SVMs on large annotated datasets.

Task 7 of SemEval 2016 (Kiritchenko et al., 2016) aimed at identifying the sentiment intensity (scores from 0, maximum negative, to 1, maximum positive) of Arabic phrases given a set of 200 common terms from Arabic tweets and their polarity. The best performing system was proposed by the iLab-Edinburgh team (Refaee and Rieser, 2016), where they developed a hybrid system that is a combination of a rule-based approach in addition to off-the-shelf lexicons, and a machine learning approach that uses Linear Regression.

3. Approach

3.1. Lexicon Preparation

In this work, we modify and extend our sentiment lexicon SLSA, a large-scale Sentiment Lexicon for Standard Arabic (Eskander and Rambow, 2015). The construction of SLSA is based on linking the lexicon of the Standard Arabic morphological analyzer AraMorph (Buckwalter, 2004) with the English sentiment lexicon SentiWordNet (Baccianella et al., 2010) along with some heuristics and back-off techniques.

One key advantage of SLSA is its high coverage as it contains 34,852 lemma-POS pairs, which makes it the largest of its type (for Standard Arabic). Each lemma-POS pair is associated with three sentiment scores (positive, negative and objective), in addition to the English gloss. Another advantage is it richness as it is a lemma-based resource that attaches the POS and English gloss information to lemmas, where the sentiment of a lemma is applicable to its corresponding inflected forms. Additionally, the creation of SLSA was fully automated based on previous resources. This makes the generation of future updates straightforward upon improving the quality of the resources.

The overall projected accuracy of the entries in SLSA is 80.1%. About 93% of the erroneous entries are cases where the sentiment scores are doubtful in SentiWordNet, while the other errors are due to incorrect glosses in AraMorph.

Since our proposed system is lexicon-based, the higher accuracy the lexicon has, the better the classification of the sentiment polarity is. Accordingly, we develop a new extension of SLSA where we manually correct SLSA entries that correspond to the most frequent 4,000 lemmas in the Penn Arabic Treebank (PATB), Part 3 (Maamouri et al., 2004), Moreover, 220 new entries are added into SLSA. Those entries correspond to lemma-POS combinations that are absent in Aramorph.

Additionally, instead of having sentiment scores ranging from 0 to 1, the scores are better normalized by rounding to their 0.25 ceilings. This means the new positive and negative scores are of the values 0.0, 0.25, 0.5, 0.75 and 1.0. The reason for the rounding is to make the system less sensitive to the little differences in the automatically generated sentiment scores. Table 1 shows examples of SLSA entries after manual correction and score normalization, where the Arabic text is transliterated using the Buckwalter scheme (Buckwalter, 2004).

Lemma	POS	English Gloss	+ve Score	-ve Score	Obj. Score
niEom_1	NOUN	wonderful	1	0	0
mubArak_1	ADJ	blessed/fortunate	0.75	0	0.25
tawaE∼aY_1	VERB	be attentive/cautious	0.5	0	0.5
AiHotiwA'_1	NOUN	inclusion/content	0.25	0	0.75
$ahoriy∼_1	ADJ	monthly	0	0	1
katab_1	VERB	write	0	0	1
munAqaDap_1	NOUN	contradiction/contrast	0	0.25	0.75
lawom_1	NOUN	blame/censure	0	0.5	0.5
dana>_1	VERB	be vile/be despicable	0	0.75	0.25
kamod_1	NOUN	swarthiness/sadness	0	1	0

Table 1: Examples of SLSA entries after manual correction and score normalization. The first column represents the lemma written in the Buckwalter transliteration. The second column is the POS tag of the lemma. The third column is the corresponding English gloss. The other three columns represent the positive, negative and objective scores, respectively, where the objective score is defined as *1 - (positive score + negative score)*.

3.2. Corpus Development

We develop a new corpus of Standard Arabic text where each sentence is manually annotated for its contextual polarity (positive, negative or neutral). The corpus contains 4,000 sentences generated from news websites. The context of the sentences is varied to cover several genres such as politics, arts, sports, fashion, religion and medicine. In the case where a sentence has both a positive and a negative sentiment, the sentiment that is more dominant within the context is assigned. Table 2 lists example sentences of different polarities and genres from the corpus.

The corpus is split into three sets for training (3,200 sentences), development (400 sentences) and testing (400 sentences), namely; SentiTrain, SentiDev and SentiTest, respectively, where each set has examples of the different genres. SentiTrain is used for the supervised training of the system, where the text is linked to entries in the sentiment lexicon through the lemma and POS information. SentiDev is then used for tuning the system, while SentiTest is a blind set that is used solely for testing.

Sentence	Polarity
لا تقلق، أنا بخير. Do not worry, I am fine.	Positive
اعتادت أن تسحرنا بصوتها الملائكي الحالم. She used to charm us with her dreamy angelic voice.	Positive
لكنهم في وقت لاحق سيندمون. But they will later regret.	Negative
لكننا لم نتمكن من تحقيق الفوز، للأسف. But we were unable to win, unfortunately.	Negative
جميعنا يستقل المصعد بشكل يومي تقريباً. We all take the elevator almost daily.	Neutral
ومن المتوقع أن يعلن عن الصفقة الاثنين. The deal is expected to be announced on Monday.	Neutral

Table 2: A sample of corpus sentences annotated for their contextual polarity.

3.3. Sentiment Classification

The purpose of the proposed system, SentiArabic, is to determine whether a given text expresses a positive or a negative sentiment, so our focus is on the sentences that have either a positive or a negative polarity. We train a supervised model based on SentiTrain, described in subsection 3.2, where a sentence represents a training unit, and the output class is the sentiment polarity.

First, for each sentence in the training corpus, we use the sentiment lexicon to extract the positive and negative scores that correspond to each word in it (zero values are used as defaults). However, since the words in the corpus are surface forms (i.e., inflected), while the lexicon has entries of lemma and POS values, it is required to run a morphological tagger that analyzes each word in context in order to obtain its lemma and POS information. One drawback in this approach is the extensive computation the morphological tagging requires due to the morphological complexity of Arabic, which violates the *lightweightness* aspect of the overall system. Instead, we obtain the POS and lemma information by looking up the word in a maximum likelihood lookup that is based on PATB, Parts 1, 2 and 3. That means, the lemma and POS for a given word are chosen to be the most frequent analysis the word has received in PATB. However, we show in section 4 that the use of a morphological tagger increases the overall performance of the system by an absolute F-score of only 0.5%, which is not a significant improvement given the computational overhead of running a morphological tagger.

After extracting the positive and negative scores for each word in a sentence, the following features are extracted and used as the sentence representation:

- *Average Positive Score per Subjective Word*
- *Average Negative Score per Subjective Word*
- *Average (Positive - Negative) Score per Subjective Word*
- *Percentage of Positive Words to Subjective Words*

- *Percentage of Negative Words to Subjective Words*

A subjective word has an objective score of zero, i.e., it is either positive, negative or both, where a word is positive if its positive score is greater than zero, and it is negative if the negative score is greater than zero. One special case is negation, where it flips the positive and negative scores of the following context, if any.

It is worth noting that all the features are simple numerical ones, where none is a lexical representation of the context. The reason behind this is to decrease the computation involved in our machine learning approach toward a lightweight sentiment analyzer.

We experiment with different machine learning techniques such as SVMs, Naive Bayes, K-Nearest Neighbors, Random Trees, and Decision Trees. The latter gives the highest accuracy and F-score on SentiDev (78.7% and 78.3%, respectively), which in turn serves our purpose of developing a lightweight system that requires minimal computation.

4. Evaluation and Results

We evaluate SentiArabic on two blind test sets; SentiTest, described is subsection 3.2, and a PATB dataset annotated for sentiment polarity by Abdul-Mageed et al. (2011). The latter is used to evaluate the sentiment analyzer by Eskander and Rambow (2015) and the state-of-the-art deep learning models for Arabic sentiment analysis by Sallab et al. (2015) (see section 2).

We first run the system using the original SLSA without the correction and normalization we applied in subsection 3.1. We call this system SentiArabic-NoExt. SentiArabic-NoExt gives an overall accuracy and a weighed F-score of 67.6% and 67.0%, respectively, when evaluated on SentiTest. These numbers increase to 81.1% and 80.8%, in order, when testing with the new extension of SLSA, which is a significant error reduction of 41.6%.

When evaluated on the PATB test set, SentiArabic achieves an accuracy of 76.7% and a weighted F-score of 76.5%. This is an absolute F-score increase of 3.0% over the state-of-the-art analyzer by Sallab et al. (2015) and 7.9% over our previous analyzer (Eskander and Rambow, 2015). Although the system by Sallab et al. (2015) performs heavy computations for the deep learning models, while SentiArabic relies on lookups and a simple decision tree, the efficiency of SentiArabic is highly supported by the high accuracy of the new extension of SLSA. This is no surprise given the performance of SentiArabic versus SentiArabic-NoExt, while the original SLSA lexicon is highly efficient compared to its counterparts (Eskander and Rambow, 2015).

Table 1 summarizes the results for SentiArabic, SentiArabic-NoExt, the system by Eskander and Rambow (2015) and the system by Sallab et al. (2015) on both SentiTest and PATB test sets (when possible).

Finally, we run an additional variation of the system where the lemma and POS information of each word is generated in context by running the state-of-the-art Arabic morphological tagger MADAMIRA (Pasha et al., 2014) instead of looking up the lemma and POS information in the PATB-based lookup (see subsection 3.3). This is in order to compare to a variation of the system that runs rich in-context computation. Using MADAMIRA increases the overall weighted F-score by an absolute 0.5%. However, we find the increase insignificant given the computational overhead incurred by MADAMIRA, which is in favor of the lightweight SentiArabic analyzer.

We conducted an error analysis on all the entries in SentiTest that are misclassified by the analyzer. About 15% of the erroneous cases are sentences that involve both a positive and a negative sentiment where the system picks the sentiment that is less dominant in the context.

System	SentiTest		PATB	
	Acc. %	**F1 %**	**Acc. %**	**F1 %**
Eskander and Rambow (2015)	–	–	–	68.6
Sallab et al. (2015)	–	–	74.3	73.5
SentiArabic-NoExt	67.6	67.0	70.4	70.8
SentiArabic	**81.1**	**80.8**	**76.7**	**76.5**

Table 3: Sentiment analysis results of SentiArabic (with and without lexicon extension) on two test sets; SentiTest and PATB, compared to two state-of-the-art systems by Eskander and Rambow (2015) and Sallab et al. (2015). SentiArabic outperforms both systems by absolute F-scores of 7.9% and 3.0%, respectively. *(Unreported results are marked as "–" .)*

5. Conclusion and Future Work

We proposed SentiArabic, a new lightweight lexicon-based sentiment analyzer for Standard Arabic. SentiArabic avoids running heavy computation by exploiting a morphology lookup along with a simple decision tree for classification As part of developing the system, we built a new extension of the SLSA lexicon that has a higher quality and better normalization of the polarity scores. We also created a new corpus that is tagged for contextual polarity, where the text covers a wide range of genres.

SentiArabic achieves an F-score of 76.5% when tested on a blind test set annotated by Abdul-Mageed et al. (2011), which is the highest result reported for that set, and an absolute 3.0% increase over a state-of-the-art system that uses deep learning models (Sallab et al., 2015).

The future plans include working on the SLSA extension to manually check all the entries and correct the erroneous cases, in addition to augmenting the lexicon with multiword expressions. We also plan to extend our work to cover other Arabic dialects, Egyptian, Levantine and Gulf in particular.

6. Bibliographical References

Abdul-Mageed, M., Diab, M., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon.

Abdul-Mageed, M., Diab, M., and Kubler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech and Language*, 28:20–37.

Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 11th International Workshop on the ACL2 Theorem Prover and Its Applications*, Laramie, Wyoming.

A.M. Alayba, V. Palade, M. E. R. I. (2017). Arabic language sentiment analysis on health services. In *Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, Nancy, France.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

T. Buckwalter. (2004). *Buckwalter Arabic Morphological Analyzer (Aramorph) Version 2.0*. Linguistic Data Consortium (LDC), 2.

El-Beltagy, S. and Ali, A. (2013). Open issues in the sentiment analysis of arabic social media: a case study. In *Proceedings of the 9th International Conference on Innovations in Information Technology (IIT)*, Al Ain, AUE.

El-Makky, N., Nagi, K., El-Ebshihy, A., and Ibrahim, S. (2014). Sentiment analysis of colloquial arabic tweets. In *Proceedings of the 3rd ASE International Conference on Social Informatics*, Boston, Massachusetts.

Eskander, R. and Rambow, O. (2015). Slsa: A sentiment lexicon for standard arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal.

Kiritchenko, S., Mohammad, S., and Salameh, M. (2016). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the 2016 International Workshop on Semantic Evaluation (SemEval)*, San Diego, California.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). Arabic treebank: Building a large-scale annotated arabic corpus. In *Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Atlanta, Georgia.

Nabil, M., Aly, M., and Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Refaee, E. and Rieser, V. (2016). ilab-edinburgh at semeval-2016 task 7: A hybrid approach for determining sentiment intensity of arabic twitter phrases. In *Proceedings of the 2016 International Workshop on Semantic Evaluation (SemEval)*, San Diego, California.

Sallab, A. A., Baly, R., Badaro, G., Hajj, H., and W. El-Hajj, K. S. (2015). Deep learning models for sentiment analysis in arabic. In *Proceedings of the EMNLP 2015 Workshop on Arabic Natural Language Processing (ANLP)*, Lisbon, Portugal.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada.