============================================================================  
ArabicNLP 2015 Reviews for Submission #8  
============================================================================  
  
Title: Deep Learning Models for Sentiment Analysis in Arabic  
  
Authors: Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El Hajj and Khaled Bashir Shaban  
============================================================================  
                            REVIEWER #1  
============================================================================  
  
  
---------------------------------------------------------------------------  
Reviewer's Scores  
---------------------------------------------------------------------------  
  
                         Appropriateness: 5  
                                 Clarity: 3  
            Originality / Innovativeness: 3  
                 Soundness / Correctness: 2  
      References / Meaningful Comparison: 2  
              Impact of Ideas or Results: 4  
                     Impact of Resources: 1  
                          Recommendation: 3  
                     Reviewer Confidence: 5  
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
This paper has a very promising but ambitious premise: to evaluate different  
deep-learning models for sentiment analysis in Arabic. Furthermore, given the  
useful resources that the authors make use of (namely the ArSenL lexicon), this  
paper had the potential of becoming one reference paper for sentiment analysis  
in Arabic.  
  
Unfortunately, the execution of the paper falls short to deliver its promises,  
and after reading it, I'm left with more questions than answers. The good news  
is that the authors show that RAE can be used to improve the sentiment analysis  
task for Arabic. Sadly, this algorithm makes no use of the ArSenL, and the  
authors do not propose a method to incorporate such rich information to their  
learning framework. However, I'm confident that if the authors fix some  
methodological issues, this paper could provide great contributions.  
  
Below, I highlight my main worries with this paper.  
  
1) input representation and architecture for models 1..3  
In my view, there is a problem with the input representation chosen for the  
deep neural models. First, it is unusual to use conditional features as input  
(here they use p(Sentiment|word) for each word in  the vocabulary. Usually,  
researchers use a one-hot binary vector model or use some pre-trained  
embeddings as input. The conditional probabilities in ArSenL could more fit to  
pre-train the embeddings, or as additional skip-arc features that feed the  
output layer directly.  
Furthermore, the architecture of a DNN needs to be motivated by the task we  
want to solve. Using 3 layers of 40 neurons each as a "recipe" from every  
problem, is not recommended. Every task has its own levels of abstraction  
(think of each layer representing something meaningful such as POS, syntax,  
semantics).

The results are more explained in the direction why the DAE and DBN are inferior. Also, the usage of ArSenL in word embedding is mentioned as future work due to limited experiments time.

The architecture of DNN and DBN is obtained empirically as mentioned, so the one in the paper is the one that worked better for both cases.  
  
2)There is a slight misrepresentation of how the RAE works. In fact, one of the  
main properties is a binary encoder, i.e. it encodes two words at the same time  
(Socher et al, 2011). This allows to create a parse tree for pairs of  
constituents that have the least reconstruction error.        However, the RAE  
presented here, does not seem to have this property. Perhaps I missed the  
explanation, but I went back to the text several times, and this was not clear.

The basic encoder used is indeed a binary encoder. An explicit mention of that is added in section 3. Also, in Figure 5, the inputs to the RAE blocks are pair of word indices.

3)There is a lack of a much needed analysis of the results. The authors only  
provide a table with the performance on the test set, and nothing more. This is  
rather unsatisfactory because if a researcher is to refer to this paper, it  
should be left with intuitions on why certain method works best in certain  
situations, and why not.  
  
Again, this paper does not do justice to the ArSenL lexicon, because a simple  
linear SVM can exploit these features accurately and provide very competitive  
results. But why they do not work well for Deep Nets? This question needs to be  
answered. Additionally other simpler architectures need to be tested and  
reported on. As a reader, I don't know if the low performance is due to  
problems on training, or in the lack of feature expressivity, etc.

The results are more explained in the direction why the DAE and DBN are inferior. Also, the usage of ArSenL in word embedding is mentioned as future work due to limited experiments time.  
  
\*\*\*Other minor issues:  
  
-In section 3 it is not clear what is the cost/error function that is being  
minimized.

The AE paragraph mentions:

“At each step of parsing, the weights of the basic AE are updated so as to minimize the reconstruction error”

This is also clarified in Figure 5. A reference to Figure 5 is added as well.

-There seems to be  a confusing statement about activations "the activation  
functions of each sigmoid is taken as  hyperbolic tangent activation". This is  
incorrect. Activations are either sigmoids or tanh.

Sigmoid is replaced by neuron

- Your problem is a binary classification problem. You don't need to do a  
softmax, but a simple logistic function (sigmoid) is enough. This cuts  
calculations by half. By the way, a logistic regression algorithm is a baseline  
you should include in the future version of this paper.

Agree, this would reduce a neuron at the output. For logistic regression, it is not expected to do better than DNN, but it can be included as one of the reference algorithms.

============================================================================  
                            REVIEWER #2  
============================================================================  
  
  
---------------------------------------------------------------------------  
Reviewer's Scores  
---------------------------------------------------------------------------  
  
                         Appropriateness: 5  
                                 Clarity: 4  
            Originality / Innovativeness: 4  
                 Soundness / Correctness: 3  
      References / Meaningful Comparison: 3  
              Impact of Ideas or Results: 4  
                     Impact of Resources: 1  
                          Recommendation: 4  
                     Reviewer Confidence: 4  
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
This paper proposes a deep learning approach for the sentiment classification  
problem on Arabic text. Three architectures were proposed and derived for: DNN,  
DBN and Deep Auto Encoders. Deep Auto encoder model provided better  
representation of the input sparse vector. The RAE model was the best deep  
learning model according to the obtained results, although it requires no  
sentiment lexicon.  
  
Although the approach is not specially novel, the paper is oerall well written  
and clearly exposes the proposed architectures and the evaluation results are  
very promising. Indeed, the results show around 9% improvement in average F1  
score over the best reported results in literature on the same LDC ATB dataset  
in the sentiment classification task for Arabic.  
  
The paper is overall clearly written, but some revision is needed :  
- References should be updated (only one recent reference from 2014)  
- English style of the conclusion has to be revised

Fixed  
- fig 6 is too big however Fig 5 is too small

Fixed  
- References formatting has to be revised

Fixed

============================================================================  
                            REVIEWER #3  
============================================================================  
  
  
---------------------------------------------------------------------------  
Reviewer's Scores  
---------------------------------------------------------------------------  
  
                         Appropriateness: 5  
                                 Clarity: 4  
            Originality / Innovativeness: 4  
                 Soundness / Correctness: 4  
      References / Meaningful Comparison: 3  
              Impact of Ideas or Results: 4  
                     Impact of Resources: 1  
                          Recommendation: 4  
                     Reviewer Confidence: 3  
  
  
---------------------------------------------------------------------------  
Comments  
---------------------------------------------------------------------------  
  
The paper proposes four deep learning models for sentiments classification on  
Arabic text: DNN, DBN, DAE, and RAE. It is well-written, diagrams explain the  
different NN architectures, and evaluation compares the four models to each  
other and to a state-of-the-art SVM model.  
  
We know that Buckwater morphological analyzer has about 40,000 lemmas, but the  
LDC ATB dataset and ArSenL have only 3,795 lemmas in common which are fed to  
the first three models. Is this small number because:  
A) LDC ATB is small in size  
B) ArSenL does not represent enough lemmas  
C) or, some lemmas are not important for sentiment analysis?  
If the answer is B, then there are two challenges to your conclusion of RAE  
being better than the other models. First, there is a room for improvement in  
ArSenL which will lead to improving the first three models (of course, as you  
pointed out, this approach is costly in terms of time and money). Second, this  
might point out another reason for why the fourth model (RAE) outperformed the  
previous three: the use of raw words is not a challenge; instead it could be an  
advantage: RAE is looking at more information while the first three models are  
omitting any word that does not match an ArSenL lemma. Please discuss.

The answer is A since the part of LDC ATB that is being used for testing is small. It is the one annotated by Abdul-Mageed.

The paper needs more discussion of related work.

Related work added  
  
  
=-=-=-=-=-=  
Minor suggestions and typos:  
  
Page 1: “These challenges add the complexity” should be “These challenges  
add to the complexity”

Fixed

Page 2, line 6: “in Arabic For the” should be “in Arabic. For the”.

Fixed

Page 2: “the input are the raw” should be “the input is the raw”.

the input are the raw  
In Evaluation, line 5: “Al-” should be “Ar-”.

Fixed

In Evaluation: “It is also noticeable that the deep auto…” this sentence  
is either not finished or re-phrased in the following sentence.

Fixed

In Conclusion: “Deep Auto encoder model provided get better”, use either  
“provided” or “got/gets”.

Fixed