Adversarial Examples in Arabic

1. **What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?**

The authors in this paper proposed the first-ever black-box adversarial examples/attack on sentiment analysis classifiers in the Arabic language. These black-box adversarial attacks were crafted to perturb Arabic textual inputs by intentionally violating the noun-adjective agreement in Arabic. The authors successfully used two state-of-the-art DNN architectures, which were fooled in the sentiment analysis task, resulting in reduced classification accuracy by an average of 52.97% for the word-level bidirectional-LSTM model and 50.44% for the word-level CNN model. Additionally, the authors were able to determine the most important token (word) in the input space, using its effect on the targeted classifier prediction, then they attacked it by changing the adjective's morphological form, which violates the noun-adjective agreement in Arabic. These perturbations include: 1) converting indefinite adjectives to definite adjectives and vice versa; 2) converting masculine adjectives to feminine adjectives and vice versa; 3) converting singular adjectives to dual adjectives and vice versa; 4) converting singular adjectives to plural adjectives and vice versa.

Interestingly, the authors found that violating this noun-adjective agreement negatively affects the performance of DNN-based classifiers. The proposed "*methods could successfully flip the classifiers' output on these examples and dramatically reduce the word-level BiLSTM classification accuracy from 100.00% to an average of 47.02% for the HARD corpus, and the word-level CNN from 100.00% to an average of 49.56% for the BRAD corpus*"—the authors. To show that these proposed adversarial attacks are hardly detectable by humans, the authors conducted a human study where they recruited ten Arabic native speakers and asked them to label 200 textual samples produced by their attacks. They also found that non-neural models like Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) are more robust against these adversarial attacks than DNNs, where the accuracy decreased in the bidirectional-LSTM model to 52.97%. In contrast, with the SVM and XGBoost, the accuracy decreased to 33.33% and 49.67%, respectively.

2. **Could I have done this work if I had the idea why or why not?**

I think I could have done this work if I had the idea. The high-level idea of the authors' work in this paper is to exploit the relationship between nouns and adjectives in Modern Standard Arabic (MSA), i.e., the authors intentionally violate the noun-adjective agreement in MSA. The authors used publicly available: two datasets (Hotel Arabic Reviews and Books Reviews in Arabic), two DNN models (word-level bidirectional-LSTM and word-level CNN), an Arabic morphological analyzer (MADAMIRA), and Word2Vec with Continuous bag-of-words to produce the word embeddings.

3. **Is there anything I could do to repeat or validate? (Make notes of materials available from the paper- data sets, source code).**

No, the authors did not open-source their adversarial attacks. Yet, the authors used two publicly available datasets: Hotel Arabic Reviews Dataset (HARD) [1] and Books Reviews in Arabic Dataset (BRAD) [2]. The authors also used two DNN famous models: word-level bidirectional-LSTM [3] and word-level CNN [4], and they used an Arabic morphological analyzer called MADAMIRA [5].

[1] HARD: Hotel Arabic-Reviews Dataset, https://github.com/elnagara/HARD-Arabic-Dataset.
[2] BRAD: Books Reviews in Arabic Dataset, https://github.com/elnagara/BRAD-Arabic-Dataset.
[3] Word-level bi-LSTM is part of 'DeepWordBug': https://github.com/QData/deepWordBug.
[4] Word-level CNN: https://github.com/yoonkim/CNN_sentence.
[5] MADAMIRA: https://camel.abudhabi.nyu.edu/madamira/.

4. **What is my best idea for follow on work that I could personally do?**

   My best idea for the follow-on work that I could personally do is to first repeat the same work using the authors' Modern Standard Arabic (MSA) datasets to develop a better understanding of how the attacks work. Secondly, I could repeat the same work using Dialectal Arabic (DA) sentiment analysis datasets to examine the generated results and test if these adversarial text attacks work in Arabic dialects like Gulf, Egyptian, Moroccan, etc.

5. **What is my best idea for follow on work that I'd like to see the authors or others do?**

   My best idea for the follow-on work that I would like to see the authors or others do is to "*evaluate more DNN-based models in an adversarial setting for a variety of NLP tasks, such as Machine Translation and Dependency Parsers, with languages beyond English*"—the authors.

6. **Any logistical experimental lessons I learned (this is a little different than #1 here you are looking for techniques)?**

   I liked how the authors creatively utilized the Arabic morphology rules to their advantage when crafting their attacks, searching and perturbing for the most important adjective in a sentence, i.e., the input, and to do so, they used an Arabic morphological analyzer called MADAMIRA to perform a part-of-speech tagging task for the input to help them programmatically identify adjectives.

7. **How does this compare to the other papers we read? Most similar? How different? Other comparisons?**

   This paper by Alshemali and Kalita (2019) is the first-ever paper that discusses the adversarial example/attacks in the Arabic language to fool sentiment analysis classifiers. The authors crafted their black-box attacks to exploit the relationship between nouns and adjectives in Modern Standard Arabic (MSA). The authors also used the *Replace-1* scoring function introduced by Gao et al. (2018) and used one of their models (the word-level bidirectional-LSTM model) in their experiments.

8. **What is your biggest criticism of the paper?**

   There is one big criticism of the paper. Although the paper discusses the adversarial attacks/examples in the Arabic language, the authors did not include any adversarial Arabic examples in the paper. Instead, they hosted them in supplementary material on GitHub instead of attaching them at the tail of the paper (https://github.com/Basemah/AE4Arabic/blob/master/CSCI2019%20-%20Adversarial%20Examples%20in%20Arabic%20-%20Supplementary%20Material.pdf).

   The authors repeatedly discussed the related work in Sections II and III. I think the mentioned related work in Section III was unnecessary, and replacing it with a few successful adversarial examples in Arabic would have been more relevant.

9. **List 3 cited references or terms/concepts that you would be most interested in reading/learning more about.**

   The authors of this paper found that non-neural networks models like Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) are more robust against adversarial attacks than the Deep Neural Networks (DNNs), which supports the conclusions of these two references below:

   M. Wang, B. Chu, Q. Liu, and X. Zhou, ***YNUDLG at SemEval-2017 task 4: A GRU-SVM model for sentiment classification and quantification in Twitter****, in The International Workshop on Semantic Evaluations, 2017, pp. 713–717.*

   V. Golem, M. Karan, and J. Šnajder, ***Combining shallow and deep learning for aggressive text detection****, in The International Conference on Computational Linguistics, 2018, pp. 188–198.*