Character-level 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 introduced the first set of Arabic character-level adversarial attacks designed for models trained in Arabic. They proposed an efficient method to produce Arabic character-level adversarial examples against Deep Neural Network (DNN) text classifiers. The authors' method in this work relied on the flip operations of one or two Arabic characters that were designed based on the most common spelling mistakes that non-native Arabic learners make.

   The authors of this work found that only a few manipulations are needed to fool powerful and popular DNN-based text classifiers trained on Arabic corpora, such as the Recurrent Neural Networks (bidirectional-LSTM) models and the word-level Convolutional Neural Networks models. As a result, for the bidirectional-LSTM model, the authors' method reduced the model's accuracy from 80.00% to up 54.50% under the Ar-Flip attack and 50% under the Ar-Flip2 attack. Their proposed method reduced its accuracy for the CNN model from 82.00% to 61% and 56.40% under the Ar-Flip and Ar-Flip2 attacks, 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 most common spelling mistakes non-native Arabic learners make. The authors used two publicly available Arabic datasets (BRAD and SANAD), two DNN models (word-level bidirectional-LSTM and word-level CNN), and two Language Models (BERT and XLNet), and Word2Vec algorithm 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: The books Reviews in Arabic Dataset (BRAD) [1] and the Single-labeled Arabic News Articles Dataset (SANAD) [2]. The authors also used two famous DNN models: word-level bidirectional-LSTM [3] and word-level CNN [4], and two Language Models (BERT and XLNet).

   [1] BRAD: Books Reviews in Arabic Dataset, https://github.com/elnagara/BRAD-Arabic-Dataset.
   [2] Single-labeled Arabic News Articles Dataset, https://data.mendeley.com/datasets/57zpx667y9.
   [3] Word-level bi-LSTM is part of 'DeepWordBug': https://github.com/QData/deepWordBug.
   [4] Word-level CNN: https://github.com/yoonkim/CNN_sentence.

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 repeat the same work using the authors' Modern Standard Arabic (MSA) datasets to understand better how the attacks work. Secondly, I could repeat the same work using Dialectal Arabic (DA) and Sentiment Analysis (SA) 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 question answering systems, 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 used the most common spelling mistakes non-native Arabic learners make as adversarial attacks. These spelling mistakes are due to the high visual similarity between a few groups of Arabic letters.

I also liked how they studied different Arabic text classification tasks (sentiment analysis task and news articles classification task) with different depth levels (one and two Arabic characters in one or more tokens) using different Arabic datasets (BRAD and SANAD).

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

This paper by Alshemali and Kalita (2021) is the second-ever paper that discusses the adversarial example/attacks in the Arabic language, whereas the first paper was for the same authors in 2019. The authors crafted their black-box attacks to exploit the most common spelling mistakes non-native Arabic learners make. Both papers 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. Yet, this paper used fewer datasets (two datasets), different DNN models, such as word-level CNN, and Language Models like BERT and XLNet.

8. What is your biggest criticism of the paper?

There is *no* big criticism of the paper.

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

I am interested in reading these six references because they are very relevant. The authors of this paper discussed them in the related work section, and I think reading them is very important.

E. Jones, R. Jia, A. Raghunathan, and P. Liang, ***Robust encodings: A framework for combating adversarial typos****, in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2752–2765.*

A. Kumar, P. Makhija, and A. Gupta, ***Noisy text data: Achilles' heel of BERT, in Proceedings of the Workshop on Noisy User generated Text****. Association for Computational Linguistics, 2020.*

C. Aspillaga, A. Carvallo, and V. Araujo, ***Stress test evaluation of Transformer-based models in natural language understanding tasks,*** *in Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, 2020, pp. 1882–1894.*

G. Heigold, G. Neumann, and J. van Genabith, ***How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse?*** *In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, 2018, pp. 68–80.*

A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neu big, ***Stress test evaluation for natural language inference,*** *in Proceedings of the International Conference on Computational Linguistics, 2018, pp. 2340–2353.*

L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, and C. Xiong, ***Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT,*** *in arXiv preprint arXiv:2003.04985, 2020.*