Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

1. **What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?**
   The authors in this paper introduced a framework, CheckList, for behavioral testing of NLP models. This framework includes a *matrix* of general linguistic *capabilities* (Vocabulary, Named Entity Recognition, and Negation) and *test types* (Minimum Functionality test, Invariance test, and Directional Expectation test) that facilitate comprehensive test ideation and a software tool to generate a large and diverse number of tests cases quickly.

2. **Could I have done this work if I had the idea why or why not?**
   I do *not* think I could have done this work if I had had the idea, given my limited knowledge of various software engineerings testing approaches such as behavioral testing and unit tests.

3. **Is there anything I could do to repeat or validate? (Make notes of materials available from the paper- data sets, source code).**
   Yes, the authors open-sourced their testing framework (CheckList) here on GitHub: https://github.com/marcotcr/checklist#table-of-contents.

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 experiment with this testing framework (CheckList) using English datasets to understand better how it works. Secondly, I could experiment with this framework using Arabic datasets, examine the generated results, and test if it supports multi-languages like the Arabic language.

5. **What is my best idea for follow on work that I'd like to see the authors or others do?**
   I do *not* have an idea for the follow-on work I would like to see the authors or others do.

6. **Any logistical experimental lessons I learned (this is a little different than #1 here you are looking for techniques)?**
   I have *not* learned any logistical experimental from this paper.

7. **How does this compare to the other papers we read? Most similar? How different? Other comparisons?**
   This paper by Ribeiro et al. (2020) is the first paper that proposed a testing framework (CheckList) for comprehensive behavioral testing of NLP models. CheckList testing framework guides users on what to test by providing a list of linguistic capabilities which apply to most tasks.

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 would like to learn more about BERT and RoBERTa models, especially how to fine-tune them on custom corpora. Another thing I would like to learn is the Masked Language Models (MLM) tasks. I would also like to read the RoBERTa paper to understand the model and its usages in depth.

   Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ***RoBERTa: A Robustly optimized BERT pretraining approach****. arXiv preprint arXiv:1907.11692.*