**Analysis: “Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment**

Paper by: Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits

Review by: Rajat Sethi

**Paper Summary:**

There exist machine learning (ML) algorithms designed to determine the connotation of a sentence, either positive or negative. While these ML algorithms are mostly accurate, some of the important words can be subtly changed or manipulated in ways that provide false positives or negatives. For example, the negatively-connotated sentence (Page 7):

“*The characters, cast in impossibly contrived situations, are totally estranged from reality.”*

Is changed to the following sentence (Page 7):

“*The characters, cast in impossibly* ***engineered******circumstances****, are* ***fully*** *estranged from reality.”*

While a human will still view the sentence as negatively-connotated, the paper reported that the ML algorithms they tested viewed the statement as positively-connotated. In essence, the sentence has been poisoned to provide a false positive.

**Novelty of the Implementation:**

Their main approach to this problem was a partial brute-force algorithm. For every important, changeable word in the list, test a list of synonyms (candidates). Then, determine if the ML algorithms issued a false positive/negative in response to the changed word. This approach, while rather simplistic, has the potential to do its job correctly and accurately. Essentially, this project used Occam’s Razor to create “TextFooler.” As expressed by the paper’s results, the TextFooler had outperformed other word perturbers (Page 6) and achieved noticeable outcomes (Page 5). That being said, I believe there are a few issues with their evaluation methods.

**Possible Problems with Implementation:**

One area that needed a clearer and longer explanation was the “Automatic Evaluation” section (Page 4), specifically with how they calculated the accuracy. Were they using another program to automatically calculate the accuracy of BERT, or was that done by a human? Was a result deemed “inaccurate” if it switched responses? I am not familiar with training neural networks, so this problem might just be a misunderstanding.

A problem that is not a misunderstanding is the sample size for human evaluators. As stated in their paper, only two human judges helped conduct the manual evaluation (Page 4). That number is simply too small to make an accurate conclusion with such an analysis.

**Other Criticisms:**

One aspect that should have been mentioned in this paper is the consequences of letting this problem go untreated. How can false negatives/positives affect businesses or research? Is it possible for a bad actor to create and post an indefinite number of false positives, thus not alerting a company that its product is being bombarded with negative reviews? What about false negatives, how can faking negative reactions be used to interfere with a comment section?

**Overall Suggestions:**

* Elaborate and expand the evaluation methods performed in the trial.
* Provide more background on how this problem can cause damage.
* Add a list of possible future endeavors or applications for TextFooler.
* Reduce or explain more of the technical jargon.
  + This was especially a problem with how they set up the neural network and their final statistical analysis.
* Explain the “Transferability” section in more detail.
* Offer solutions to improve ML algorithms that can be poisoned in such a manner.