Ishaaq Razack

Professor Mazidi

CS 4395.001

13 November 2022

Portfolio: Reading ACL Papers

For this portfolio assignment, I chose to read “HateCheck: Functional Tests for Hate Speech Detection Models,” which was published by the Association for Computational Linguistics and is affiliated with the University of Oxford, The Alan Turing Institute, Utrecht University, and the University of Sheffield. The authors of this paper and their respective number of citations are, in the order they are listed on the paper: Paul Röttger (104), Bertie Vidgen (881), Dong Nguyen (2,447), Zeerak Waseem (N/A), Helen Margetts (13,492), and Janet Pierrehumbert (28,294). Previously, most of these authors published papers regarding NLP, and all of them published at least one paper within the realm of machine learning. The paper they wrote which I read for this assignment focuses on the issue of how current tests for hate detection models can be flawed, and proposes a new series of tests to better validate a model’s performance.

To begin, the authors of this paper explain what hate speech is. In their definition, hate speech is “abuse that is targeted at a protected group or at its members for being a part of that group.” They go on to define protected groups as groups discriminated based on their “age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation.” The goal of a hate detection model is to perform a binary classification task on speech to predict whether it should be considered as hate speech or not.

The paper continues by explaining why current tests for hate detection models are flawed. Currently, a hate detection model would be evaluated by measuring its performance using cross

validation or the validation set approach on test data using metrics such as accuracy, precision, and recall. However, relying on such metrics makes it difficult to identify specific weak points in a model. Also, many of these test datasets contain certain biases due to how they were generated. Both of these issues make it difficult to properly assess how well a hate detection model can identify hate speech.

To solve these issues, the solution that the authors of this paper suggest is a series of tests called HateCheck. HateCheck consists of 29 tests that are of two types: distinct expressions of hate and contrastive non-hate. Distinct expressions of hate tests try to categorize hate speech into different types and check how well the model detects each type of hate speech individually. For example, the model's performance when detecting hate speech against a certain race would be measured separately from how the model performs with hate speech against members of a certain religion. As for contrastive non-hate, the tests were generated by taking positive statements about protected groups and converting them to negative statements. For example, "I love immigrants" would be converted to "I hate immigrants." Overall, these tests do a much more comprehensive evaluation of a hate detection model than standard evaluation metrics can.

To demonstrate the capabilities of HateCheck, a few popular hate detection models were tested. The researchers found that HateCheck did a good job finding biases that were known to exist in each of the models. In the future, HateCheck can easily incorporate more tests in order to expose other biases in models towards certain protected groups. For now, though, HateCheck can be used as an alternative to standard results metrics when measuring the performance of hate detection models.