Zach Allen & Nebil Weber
Portfolio Assignment: Reading ACL Papers
CS 4395.001
Dr. Karen Mazidi

Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech (Summary)

"*Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech*" is a paper written by Margherita Fanton, Helena Bonaldi, Serra Sinem Tekirog˙lu, and Marco Guerini of the University of Trento, Italy. The paper details the research done by the authors in the realm of online hate speech combat, particularly by creating a hate speech, counter narrative (HS/CN) database that generates training data in loops, with each iteration of the machine's new training data being reviewed and post-edited by experts in the field. The experiments done by the authors comprise of several loops including dynamic variations. With the rapid growth of social media platforms, there is no surprise that hate speech is a problem online that many feel the need to combat. Due to this desire to combat online hate speech, there have been many prior works involving both hate speech and counter narratives in the natural language processing domain. A popular study classifying hate speech on Twitter is "*Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93.*", by Zeerak Waseem, which details a program classifies data as hate speech based on a series of metrics but does not propose a way for the program to combat this hate speech. On the other hand, there are counter narrative research papers such as "*Analyzing the hate and counter speech accounts on twitter*." By Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. This paper details a web crawler on Twitter that analyzes counter speech accounts/posts, compiling the posts into a counter narrative database. The research done in "*Human-in-the-Loop for Data Collection:*

a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech" is the combination of the two topics – a hate speech, counternarrative database. When discussing the NLP program written by the authors, they stated "To our knowledge, the resulting dataset is the only expert-based multi-target HS/CN dataset available to the community." The fact that the author's NLP program is reviewed and post-edited by experts in the field of linguistics and hate speech, along with the join-compilation of hate speech and counter narratives, make it the most holistic program in the field of hate speech combat. The work is evaluated by linguists at each iteration of the program running, as it runs many times in a loop, creating new training data in each loop. The training data is then looked over by the experts for accuracy and completeness, and then put back into the database to be evaluated for the next training session in the following loop. Each experiment is comprised of several loops. The authors of "*Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech*" have received 14 citations on Google Scholar. I believe their work is important because of the time of the paper's publishing, they developed the only expert-based multi-target HS/CN dataset available to the community. As stated previously in this summary, other work done in these fields focus solely on one topic, whether it be hate speech classification or counter narrative generation. The combination of the two subfields into one dataset and the expert editing of the linguists distinguishes this research from others in its field. As for the individual citations of the authors, Margherita Fanton has been cited 17 times, Helena Bonaldi has been cited 17 times, Serra Sinem Tekirog˘lu has been cited 276 times, and Marco Guerini has been cited 1894 times. This makes sense as the first authors are PhD students, and Marco Guerini is the advisor of the research and has published many other research papers.