# ABSTRACT

Michael Decker, Advisor

In this thesis, we study commented-out code in software and its automatic identification. Commented-out code can inhibit program comprehension, it can have legal ramifications, and may degrade the performance of text analysis techniques. In order to address these issues, we develop a taxonomy on comments and commented-out code. Using this taxonomy, we manually classify nearly 3,000 lines of comments as commented-out code or English Prose. Using the resulting gold set, a model to automatically classify commented-out code based on character frequencies and decision trees is developed. The generated model achieves a F1 score of 87.51 on a testing set drawn from 50 projects not used to form the gold set. The model is additionally applied to these 50 open-source projects to analyze the prevalence of commented-out code. Results show that on average, 4.46% of lines of comments are commented-out code (median 2.76) with only five of the 50 projects containing no commented-out code.

This thesis is dedicated to Arthur Grills and the Brothers and Sisters
who have fallen in service to their country.

# ACKNOWLEDGMENTS

I would be remiss to not mention a few people who have changed both my academic career and my life as whole. To the Bowling Green State University computer science department faculty and staff, you have all changed my life in ways beyond counting, we have shared many laughs and tears through the trials and tribulations that was my time here. To Dr. Green and Dr. Decker, both of you, whether or not you realize it, saved my life during some of the hardest and darkest times coming out of the service. But, neither of you were willing to stop at just that, you both pushed me to continue in my academic career, you showed me just what I could become, and I am proud to say that I am here now because of the two of you. I think of both of you as more than family or friends, you are both heroes in my eyes and will always hold a special place in my heart.