Martin Rehak
Resistant AI s.r.o.
martin.rehak@resistant.ai
+420 737 113 153

Diploma Thesis Assessment

I had a pleasure to read and assess the diploma thesis of Mr. David Herel, entitled "Adversarial Attacks on Text Classifiers".

Summary: The thesis improves on the existing adversarial attacks targeted at NLP classifiers by defining a new, semantic-sensitive metric to better assess the similarity between the targeted text and the modified text. It also defines a new adversarial attack. The work clearly shows that the student understands the topic in depth, can perform independent research on the level that is more appropriate for an early-stage Ph.D. student and can position his research correctly in context of prior art. The quality of student's writing is very good. **Therefore, I grade the thesis as "Excellent" (numeric grade 1), recommend its acceptance as a Diploma thesis and suggest its nomination for Dean's award.**

Relevance: With the growing volume of social media and discussion forum posts, their manual moderation is next to impossible and NLP methods are often deployed to identify and remove hate speech and other behaviors that violate the terms of use. As the attackers aim to circumvent these systems, they can design automated adversarial attacks and use the rejection by the moderator algorithm as an input for an adversarial classifier, crafting the test so as to conserve the meaning for a human, while becoming admissible to moderation algorithm. This makes the research highly relevant.

Quality: The research conducted by the student follows a solid methodology and delivers original results that outperform SOTA approaches. While have small reservations about some of the design decisions (e.g. lack of discussion of parameter values and their impact), the work is clearly beyond the level requested by our School in terms of research quality and thoroughness.

Detailed comments for the student:
(Not to be discussed during the State exam)

Page 33 and Section 6.1.2: "created numerous combinations, and comparisons of different aspects of SPE to find the best performing combination". I fully understand the constraints of Diploma thesis scope, but Section 6.1.2 does not really explain the parameter selection in detail.

Figure 6.1 - This figure could ave been selected better, as, arguably (and the author agrees in the text), is suggests that 4 models would provide better combination of performance and accuracy. Possibly, the author should have selected other dataset for illustration figures, where the impact of 7 classifiers is positive.

In Seattle, on May 30, 2022

Martin Rehak