

Examination Form 2: Independent Report

MPhil in Machine Learning and Machine Intelligence 2022

Candidate's Name Shane Weisz	Assessor's Name Marcus Tomalin
Assessor's phone/mobile number:	1223762860

This is for your **independent report** on the project (around 200-400 words). Assessors are reminded that this independent report will be made available to the candidate by the Degree Committee for the Department of Engineering.

Shane's thesis is extremely impressive. He has accomplished a significant amount of research over the last few months and he has created a system that is a substantial contribution to the field.

Before he was even able to start developing a state-of-the-art dialogue system that could produce more effective counterspeech, he had to create an experimental framework from the ground up. This involved creating a set of scoring metrics, test sets, and so on. It is unusual for MPhil students to have to do this. Usually, they are able to work within an existing experimental framework (e.g., WER for ASR tasks etc). However, Shane was completely undaunted by this undertaking, and, having created a robust experimental framework, he then focused on the main task – that of fine-tuning DialoGPT so that it could generate linguistic outputs that were more effective as counterspeech. During this process he exhaustively explored various methods, and made good decisions about the ones to include in his final system. Although the automated metrics indicated that his system performed effectively, Shane also obtained results from experiments involving human participants, and these confirmed that, on this whole, his system produced outputs that were comparable to the outputs humans would produce in response to hate speech.

The thesis shows substantial evidence of detailed extra-curricular academic reading, critical thought, and original interpretation. Shane's depth of understanding of the technical details of the systems he created is manifest both in their design and construction as well as in the write-up which describes them.

In addition to this, Shane created a web-based interface so that it is possible to try out his system: users can input hate speech and see the responses that the system generates. Again, this demo is extremely helpful in enabling naïve-users to appreciate the significance of the system he has created.

This was undoubtedly a very challenging project, and Shane coped with it all with confidence and conviction, and he as consequently produced substantial deliverables.

Given all of this, it is self-evident that the work merits the award of an MPhil degree.

Please send this form to the Course Administrator (mlmi-mphil-admin@eng.cam.ac.uk) no later than 12:00 noon on Monday 5th September 2022.

Assessor's Signature

A handwritten signature in black ink, appearing to read 'M. Tuli', is written over the 'Assessor's Signature' text.

Date

29/08/22

Examination Form 2: Independent Report

MPhil in Machine Learning and Machine Intelligence 2022

Candidate's Name	Assessor's Name
Shane Weisz	Professor William Byrne
Assessor's phone/mobile number:	

This is for your **independent report** on the project (around 200-400 words). Assessors are reminded that this independent report will be made available to the candidate by the Degree Committee for the Department of Engineering.

The title of this thesis is 'Automating Counterspeech in Dialogue Systems'. Overall, it is a very impressive effort. Writing, organisation, and presentation are admirably clear throughout, and there is clear evidence of extensive reading of the recent literature. The experimental results are clearly presented against an appropriate baseline, with interesting results. A very complete effort with almost no 'loose ends'.

Following the Introduction, Chapter 2 gives the background to the thesis, reviewing why Counterspeech is needed and what is needed for its automation. Transformers and dialogue systems are reviewed to the extent needed for this work. There is also a survey of 'NLP for counter speech' reviewing available data (created by experts and through crowd sourcing) for Counterspeech. Chapter 3 presents the 'Methodology', including data sets, the baseline dialogue system (DialogPT) and how it is refined for the Counterspeech task, and automatic and human evaluation. Chapter 4 continues with the 'Experimental Setup' – it is worth noting that the separate presentation of methodology ahead of the experimental setup works particularly well in that it allows for a general discussion of issues related to models, data, metrics, etc, ahead of the detailed description of what was actually done for the thesis. Chapter 5 presents the results of automatic and human evaluation of dialogue systems to assess their inherent toxicity and the effectiveness of fine-tuning strategies using the counterspeech data sets at improving their responses. Chapter 6 concludes with a summary and suggestions for future work.

Some comments:


- In the discussion of fine tuning DialogPT (3.2.1) there is only one paragraph describing the training data and procedures used in the baseline. More detail on how the baseline was trained would make the choice of linearisation scheme of the counterspeech data more understandable. In other tasks it has been observed that linearisation schemes for the fine-tuning data can affect how the baseline system adapts. It's a minor point, but possibly relevant to subsequent results showing degradations in general dialogue quality after finetuning.
- In the discussion of the metric suite (3.3.2), it would be interesting to know how well these automatic metrics (which were developed in other domains) carry over to the types of dialogue responses needed for counterspeech. For example, is a 'fluency measure' trained on CoLA a good fit for this task, or would it be better to fine-tune some of these model-based metrics on in-domain data, as well. The metrics chosen do seem appropriate, though, and

(impressively) seem to be largely consistent with each other in the experimental results.

- The GPS system appears to be an appropriate choice of baseline for this task, as a recently published point of comparison. However its not surprising that the transformer-based approaches do better overall, particularly under these automatic metrics.
- A more detailed investigation into the value of expert annotations (as in the MultiCONAN dataset) relative to crowdsourced annotations (as in the Gab dataset) would be very interesting. Table 5.1 suggests that using Gab with MutliCONAN yields the best results in terms of toxicity. But how does crowdsourced data do alone?
- The degradation in 'general conversational ability' (5.1.2) is unfortunate, but the suggested explanation that 'counterspeech fine-tuning implicitly changes the general conversational style of the responses produced by the systems to be more geared towards disagreement' is convincing. This suggests a further interesting line of work, in which the system learns to change its behavior only when counterspeech is needed.

This work clearly merits the award of the MPhil degree.

Please send this form to the Course Administrator (mlmi-mphil-admin@eng.cam.ac.uk) no later than 12:00 noon on Monday 5th September 2022.

Assessor's Signature 	Date 1 Sep 22
---	------------------