# **Detailed Project Proposal**

|  |  |
| --- | --- |
| First Name: | Clemence |
| Last Name: | Weiss |
| Student Number: | 1804825 |
| Supervisor: | Chris McDermott |

## **Defining your Project**

**1.1 Project title**

***Help:*** *a brief statement about what you are going to do.*

|  |
| --- |
| Using Machine Learning to Detect Hate Speech on Social Media |

**1.2 Background**

***Help:*** *Provide the background to your project. This section should highlight the main topics in the area you are going to research. Essentially what is the project about, what has been done before and why is this project important? ~500 words*

|  |
| --- |
| Hate speech is becoming more and more common on social media platforms and the consequences of these can be very severe. Hate speech threatens peace, as it often precedes real world violence [1], and may lead to hate crimes as it has been linked to attacks on minorities including mass shootings and genocides [3]. Social media can increase hate speech as it makes these words very easy to spread rapidly and with little regulation [3] and correlations have been found between hate speech on social media and hate crimes in the person, such as attacks on refugees in Germany [3].  The goal of this project is to create a functioning software that analyses tones in social media posts that can be used to help detect hate speech, and creating a unique training dataset to be used by the system. It is often difficult to identify hate speech, as definitions vary and something that would be considered offensive to one person or culture may not be to another. This is something that must be considered when creating the software and is one of the big issues in hate speech detection programs [6].  As the amount of content uploaded to social media steadily increases it has become impossible to manually monitor the rise of hate speech and hateful content, therefore it is important to have effective content detectors that can flag hate speech. This is also helpful in reducing the psychological burden of human moderators who have to look through hateful and violent speech to remove them from websites [5]. However there are difficulties with using machine learning, as there are often biases coded into the software by the developers, such as a 2019 study that has found that the current leading AI hate speech detectors were biased against black people [4]. It is therefore important to create datasets that are as unbiased as possible to get the most accurate detections.  Using machine learning is a popular method of detection as it will be more reliable than simple word filters and more efficient than manual flagging. Creating datasets is important in machine learning as these are what the software will be trained on to know what is needed to be identified. Creating a dataset that can detect nuance of speech is one of the biggest difficulties in speech pattern detection, for example a tweet saying ‘I was told women belong in the kitchen’ might be flagged as hate speech despite the fact it is speaking about hate speech. This is a problem as it can lower the user’s trust in the system and if it happens often enough may lead to them ignoring any of the hate speech flags being raised.  [1] UNITED NATIONS, 2019. *Hate speech is rising around the world.* [online]. United Nations. Available from: <https://www.un.org/en/hate-speech> [Accessed 25/09/2022]  [2] ECRI, 2020. *Hate speech and violence.* [online]. Council of Europe. Available from: <https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence> [Accessed: 25/09/2022]  [3] CFR, 2019. *Hate Speech on Social Media: Global Comparisons.* [online]. Council on Foreign Relations. Available from: <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons> [Accessed: 25/09/2022]  [4] Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. 2019. *The Risk of racial Bias Hate Speech Detection.* University of Washington, Carnegie Mellon University, Allen Institute for Artificial Intelligence. <https://maartensap.com/pdfs/sap2019risk.pdf> [Accessed: 25/09/2022]  [5] THE ALAN TURING INSTITUTE, 2021. *Finding critical weaknesses in AI models for hate speech detection with HateCheck.* [online]. The Alan Turing Institute. Available from: <https://www.turing.ac.uk/blog/finding-critical-weaknesses-ai-models-hate-speech-detection-hatecheck> [Accessed: 25/09/2022]  [6] Kovaks, G., Alonso, P., Saini, R., *Challenges of hate Speech in Social Media.* Lulea University of Technology, Sweden. Available from: <https://link.springer.com/article/10.1007/s42979-021-00457-3> [Accessed from: 25/09/2022] |

**1.3 Aim & Objectives**

***Help:*** *Outline what are the main things your project is going to do and what steps or milestones will be used to achieve this aim. The Aim is unlikely to change throughout your project; however, the objectives are likely to adapt to your ongoing research and development.*

**Example**:

Aim: To create a functioning attendance application that efficiently automates the taking of class registers.

Objective 1: Study existing register system in place at RGU and identify weaknesses

Objective 2: Research existing automation technology’s and identify and evaluate those that may be appropriate to taking in class registers

Objective 3: Implement chosen technologies to create prototype application

Objective 4: Conduct user trials to evaluate capabilities of prototype application

Objective 5: Create a refined application incorporating feedback from user trials

|  |
| --- |
| Aim: To create a dataset and use machine learning methods to create a software capable of detecting hate speech on social media  Objective 1: Research existing detection systems  Objective 2: Research machine learning methods and identify the best method for creating datasets and identifying hate speech  Objective 3: Collect data for dataset  Objective 4: Prepare and format dataset  Objective 5: Implement chosen techniques and use dataset to create detection software  Objective 6: Conduct trials to evaluate effectiveness of software  Objective 7: Create final software using trial results  Objective 8: Evaluate the final product and its effectiveness |

**1.4 Tools & Technologies**

***Help:*** *Perform some initial research into the area and outline what tools and techniques you expect to be using in your project.*

|  |
| --- |
| The language used for machine learning will be python.  A web scraping tool will be used to collect data for the dataset.  Trello will be used to keep track of project progression and tasks.  More tools and technologies to use will be decided after the initial research phase where the best tools to use will be found. |

**1.5 Project Plan**

***Help:*** *This is the project plan as to how you will go about achieving your objectives over the timescale of the Honours Project. At a minimum this can be a month-by-month plan.*

|  |
| --- |
|  |

**1.6 Ethics Form**

***You must include in your signed ethics form in this submission, or you will not be able to continue the project.***