# 36118 Applied Natural Language Processing

Assessment task 2A:

**Peer feedback 2**

Banter Block
Abuse Detection

**Student name:** Scott Hamilton
**Student number**: 14325512
**Submission date**: 20 March 2023

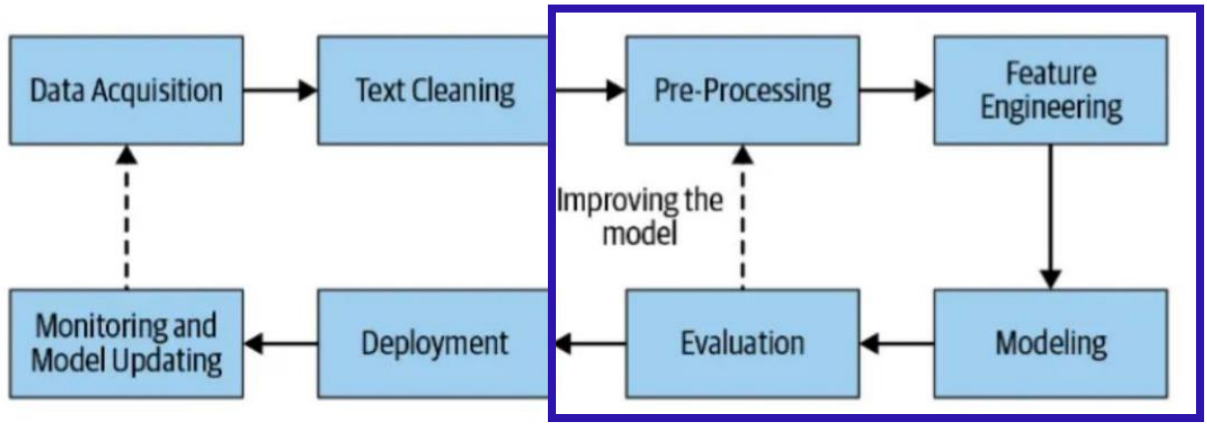
Thank you for the opportunity to review Banter Block's Abuse Detection project.

**1. Summary**

Banter Block's project goal is to identify harmful online text—abusive, racist, sexist, or alike with support vector machine, regression, and neural networks, named entity recognition models including conditional random fields and hidden Markov models, topic modelling and sentiment analysis.

Data acquisition is well progress utilising the Kaggle data set referenced in the appendix, while further web scrapped twitter comments will form the test data.

What perhaps sets this project apart is the intent to use feature engineering to compare model results, and thus remain within the pre-processing-feature engineering-modelling-evaluation loop for a period while features are progressively added to the data for modelling.




What is not clear, or which may be yet to explored, is whether features from one model may be utilised in another to iteratively improve results.

It is anticipated the project will achieve the outcomes sought to incrementally remedy the important issue identified.

**2. Novelty**

Neutral

**3. Suggestions**

Perhaps these thoughts have been considered and obscured within the condensed project summarisation.

Context. From ANLP AT1 literature review, it is recalled that some words considered mainstream in one culture are considered profanity or sexist in another. An example to a lesser extent from my experience; I was brought up in an environment and thus culture where the term 'sheila' was a perfectly normal and accepted as the female equivalent of 'bloke' out in the Queensland bush. I came to Sydney and found quite quickly this not to be the case and calling someone a 'sheila' is seen as derogatory.

Deployment. Deployment and operation of the model into a feature where abusive text is prevented at the time of input and a real-time or near real-time manner. Perhaps at the output, users may customise their tolerance for level of profanity they accept. For example, on a scale of 1 to 10, a prude may set the level at 1, which would redacts *Gone With the Wind's* provocative at the time 'Frankly my dear, I don't give a damn', while at level 10 the user receives text unvetted.

**Appendix**:
1. Banter Block's Abuse Detection – project summary

**Appendix 1**

36118 Applied Natural Language Processing – Assignment 2A: Project Summary
Abuse Detection

# Banter Block

## PROJECT OBJECTIVES

Online abuse and toxicity are the most frequent issues users encounter when using online forums. The major goal of this research is to identify the phrases and sentences that indicate that someone or a group of individuals is being abusive, toxic, racist, or sexist.

## DATA ACQUISITION

The Dataset has been chosen from Kaggle repository [1]

Web Scraping twitter comments for testing our model

## NLP TECHNIQUES

- Text Classification models such as Support Vector Machines, Logistic Regression, and Neural Networks
- Sentiment Analysis
- Named Entity Recognition Models such as CRFs and HMMs
- Topic Modeling
- Sequence Labeling
- Feature Engineering to compare the results of the models implemented

## PROJECT OUTCOMES AND INSIGHTS

The project is expected to flag abusive comments, words, and phrases used on various online platforms. This could be helpful to multiple groups of people who undergo abuse online. This includes groups such as women, children and teenagers, people of the LGBTQ+ community, Journalists and activists, etc.

The significance of the project can be drawn from the fact that every human faces abuse online to a certain extent. This model could help reduce the derogatory phrases that we come across on the internet which could greatly benefit factors such as inclusivity and mental health.

## PROJECT PROGRESS

The dataset has been chosen from Kaggle and cleaned.
The Machine learning models to implement have been discussed
A GitHub repository has been made to track project progress. [2]

## TEAM MEMBER ROLES

For this particular assignment every team member will be working on a round robin basis.
The overview of the roles would be as follows:

Prinston Mascarenhas (24587331) - Technical Lead.
Akanksha Kamath (24683498) - Report Writing Lead.
Hema Vajravelu ( 24613503) - Process Implementation
Ronik Jayakumar (24680264) - Project Manager

## APPENDIX

[1] : https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset
[2] : https://github.com/prinston27/NLP_AT2