NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Toxic Comment Detection

Sentic Computing for Social Good

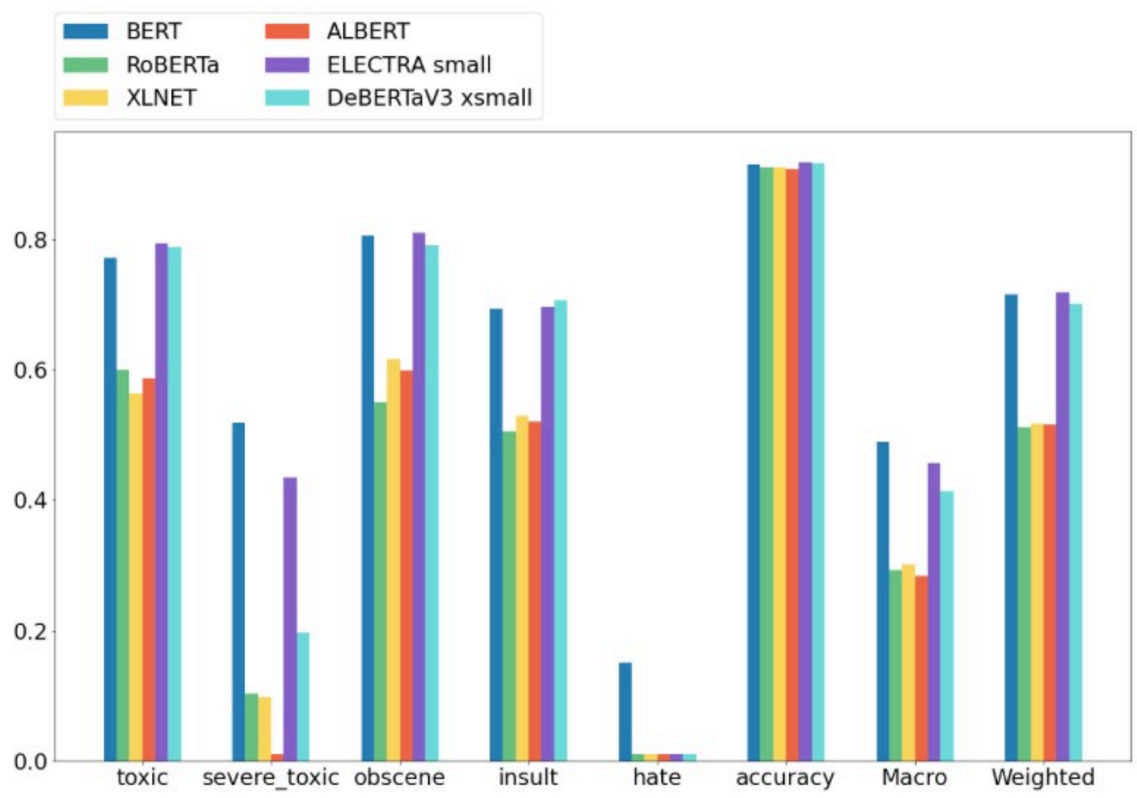
Student: Wang Jingtan

Supervisor: Assoc Prof Erik Cambria

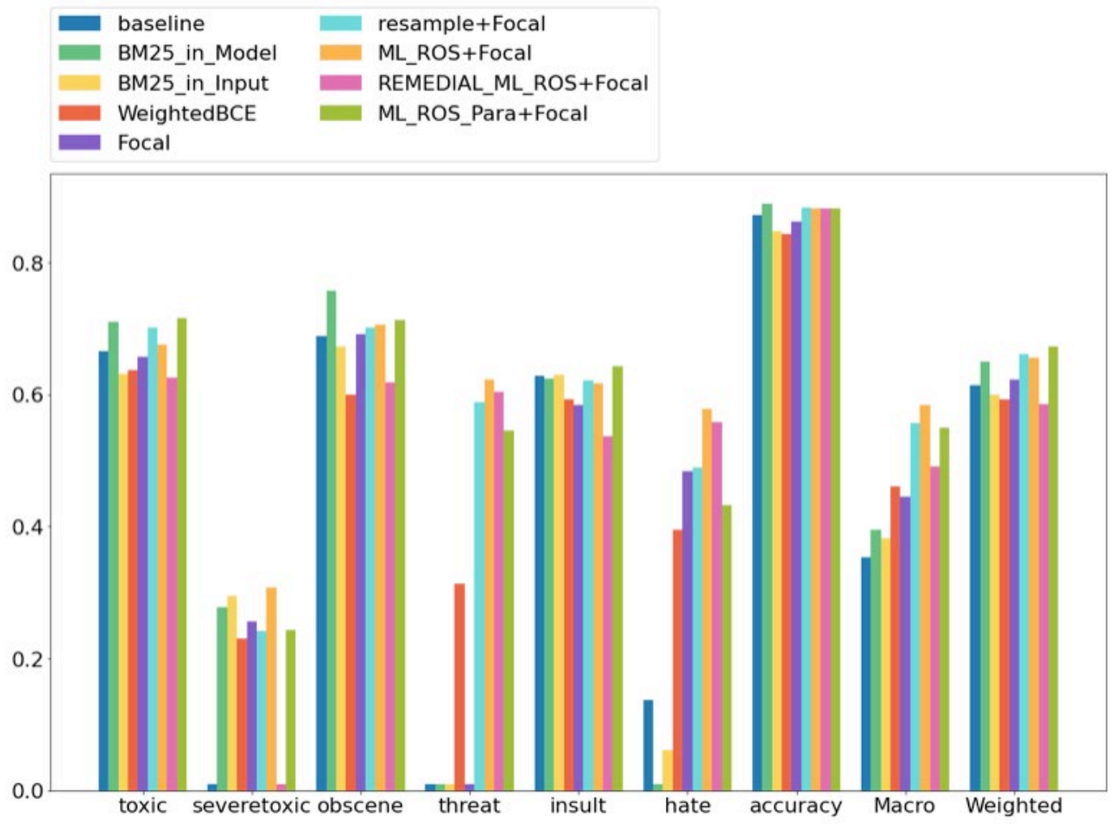
Project Objectives:

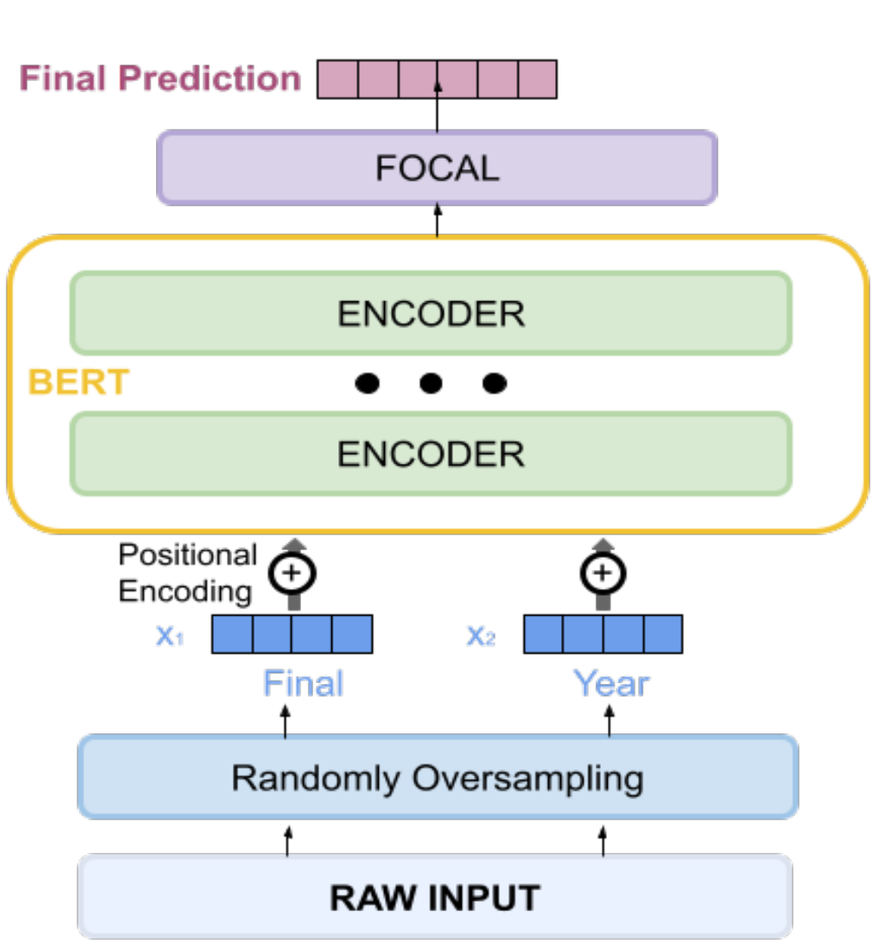
Surrounded by a massive amount of information and hurt by the toxic comments hidden in them, it is critical to filter these inappropriate messages, preventing people from verbal violence. Our project adopted deep learning models for toxic comment detection: predicting the belongs of a raw textual input into six categories of inappropriate behaviors: namely Toxic, Severe toxic, Obscene, Threat, Insult, and Identity Hate. The designed experiment and results revealed a possible approach for similar problems and served as an inspiration for future studies.

Deep learning Models Comparison:



Address Minority Labels:





- **Dataset**
  Real-word Multilabel Dataset
- **Pre-trained Model Comparisons**
  BERT, ALBERT, RoBERTa
  ELECTRA XLNET, DeBERTaV3
  BERT performed the best
- **Address Imbalance**
  External Feature, Loss Metrics Modification
  Resampling Method
  BERT + Focal Loss +
  Random Oversampling Method (ML_ROS)