Detecting Cyberbullying in Localized Text Messages using a Novel Genetic Noisy Student Training Technique

Prannaya Gupta

Motivation and Background

- Cyberbullying on the rise since the start of the Covid-19 pandemic
- **27%** of Children (8-12 yrs old) were affected by Cyberbullying worldwide
- Singapore ranked **13th** in a list of 30 countries

Current Problems in Field

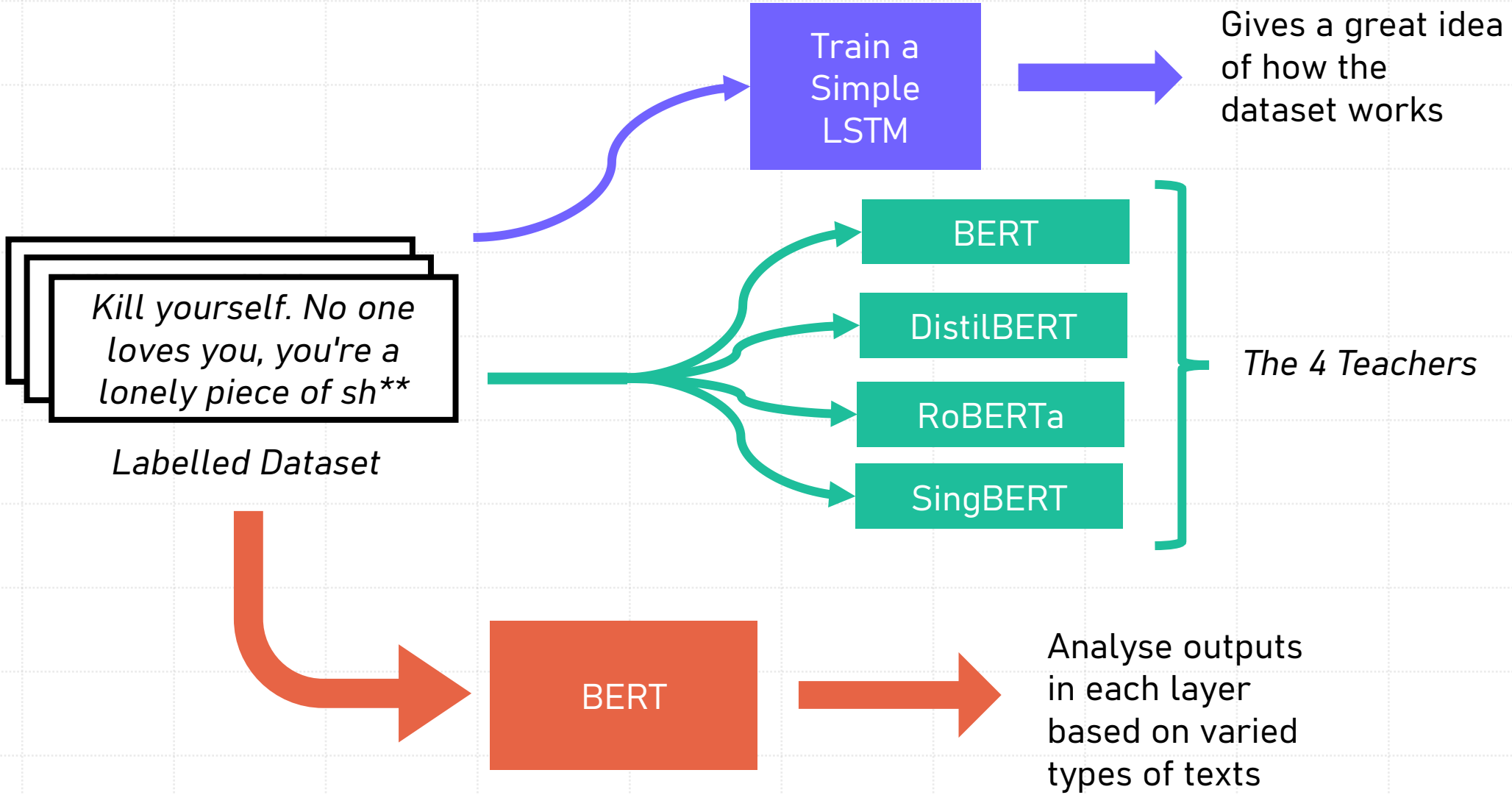
- Much of Cyberbullying is attacks based on Race, Religion, Gender
  - Models are often **biased**
- We are prone to speaking in multiple languages
  - Most language models are unable to support this
  - multilingual text is **ignored**

Key Idea: We need models to be interpretable

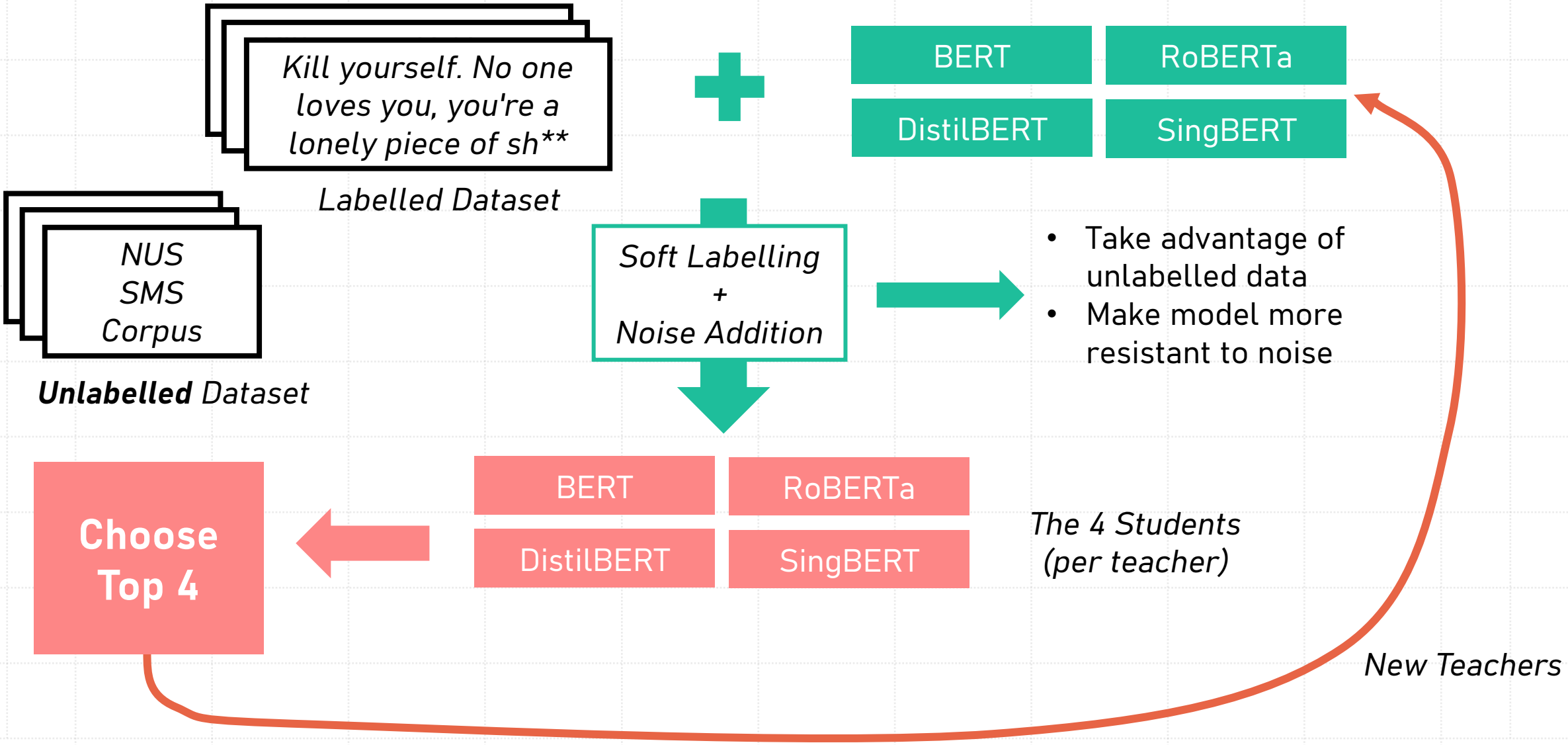
Proposed Project

- We propose observing how models train on the following:
  - Mendeley's Cyberbullying Dataset – Separated samples of aggressions, attacks
  - SOSNet's Cyberbullying Dataset – Simple Binary Classification
- Research Questions:
  - What happens when specific texts are fed into the model? Is the model biased towards texts talking about Race, Religion, Gender?
  - How does the accuracy change when more noise is fed into the dataset, and the Noisy Student Approach is conducted?

Proposed Methodology



Proposed Methodology (2)



Novelty

- Noisy Student is a recently introduced structure (2018)
- We introduce a Classroom Structure:
  - Each “teacher” teaches their “*styles*” to the “student”
  - Natural Selection to decide new teachers
  - More realistic, similar to real life
- We are also investigating the interpretability of a Cyberbullying-related model
  - Helps debias models and makes sure that specific flaws are not introduced

Thank you

for listening!