Jigsaw Unintended Bias in Toxicity Classification

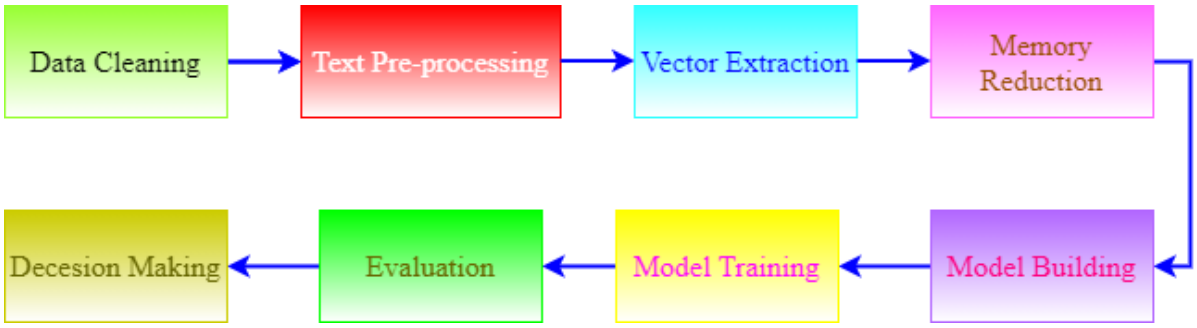
Author: Khadija Akter Lima

1. Introduction

Toxic comment classification has become an active research field with many recently proposed approaches. Keeping online conversations constructive and inclusive is a crucial task for platform providers. Automatic classification of toxic comments, such as hate speech, threats, and insults, can help in keeping discussions fruitful. For this purpose, I have developed a deep learning model utilizing Bidirectional LSTM. My proposed model achieved accuracy of 93.38%. This project focuses on utilizing less resource to get optimal solution. This is a data centric project where best solution is being tried to achieve through exploratory data analysis, text pre-processing.

2. Methodology

Work Flow Diagram >>



Model Architecture >>