Capstone Project 2 Proposal

Wikipedia Toxic Comment Classification

Data Science Career Track, By Minh Ngoc Pham

Introduction

Dataset: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
This dataset contains Wikipedia comments that have been rated for toxic behaviour. There are 6 classes for 6 types of toxic behaviour including: toxic, severe_toxic, obscene, threat, insult, identity_hate. There are 395,667 comments in the training set (which also includes comments that are rated as neutral, meaning being scored 0 for all classes). The test set include 381,580 comments.

The problem

There is an increasingly high number of toxic behaviour and comments over the internet which are making it difficult to have meaningful discussion on the net. As the world is becoming more and more technologically involved and people becoming more present on the internet as opposed to in real life, it is important to make it a healthy and safe place to express opinions. One does not need to look far but at Youtube, where presence of toxic and hatred comments in ubiquitous in almost any video. This has led many Youtube channels to disable the comment section, which in itself might be a bit of an extreme measure given the fact that genuine commentator might be interested in the content and want engage in meaningful discussion. It is important to come up with a good way to filter out comments that are outright toxic and not healthy on the net.

Potential clients

Many organisations are now concerned with tools to improve online conversation. This project will be of interest to:

• Social media platforms such as youtube, facebook, instagram, reddit and twitters which might be interested in bettering their toxic comment filtering tools

• News agency and other online discussion forums that are concerned with their comment sections.

Problem Solving Steps

• This is a supervised classification problem

• This is a **multi-label classification** problem because each comment can belong to more than 1 class simultaneously (there are 6 labels in total)

• The dataset will still need to be clean up using stop word,

Deliverables

• Final Codes

• Final Report on the toxic comment classification

• A blog post

• Presentation slides