Unintended Bias in Toxicity Classification

Capstone 2 Project for Data science Career Track

Natural Language Processing is a complex field which is hypothesised to be part of AI-complete set of problems, implying that the difficulty of these computational problems is equivalent to that of solving the central artificial intelligence problem of making computers as intelligent as people. With over 90% of data ever generated being produced in the last 2 years and with a great proportion being human generated unstructured text there is an ever increasing need to advance the field of Natural Language Processing.

Recent UK Government proposal to have measures to regulate social media companies over harmful content, including "substantial" fines and the ability to block services that do not stick to the rules is an example of the regulamentary need to better manage the content that is being generated by users.

Other initiatives like Riot Games' work aimed to predict and reform toxic player behaviour during games is another example of this effort to understand the content being generated by users and moderate toxic content.

However, as highlighted by the Kaggle competition Jigsaw unintended bias in toxicity classification, existing models suffer from unintended bias where models might predict high likelihood of toxicity for content containing certain words (e.g. "gay") even when those comments were not actually toxic (such as "I am a gay woman"), leaving machine only classification models still sub-standard.

The outcome of our analysis is the type of algorithm that companies will use to define what is free speech and what shouldn't be tolerated in a discussion. This challenge actually starts with how the training dataset was produced: Multiple people (annotators) read thousands of comments and defined if those comments were offensive or not. Where is the trick? They disagreed in many of them. Having tools that are able to flag up toxic content without suffering from unintended bias is of paramount importance to preserve Internet's fairness and freedom of speech.

Dataset :

At the end of 2017 the Civil Comments platform shut down and chose make their ~2m public comments from their platform available in a lasting open archive so that researchers could understand and improve civility in online conversations for years to come. Jigsaw sponsored this effort and extended annotation of this data by human raters for various toxic conversational attributes.

In the data supplied for this competition, the text of the individual comment is found in the comment_text column. Each comment in Train has a toxicity label (**target**), and models should predict the **target toxicity** for the Test data. This attribute (and all others) are fractional values which represent the fraction of human raters who believed the attribute applied to the given comment.

For evaluation, test set examples with target >= 0.5 will be considered to be in the positive class (toxic).

The data also has several additional toxicity subtype attributes. Models do not need to predict these attributes for the competition, they are included as an additional avenue for research. Subtype attributes are:

- Severe_toxicity
- Obscene
- Threat
- Insult
- Identity_attack
- sexual_explicit

Approach:

Data Pre-Processing and EDA:

> initially basic exploration will be performed based on number of word, character, uppercase word etc
> later data will be analysed based on race or ethnicity, gender, sexual orientation, religion and disability based on Toxicity subtype attributes
> Also analysis based on metadata features like rating, funny, wow and creation date can have some important information
> Also understanding of comments will be done after some preprocessing like Lower casing, Punctuation removal, Frequent words removal, Spelling correction, Tokenization, Lemmatization
> Word vectorization and Sentence Vectorization with feature engineering will be performed

Modelling :

> Firstly standard classification model will be implemented on word and sentence vectors
> later deep learning models like LSTM will be implemented to gain better results
> all implemented models will be evaluated and report will be created

Writeup:

The deliverable will include a powerpoint presentation and python code. Powerpoint will include related EDA Graph, any correlation analysis , final model , confusion matrix , related statistics summary, recommendations