CASE STUDY – TOXIC COMMENT CLASSIFICATION

Website: www.analytixlabs.co.in
Email: info@analytixlabsl.co.in

BUSINESS CONTEXT:

One of the leading digital technology company is working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The objective of this case study is to build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate.

Data Availability: You'll be using a dataset of comments from Wikipedia's talk page edits. Improvements to the current model will hopefully help online discussion become more productive and respectful

Data Description:

You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

You must create a model which predicts a probability of each type of toxicity for each comment.

File descriptions:

train.csv - the training set, contains comments with their binary labels

test.csv - the test set, you must predict the toxicity probabilities for these comments (which are not included in train).

sample_submission.csv - a sample submission file in the correct format

Note:

This problem can be solved using any machine learning algorithm. However, you required to use any deep learning technique to solve the problem.

This case study is sourced from Kaggle competition (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data), you may go through different discussions to get more understanding about different approaches and data sets.