# Report of Toxic Span Detection

Name: Sufyan Shaha

## Introduction

Background of the Task

The Toxic Spans Detection task concerns the evaluation of systems that detect the spans that make a text toxic, when detecting such spans is possible. Toxic Span Detection is used to classify foul language which can then be moderated on online platforms as these platforms have users ranging from all walks of life.

Description of the Task

A model is used to identify a span of toxic words in the given sentence from the dataset. The dataset contains sentences similar to those which can be found online in a forum or a discussion. The task is to classify the toxic span of words in the sentence which may or may not include any toxic elements at all.

Application

The primary application of this is moderation. Moderation in online platforms and negativity filtering is key to maintaining a positive and progressive forum/discussion. Such models can be used to identify any potential comments from users which might lower the quality of the platform.

Challenges faced

The main challenge which a model can face is the language barrier. Online communities are often diverse and the model might be completely useless when a different language is being used. Another challenge is the different ways a toxic sentence can be presented. Human languages can be ambiguous at times and the model might not be able to predict a toxic span even though it may be highly offensive.

## Problem Formulation

Input

A dataset containing sentences from online forums and comments was used. The comments may have one or more toxic spans. Some sentences might even need processing as they could contain special character like a URL

Output

The output of the task is a classification of whether the comment contains a toxic span or not.

Task Type

Sequence Labelling is used to classify a toxic span. The model was trained on the given dataset to recognize the toxic spans and was then used on a labelled dataset to validate it. The toxic words, if found are labeled as toxic.

## Method

Data Preprocessing

Since the data is from online comments, processing it is a vital task before we can even begin with our model. Spacy, nltk and tokenizer libraries were used to clean the data for the model. It was then made into bert tokens using huggingface’s bert pretrained model. The data was then padded and masked.

Model Design

A pretrained model of Bert from huggingface was used. Huggingface offers a lot of pretrained models like Roberta,distillabert and bert. In this project, a pretrained Bert classifier was used as it was the model I am most familiar with. The optimizer used was Adam. For loss, Binary Cross Entropy was used.

## Experiments

Hardware

The model was run on Colab on a Tesla GPU for three epochs.

Text

Description automatically generated

Evaluation

The dataset had a high number of toxic span distribution as can be seen from the histogram.

Chart, histogram

Description automatically generated

The f1 scores curve was

Chart

Description automatically generated

The loss curve was

Chart, histogram

Description automatically generated

The F1 score calculated was 0.6536. It took approximately 8 minutes for each epoch.

## Conclusion

Toxic Span detection is a complex task and can be improved with using more data by enlarging the dataset. Using complex models like Roberta and DistillaBert can also improve the performance of the task.