Toxicity Classification Model

Project Report

Submitted by

Hemang Sharma (Student ID: 24695785)

Nusrat Zahan (14367472)

Rajveer Singh Saini (14368005)

**ABSTRACT**

The prevalence of toxic and abusive language in online communication has become a significant problem in the digital age. To address this issue, researchers and developers have been working on creating automated tools and models that can detect and classify toxic language in text data. One such tool is the Toxicity Classification Model, which is designed to identify and flag abusive language in online communication.

Toxicity Classification Model is a type of machine learning model that is designed to identify and classify toxic or abusive language in text data. The model is trained on a dataset of labeled examples of toxic language, and learns to recognise patterns and features in the text that are associated with toxic or abusive content.

The goal of a Toxicity Classification Model is to accurately predict whether a given piece of text contains toxic language, and to assign a score or label that indicates the degree of toxicity. There are many different approaches to building a Toxicity Classification Model, but most involve using a combination of natural language processing (NLP) techniques and machine learning algorithms.

This report provides a comprehensive overview of the Toxicity Classification Model, its applications, and the challenges involved in building and deploying this type of model in the real world.Ultimately, this report highlights the importance of the Toxicity Classification Model in promoting healthy online interactions and creating a safer and more inclusive online environment.

**ACKNOWLEDGEMENT**

I would like to acknowledge the Gadigal people of the Eora Nation upon whose ancestral lands our City campus now stands. I would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for this land. I would also like to convey my heartfelt thanks to, Dr. Shibani Aileen for her invaluable advice, constant encouragement, prompt assistance and provision of an ideal research environment. Throughout the project, despite her hectic schedule, she has been a cordial supporter of my efforts to complete this project.

Hemang Sharma (24695785)

Nusrat Zahan (14367472)

Rajveer Singh Saini (14368005)

**TABLE OF CONTENTS**

## ABSTRACT

## ACKNOWLEDGEMENT

## LIST OF TABLES

## LIST OF FIGURES

## ABBREVIATIONS

## INTRODUCTION

## SYSTEM ANALYSIS:

## Data description

## MODEL SPECIFIC DETAILS

## SYSTEM DESIGN:

## ARCHITECTURE

## RESULTS:

## DATA (Explanation)

## CODE & EXPLANATION OF CODE

## OUTPUTS/GRAPHS

## CONCLUSION

## FUTURE ENHANCEMENTS

**1. INTRODUCTION**

In recent years, the prevalence of toxic and abusive language in online communication has become an increasingly urgent issue. From social media to online gaming and beyond, people of all ages and backgrounds are exposed to a range of harmful content that can have a profound impact on their well-being and mental health. To address this problem, researchers and developers have been working to create automated tools and models that can detect and classify toxic language in text data. One such tool is the Toxicity Classification Model, which is designed to identify and flag abusive language in online communication. In this report, we will explore the workings of a Toxicity Classification Model, its applications, and the challenges involved in building and deploying this type of model in the real world.

We are building a model that recognises toxicity and minimises bias with respect to mentions of identities. We will be using dataset labeled for identity mentions and optimising a metric designed to measure unintended bias. The reason for choosing this project is to create a tool that could help moderators in identifying and removing such comments, improve online discourse and create a safer environment for people to engage in online conversations.

We will also discuss some of the ethical and societal considerations that arise when using automated tools to moderate online content, and suggest possible ways to address these concerns.