**CSCE 5290: Natural Language Processing**

**Project Proposal**

**Fight Online Abuse**

**Toxic Comment Classification**

**Team Members**

1. Chakravarthy Akula

2. Amith Reddy Nomula

3. Dheeraj Peddi

4. Rohith Sirisanagandla

( Github : <https://github.com/Dheeraj2326/NLP> )

**1. Goals and Objectives**

**1.1 Motivation**

Internet is a wide source of communication globally where people freely express their thoughts, feelings and opinions. While this situation contributes significantly, it involves great dangers since these written texts can be highly toxic causing offense, abuse, harassment and bullying behaviors. Online sites struggle to promote discussions effectively, leading many communities to limit or shut down user comments altogether. By analysing the text and restricting users to post toxic comments, we could promote safer discussions on various online portals.

**1.2 Significance**

Toxic comments passed by a user on the internet portals are offensive, rude and disrespectful. It may also take an extensive verge, where the users send threatening, degrading, or sexually explicit messages. As a result, detecting and reporting such toxic comments from online portals is crucial. Analysing user comments over the internet is crucial for maintaining civility in online communities.

**1.3 Objectives**

We aim to create a machine learning model which detects the level of toxicity a piece of text contains. This model will be used for fighting online abuse making online communities grow better and healthier in user discussions and conversations openly. We hence intend to help online communities detect and restrict toxic comments which any user tries to post.

**1.4 Features**

The types of toxicity the text can contain are toxic, severe toxic, obscene, threat, insult and identity hate. We propose to develop a machine learning model which uses natural language processing techniques for detecting the type of toxicity a comment contains. This model provides, as a result the probability of each type of toxicity for a given comment.

**References**

1. Navoneel Chakrabarty, “A Machine Learning approach to comment classification”, <https://arxiv.org/ftp/arxiv/papers/1903/1903.06765.pdf>
2. Iarjset Jornal, “Toxic comment classification using neural networks and machine learning”, <https://iarjset.com/papers/toxic-comment-classification-using-neural-networks-and-machine-learning/>
3. Darko Androcec, “Machine Learning methods for toxic comment classification: A systematic review”, <https://www.researchgate.net/publication/349929587_Machine_learning_methods_for_toxic_comment_classification_a_systematic_review>