"Linguistic Model for Analyzing and Detecting Inappropriate Comments“

Chennuru Giridhar (21CS002368) ,Syed Roshan (21CS002424),
Thotamadugula Gowtham Reddy (21CS002426)

SPSU
Sir Padampat Singhania University
a JK Cement Initiative

Abstract/Intro/Motivation

This project aim develops linguistic model identifying and analyzing inappropriately online comments, driven by critical need for better content moderation tools. Methods exist often fall short detect nuanced inappropriate language, impacting user experience and community well-being. Leveraging natural language processing techniques, this model enhance content moderation, fostering healthier online communities and enabling platforms maintain positive interactions!

Background

Within digital spaces, spotting and rectifying unsuitable remarks hold significance in sustaining favorable user encounters and communal well-being. Present content oversight strategies frequently battle with intricate linguistic subtleties, underscoring the necessity for superior language models to boost online security and involvement.

Objectives

The objective of this project is to develop a robust linguistic model capable of accurately identifying and analyzing inappropriate comments in online text, with the goal of enhancing content moderation practices and fostering healthier and more respectful online interactions.

Methods

The project will utilize binary classification methods, incorporating TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction and part-of-speech (POS) tagging to enhance linguistic analysis. Through a series of NLP techniques including text preprocessing, data will be collected and annotated to train the model on various forms of inappropriate language. Following this, supervised machine learning algorithms will facilitate model training and evaluation using specific metrics to gauge its proficiency in detecting and analyzing inappropriate comments. Iterative refinement will play a crucial role in optimizing the model's performance and ensuring its robustness across diverse online platforms and contexts.

Results/Discussion

The model demonstrated effective performance in detecting inappropriate comments across diverse online platforms and contexts. It exhibited strong accuracy and precision, significantly improving content moderation efforts and enhancing user experience. Iterative refinement further optimized its effectiveness, highlighting its value in maintaining online community health and fostering respectful interactions.

Future Directions

Looking ahead, this project will focus on refining the model's ability to adapt to evolving forms of inappropriate language by updating and expanding the dataset. Exploring advanced NLP techniques, like deep learning architectures such as transformers, could enhance detection accuracy. Integrating real-time feedback mechanisms will enable adaptive learning and timely responses to emerging online trends. Collaborations with platforms will facilitate deployment and scaling, promoting safer digital environments.