Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

INAPPROPRIATE LANGUAGE AND HATE SPEECH RECOGNITION

Mrs. S.Suganya, Assistant Professor, Dept. Of Computer Science, SSM Institute of Engineering and Technology, Dindigul.
G.Y.Asmetaa, Student 1, Dept. Of CSE, SSM Institute of Engineering and Technology.
S.Harini, Student 2, Dept. Of CSE, SSM Institute of Engineering and Technology.
S.Jeyashree, Student 3, Dept. Of CSE, SSM Institute of Engineering and Technology.
J.Sriram, Student 4, Dept. Of CSE, SSM Institute of Engineering and Technology.

Abstract
The problem of online abuse, harassment, and discrimination is a growing concern in today's digital world. To address this issue, the project "Identify inappropriate language and hate speech" aims to develop algorithms and models that can automatically detect and flag instances of inappropriate language and hate speech in text-based content. The project uses natural language processing (NLP) and machine learning techniques to analyze text and identify patterns that are associated with inappropriate language and hate speech. Large datasets of annotated text are used to train and test the models, which are designed to be applied to a wide range of online content, including social media posts, comments, and messages. The ultimate goal of the project is to create tools and technologies that can help to reduce the prevalence of online abuse and promote a more respectful and inclusive online community.

Keywords: Artificial intelligence, Natural Language Processing, Text Processing, Deep Learning, API's.

I. Introduction

The rise of social media has given everyone a platform to express their opinions and thoughts. While this has created a space for free speech, it has also led to an increase in hateful and offensive language. The internet has become a breeding ground for trolls, bullies, and hate speech. This type of content not only harms individuals but can also have a negative impact on society as a whole. The current methods used to moderate online content are not fool proof and require a lot of manual work. This leads to delays in identifying harmful content and removing it from the platform. Hence, there is a need for an automated system that can identify and flag hateful and offensive language in real-time.

The Hate Speech Recognition mini-project aims to develop a model that can detect offensive language and hate speech in social media text data. The model is trained on a dataset of Twitter data, where each tweet is labeled as either hate speech, offensive language, or non-offensive text.The project utilizes a Twitter dataset containing tweets along with their corresponding labels. The dataset is loaded using the pandas library and preprocessed to clean the text. The preprocessing steps include converting text to lowercase, removing URLs and hashtags, tokenizing the text, removing stopwords and punctuation, and lemmatizing the tokens.The preprocessed text data is transformed into numerical features using the CountVectorizer, which creates a matrix of token counts. The dataset is split into training and testing sets using the train_test_split function. A Decision Tree Classifier is then trained on the training data. The trained model is evaluated on the testing data to measure its performance in detecting hate speech and offensive language.

II. Literature

Guanyi Mou and Kyumin Lee proposed An Effective, Robust and Fairness-aware Hate Speech Detection Framework with the widespread online social networks, hate speeches are spreading faster and causing more damage than ever before. Existing hate speech detection methods have limitations