

Offensive Text Detection

09.03.2023

Text Guardians Team

Machine Learning System Design - Course Project

Overview

Offensive text detection is a growing concern in today's online world. The objective of this project is to develop an offensive text detection system for English text, using Natural Language Processing (NLP) techniques. The system will be designed to identify and flag offensive or inappropriate English text in online platforms, including social media sites and chat applications. The system aims to enhance online safety by automatically detecting and filtering inappropriate text, preventing potential harm or offense to users.

Goals

1. To develop a robust offensive text detection system that can accurately identify offensive or inappropriate text in online platforms.
2. To create a user-friendly interface for the system that can be integrated into various online platforms to filter offensive text in real-time.


Specifications

Requirements

1. Access to a large and diverse dataset containing offensive and non-offensive English text data for training and testing the model.
2. Access to computational resources such as high-performance GPUs to train machine learning models.
3. Expertise in NLP techniques such as text pre-processing, feature engineering, and classification algorithms.
4. A cloud-based server or high-performance computer for deploying the system.

Objectives

1. Develop a machine learning model using NLP techniques to classify offensive and non-offensive text in various contexts.
2. Enhance online safety by identifying and filtering inappropriate text in real-time.

- 
3. Develop an API to interface with the system and integrate it with various online platforms.
 4. Design a user-friendly interface that displays the results of the offensive text detection system.
 5. Continuously improve the system's performance through data-driven analysis and optimization techniques.

Constraints

1. Limited resources for data collection and processing may limit the scope of the project.
2. Time constraints for developing and deploying the system.
3. Limited availability of computational resources for training machine learning models.
4. Legal and ethical constraints related to the use of user data and privacy concerns.
5. The model may not be able to accurately detect all instances of offensive language, as language use can vary widely and may be context-dependent.
6. The model should have a high accuracy rate and low false positive rate.
7. The model should not be biased towards specific races, genders, or religions.

Features

1. The system will be able to detect a wide range of offensive text, including hate speech, cyberbullying, and profanity.
2. The system will be able to operate in real-time to detect and flag offensive text as it is posted.
3. Integration with various online platforms, including social media sites and chat applications.
4. Customizable parameters for the model, such as the level of sensitivity to offensive language.

Jobs that influence the project:

1. Data gathering and preprocessing: Collecting and preprocessing diverse and representative offensive and non-offensive text datasets.
2. Model development and training: Developing and training machine learning models using NLP techniques to accurately detect offensive text.

3. System integration and interface development: Developing a user-friendly interface for the offensive text detection system and integrating it into various online platforms.
4. Testing and optimization: Testing the system's performance and optimizing it based on data-driven analysis.
5. Deployment and maintenance: Deploying the system and maintaining its performance and functionality.

Estimations

- Data gathering and preprocessing: 2 weeks
- Model development and training: 4-6 weeks
- System integration and interface development: 2-3 weeks
- Testing and optimization: 2-3 weeks

Total estimated time: 10-14 weeks

Dataset

The following dataset will be used as training data

- Hammer, H. L., Riegler, M. A., Øvrelid, L. & Veldal, E. (2019). "THREAT: A Large Annotated Corpus for Detection of Violent Threats". 7th IEEE International Workshop on Content-Based Multimedia Indexing.

This dataset consists of a total of around 30,000 English sentences from around 10,000 YouTube comments. Each sentence is manually annotated as either being a violent threat or not.

Members

1. Hamed Saadati, hamed.saadati078@gmail.com (Team Head and ML engineer)
2. Behnam Saedi, saedi.behnam.336@gmail.com (DevOps and ML engineer)