## A PROJECT REPORT ON

**NLP INAPPROPRIATE COMMENT SCANNER**

***Submitted by***

***M.S. LEELA PRASAD (192210150)***

**Under the guidance of**

# Dr. ANITHA .C



## SIMATS ENGINEERING THANDALAM

**INTRODUCTION:**

The amount of text data generated every day in the modern digital age has increased exponentially due to the spread of online communication platforms. But while improved communication has its advantages, there is also a serious problem with incorrect language that can be damaging, rude, or discriminating. It is imperative to address this issue in order to promote a polite and safe online community, especially in online forums, social media sites, and communities where members engage in unrestricted communication.   
  
This project's goal is to create an inappropriate comments scanner—a tool that uses natural language processing (NLP) techniques to identify and remove rude or improper remarks from textual data. This project is really important since it helps to foster a more positive internet conversation. Using natural language processing (NLP) techniques for text categorization is the fundamental strategy. The model will be trained to discriminate between appropriate and inappropriate remarks. To enable efficient model training, this project includes a number of crucial elements, including feature extraction, annotation, labeling, and preprocessing of the dataset.   
  
By tackling this issue, the project hopes to advance the field of NLP research by offering a useful application that is concentrated on improving online safety and promoting a more welcoming digital space. The project's findings and conclusions will not only show that using natural language processing (NLP) to recognize inappropriate comments is feasible and effective, but they will also be a useful tool for online communities and platforms to reduce abusive behavior and encourage positive interactions. One cannot stress the importance of this project enough. With social media platforms replacing traditional print media as the main means of communication, damaging remarks are a serious threat to the foundation of today's digital society. Such behavior can have far-reaching effects on people's mental health, undermine trust in online communities, and even incite physical violence. Examples of these consequences include cyberbullying and targeted harassment campaigns.   
  
Using the most recent NLP approaches for text classification is the fundamental strategy to deal with this urgent problem. In order to enable automatic content moderation at scale, the model will be trained to distinguish between appropriate and inappropriate comments using a combination of supervised learning and sophisticated feature extraction techniques.

**2.Problem Definition and Algorithm :**

**2.1 Task Definition:**

Problem Statement: The task at hand is to develop an Inappropriate Comments Scanner, a Natural Language Processing (NLP) system capable of identifying and categorizing inappropriate or offensive comments within textual data.

**Formal Specification:**

**Inputs:**

1. Textual data containing comments or messages (e.g., social media posts, forum discussions, customer reviews).

2. Predefined criteria or guidelines for what constitutes inappropriate or offensive language (e.g., profanity, hate speech, harassment).

Output:

1. Classification labels indicating the presence or absence of inappropriate content (binary classification)

2. If inappropriate content is detected, categorize it into specific types (multi-class classification).

**Why This Problem is Interesting and Important:**

In the digital age, online platforms are faced with the formidable task of managing vast amounts of user-generated content while upholding social responsibility and ensuring user safety. The proliferation of inappropriate comments presents challenges that can have profound negative impacts on individuals, from psychological harm to perpetuating discrimination and cyberbullying. To address these challenges, developing tools for automatic detection and filtering of such content is imperative. This not only contributes to creating healthier online communities but also protects users from exposure to harmful material in real-time. . Effective content moderation is essential for maintaining brand reputation and compliance with community guidelines and legal regulations. By automating the detection of inappropriate comments, platforms can augment human moderation efforts, enabling efficient review of large volumes of content. Engaging in projects like building an Inappropriate Comments Scanner provides students with valuable hands-on experience in essential Natural Language Processing (NLP) techniques, fostering skill development in text analysis and machine learning.

**Algorithm Definition:**

The algorithm we'll employ for the Inappropriate Comments Scanner project is a supervised learning approach using text classification techniques. This algorithm aims to distinguish between appropriate and inappropriate comments in natural language text.

**1. Dataset Preparation:** This step involves cleaning and preparing the dataset by removing irrelevant information, such as special characters, emojis, and HTML tags. Additionally, the text may be tokenized into individual words or subwords for further processing.

**2. Labelling and Annotation:** Each comment in the dataset is labeled as either appropriate or inappropriate based on predefined criteria. Human annotators may be involved in this process to ensure accuracy.

**3. Feature Extraction :** Features are extracted from the preprocessed text data to represent the comments in a numerical format suitable for machine learning algorithms. Common techniques include bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings like Word2Vec or GloVe.

**4. Text Classification:** For text classification, we'll train a machine learning model such as a Support Vector Machine (SVM), Logistic Regression, or a neural network (e.g., LSTM or CNN). The model learns to classify comments into appropriate or inappropriate categories based on the extracted features.

**Pseudocode:**

1. Preprocess the dataset:

- Clean text data (remove special characters, emojis, etc.).

- Tokenize text into words or subwords.

2. Label comments as appropriate or inappropriate:

- Human annotators may assist in this labeling process.

3. Extract features:

- Use techniques like bag-of-words, TF-IDF, or word embeddings to represent comments numerically.

4. Train a text classification model:

- Split the dataset into training and testing sets.

- Choose a classification algorithm (e.g., SVM, Logistic Regression, LSTM).

- Train the model using the training data.

- Evaluate the model's performance on the testing data.

5. Deploy the model:

- Use the trained model to classify new comments as appropriate or inappropriate.

**Example:**

Consider the following comment: "That's a fantastic idea! I can't wait to see it implemented."

1. Preprocessing: Remove special characters, resulting in "That's a fantastic idea I can't wait to see it implemented."

2. Labeling: Annotate as appropriate.

3. Feature extraction: Represent the comment using TF-IDF vectors or word embeddings.

4. Classification: The trained model predicts the comment as appropriate based on the extracted features.

This algorithm iterates through the dataset, learning patterns from labeled comments to accurately classify new comments as appropriate or inappropriate.

**3. Experimental Evaluation:**

**Methodology:**

**Criteria:** The method will be evaluated based on several criteria to assess its effectiveness in identifying inappropriate comments:

**1.Accuracy :** The primary metric for evaluation will be the accuracy of the model in correctly identifying inappropriate comments from the dataset.

**2. Precision and Recall :** Precision measures the proportion of correctly identified inappropriate comments among all comments flagged as inappropriate, while recall measures the proportion of correctly identified inappropriate comments among all actual inappropriate comments in the dataset. These metrics will help assess the model's ability to minimize false positives and false negatives.

**3. F1 Score:** The F1 score, which combines precision and recall, will provide a balanced assessment of the model's performance.

**4. Efficiency:** The computational efficiency of the method will also be considered, including training time and inference speed, to ensure scalability and practicality.

**Experimental Hypotheses:**

The experiment aims to test the following hypotheses:

1. The proposed method can accurately classify inappropriate comments from a diverse dataset with high precision and recall.

2. The method demonstrates robustness in identifying inappropriate comments across different linguistic variations and cultural contexts.

**Experimental Methodology:**

**1. Dataset Preparation:** Curating a diverse dataset containing a mix of appropriate and inappropriate comments from various sources, including social media platforms, forums, and online communities.

**2. Preprocessing and Annotation**: Preprocessing the dataset to remove noise, tokenize the text, and annotate comments as appropriate or inappropriate.

**3. Feature Extraction:** Extracting relevant features from the text data, such as word embeddings or TF-IDF vectors, to represent the comments numerically.

**4. Model Training:** Training a machine learning or deep learning model using the labeled dataset to classify comments as appropriate or inappropriate.

**5. Evaluation:** Evaluating the trained model on a separate test dataset to measure its performance based on the defined criteria.

**Dependent and Independent Variables:**

**Dependent Variable:** The performance metrics (accuracy, precision, recall, F1 score) obtained from evaluating the model.

**Independent Variables:** The method used for text classification, dataset composition, and preprocessing techniques.

**Training/Test Data:**

The training data consist of labeled comments categorized as appropriate or inappropriate. The test data are separate comments that were not used during the training phase, ensuring an unbiased evaluation of the model's performance. The dataset reflects realistic online comment data, making the experiment relevant to real-world scenarios.

**Performance Data Collection and Analysis:**

Performance data include accuracy, precision, recall, and F1 score obtained from evaluating the model on the test dataset. These metrics will be presented using visualizations such as confusion matrices, precision-recall curves, and ROC curves to analyze the model's performance comprehensively. Additionally, comparisons will be made with existing methods or baseline models to assess the proposed method's superiority or equivalency in addressing the problem of identifying inappropriate comments.

**Results:**

Quantitative results from the experiments conducted on the Inappropriate Comments Scanner project reveal several key insights. The model achieved an overall accuracy of 85% in classifying comments as either appropriate or inappropriate, with precision and recall scores calculated at 0.87 and 0.82 respectively, indicating a good balance between correctly identifying inappropriate comments and minimizing false positives. Through thorough preprocessing techniques, including tokenization, stop-word removal, and lemmatization, the dataset was effectively cleaned and made suitable for model training. The resulting dataset contained 10,000 comments, evenly distributed between appropriate and inappropriate categories. Feature extraction techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) were applied to convert text data into numerical vectors, capturing the importance of each word in the comments.

**Discussions:**

The "Inappropriate Comments Scanner" project aims to develop a system capable of identifying and flagging inappropriate comments within text data, falling within the domain of text classification in natural language processing (NLP). The project involves several key stages, including dataset preprocessing, labeling and annotation, and feature extraction, culminating in the classification of text as either inappropriate or not. This project's success depends on a number of variables, starting with the caliber and volume of the training dataset. A good dataset should have a wide range of comments including both acceptable and unsuitable material so the model can reliably distinguish between the two groups.

**Related work:**

An automated approach for identifying offensive or hateful remarks on social media sites was created as part of a research study that tackled a related issue. Their goal was to address the issue of bad information spreading widely on the internet, which may have detrimental impacts on people and communities. Their approach includes the analysis of text data using machine learning techniques, namely deep learning models, to categorize comments into groups such as hate speech, inflammatory language, or neutral material. For feature extraction and classification, they used methods including word embedding, recurrent neural networks (RNNs), and convolutional neural networks (CNNs).Building a content moderation system for online forums with the goal of removing offensive or spammy remarks was the subject of another similar project. Their focus was on the issue of effective moderation—which is necessary to keep the user base safe and happy. They employed sentiment analysis, text parsing, keyword filtering, and other natural language processing techniques to find potentially hazardous remarks. Additionally, they included user input systems to gradually increase the system's accuracy.

**Future work:**

One major shortcoming of the current method is its limited scope in detecting inappropriate comments. While the project focuses on recognizing similar texts, it may not cover the full spectrum of inappropriate language, including nuances, cultural references, or evolving slang. To address this, future enhancements could involve incorporating more diverse datasets that capture a broader range of inappropriate language across various contexts. Additionally, the current method may struggle with detecting subtle forms of inappropriate comments or sarcasm, which often require a deeper understanding of context. Integrating sentiment analysis or context-aware techniques could improve the model's ability to accurately identify such instances.

**Conclusion:**

In conclusion, this project has demonstrated the feasibility of building a text classification system for detecting inappropriate comments using Natural Language Processing (NLP) techniques. The key results highlight the effectiveness of the approach in recognizing similar texts, performing dataset preprocessing, labeling, annotation, and feature extraction. However, the study also revealed certain limitations, including the need for a broader dataset encompassing diverse forms of inappropriate language, the challenge of detecting subtle nuances, and the reliance on manual labeling. The most important points illustrated by this work include the potential of NLP in addressing the issue of inappropriate comments, the importance of dataset diversity and context-awareness, and the opportunity for further advancements in semi-supervised or unsupervised learning techniques.

**References:**

1. Kasbekar, Ameya, et al. "Detecting Offensive Text on Facebook Using Natural Language Processing and Machine Learning." *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*. Springer Singapore, 2020.

2. Engman, Love. "Automatic detection of cyberbullying on social media." (2016).

3. Kim, Allen, et al. "Cleaning dirty books: Post-OCR processing for previously scanned texts." *arXiv preprint arXiv:2110.11934* (2021).

4. Piotrowski, Michael. *Natural language processing for historical texts*. Morgan & Claypool Publishers, 2012.

5. Liu, Zhongxin, et al. "Automating just-in-time comment updating." *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 2020.

6. Tan, Lin, et al. "/\* icomment: Bugs or bad comments?\*." *Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles*. 2007.

7. Kirk, Hannah Rose, et al. "Handling and presenting harmful text in NLP research." *arXiv preprint arXiv:2204.14256* (2022).

8. Wadud, Md Anwar Hussen, et al. "How can we manage offensive text in social media-a text classification approach using LSTM-BOOST." *International Journal of Information Management Data Insights* 2.2 (2022): 100095.

9. Boucher, Nicholas, et al. "Bad characters: Imperceptible nlp attacks." *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

10. Bajaj, Peehu, et al. "Developing an Efficient Toxic Comment Detector Using Machine Learning Techniques." *International Advanced Computing Conference*. Cham: Springer Nature Switzerland, 2023.