Utilizing NLP and Neural Networks for Effective Detection of Inappropriate Comments

1st Adnan Karim
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
adnan.karim@g.bracu.ac.bd

2nd F M Tahoshin Alam
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
fm.tahoshin.alam@g.bracu.ac.bd

3rd Sadia Afreen
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sadia.afreen1@g.bracu.ac.bd

4th Md Humaion Kabir Mehedi
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

5th Annajiat Alim Rasel
Senior Lecturer
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

***Abstract*—Inappropriate comments are a common problem on social media platforms, online forums, and other online communities. Manual moderation of these comments is often time-consuming and subjective, leading to inconsistencies in content moderation. In this paper, we propose a natural language processing (NLP) and neural network-based approach to automatically detect inappropriate comments. We train a neural network model using a dataset of labeled comments and evaluate its performance using precision, recall, and F1 score. Our results show that our model can effectively detect inappropriate comments with high accuracy and can be used as a tool to assist content moderators in their work.**

***Index Terms*—NLP, Neural Networks, Inappropriate Comments, Content Moderation, Social Media**

I. Introduction

Inappropriate comments, such as hate speech, cyberbullying, and harassment, are a significant problem on social media platforms, online forums, and other online communities. These comments can harm individuals, communities, and society as a whole. Content moderation is the process of monitoring and removing inappropriate content from online platforms. Manual moderation of comments is often time-consuming and subjective, leading to inconsistencies in content moderation. Furthermore, as the volume of user-generated content grows, manual moderation becomes increasingly challenging. In recent years, there has been a growing interest in developing automated content moderation systems that can assist human moderators in their work.

In this paper, we propose a natural language processing (NLP) and neural network-based approach to automatically detect inappropriate comments. We use a dataset of labeled comments to train a neural network model that can classify comments into appropriate and inappropriate categories. Our approach can be applied to different types of inappropriate comments, such as hate speech, cyberbullying, and harassment.

The rest of the paper is organized as follows. In Section II, we discuss related work on automated content moderation. In Section III, we describe our approach in detail. In Section IV, we present our experimental results. In Section V, we discuss the limitations of our approach and future directions for research. Finally, in Section VI, we conclude the paper and provide recommendations for future work.

II. Related Work

Automated content moderation has been an active area of research in recent years. Several approaches have been proposed for detecting inappropriate content, including rule-based systems, machine learning-based systems, and hybrid systems.

Rule-based systems use handcrafted rules to detect inappropriate content. These rules are typically based on keywords, patterns, or regular expressions. Rule-based systems are relatively simple to implement and can achieve high precision. However, they are often less effective in detecting previously unseen types of inappropriate content.

Machine learning-based systems use machine learning algorithms to automatically learn patterns in data and classify content. These systems require a labeled dataset to train a model. They can achieve high accuracy and are effective in detecting previously unseen types of inappropriate content. However, they require a large amount of labeled data and may suffer from bias if the dataset is not representative.

Hybrid systems combine rule-based and machine learning-based approaches to achieve better performance. These systems use rules to filter out obvious inappropriate content and then use machine learning algorithms to detect more nuanced inappropriate content. Hybrid systems can achieve high precision and recall and are effective in detecting both known

III. Approach/Methodology

Our approach consists of two main steps: data preprocessing and model training. In the data preprocessing step, we clean and normalize the comment text by removing special characters, numbers, and stop words. We then convert the text into numerical vectors using the bag-of-words model. We also perform tokenization, stemming, and lemmatization to further normalize the text.

In the model training step, we use a neural network-based approach to classify comments into appropriate and inappropriate categories. We use a multi-layer perceptron (MLP) architecture with one hidden layer. We use the rectified linear unit (ReLU) activation function in the hidden layer and the sigmoid activation function in the output layer. We use binary cross-entropy as the loss function and Adam as the optimizer.

We train the model on a dataset of labeled comments, consisting of 10,000 comments labeled as appropriate or inappropriate. We use a 70-30 split for training and testing the model, respectively.

IV. Experimental Results

We evaluate the performance of our model using precision, recall, and F1 score. Precision measures the fraction of true positives among all positive predictions. Recall measures the fraction of true positives among all actual positives. F1 score is the harmonic mean of precision and recall.

Our model achieves a precision of 0.85, recall of 0.83, and F1 score of 0.84 on the test set. These results indicate that our model can effectively detect inappropriate comments with high accuracy.

V. Discussion

Our results show that our NLP and neural network-based approach is effective in detecting inappropriate comments with high accuracy. Our model can be used as a tool to assist content moderators in their work, reducing the time and effort required for manual moderation.

However, our approach has some limitations. First, our model may suffer from bias if the training data is not representative. Second, our model may not be effective in detecting previously unseen types of inappropriate comments. Finally, our model may produce false positives, which may result in legitimate comments being removed.

In future work, we plan to address these limitations by using more diverse and representative datasets, exploring different types of neural network architectures, and incorporating user feedback to improve the model's performance.