# Introduction

## Background

With the advent of the digital age, the number of electronic documents is increasing. One of the more typical phenomena is that with the growth of online communities on the Internet, a large amount of cluttered textual information, such as posts, comments, etc., floods the entire online environment. Most normal and virtuous netizens are spontaneously maintaining the sustainability and usability of the Internet, but in contrast, there are some users who post anti-social and malicious comments on online platforms in an attempt to undermine the usability of the Internet [1]. The emergence of malicious comments violates the legitimate rights of netizens and can cause serious mental and psychological harm to them. Therefore, the efficient management and processing of large volumes of comment texts have become a target of interest for researchers [2].

Text classification is one of the effective ways to locate and triage information efficiently and accurately, solving the problem of information clutter as much as possible [3]. Text classification (TC), also known as text categorization, is an extensive area of current research in linguistic text mining and processing. TC is a process that uses deep learning algorithms to categorize text content into pre-given sets of labels [4]. Deep learning-based text classification techniques have been developed and matured since the 1990s. Compared to text classification systems based on knowledge engineering and expert systems, classification techniques using deep learning provide better classification results and flexibility and have become the main techniques used in related fields [5]. Among the techniques for text classification by different criteria, sentiment analysis (SA), also known as opinion mining, is a branch of text classification. Its main function is to identify and analyze the sentiment in a text by using pre-given labels with human sentiment colors and sentiment tendencies, such as positive, negative, neutral, etc [6].

The report is divided into six sections and the structure of the report with the main content of each section is organized as follows. The first section introduces the basic concepts of text classification, the purpose and significance of this research, an analysis of the problems addressed by text classification, and an overview of the research on the topic. Finally, the overall structure of the report is given. Section 2 introduces the research background of text classification and summarizes the current state of research and the main features of text classification. Section 3 presents the main research methods chosen for the topic, the techniques used, the data and the model testing strategy. Section 4 gives the results of each experiment, while Section 5 shows the project management plans including the activities, schedules, data, and version management plans with the potential risks which may appear during the process of the research and the relevant legal, social, ethical, and environmental issues in the context of the project. The final section presents the achievements accomplished and future works of the project.

## Aim

The main goal of this project is to develop different deep learning-based models for the detection and classification of toxic comments automatically.

## Objectives

The objectives of this text classification project are as follows:

* Conduct background research on text classification, understand the field of Natural Language Processing and the corresponding technologies.
* Collect usable dataset from the Internet.
* Clean and pre-process the data for modeling.
* Extract features from the text in the cleaned datasets.
* Train different models using datasets and assess the quality of the models.
* Analyze the quality of the models and compare the strengths and weaknesses of each model.
* Develop data and model testing and evaluation strategy.
* Analysis risks based on current project.

## Project Overview

### Scope

The project is designed to analyze the sentiment of comments made by users on the talk page of an online encyclopedia website named Wikipedia, filtering out malicious comments and classifying them into different categories, such as hate speech, personal attacks, pornography, or violence, etc. The project helps social network staff to automatically screen out therefore manage malicious comments, reducing labor and time costs, while also helping to clean up the online environment.

### Audience

Text classification is one of the effective managements to helps to locate and triage information efficiently and accurately, solving the problem of information clutter as much as possible[3].

# Background Review

| **Model** | **Recall Ratio** | **Precision Ratio** | **F1** | **Data processing** |
| --- | --- | --- | --- | --- |
| Bi-LSTM + Word2Vec | 97.00% | 89.00% | 92.00% | Word2Vec |
| Bi-LSTM | / | / | 92.79% | Attention selection mechanism + Fine-grained text classification |
| AC-BiLSTM | 87.81% | 87.28% | 86.45% | Attention mechanism + Convolutional Layer |

Table 2 Existing Approaches and Features

The table illustrates the features of existing approaches of implementation using recurrent neural network for text classification.

To date, several models have been used in research on sentiment classification. Among them, Kong F and Chen G have presented a neural network model combining Word2Vec [7] with Bi-LSTM[8] to learn the spatial representation of word vectors through Word2Vec, transforming the text into a sentence representation in the input layer feature space, and improve the network using constant mapping covariance theory. The model using the improved Bi-LSTM is able to present excellent improvements in the dataset.

Ding Y have proposed a classification model based on an attentional mechanism called ON-LSTM [9]. The method is mainly based on a transfer learning approach through feature extraction, where the performance of the model is tuned to the best in the source dataset and then applied to the test set. A multi-level embedding model under the attentional selection mechanism is also proposed. Through the embedding representation at the character level and sentence level in addition to the word level, information that is more conducive to classification in the text can be extracted[10].

Li G and Guo J have proposed a new architecture of bidirectional LSTM (Bi-LSTM) [8] with attention mechanism and convolutional layer, which can more precisely extract text semantics and achieve better text classification results. In AC-BiLSTM model, the convolutional layer is used to retrieve higher-level phrase representations from word embedding vectors, and Bi-LSTM layer is used to gain access to the forward and backward contexts. The attention mechanism is applied to put more focus on the important information in the output of the hidden layer. A SoftMax classifier is ultimately used to categorize the text information after processing. The strength of the AC-BiLSTM model lies in its ability to both extract partial features of phrases and to understand the semantics of phrases within a sentence [11].

# Conclusion

Due to the continuous deterioration of the online environment, malicious speech appearing in public social platforms needs to be screened and blocked. However, as manual review of malicious comments consumes excessive social resources, this paper proposes a deep learning-based solution that can effectively reduce the waste of resources. This paper focuses on text classification based on two different RNN variant models on users' comments on the Internet. The experiments use the same training and testing sets on two different deep learning models, LSTM and GRU, to compare the effectiveness of the two models in classifying malicious text. The results show that both deep learning models perform well on the Wikipedia dataset, recognizing and classifying the sentiment of text with high accuracy, but both models do not perform satisfactorily on the “Threat” type of text. On a general level, when evaluating the results of the two models, it is found that the model using the GRU method performed better than the model using the LSTM method, with higher Accuracy and Recall.

Future research should concentrate on ways to increase the precision of models. The accuracy of models can be increased by using attentional mechanisms to improve semantic understanding. The following areas will be the focus of follow-up work: (1) employing attention mechanisms to enhance the model further; (2) analyzing how attention mechanisms affect the model's performance; (3) developing a new attention mechanism and network architecture.