**SYSTEM ANALYSIS**

**EXISTING SYSTEM:**

* Waseem employed a dataset of 16K tweets to categorize them as gender stereotypes, racial prejudice, or none of the above. He did the best, using the LR algorithm when compared to other techniques like character and word n-grams.
* Gaydhani et al. used a mix of three datasets to perform logistic regression. She discovered that using logistic regression and a term frequency and inverse document frequency vectorizer resulted in a 95.6 percent accuracy.
* Davidson, he analyzed a 24k tweet corpus that he divided into three categories: hate, offensive, and neither. He then used NLP techniques on the tweets, such as extracting the base form of word, text clustering, term frequency vectorization, and sentimental techniques of vader, before running various supervised learning algorithms, the best of which was LR with L-2 regularization. However, they determined that using lexical approaches, it was impossible to discern between Hate and Offensive material.

**DISADVANTAGES OF EXISTING SYSTEM:**

* The existing system accuracy depends on the quality of the data.
* With large data, the prediction stage might be slow.
* The existing system is sensitive to the scale of the data and irrelevant features.
* The existing system requires high memory – need to store all of the training data.
* As in the existing system given that it stores all of the training, it can be computationally expensive.

**PROPOSED SYSTEM:**

* The purpose of the project is to put a set of comments into one of six categories, which are:
* Toxic: A toxic comment is one that is unpleasant, disrespectful, or irrational and is likely to drive other users away from a conversation. Toxic comment classification is a subtask of sentiment analysis.
* Severe Toxic: Adverse effects that arise following the repeated or continuous administration of a test sample for a significant portion of one's life span are referred to as Severe Toxic.
* Obscene: When you say something is obscene, you're implying that it offends you because it involves sex or violence in a way that you find offensive and disturbing.
* Insult: An insult is a purposefully rude action or manner of speech, as well as a lack of regard, esteem, or courteous behavior.
* Threat: It refers to a threatening message in a terrifying manner or a threatening message in a frightening manner.
* After preprocessing our next step is to train the model by Random Forest Classifier. The predictive approach is another name for this concept. This procedure is based on a dataset that has been coded with or without guidance and may be utilised for preparing the model. This labelled dataset is used to train the model by Random Forest Classifier in order to categorise the comments into different levels.
* Our proposed system model is developed using Random Forest Classifier which is a supervised learning algorithm that is employed to learn from the training data and predict the output for the test data. This approach generates a clustering model with more accuracy and less overfitting cases of each feature. It is the most used algorithm and employed in every ML related real-time problem. It contains a number of decision trees from the subsets of the datasets and takes the average to improve the model accuracy. It also supports larger datasets with great dimensionalities.
* We have developed the system using Flask web framework where the user is asked to enter the comment. The commented values are passed into a machine learning model Random Forest Classifier. From the above input values our system predicts the toxic class and shows the user with the level of toxic.

**ADVANTAGES OF PROPOSED SYSTEM:**

* The Accuracy of our proposed system model is generally very high.
* Its efficiency is particularly notable in Large Data sets.
* Provides an estimate of important variables in classification.
* Forests Generated can be saved and reused.
* Unlike other models it doesn’t over fit with more features
* It provides an effective way of handling missing data.
* Random Forest is comparatively less impacted by noise.
* Random Forest is usually robust to outliers and can handle them automatically.