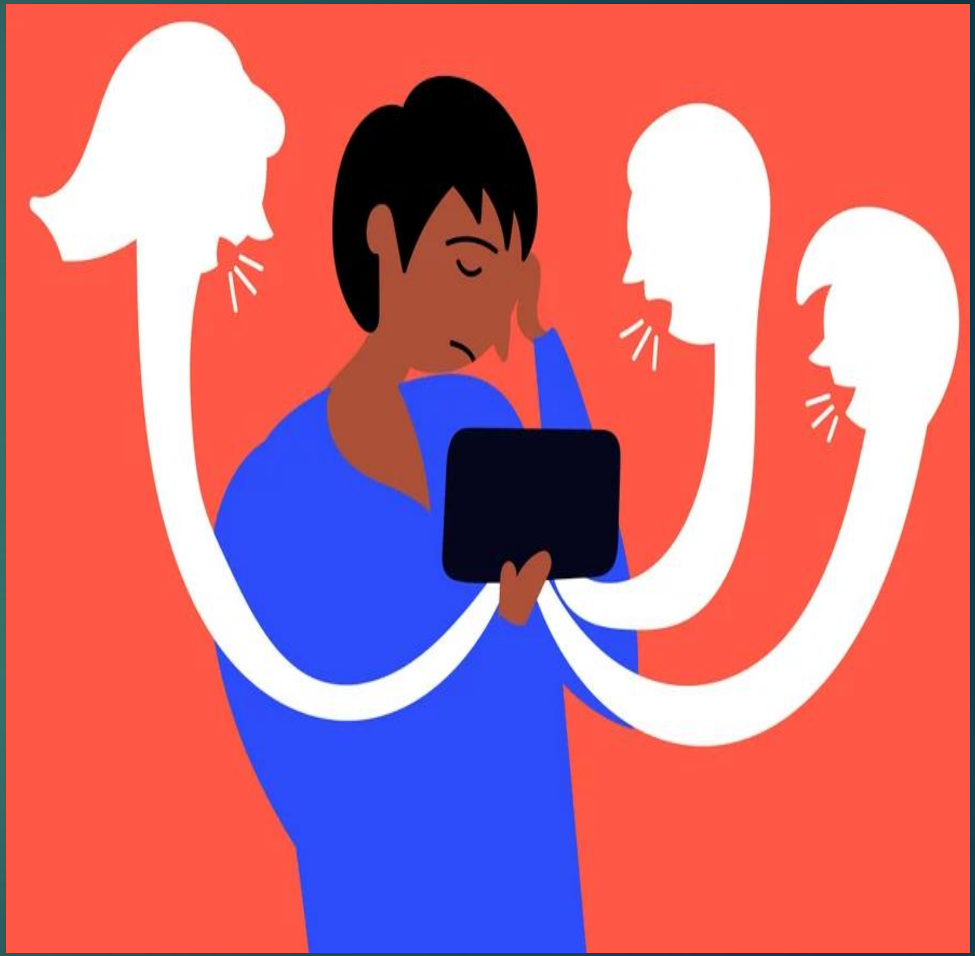
# Toxic Speech Detection: Design, Implementation, and Evaluation

**STUDENT NUMBER: 2310348**

**STUDENT NAME: SHEIKH SAQIB**

# Introduction

- Due to the penetration of the internet in all domains of life which has led to an increase of people's participation actively and remarks as an issue of communicating their opinion in various online forums. Although most of the time these comments are helpful for the creator to extemporize the substance that is being provided to people, sometimes these may be abusive and create hatred-feeling among the people. Thus as these are openly available to the public and being viewed from various sections of society, people in different age groups, different communities, and different socio-economic backgrounds, it becomes the prime responsibility of the content-creator (the host) to filter out these comments to stop the spread of negativity or hatred within people.
- Detecting Toxic comments has been a great challenge for all the scholars in the field of research and development. This domain has drawn a lot of interest not just because of the spread of hate but also because people refraining people from participating in online forums which diversely affects all the creators/content-providers to provide a relief to engage in a healthy public interaction which can be accessed by public without any hesitation.



# Objective

- Design a classifier for a specific task, focusing on practical implementation considerations.
- Implement the designed classifier, ensuring functionality and effectiveness in detecting patterns in data.
- Execute training, validation, and evaluation processes on the provided dataset to assess classifier performance.
- Justify modeling choices, including algorithm selection, feature engineering, and hyperparameter tuning, to enhance model effectiveness.
- Explore the potential of using additional datasets during training, with clear justification for their relevance and contribution to model improvement.
- Foster understanding of scientific evaluation methodologies, including appropriate metrics selection and result interpretation.
- Aim for comprehensive evaluation of classifier performance, considering aspects like accuracy, precision, recall, and F1 score.
- Emphasize the importance of transparency and accountability in model development, including documentation of processes and decisions m

# Methodology

- Dataset Analysis:
  - Three datasets are available: Train, Valid, and Test.
  - Initial analysis of the data is conducted to understand its characteristics, including size, features, and class distribution.
  - Insights from the analysis guide further preprocessing and modeling decisions.
- Visual Representation:
  - Visualizations such as histograms, bar charts, or word clouds are employed to gain insights into the distribution of classes, feature frequencies, and potential patterns in the data.
  - Visual representations help identify trends, outliers, and areas requiring attention during preprocessing.
- Data Cleaning and Preprocessing:
  - Data cleaning techniques are applied to address inconsistencies, missing values, and outliers in the datasets.
  - Preprocessing steps involve text normalization, tokenization, stop-word removal, and stemming or lemmatization to prepare the text data for modeling.
  - Techniques like TF-IDF transformation may be used to convert text data into numerical representations suitable for modeling.
- Implementation of Models:
  - Generative and discriminative models, such as Naive Bayes and Decision Trees, are implemented to make predictions on the preprocessed data. In addition to these other models like Random Forest Classifier, Logistic Regression Model, and Gradient Boosting Classifier were also implemented.
  - The models are trained using the Train dataset and validated using the Valid dataset.
  - Model performance metrics, such as accuracy, precision, recall, and F1 score, are calculated to evaluate the effectiveness of the models in predicting toxicity.

Let’s now go to the code to discuss the results.

# Conclusions

- **Model Selection and Performance Evaluation:**
  - Explored Decision Tree Classifier and Naive Bayes models for toxicity analysis, considering their distinct theoretical foundations.
  - Evaluated model performance, highlighting challenges like class imbalance and mislabeling in the dataset.
  - Found that both models had comparatively small F1 scores compared to state-of-the-art models, suggesting areas for improvement.
- **Insights Gained and Challenges Identified:**
  - Gained insights into the effectiveness of different models for toxicity analysis, emphasizing the importance of rigorous model selection and evaluation metrics like the F1 score.
  - Identified challenges such as class imbalances and dataset mislabeling, which could impact model performance, and highlighted the necessity of addressing these issues.
- **Room for Improvement and Lessons Learned:**
  - Acknowledged the study's contributions while recognizing opportunities for improvement.
  - Recommended future research directions, including correcting class imbalances using oversampling approaches and exploring advanced algorithms to enhance model accuracy.
  - Reflected on lessons learned, such as the potential expansion of model selection to include deep learning architectures and the importance of comprehensive evaluation criteria.
- **Future Directions and Overall Impact:**
  - Suggested future directions for research, such as incorporating advanced techniques like word embedding or topic modeling and conducting a more thorough analysis of errors.
  - Concluded by emphasizing the study's contributions to understanding text categorization tasks and the need for further investigation to enhance the rigor and impact of the results.

# References

*1] G.M. Cramer, R.A. Ford, and R.L. Hall. Estimation of toxic hazard—a decision tree approach. Food and Cosmetics Toxicology, 16(3):255–276, 1976. ISSN 0015-6264. doi: https://doi.org/10.1016/S0015-6264(76)80522-6. URL https://www.sciencedirect.com/science/article/pii/ S0015626476805226.*

*[2] Muhammad Husnain, Adnan Khalid, and Numan Shafi. A novel preprocessing technique for toxic comment classification. 06 2021.*

*[3] Ziqi Zhang Zhixue Zhao and Frank Hopfgartner. A comparative study of using pre-trained language models for toxic comment classification. In Proceedings of the Web Conference 2021 (WWW '21 Companion), page 8, 2021.*

*Thank You* 😊