Name: Sai Nikhil Guptha M

Roll No: AM.EN.U4AIE21142

NLP Project

Study of Existing System

Title of your project: Toxic Comment and Text Classification

Domain: Primarily Natural Language Processing (NLP) within the broader field of Artificial Intelligence (AI) and Machine Learning (ML)

(This is individual submission. But each team member should separate out your work as asked below)

Team members

Sl No	Name	Roll No
1	Sai Nikhil Guptha M	AM.EN.U4AIE21142
2	Mohit Yadav G	AM.EN.U4AIE21128
3	Khitish Behara	AM.EN.U4AIE21121
4	Sai Varun A	AM.EN.U4AIE21120
5	Charan Sai Reddy P	AM.EN.U4AIE21149

PAPER 1: [BERT LLM]

1. Team Member Name : Mammula Sai Nikhil Guptha

2. Title: **DeTox at GermEval 2021: Toxic Comment Classification**
   Name of journal/conference: Association for Computational Linguistics
   Year of publication: September 2021

3. Weblink of the paper: https://aclanthology.org/2021.germeval-1.8/

4. **Problem addressed**: The paper "DeTox at GermEval 2021: Toxic Comment Classification" addresses the problem of classifying German comments as toxic or non-toxic. The primary goal was to develop models that could accurately identify and categorize toxic comments to mitigate harmful online behavior.

   Input for Training:
   Text Data: Comments or tweets in German.
   Linguistic Features: Various features extracted from the text, including:
   - Word count
   - Punctuation count (e.g., number of exclamation marks, question marks)
   - Hate word count
   - Sentiment scores
   - Emoji-related features
   - Other textual features relevant to toxicity detection

   Output for Training: Binary labels indicating whether a comment is toxic (1) or non-toxic (0).

   Input for Testing:
   Text Data: New comments or tweets in German that need to be classified.
   Linguistic Features: The same set of features extracted from the new text data.

   Output for Testing:
   Predicted Labels: Binary predictions indicating the classification of each comment as toxic (1) or non-toxic (0).

	Toxic	Not Toxic	Total
Train	1122 (35.6%)	2122 (64.4%)	3244
Test	350 (37.1%)	594 (62.9%)	944
Total	1472	2716	4188

   User-level functionalities:

   - Can be integrated into social media platforms, online forums, and comment sections to filter out toxic comments and improve the quality of online interactions.
   - Provides accurate model for German language processing tasks, especially in the context of toxicity detection

5. Which LLM pretraining model did they use and why?

   Model Used: The paper employs the German BERT model for pretraining.

   Reason for Choice:

   - Pretraining Benefits: BERT's architecture, which relies on bidirectional context, allows for a deeper understanding of language nuances, crucial for accurately detecting toxicity in comments.
   - Language Specificity: German BERT is tailored for the German language, making it particularly effective for processing German text.

6. Any fine-tuning done on the new dataset? What dataset- format? How did they procure the dataset?

   The model was fine-tuned on a dataset specifically curated for the GermEval 2021 Shared Task.

   Dataset Format: The dataset consists of German comments labeled as toxic or non-toxic.

   Procurement: The dataset was provided as part of the GermEval 2021 Shared Task, which included publicly available German text data from various social media platforms.

7. Fine-tuning Methodology :

   Approach: The German BERT model was fine-tuned using supervised learning. The training involved updating the model weights based on the labeled toxic and non-toxic comments.
   Why: This approach leverages pre-existing language understanding from BERT and adapts it to the specific task of toxicity detection, improving accuracy and relevance to the task

8. Did they use any prompt engineering strategies like zero shot, one shot ,few shot? Brief the prompts? Why?
   The paper does not explicitly mention the use of prompt engineering strategies like zero-shot, one-shot, or few-shot learning. The focus was primarily on supervised fine-tuning

9. Did they use RAG concept? Brief
   The paper does not utilize the RAG concept. The approach is centered on classification rather than generating responses based on retrieved documents.

10. Did they use RLHF technique? Brief
    The paper does not mention using RLHF. The methodology is based on supervised learning using labeled datasets.

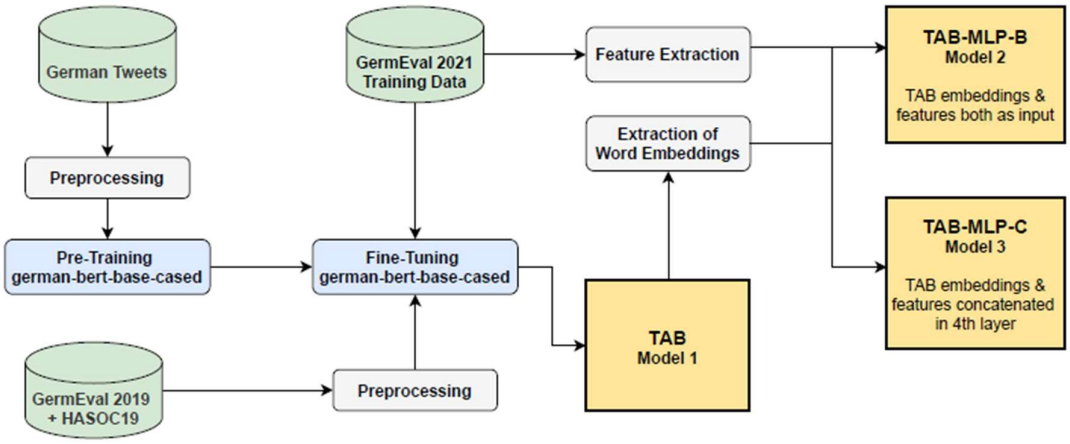
11. Any other techniques adopted to the work?

Feature Engineering: The model incorporates various linguistic features such as word count, punctuation count, hate word count, sentiment scores, and emoji-related features to enhance classification accuracy.

12. Methodology in brief- Explain the overall working of the system as a whole

Overall System Working:

I. Data Preprocessing: Text data is cleaned and various linguistic features are extracted.
II. Model Fine-Tuning: The German BERT model is fine-tuned on the labeled dataset, incorporating both text data and extracted features.
III. Training: The model learns to classify comments as toxic or non-toxic based on the training data.
IV. Prediction: For new comments, the fine-tuned model predicts whether they are toxic or non-toxic.



13. Out of all of the above details What was their exact innovation/research contribution in the whole work. Give in bullets each innovation

- Enhanced German BERT: Fine-tuning the German BERT model specifically for toxicity detection in German comments.
- Feature-Rich Approach: Incorporating various linguistic features (e.g., punctuation, sentiment scores) alongside the text data to improve classification performance.
- Dataset Contribution: Utilizing and contributing to a robust dataset for the GermEval 2021 Shared Task, advancing the field of toxic comment classification in the German language.

PAPER 2: [CNN-LSTM]

1. Team Member Name : Mammula Sai Nikhil Guptha

2. Title: **Predicting Different Types of Subtle Toxicity in Unhealthy Online Conversations**
   Name of conference: 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare
   Year of publication: November 2021

3. Weblink of the paper:
   https://www.sciencedirect.com/science/article/pii/S1877050921024935

4. Problem addressed in the paper with all the functionalities

   - The paper discusses a system for detecting toxic online comments.
   - The system uses a deep learning model (CNN-LSTM) to classify comments into healthy and various types of unhealthy categories.
   - The system's functionalities include the ability to analyze sentiments in comments and classify them into categories such as antagonize, condescending, dismissive, generalization, generalization unfair, hostile, and sarcastic.

   Input and output for training and testing.

   - Training Input: Text comments from an online platform.
   - Output: Predicted class labels (e.g., healthy, hostile) for each comment.
   - Testing Input: New text comments.
   - Testing Output: Classified labels and associated sentiment scores for the new comments.

   [Towards these goals, we leveraged a public dataset of 44K online comments, finding that most unhealthy comments contained negative sentiment, and healthy and unhealthy comments were distinguishable from each other with micro and macro F1-scores of nearly 89% and 68%, respectively.]

   User-level functionalities:

   - Toxic Comment Detection: The system can detect and classify different types of toxic comments, such as antagonizing, condescending, dismissive, generalization, hostile, and sarcastic comments.
   - Sentiment Analysis: The system performs sentiment analysis to detect the general sentiment (positive, negative, neutral) associated with healthy and unhealthy comments.

5. Which LLM pretraining model did they use and why?

    - The paper compares the performance of traditional classifiers with the CNN-LSTM model.
    - A reference is made to a study using BERT (Bidirectional Encoder Representations from Transformers), but due to high training time and resource requirements, CNN-LSTM was preferred for its balance between performance and training efficiency.

6. Any fine tuning done on new dataset? What dataset- format? How did they procure the dataset?
    - Dataset: The dataset used includes comments from a Canadian newspaper website, labeled with various forms of toxicity.
    - Fine-tuning: The paper mentions the potential use of data augmentation techniques to enhance the dataset and improve model performance in future work.

7. Fine-tuning Methodology : What method did they use for finetuning the LLM. Why?

    The current work did not involve fine-tuning a pre-trained LLM like BERT due to resource constraints. Instead, the focus was on training the CNN-LSTM model with the available dataset.

    - **CNN Layer**: Convolutional Neural Networks (CNN) were used to capture local features and patterns within the text data. The CNN layer applied multiple filters to the input embeddings, generating feature maps that highlighted significant n-grams and phrases.

    - **LSTM Layer:** Long Short-Term Memory (LSTM) networks were employed to capture the sequential dependencies and contextual information within the text. The LSTM layer processed the feature maps from the CNN layer, maintaining the temporal order of the words and phrases.

    - **Fully Connected Layer:** The output from the LSTM layer was passed through a fully connected layer to map the extracted features to the final output classes, representing different types of toxicity.

    - **Training and Optimization**: The model was trained using a labeled dataset of toxic and non-toxic comments. The training process involved optimizing the model parameters using backpropagation and techniques like Adam or RMSprop optimizers to minimize the loss function.

8. Did they use any prompt engineering strategies like zero shot, one shot ,few shot? Brief the prompts? Why?

    The paper did not explicitly mention using prompt engineering strategies like zero-shot, one-shot, or few-shot learning.

9. Did they use RAG concept or RHLF techniques? Brief

   No mention of using Retrieval-Augmented Generation (RAG) or Reinforcement Learning from Human Feedback (RLHF) techniques in the work.

10. Any other techniques adopted to the work? No

11. Methodology in brief- Explain the overall working of the system as a whole

    I. **Data Collection**: A large dataset of comments labeled as toxic or non-toxic was collected. The dataset included various types of toxic behaviors, such as hate speech, harassment, and bullying.

    II. **Data Preprocessing:** The collected data underwent preprocessing to clean and normalize the text. This step included tokenization, removing special characters, converting text to lowercase, and removing stopwords.

    III. **Embedding Representation**: Pre-trained word embeddings were used to convert the textual data into numerical vectors. These embeddings captured the semantic meaning of the words and served as the input to the CNN-LSTM model.

    IV. **Model Architecture:** The system utilized a CNN-LSTM architecture. The CNN layers extracted local features and patterns from the word embeddings, while the LSTM layers captured the sequential dependencies and contextual information.

    V. **Training the Model**: The CNN-LSTM model was trained on the preprocessed and embedded dataset. The training process involved iteratively adjusting the model parameters to minimize the loss function and improve classification accuracy.

    VI. **Evaluation and Testing**: The trained model was evaluated on a separate test dataset to assess its performance. Metrics such as accuracy, precision, recall, and F1-score were used to measure the model's effectiveness in detecting toxic comments.

    VII. **Deployment**: The final model was deployed as a service that could be integrated into online platforms. Users could input comments, and the system would provide real-time feedback on the toxicity of the comments.

12. Out of all of the above details What was their exact innovation/research contribution in the whole work. Give in bullets each innovation.

- Novel CNN-LSTM Architecture: The research introduced a novel combination of CNN and LSTM layers to effectively capture both local patterns and sequential dependencies in text data.

- Real-time Toxic Comment Detection: The system provided real-time feedback on the toxicity of comments, making it practical for deployment in online platforms for monitoring and moderating user interactions.

- Comprehensive Evaluation: The research included a comprehensive evaluation of the model's performance using various metrics, demonstrating the effectiveness of the CNN-LSTM architecture in detecting toxic comments.

Table 1. Classification results.

Model	Average Micro F1	Average Macro F1	AUC-ROC
Logistic Regression	57.54%	48.31%	0.51
SVM	69.15%	61.29%	0.62
CNN LSTM Network	**88.76%**	**67.98%**	**0.71**

PAPER 3: [ML]

1. Team Member Name : Mammula Sai Nikhil Guptha

2. Title: **Toxic Comments Classification**
   Name of journal/conference: International Journal for Research in Applied Science & Engineering Technology
   Year of publication: June 2022

3. Weblink: https://www.ijraset.com/best-journal/toxic-comments-classification

4. **Problem addressed**: The project addresses the problem of conversational toxicity on social media, which can hinder free expression and foster negative interactions. The goal is to detect and classify toxic comments to prevent antisocial behavior online.

   User-level functionalities:

   - Classify comments as toxic or non-toxic.
   - Alert users before transmitting potentially toxic messages.

5. Which LLM pretraining model did they use and why?

   Model Used:

   The document does not specify a particular LLM pretraining model (like BERT). Instead, it mentions using various machine learning algorithms for text classification, including logistic regression, random forest, SVM, Naive Bayes, and XGBoost classifiers.

   Reason for Choice:

   These models were chosen to evaluate their performance in classifying toxic comments. The model with the highest accuracy was selected for toxicity prediction on unseen data.

6. Any fine-tuning done on the new dataset? What dataset- format? How did they procure the dataset?
   - The document does not explicitly mention any fine-tuning of pretrained language models. It focuses on using machine learning algorithms.
   - Dataset Format: The dataset used is in a format suitable for machine learning models, typically involving labeled text data for training and testing.
   - Dataset Procurement: The specific method of procuring the dataset is not detailed, but it involves data from social media networks containing toxic and non-toxic comments

7. Fine-tuning Methodology: Not applicable as there is no mention of fine-tuning pre-trained language models in the paper. Methodologies used are

- Logistic Regression: A statistical method for binary classification that predicts the probability of a binary outcome based on one or more predictor variables. It uses the sigmoid function to output probabilities between 0 and 1.

  [the probability is greater than 0.5, the predicted class is 1. If the probability is less than 0.5, the predicted class is 0.]

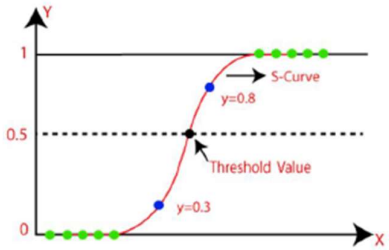
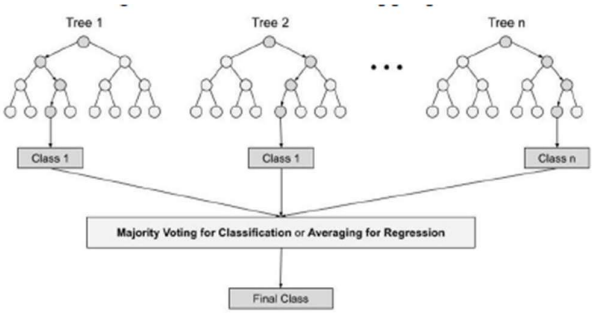
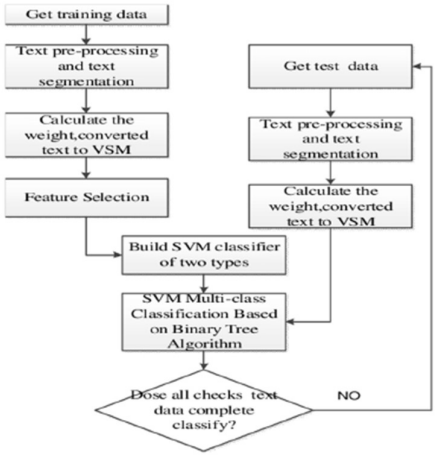


Fig : Logistic Regression curve

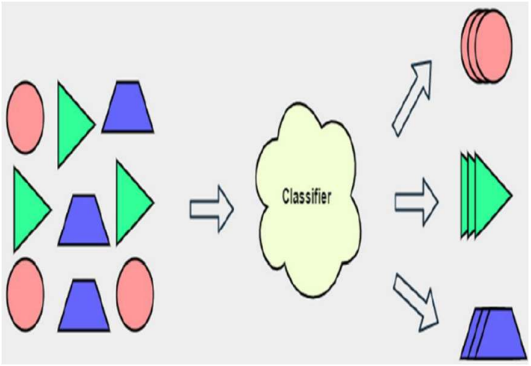
- Random Forest: An ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



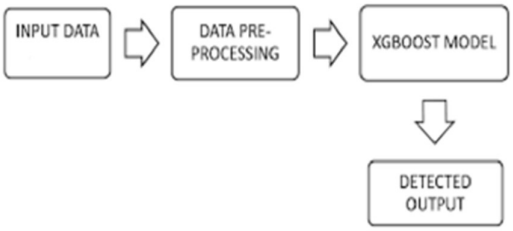
- Support Vector Machine (SVM) Classifier: A supervised learning model used for classification and regression tasks. SVM works by finding the hyperplane that best divides a dataset into classes.



- Naive Bayes: A classification technique based on Bayes' Theorem with an assumption of independence among predictors. It is particularly effective for large datasets and is used in text classification due to its simplicity and efficiency.



- XGBoost Classifier: An optimized gradient boosting machine learning library designed for speed and performance. It is used for supervised learning problems and is known for its efficiency and accuracy in classification tasks.



8. Did they use any prompt engineering strategies like zero shot, one shot ,few shot? Brief the prompts? Why?
   There is no mention of using prompt engineering strategies such as zero-shot, one-shot, or few-shot learning.

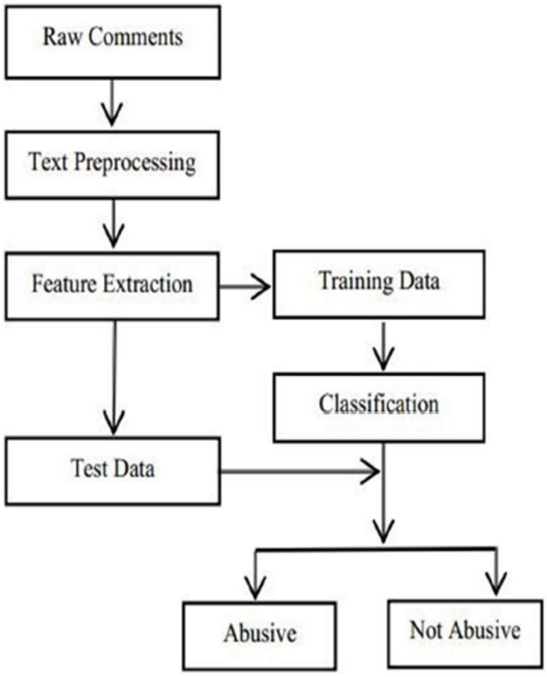
9. Did they use RAG concept? Brief
   The paper does not mention the use of Retrieval-Augmented Generation (RAG).

10. Did they use RLHF technique? Brief
    The document does not mention the use of Reinforcement Learning from Human Feedback (RLHF).

11. Any other techniques adopted to the work?
    No other techniques were adopted in the work.

12. Methodology in brief- Explain the overall working of the system as a whole?
    Overall Working of the System:

    I. Data Collection and Preprocessing: The dataset containing comments is collected, and preprocessing steps such as stemming and removing stop words are applied.
    II. Model Training: Various machine learning models (logistic regression, random forest, SVM, Naive Bayes, XGBoost) are trained on the preprocessed data.
    III. Model Evaluation: The accuracy of each model is evaluated.
    IV. Model Selection: The model with the highest accuracy is selected.
    V. Prediction: The chosen model is used to predict the toxicity of unseen comments.



13. Out of all of the above details What was their exact innovation/research contribution in the whole work? Give in bullets each innovation.

- Application of multiple machine learning algorithms to classify toxic comments and comparison of their accuracies.
- Development of a preprocessing pipeline that includes stemming and stop word removal for text classification.
- Creation of a system to alert users about potentially toxic comments before they are posted, thus preventing antisocial behavior on social media platforms.

PAPER 4: [CapsNet & BERT]

1. Team Member Name : Mammula Sai Nikhil Guptha

2. Title: Evaluating The Effectiveness of Capsule Neural Network in Toxic Comment Classification using Pre-trained BERT Embeddings

   Name of journal/conference: TENCON 2023 - 2023 IEEE Region 10 Conference
   Year of publication: November 2023

3. Weblink of the paper: https://ieeexplore.ieee.org/document/10322429

4. **Problem addressed**: The project aims to evaluate Capsule Neural Networks (CapsNet) against alternative neural network architectures for detecting potential toxicity in text data. The goal is to improve automated systems' ability to detect toxic messages on social media platforms.

   User-level functionalities: The system provides a mechanism to automatically analyze and detect toxic messages in online platforms. It can quantify the toxicity of comments and take necessary actions, such as flagging or removing inappropriate content.

5. Which LLM pretraining model did they use and why?

   Model Used: Capsule Neural Networks (CapsNet) along with pre-trained BERT embeddings

   Reason for Choice: BERT is known for its state-of-the-art performance in natural language processing tasks, including text classification. By leveraging BERT embeddings, the model can capture intricate linguistic patterns and context, enhancing the performance of CapsNet for toxic text classification.

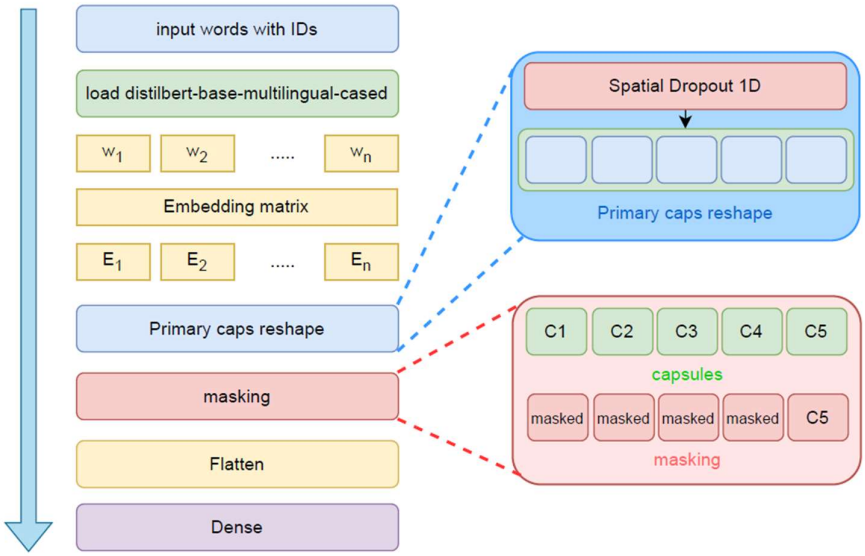
6. Any fine-tuning done on the new dataset? What dataset- format? How did they procure the dataset?

   - Yes, fine-tuning was done on the new dataset, which was obtained from Jigsaw via a 2020 Kaggle competition. The dataset is in a multiclass format for classifying toxic comments, containing over 223,000 comments with classes including toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset was found to be extremely unbalanced, with the majority of comments being in English.

   - The dataset was cleaned to remove unnecessary elements such as punctuation, timestamps, and user information, as it consisted of public comments. The data was then tokenized, which involved splitting the sentences into the smallest possible strings, with the aim of further text cleaning, such as lemmatization or converting the output to a data frame for improved use in a model. The study used a pre-trained BERT tokenizer, which helped speed up the work.

7. Fine-tuning Methodology :

   In the fine-tuning process, the dataset undergoes preprocessing steps to clean up unnecessary elements like punctuation and timestamps. The texts are then tokenized, and BERT embeddings are applied to represent the words in a meaningful way. The model architecture includes a capsule layer with 5 capsules and 5 dimensions, which helps capture the spatial relationships between words. This layer is followed by a flatten layer, which reshapes the data, and a Dense layer with 128 neurons, which learns higher-level representations of the text.



8. Did they use any prompt engineering strategies like zero shot, one shot,few shot? Brief the prompts? Why?

   The project does not explicitly mention the use of prompt engineering strategies like zero-shot, one-shot, or few-shot learning.

9. Did they use RAG concept? Brief
   The project does not mention the use of the Retrieval-Augmented Generation (RAG) concept.

10. Did they use RLHF technique? Brief
    The project does not mention the use of the Reinforcement Learning from Human Feedback (RLHF) technique.

11. Any other techniques adopted to the work?
    The project includes the use of Convolutional Neural Networks (CNN) and DistilBERT for comparison with CapsNet. They also use the googletrans library for translating non-English texts.

12. Methodology in brief- Explain the overall working of the system as a whole?

I. Data Collection: The system collects data from various sources, such as online platforms or databases, related to toxic comments.

II. Data Preprocessing: The collected data undergoes preprocessing to clean and prepare it for analysis. This includes removing irrelevant information, handling missing data, and standardizing the format of the text.

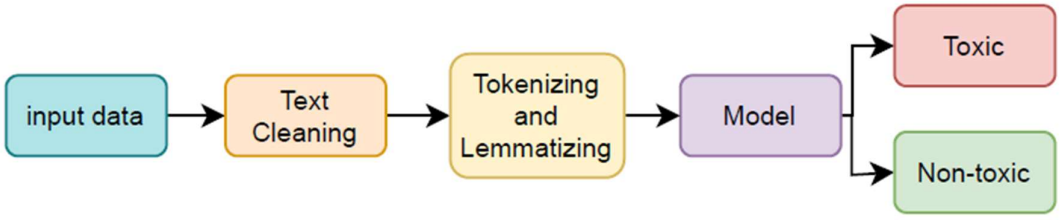
III. Tokenization: The text data is tokenized, which means breaking down the text into smaller units like words or subwords. This helps in preparing the text for further processing.

IV. Model Selection: A deep learning model, such as a CNN-LSTM model, is selected for its effectiveness in handling text data and capturing contextual information.

V. Model Training: The selected model is trained on the preprocessed data. During training, the model learns to identify toxic comments based on the patterns and features present in the data.

VI. Validation and Testing: The trained model is validated using a separate dataset to ensure its performance. It is then tested on new data to evaluate its effectiveness in identifying toxic comments accurately.

VII. Fine-tuning: The model may undergo fine-tuning, where its hyperparameters are adjusted to improve its performance further.



13. Out of all of the above details What was their exact innovation/research contribution in the whole work. Give in bullets each innovation

- Utilizing CapsNet for toxic text classification, demonstrating its potential in comparison to other architectures.
- Employing pre-trained BERT embeddings to enhance CapsNet's performance in understanding text context.
- Investigating and characterizing sentiment, polarity, and readability of the dataset to provide deeper insights into the nature of toxic comments.
-

PAPER 5: [CNN]

1. Team Member Name : Mammula Sai Nikhil Guptha

2. Title: **Convolutional Neural Networks for Toxic CommentClassification**
   Name of journal/conference: ARIXV
   Year of publication: February 2018

3. Weblink of the paper: https://arxiv.org/abs/1802.09957

4. **Problem addressed**: The problem addressed in the paper is the classification of toxic comments in online communication. Toxic comments can lead to personal attacks, online harassment, and bullying behaviors, thus affecting the quality of online interactions. The paper aims to explore Convolutional Neural Networks (CNNs) as an approach to efficiently classify toxic comments, comparing their performance against traditional Bag-of-Words (BoW) methods combined with various classification algorithms.

   User-level functionalities:
   - Identifying toxic comments in online communication platforms.
   - Enhancing safety and fostering positive online interactions.
   - Providing insights into effective text classification methodologies for toxic comment detection.

5. Which LLM pretraining model did they use and why?
   The paper doesn't explicitly mention the specific LLM (Large Language Model) pretraining model used. Since the paper discusses Convolutional Neural Networks (CNNs) and Bag-of-Words (BoW) approaches for text classification, it's likely that they didn't directly utilize an LLM pretraining model for their classification task. Instead, they focus on comparing CNNs with traditional BoW methods.

6. Any fine-tuning done on the new dataset? What dataset- format? How did they procure the dataset?

   The paper doesn't mention fine-tuning on a new dataset.
   - Dataset Format: The dataset used in the study is from a Kaggle competition regarding Wikipedia's talk page edits. The dataset contains comments labeled by human raters for toxic behavior. Although the original dataset includes six types of indicated toxicity (e.g., 'toxic', 'severe toxic', 'obscene', 'threat', 'insult', 'identity hate'), all these categories were considered as toxic for this study, converting the problem to binary classification.
   - Dataset Procurement: The dataset was obtained from Kaggle

7. Fine-Tuning Methodology:

   - Two variants of CNN: CNNfix and CNNrand, where CNNfix uses pre-trained word embeddings from word2vec and CNNrand initializes word representations randomly and tunes them during training.
   - Employed three convolutional layers simultaneously with filter sizes of 3, 4, and 5, and applied max-over-time pooling after each layer.
   - Trained the CNN using Stochastic Gradient Descent (SGD) with mini-batch size 64 and learning rate 0.005.

8. Did they use any prompt engineering strategies like zero shot, one shot ,few shot? Brief the prompts? Why?

   The paper doesn't mention the use of prompt engineering strategies like zero-shot, one-shot, or few-shot learning.

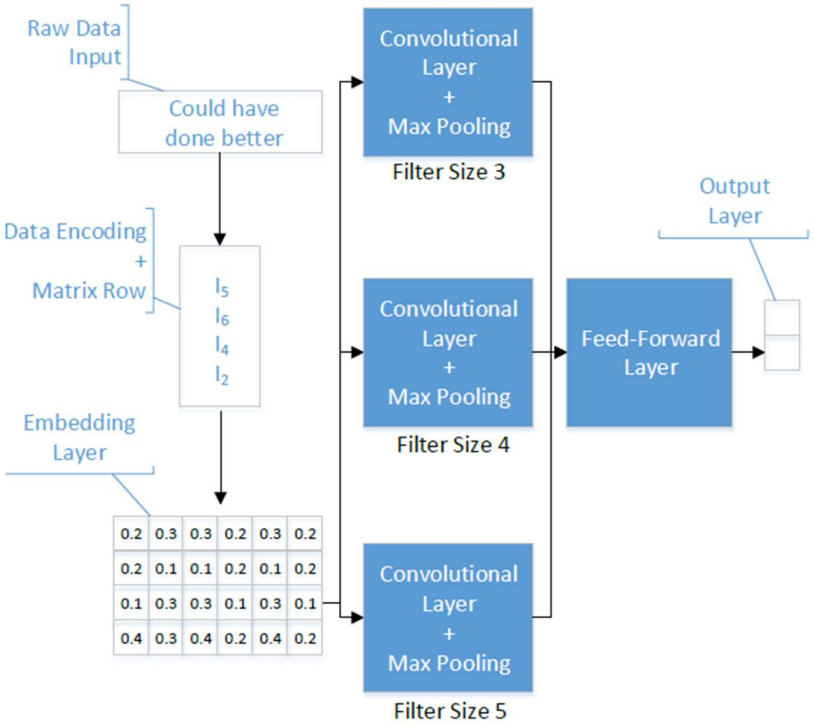
9. Did they use RAG concept? Brief
   The paper doesn't mention the use of the Retrieval-Augmented Generation (RAG) concept.

10. Did they use RLHF technique? Brief
    The paper doesn't mention the use of Reinforcement Learning from Human Feedback (RLHF) technique

11. Any other techniques adopted to the work?
    Other techniques adopted in the work include Convolutional Neural Networks (CNNs) for text classification, Bag-of-Words (BoW) representation, and various traditional classification algorithms like Support Vector Machines (SVM), Naive Bayes (NB), k-Nearest Neighbor (kNN), and Linear Discriminated Analysis (LDA)

12. Methodology in brief- Explain the overall working of the system as a whole?

    - Utilized Convolutional Neural Networks (CNNs) for toxic comment classification, compared with Bag-of-Words (BoW) approach combined with traditional text classification algorithms.
    - Constructed Document-Term-Matrix (DTM) using TF-IDF for BoW approach.
    - Experimented with SVM, Naive Bayes, k-Nearest Neighbor, and Linear Discriminated Analysis on BoW approach, and CNNfix and CNNrand variants of CNN.
    - Balanced dataset by sub-sampling non-toxic texts and evaluated methods 20 times on random separations.
    - Conducted statistical analysis using confusion matrix, evaluating metrics like Accuracy, Precision, Recall, F1-score, False Discovery Rate, and Specificity.



13. Out of all of the above details What was their exact innovation/research contribution in the whole work? Give in bullets each innovation
    - Exploration of Convolutional Neural Networks (CNNs) for toxic comment classification.
    - Comparison of CNNs with traditional Bag-of-Words (BoW) methods and various classification algorithms for text classification.
    - Demonstration of CNNs' effectiveness in enhancing toxic comment classification performance