**CSCE 5290: Natural Language Processing**

**Project Proposal**

**Title: Hate Speech Identification and Classification**

**Github Link: https://github.com/kritinkesavkts/Nlp-project/tree/main**

1. **Motivation**

In this digital world, where we have many online social media communities, people leave their opinion online on various topics. These sometimes can be appreciating and useful as feedback but most of the times it is toxic and hateful that can have a bad and lasting effect on the user and the fellow browsers of the online platforms and make it an unsuitable and unpleasant environment for various age groups online. This is a growing concern since it is easy to slander or abuse online while being anonymous and it’s a place where growing children usually refer for information and knowledge. Also, social media has become a common ground for online disputes between admirers of certain groups or people or for them to post hateful comments and criticize or scrutinize famous personalities or companies which may lead to bad reputation or sometimes lead tragic consequences.

1. **Significance**

The main aim of this project is to detect hateful comments that are being posted online so that it can be removed to provide a safer and friendly online environment for a diverse range of people. The detected hateful comments can then be analyzed to know the level of toxicity and report them so that appropriate action is taken against them.

1. **Objectives**

The main objective of this project is to be able to classify hate speech in multiple languages and then detect the severity or the toxic level of the hate comment. Some of the major milestones we aim to accomplish are to train a pre-trained model on our multi-lingual hate comments dataset. Next, we train a language understanding model to find the toxicity of the hate comment detected and get the major keywords in it. We then try to fine tune the models for efficient classification. Following would be the measures and the enhancement with barometer of toxic levels.

1. **Features**

The first milestone of the project is text classification, and the main aim is to train a pre-trained model like BERT on the provided OLID and hate speech datasets. Then the second milestone is language understanding and to train language understanding model on toxic comments on Kaggle dataset to determine the toxicity level of the hate comment. These two models can then be fine-tuned to effectively classify the hate comments to block such content and detect the severity of it and get its key words so that it can be reported, and proper action is taken against it. Also there would be considerations in ethical procedures so that bias would be restricted and would follow content moderation policies responsibly. Coming to the evaluation metrics which is a vital part to say how good a model is we use performance metrics like recall,precision,F-1 score to get to know the effectiveness.

Now coming to the add on is the level of toxic Ness is a further enhancement for that we develop a function with the limit or criterion of the words used to measure with the help of content, sentiment analysis and context.This can be integrated with the model that we already have by combining the scores with the classification results.

1. **Dataset.**

The datasets we intend to use are OLID and hate speech datasets for text classification and toxic comments on Kaggle for language understanding. The dataset size is 14000 tweets. For preprocessing, we plan to remove unnecessary fields which would not have meaningful contribution while training the model and also handle missing value either by having a default value or removing the entry if the dataset is huge enough and doesn’t affect the model much while training.

About dataset:

Type: The dataset is labeled for various categories related to offensive language:

subtask\_a: Offensiveness classification (NOT, OFF)

subtask\_b: Targeted/off-targeted classification (TIN, UNT)

subtask\_c: Hate speech category classification (IND, GRP, OTH)

Basic preprocessing steps:

Text Cleaning: We remove website links and any symbols that are not letters or numbers from the tweets. This helps in getting rid of unnecessary elements and focusing on the actual text content.

Tokenization: We break down the tweets into separate words or tokens. This step is important for analyzing the text because it allows us to work with individual words and understand their meanings.

Normalization: We convert all the text to lowercase letters and remove any accents or special marks from letters. This makes the text consistent and standardized, ensuring that similar words are treated the same way regardless of their accents.

**6. Visualization**

[[Make tables and workflow diagrams as visual aids to improve comprehension. In order to facilitate clear communication of the project's overall structure and procedures, these figures should be used as a visual representation of the project. Then, expand on these illustrations to give a thorough description of your idea.]]

Workflow Description:

1. **Data Collection & Preprocessing**:
   * Gather tweets from the OLID dataset.
   * Clean and preprocess text data (remove URLs, special characters, etc.).
   * Tokenize tweets into individual words.
2. **Feature Engineering**:
   * Convert tokenized text into numerical representations using word embeddings (e.g., Word2Vec, GloVe).
   * Normalize text by converting to lowercase and removing accents.
3. **Model Development & Training**:
   * Choose a deep learning architecture (e.g., BERT, LSTM).
   * Fine-tune the selected model on hate speech detection task using OLID dataset.
   * Train the model using optimized parameters.
4. **Model Evaluation & Validation**:
   * Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
   * Validate the model on a held-out test set to assess generalization.
5. **Toxicity Scoring Integration**:
   * Develop a toxicity scoring function based on model predictions.
   * Integrate toxicity scores with hate speech detection results for comprehensive analysis.
6. **Visualization & Interpretation**:
   * Visualize toxicity scores and hate speech classifications using barometers or scales.
   * Provide insights and interpretations based on visualization for content moderation.

Barometer Description:

* The toxicity barometer represents different levels of toxicity:
  + Low: Mild offensive language.
  + Moderate: Moderate offensive language.
  + High: Severe offensive language or hate speech.

GROUP 6  
Names:  
Kaushik Reddy Tolishammagari-11726552  
Deekshitth Vegi - 11716217  
Kritin Kesav Chitturi – 11720237  
Marri Sanjana -11700133