Group name: Solitude Ensemble

Group member's details

Name:	Parthkumar Sutariya
Email ID:	parthsutariya31@gmail.com
Country:	India
College:	BSBI, Berlin, Germany
Specialization:	Data Analyst

Problem Description:

The term hate speech is understood as any type of verbal, written or behavioral communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, color, ancestry, sex or another identity factor. To address this issue, we aim to develop a Hate Speech Detection system using state-of-the-art natural language processing (NLP) techniques, specifically leveraging transformer-based models.

Hate Speech Detection is a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on data that is used to classify sentiments. So, for the task of the hate speech detection model, we will use Twitter tweets to identify tweets containing Hate speech.

Business Understanding:

Target Audience: The target audience includes social media platforms, online forums, and any digital platform seeking to maintain a safe and inclusive online environment.

Business Goals: The primary goal is to deploy an efficient hate speech detection system that can automatically identify and filter out toxic content in real-time. This will enhance user experience, promote positive engagement, and mitigate potential legal and reputational risks for platform owners.

Success Metrics: Success will be measured by the accuracy, precision, recall, and F1 score of the hate speech detection model. Additionally, the reduction in the prevalence of hate speech and toxic content on the platform will serve as key performance indicators.

Project Lifecycle:

1. **Initiation and Planning (1 week):**
   - Define project objectives, scope, and stakeholders.
   - Conduct a comprehensive review of existing literature and datasets related to hate speech detection
   - Select appropriate transformer-based models (e.g., BERT, RoBERTa, GPT, etc.) and NLP libraries (e.g., Hugging Face Transformers, TensorFlow, PyTorch).
   - Develop a detailed project plan, including milestones, deliverables, and timeline.
2. **Execution (3 weeks):**
   - Acquire and preprocess relevant datasets, ensuring data privacy and ethical considerations.
   - Annotate the datasets for hate speech and non-hate speech content.
   - Fine-tune pre-trained transformer models on the annotated datasets using appropriate training techniques (e.g., transfer learning, adversarial training).
   - Evaluate model performance using cross-validation and other validation techniques. Iteratively refine the models based on performance feedback.
3. **Monitoring & Controlling (2 weeks):**
   - Monitor model performance and identify potential issues or biases.
   - Implement mechanisms for continuous improvement, such as retraining the models with updated datasets or fine-tuning hyperparameters.
4. **Closure (1 week):**
   - Finalize the hate speech detection model and prepare it for deployment.
   - Document project findings, including methodologies, challenges, and lessons learned.
   - Conduct a knowledge transfer session to ensure the sustainability of the project outcomes.

Deadline:

The entire project lifecycle, from initiation to closure, is estimated to be completed within 7 weeks. This timeline allows for adequate planning, execution, and refinement of the hate speech detection system while ensuring timely delivery to meet business objectives.