Toxic Comment Classification with Transformer Models

PROBLEM

The primary problem in this project is the pervasive issue of toxic comments on online platforms like Wikipedia’s talk page edits, which can include rude, disrespectful, or harmful that discourage participation and stifle open discussion. Suppose the models are not sufficiently accurate and cannot differentiate between specific types of toxicity, such as threats, obscenity, insults, and identity-based hate. In that case, many online communities are forced to limit or shut down user comments to manage negative behaviors, ultimately hindering productive and respectful online conversations. The challenge is to develop a more accurate and proficient multi-headed classification to detect better and categorize these toxic comments on Wikipedia platforms' talk pages, thereby enhancing the quality of online interactions.

SOLUTION

- **Data Preparation**:
  - Preprocessed the dataset to clean and prepare the text for model training.
  - Cleaning the texts
    - Remove HTML tags
    - Replace URLs with a special token
    - Replace email addresses with a special token
    - Convert text to lowercase
    - Remove stop words
    - Remove punctuation
    - Remove special characters
    - Remove whitespaces
    - Remove text within double equals (commonly used for headings in Wiki text)
    - Normalize whitespaces, replacing multiple spaces with just one
    - Replace newline characters and carriage returns with space
  - Used techniques like tokenization and handled class imbalances through oversampling, as we had a very high-class imbalance.
- **Model Selection and Training**:
  - Tested various pre-trained Large Language Models for the task, like BERT, Roberta, and DistilBERT.

- Selected the DistilBERT model for its balance between performance and computational efficiency.
- Fine-tuned the DistilBERT model on the dataset to optimize it for the classification task.

- **Training Process**:

  - Implemented cross-validation to ensure the model's robustness and generalizability.
  - Used the FocalLoss function to handle class imbalances and improve the model's performance on less frequent classes.
  - Employed the ADAMW optimizer and the ReduceLROnPlateau scheduler to adjust the learning rate based on the validation loss, helping to prevent overfitting.
  - Utilized Google Colab with GPU acceleration to efficiently run and fine-tune the DistilBERT model, enabling faster training and evaluation processes.

- **Evaluation**:

  - Evaluated the model using the mean column-wise ROC AUC metric.
  - Achieved an AUC-ROC score of 98%, indicating a high level of accuracy in predicting the probabilities for each type of toxicity.

Results

1. **Performance**:
   - The DistilBERT model outperformed other models tested, achieving a high AUC-ROC score of 98%.
   - This performance indicates that the model can effectively detect toxic comments, improving the current Perspective API models.
2. **Impact**:
   - The model can help online platforms manage toxic comments better and tailor their moderation strategies to address the most harmful behaviors without overly restricting free expression. And they can foster more productive and respectful online discussions.