Solution Building Report

Baseline: Dictionary-Based Approach

Initially, a simple rule-based system was implemented using a dictionary of toxic terms. This approach provided a quick and straightforward method for identifying obvious cases of toxicity.

Hypothesis 1: Custom Embeddings

Believing that a more nuanced approach could capture the subtleties of language use, we trained custom embeddings using a Word2Vec model on our corpus. These embeddings aimed to reflect the specific linguistic environment of our data, capturing contextual relationships between words more effectively than general-purpose embeddings.

Hypothesis 2: RNN Enhancement

To address the limitations of capturing long-term dependencies in sentences, we experimented with adding more layers to our RNN architecture, specifically LSTM layers. The rationale was that a deeper network could potentially learn more complex patterns and relationships within the text.

Hypothesis 3: Leveraging Pretrained Embeddings

We explored the potential of leveraging pre-trained embeddings from established models such as GloVe and BERT. These embeddings are trained on vast amounts of data, providing rich semantic representations that could improve our model's understanding of language intricacies.

Results and Insights

After testing and validation, the following results were obtained:

Custom Embeddings: The Word2Vec model yielded embeddings that slightly outperformed the baseline on certain metrics but still failed to capture the full range of toxic language nuances.

Enhanced RNN Architecture: Adding more LSTM layers did not lead to the expected improvements. Instead, the model complexity increased without a proportional gain in performance, suggesting that the additional layers were not effectively utilized, possibly due to overfitting.

Pretrained Embeddings: The use of pre-trained embeddings, particularly from the BERT model, showed the most promise. The rich semantic information encoded in these embeddings led to substantial performance gains across all metrics, outshining both the baseline and the custom embeddings.

Conclusion

In conclusion, our exploration revealed that the sophisticated semantic understanding provided by pre-trained embeddings, especially those from BERT, was crucial to our model's success. The insights gained from these hypotheses steered us toward our final solution, which harnesses the power of BERT's transformer architecture to deliver a robust and nuanced text classification model.