Project Report: Explainable Detection of Online Sexism (EDOS)

Submitted By:

MD. MUSFIQUR RAHMAN SHOVON - 20301332

Submitted To:

- Dr. Farig Yousuf Sadeque
  BRAC UNIVERSITY

Introduction:

Text classification is a basic natural language processing (NLP) problem that has several uses, including subject classification, sentiment analysis, and spam detection. In this project, we employ Bi-directional LSTM (Long Short-Term Memory) neural networks for text classification. The dataset used here contains textual data along with corresponding labels for two classification tasks- Sexism classification and Category classification.

Data Loading and Preprocessing:

The dataset is loaded from a CSV file which contains two columns: 'text' and 'label category' for category classification and 'label sexist' for sexism classification. We filter the columns accordingly. The text data is tokenized using the Keras Tokenizer and padded to ensure uniform sequence length. We used an 80-20 split to divide the dataset into training and testing sets for each task.

Glove Word Embeddings:

Glove (Global Vectors for Word Representation) Pre-trained word vectors known as embeddings are used to represent the semantic connections between words. We utilize Glove embeddings to represent words in our text data. The embeddings are loaded from a pre-trained file ('glove.6B.100d.txt') and mapped to the corresponding words in our vocabulary.

Model Architecture:

For both category classification and sexism tasks, we used a neural network architecture based on long short-term memory (LSTM). The model is composed of an embedding layer, an LSTM layer for sequence modeling, a dense layer with Softmax activation function for multi-class classification, and a spatial dropout layer to prevent overfitting.

Data Splitting:

A stratified 80-20 split is used to divide the dataset into training and testing sets. This assures that the class distribution is maintained in the training and testing sets.

Results and Analysis:

We train the LSTM model independently for tasks related to category classification and sexism. After training, we evaluate the models on the testing set and analyze their performance.

Results for Sexism Classification:

Accuracy: 81.11%

Loss: 0.43%

The LSTM model achieved an accuracy of 82.85% on the testing set for sexism classification.

Results for Category Classification:

Accuracy: 78.14%

Loss: 0.70%

For category classification, the LSTM model attained an accuracy of 78.91% on the testing set.

Conclusion:

We successfully developed LSTM-based models in this project to perform text classification tasks which is sexism and category classification. On the testing set, the models performed wonderfully, obtaining excellent accuracy and comparatively low loss values. Text categorization tasks benefit greatly from LSTM's capacity to grasp long-range dependencies in sequential input. The performance of the models may be improved by more research into hyperparameters, model designs, and new preprocessing methods.