OnionOrNot: Classifying Satirical vs. Non-Satirical News

Code and materials for creating a machine learning model to categorize news headlines as satirical ("Onion") or non-satirical may be found in this repository. For text categorization, the model makes use of deep learning techniques, particularly LSTM and Bidirectional LSTM (BiLSTM), as well as natural language processing approaches.

The dataset

The OnionOrNot.csv CSV file contains the dataset used in this research. It includes news headlines with the labels "satirical" (1) and "non-satirical" (0).

Important Features of the Dataset:

Headlines: The news headlines are represented by text data.

Binary labels that indicate whether or not the headline is satirical (1) or (0) are used.

Preparation

To get the dataset ready for training, the preprocessing procedures listed below were used:

Text cleaning is the process of changing text to lowercase and eliminating extraneous characters and punctuation.

Tokenization: Using a tokenizer with a 10,000-word vocabulary to transform text into integer sequences.

Padding: Adding zeros to shorter sequences to make sure they are of the same length.

Train-Test Split: Dividing the data into subsets for testing (20%) and training (80%).

Architecture Model

Two models are used in the project:

LSTM: A unidirectional Long Short-Term Memory (LSTM) model that uses text input to identify sequential patterns.

BiLSTM: A bidirectional LSTM model that better captures contextual information by processing input sequences both forward and backward.

Important Layers:

Word indices are transformed into dense vector representations by the embedding layer.

Extract sequential patterns from the text using LSTM/BiLSTM layers.

Dense output layer: Provides binary classification probabilities.

Approach and Justification

Contextual awareness is essential for the dataset's brief text samples (headlines). The Bidirectional LSTM was selected because, in contrast to a regular LSTM, it processes the text both forward and backward, allowing for a better comprehension of the context. This method enhances the model's capacity to identify minute linguistic variations that differentiate satirical headlines from non-satirical ones.

LSTM vs. BiLSTM comparison

Although the architectures of LSTM and BiLSTM are similar, their main distinction is in the way they handle input sequences:

LSTM: Performs sequential processing of input data in a single direction, either forward or backward.

BiLSTM: Captures richer contextual information by processing input data simultaneously in both forward and backward directions.

Because BiLSTM uses bidirectional processing to better comprehend context, it was the best option for this project because to the brief and context-dependent nature of the dataset. For simpler datasets, the performance difference could be negligible.