ASSIGNMENT 8

ML project for abusive text detection.

22CS60R70 - Paneliya Yashkumar Shaileshbhai

OBJECTIVE:

In this assignment, you will learn how to classify comments as abusive or non-abusive using the TF-IDF feature extraction technique and KNN classifier. In part-2 of the assignment, you will classify the same thing using LSTM. And in part-3, you will build a text classifier using Multilingual language models like mBERT and MuRIL.

APPROACH:

1. Data preprocessing:

- Created a set of stop words in Hindi languages from several resources.
- Removed the stop words from the dataset sentences
- Removed punctuation marks from sentences
- Converted emojis to text equivalent representation using emot library
- Removed digits from text

2. TF-IDF:

- Used the preprocessed data to tokenize and calculate tf-idf values using TfIdfVectorizer
- Split dataset into 80:20 split for training and testing
- Fitted the train data and tested on test data

3. LSTM:

- Created a class for LSTM architecture with the following layers, activation function, and dimensions

  LSTM(

    (embedding): Embedding(29941, 300)

    (lstm): LSTM(300, 600, num_layers=2, batch_first=True, dropout=0.3)

    (fc): Linear(in_features=600, out_features=1, bias=True)

    (dropout): Dropout(p=0.3, inplace=False)

    (sig): Sigmoid()

)

- Created vectorized dataset using word_tokenizer()
- Padded all the sentences to a maximum length
- Split the dataset into an 80:20 ratio
- Trained the model with the below hyper-parameters

  vocab_size = len(vocab)

  embedding_dim = 300

  hidden_dim = 600

  num_layers = 2

  epochs = 10

  lr = 0.001 # learning rate

- Also embedded the logic of early stopping by maintaining a counter.

4. mBert and MURiL:
- Tokenized and encoded the dataset using hugging face's “bert-base-multilingual-cased” and “google/muril-base-cased” tokenizer.
- Fine-tuned prebuilt model for the same mBert and MURiL architecture

RESULTS:

MODEL	VALIDATION ACCURACY	MACRO F1
KNN (k=18)	64%	62%
LSTM	78.30%	78.09%
mBert	82.44%	82.17%
MURiL	85.29%	84.98%

LINK TO COLAB:
https://colab.research.google.com/drive/1_QQfZ3QD2bFQnGuHU37tPH9mNsNP4qor?usp=sharing