Comparative Analysis of Text Classification: 20 Newsgroups vs. AG News using BiLSTM with GloVe Embeddings

Saatvik Pradhan & Kashvi Prawal

Contents

1 Introduction

This report presents two projects focused on text classification using a Bidirectional LSTM (BiLSTM) network with pre-trained GloVe word embeddings. The first project uses the 20 Newsgroups dataset, and the second uses the AG News dataset. Both projects involve similar preprocessing steps—tokenization, padding, and label encoding—but differ in dataset characteristics (e.g., vocabulary size and maximum sequence length). The models are trained using early stopping, and their performance is evaluated using classification reports and confusion matrices. Finally, sample predictions are provided to demonstrate real-world performance.

2 Project 1: 20 Newsgroups Classification

2.1 Problem Definition and Dataset Curation

The goal for this project is to classify news articles from 20 Newsgroups into 20 distinct categories (e.g., `alt.atheism`, `comp.graphics`, `rec.sport.baseball`, etc.). The dataset is obtained via `fetch_20newsgroups` from scikit-learn, with headers, footers, and quotes removed. The data is split into training and test sets.

2.2 Preprocessing

- **Tokenization:** The Keras `Tokenizer` is used with a maximum vocabulary size of 20,000.
- **Padding:** Sequences are padded to a maximum length of 200 tokens to handle variable-length documents.
- **Label Encoding:** The categorical labels are one-hot encoded for multi-class classification.

2.3 Word Embeddings and Model Architecture

- **Pre-trained Embeddings:** GloVe embeddings (100-dimensional) are loaded using `gensim.downloader` and an embedding matrix is created based on the tokenizer vocabulary.
- **Model Architecture:**
  - An embedding layer that uses the pre-trained GloVe weights (set to non-trainable).
  - A Bidirectional LSTM layer with 128 LSTM units and dropout set to 0.4.
  - A dense layer with 64 units (ReLU activation) followed by a dropout layer (dropout rate 0.5).
  - A final softmax output layer for classification into 20 categories.

2.4 Training and Optimization

- The model is trained with a batch size of 32.
- Early stopping is implemented with a patience of 3 epochs (monitoring validation loss).
- A validation split of 10% of the training data is used.

2.5 Evaluation and Results

The model is evaluated on the test set using multiple metrics. The results are as follows:

- **Test Accuracy:** 57%
- **Precision (Macro):** 54%
- **Recall (Macro):** 55%
- **F1 Score (Macro):** 54%

Table 1: Classification Report for 20 Newsgroups Model

Category	**Precision**	**Recall**	**F1 Score**
alt.atheism	0.25	0.28	0.26
comp.graphics	0.46	0.54	0.50
rec.sport.hockey	0.90	0.80	0.85
sci.space	0.57	0.71	0.63
Macro Avg	**0.54**	**0.55**	**0.54**
Accuracy	**57%**		

3 Project 2: AG News Classification

3.1 Problem Definition and Dataset Curation

For the AG News project, the task is to classify news headlines into four categories: **World**, **Sports**, **Business**, and **Science/Tech**. The dataset is loaded using the `datasets` library and split into training and test sets. The integer labels are optionally mapped to their respective category names.

3.2 Preprocessing

- **Tokenization:** Using Keras `Tokenizer` with a vocabulary size of 10,000 and an out-of-vocabulary token.
- **Padding:** Headlines are padded to a fixed length of 50 tokens.
- **Label Encoding:** One-hot encoding is applied to convert integer labels into categorical format.

3.3 Word Embeddings and Model Architecture

- **Pre-trained Embeddings:** GloVe embeddings (100-dimensional) are again loaded and used to create an embedding matrix.
- **Model Architecture:**
  - An embedding layer that uses the pre-trained weights and is set as non-trainable.
  - A Bidirectional LSTM layer with 128 LSTM units and dropout set to 0.4.
  - A dense layer with 64 units and ReLU activation, followed by a dropout layer (0.5).
  - A softmax output layer with 4 units, corresponding to the four news categories.

3.4 Training and Optimization

- The model is trained using a batch size of 32.
- Early stopping is used (patience of 3 epochs) to halt training once the validation loss stops improving.
- A 10% validation split is applied during training.

3.5 Evaluation and Results

After training, the AG News model is evaluated on the test set. The results are as follows:

- **Test Accuracy:** 91%
- **Precision (Macro):** 91%
- **Recall (Macro):** 91%
- **F1 Score (Macro):** 91%

Table 2: Classification Report for AG News Model

Category	**Precision**	**Recall**	**F1 Score**
World	0.92	0.91	0.92
Sports	0.97	0.96	0.97
Business	0.88	0.87	0.87
Science/Tech	0.87	0.90	0.88
Macro Avg	**0.91**	**0.91**	**0.91**
Accuracy		**91%**	

4 Conclusion

In conclusion, both the 20 Newsgroups and AG News classification projects demonstrate the versatility of BiLSTM models combined with pre-trained GloVe embeddings for text classification tasks. The projects illustrate effective techniques in data preprocessing, model design, and training, while also highlighting areas for further experimentation—such as incorporating other types of embeddings or exploring alternative architectures—to further improve performance.