Lab 3 Report (BERT)

AYOUB FRIHAOUI G2

NATURAL LANGUAGE PROCESSING LABS (CLICK)
GITHUB

ENHANCING TEXT CLASSIFICATION: A COMPARATIVE ANALYSIS OF FEATURE EXTRACTION TECHNIQUES AND MODEL ARCHITECTURES

This report explores the advancements made in text classification by comparing the results and methodologies employed in Lab 3 with those of Lab 2. Lab 3 specifically aimed to refine and improve upon the models developed in Lab 2, addressing limitations and exploring alternative approaches.

LAB 2 RECAP: FEATURE EXTRACTION AND BASELINE MODELS

Lab 2 focused on establishing baseline models for text classification using various feature extraction techniques, including:

- **Lexical Frequency (CountVectorizer):** This method captures the frequency of word occurrences within documents, providing a basic representation of textual content.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** TF-IDF assigns weights to words based on their frequency within a document and across the entire corpus, emphasizing terms that are distinctive to specific documents.
- **Word Embeddings (Word2Vec, GloVe, FastText):** These techniques map words to dense vector representations, capturing semantic relationships and contextual information.

These extracted features were then used to train several machine learning models, primarily focusing on Multi-Layer Perceptrons (MLPs) with varying architectures and **50 epochs**.

Lab 2 Observations:

- Models trained on TF-IDF features generally outperformed those trained on lexical frequency features.
- Word embedding-based models demonstrated promising results, particularly Skip-Gram and CBOW variants of Word2Vec.
- The performance of MLP models was influenced by their architecture, with deeper networks and optimized hyperparameters yielding better accuracy.

LAB 3: REFINING THE APPROACH

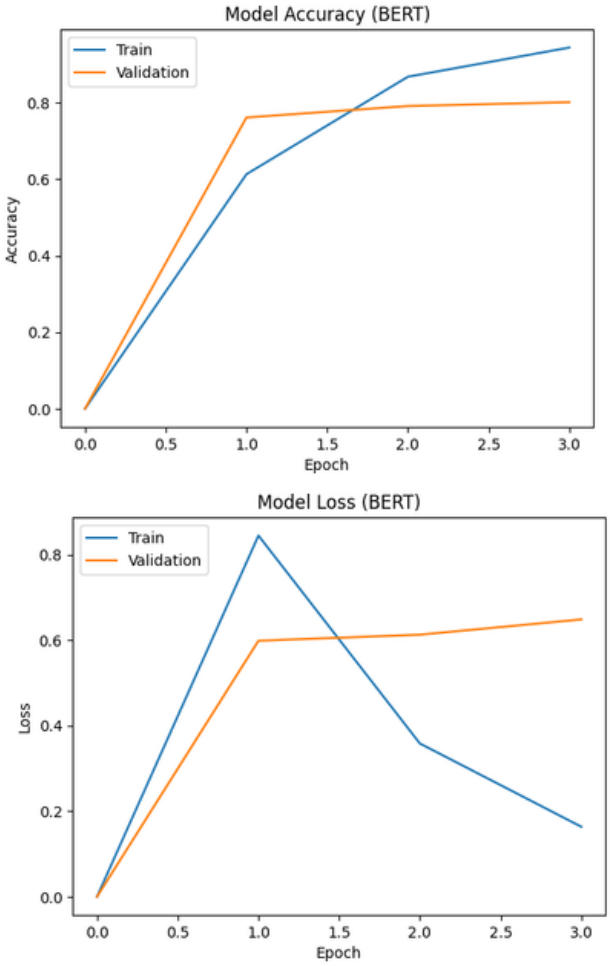
Lab 3 aimed to address the limitations of Lab 2 models by exploring the following aspects:

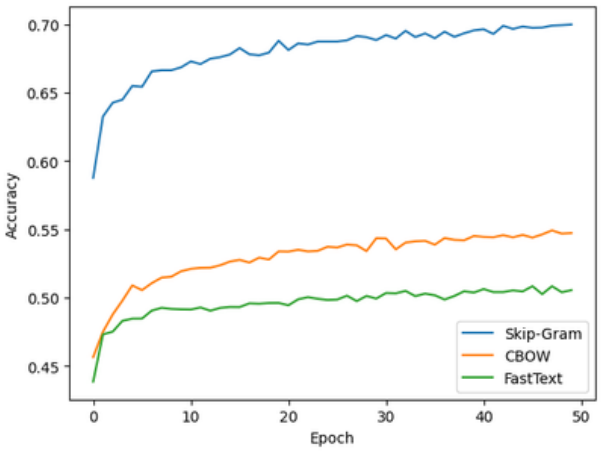
- **Fine-tuning BERT:** Instead of relying solely on extracted features, Lab 3 employed transfer learning by fine-tuning a pre-trained BERT model. This leveraged BERT's ability to capture contextualized word representations and inherent language understanding.
- **Data Sampling:** To manage computational resources and address potential class imbalances, Lab 3 implemented data sampling techniques, focusing on a subset of the original dataset.

Lab 3 Results and Analysis:

The fine-tuned BERT model significantly surpassed the performance of models developed in Lab 2 **(Skip-Gram** with an accuracy of **69.99%)**, achieving a test accuracy of **80.17%** with only 3 epochs and using a fraction of the original dataset. This improvement can be attributed to BERT's powerful contextualized word representations and the effectiveness of transfer learning that took **68 minutes on i5 1135G7 and 16GB RAM**.

However, the results also indicated potential overfitting, as the validation accuracy plateaued after the second epoch and the test accuracy fell slightly below the validation accuracy.





COMPARATIVE DISCUSSION AND FUTURE DIRECTIONS

The comparison between Lab 2 and Lab 3 highlights the effectiveness of utilizing pre-trained language models like BERT for text classification tasks. BERT's ability to capture complex language nuances and semantic relationships leads to superior performance compared to traditional feature extraction methods.

Further improvements can be explored through:

- **Regularization Techniques:** Implementing methods like dropout or L1/L2 regularization to mitigate overfitting and improve generalization.
- **Hyperparameter Optimization:** Fine-tuning hyperparameters such as learning rate, batch size, and the number of epochs to optimize model performance.
- **Error Analysis:** Analyzing misclassified examples to understand the model's weaknesses and identify areas for improvement.

- **Data Augmentation:** Expanding the training dataset by applying techniques like back-translation or text generation to improve model robustness and generalizability.

CONCLUSION

The findings from Lab 3 demonstrate the substantial progress made in text classification by leveraging the capabilities of pre-trained language models. The fine-tuned BERT model outperformed the models based on traditional feature extraction methods, achieving higher accuracy with fewer training epochs and less data. By addressing the observed overfitting and exploring further refinement techniques, the text classification model's performance and generalizability can be further enhanced.