**DIALECT IDENTIFICATION**

**1.Abstract:**

This project is targeting the challenge of dialect identification in NLP. Dialects differ in linguistic patterns and vocabulary from standard languages, which puts forward unique challenges for language processing. We used the BERT model, utilizing its contextual understanding to classify dialects into three categories: Indian English, American English, and British English. This project discusses the development, training, and evaluation of the model as well as insights gained through experiments.

**2. Introduction:**

Identifying dialect is an important task for NLP because it can lead to improvement in applications such as machine translation, speech recognition, and sentiment analysis, among others. Earlier approaches rested on handcrafted features with statistical models. However, the emergence of modern models such as BERT enable linguistic nuances to be better processed. This project involves fine-tuning a pre-trained BERT model to classify text into the Indian, American, or British English dialects. Preprocessing, feature representation of the model, fine-tuning of the model along with the evaluation process while performing the task to receive the best output are even discussed in the project.

**3. Plan to Work:**

As a part of the project divided into the following phases:

Dataset Collection and Exploration:

Collect preprocessed datasets with dialectical text samples labeled as Indian, American, or British English.  
Preprocessing: Tokenize and clean the text data and ensure that it fits well with BERT models  
Model Fine-Tuning: fine-tune the pre-trained BERT model on the dataset with the task of dialect classification  
Evaluation: Assess the model with accuracy, precision, recall, and F1 score metrics.  
Analysis and Optimization: Results must be evaluated to improve the model, using hyperparameter tuning.  
Compilation of Report: Summarize the findings, methodologies, and results into a final report.

**4. Data Explored:**

The data for this project had three dialects in the form of labeled text samples.  
Indian English: This one contains regional influence spellings and usage of vocabulary.  
American English: This is the text of the American vocabulary, spelling, and idiomatic expressions.  
British English: This possesses some unique spellings, grammar, and colloquial words.  
Data statistics:  
Number of samples: approximately 30,000 roughly (10,000 for each dialect).  
Text length: It ranges from 5 words to 150 words  
Label distribution: it is quite balanced across classes  
Source: Public datasets, news articles, social media, and conversational data.

**5. Project Milestones:**

Week 1-2 Dataset collection, cleaning, and exploration.  
Week 3 Preprocessing and first experimentation with BERT tokenization.  
Week 4-5 BERT fine-tuning along with hyperparameter tuning  
Week 6 Model evaluation and error analysis.  
Week 7: Final analysis, report writing, and presentation.

**6. Work:**

Dataset collection and preprocessing  
Tokenizing the dataset using the BERT tokenizer with padding, truncation, and sequence length adjustment  
Classifying dialects using fine-tuned BERT base model in transformers library  
Diverse batch sizes, learning rates, and epochs experimentations resulted in consistence accuracy improvements  
Generated metrics and misclassifications analyses.

**7. Analysis:**

The greatest challenge to dialectal identification was the subtlety of spelling and vocabulary differences. BERT, because it is based on contextual word embedding, is a good fit to work with such problems.  
An analysis of the type of misclassifications presented patterns that overlapped dialectially for some of the text samples. For example, some dialects will require more features or new training data to enhance accuracy further.

**8. Experiments:**

Baseline Model: Loganistic regression with TF-IDF features achieved approximate accuracy of ~70%.  
BERT Model: Fine-tuned BERT was 94% accurate over the test set.  
Hyperparameter Tuning:  
Learning Rates: Tried 2e-5, 3e-5, and 5e-5. Best found to be 3e-5.  
Batch sizes: have tried 8, 16 and 32. Best found to be 16  
Epochs: Optimum performance achieved at epoch 3.

**9. Time Consumed:**

Data collection and Preparation 2 weeks.  
Model finetuning: 2.5 weeks  
Evaluation and tuning: 1.5 weeks.  
Writing report 1 week.

**10. Report:**

This final report included sections such as the project background, methodology, experimental results, and insights. To enhance key findings, visualizations include accuracy trends, confusion matrices, and learning curves.

**11. Previous Related Work:**

The previous techniques used handcrafted features, like N-grams, POS tags, and phonetic features, to solve the problem of dialect identification. The method is reasonable in terms of classification but fails to take into account contextual effects. Recently, Word2Vec, GloVe, and BERT have provided much improvement by contextual embedding; this paper further utilizes a BERT base model in classifying dialect.

**12. BERT Model:**

We employed the BERT base uncased model from the transformers library:  
  
Preprocessing: The BERT tokenizer was used for tokenization of the text data where the max length of sequences is set at 128.  
Fine-Tuning: Using the class, fine-tunes the model into a three-class classifier, cross-entropy loss was selected as loss function and used AdamW optimizer  
Training Setup : Fine-tune 3 epochs, initial learning rate at 3e -5 with a batch size of 16 and executed in an Nvidia GPU.

**13. Results:**

Trained accuracy = 98  
Validation Accuracy = 95%  
Test Accuracy = 94%  
Confusion Matrix Analysis:  
High accuracy for distinguishing American and British English.  
Some overlap between Indian and British English due to common vocabulary.

**14. Conclusion:**

The current research did well in validating the process of fine-tuning a pre-trained BERT model for dialect identification. The system scored an impressive 94%, thus improving beyond traditional approaches. Challenges from the overlapping nature of dialectal patterns reveal a need for the development of bigger datasets with additional linguistic features. Possible future works might include enhancing the size of the dialect dataset with incorporation of speech or phonetic features.

Team members:

Shreya Talod (se22uari171)

Gouri Mandhani (se22uari215)