Emotion Intensity Prediction Project

The EmoInet project, a profound exploration in predicting emotion intensity from textual data, presents a captivating challenge within Natural Language Processing (NLP). This venture combines the intricacies of NLP with regression characteristics, aiming to discern the intensity of four distinct emotions - Anger, Joy, Sadness, and Fear - on a continuous scale from 0 to 1.

1. Project Selection and Significance:

Having previously engaged in sentiment analysis projects, the deliberate choice of the EmoInet project serves as an intentional diversification of expertise. This assignment extends beyond sentiment, introducing the complexity of predicting emotion intensity. The strategic decision to construct both a Machine Learning (ML) statistical model and a Deep Learning (DL) model using transfer learning accentuates a multifaceted approach to this challenging task.

2. Statistical Model Development (Machine learning model):

1. **Data Acquisition and Conversion:** Retrieved training, dev, test data from competition hosted website and transformed training and development data into a structured DataFrame, laying the foundation for subsequent analysis.

2**. Exploratory Data Analysis (EDA):** Conducted a comprehensive visual analysis, exploring common hashtags and employing word cloud analyses to discern prevalent themes for each emotion.

3. **Data Preprocessing and Feature Exploration**: Cleaned text data, encoded emotion labels, and utilized TF-IDF, for text representation, ensuring meaningful content focus.

4. **Feature Selection and Vectorization**: Selected both Text and Emotion has features, applied Bag of Words and TF-IDF vectorization, and created dummy variables for emotion as it is categorical column.

5. **Model Selection and Training:** Implemented multiple regression models and evaluated their performance using Mean square error, Root mean square error and Mean Absolute error Metrices, with Ridge Regression and SVM emerging as superior performers.

6**. Testing and Evaluation**: Prepared testing data, predicted intensity values using Ridge and SVM regression models, and visualized the model's performance for Ridge Regression and SVM.

7. **Final Model Training for Each Emotion:** Merged training and development data, developed separate models for each emotion, tested the model for testing data given and visualized performance for different emotions using regplots.

3. Deep Learning Model with Transfer Learning:

1. **Choice of Transfer Learning**: Opted for transfer learning for computational efficiency and for powerful language understanding capabilities.

2**. Data Preparation:** Combined training and development data, utilizing both text and emotion as input features.

3. **Text Tokenization with DistilBERT**: Tokenized text data to facilitate DistilBERT model processing.

4. **DistilBERT Model Loading and Freezing**: Loaded pre-trained DistilBERT model, freezing layers to retain pre-trained weights during task-specific training.

5. **Neural Network Model Creation:** Constructed a model architecture combining DistilBERT processing with additional layers for emotion and regression.

6. **Model Training and Testing**: Trained the model on combined data, prepared testing data, and evaluated performance metrics.

4. Approach Considerations and Insights:

The ML model demonstrated proficiency in handling structured data, while the Deep Learning model exploration involved meticulous adjustments in epochs, learning rate and dense layer configurations. Recognition of the dataset's limitations, especially its modest size, underscored the importance of a more extensive dataset for optimal DistilBERT model performance. Visualization tools like KDE plots offered nuanced insights into model behaviour.

5. Conclusion and Next Steps:

This comprehensive exploration showcases a deliberate and multifaceted approach to the EmoInet project. The strategic combination of statistical and deep learning models highlights adaptability and a commitment to nuanced problem-solving. As the dataset size significantly influences the model's performance, future steps involve exploring avenues for acquiring more extensive data and potentially experimenting with other DL architectures like LSTM for enhanced understanding.

This report not only highlights technical proficiency but also emphasizes a strategic and thoughtful approach to challenging problems within NLP. The presented models offer robust solutions to emotion intensity prediction, laying the groundwork for future refinements and enhancements