A Deep Learning Approach to Salary Frequency Classification

Using Bidirectional LSTMs and Pre-trained Embeddings

The Challenge: Structuring Nuanced Salary Data

Goal: Automatically classify a salary sentence into one of **11 distinct categories**.

Why? Salary data is often buried in unstructured text. An automated classifier allows for large-scale, detailed data analysis.

Dataset:

- A custom dataset of 1,014 salary sentences manually curated from Indeed job postings.
- **11 Classes:** `hour`, `day`, `week`, `biweekly`, `month`, `year`, `task`, `extra`, `sne` (not specified), `nd` (no digits), and `other`.
- Data showed significant class imbalance, with 'hour', 'year', and 'month' being the most frequent.

A Deep Learning Approach: Bidirectional LSTMs

Why a Recurrent Neural Network? Salary context depends on word order (e.g., "per hour" vs. "hourly wage"). A **Bidirectional LSTM (BiLSTM)** was chosen because it processes text forwards and backwards, capturing context from the entire sentence.

Pre-trained Embeddings: Used **GloVe embeddings** (100-dimensional) to give the model a foundational understanding of word relationships, which is highly effective for smaller, custom datasets.

Model Architecture:

1. Non-trainable Embedding Layer (GloVe)
2. Two stacked BiLSTM layers (64 and 32 units) to process sequences.
3. A Dense layer with a Dropout layer (50%) to prevent overfitting.
4. A final Softmax layer to output probabilities for the 11 classes.

Optimizing Performance with KerasTuner

Challenge: The initial model performed well, but could it be better? The performance of a neural network is highly dependent on its configuration.

Solution: Used **KerasTuner** with a `RandomSearch` strategy to systematically find the best hyperparameters.

Tuned Parameters:

- Number of units in both LSTM layers.
- Number of units in the Dense layer.
- Dropout rate.
- Learning rate of the Adam optimizer.

Result: The search identified an optimal configuration that significantly improved performance, leading to the final model. The best learning rate was 0.001 and the optimal dropout was 0.4.

High Accuracy with Key Insights

The final, tuned model achieved an excellent **test accuracy of 93.1%**.

Early Stopping was critical: The training plots showed the model began to overfit around epoch 5. Early stopping ensured we saved the model at its peak performance, preventing a drop in generalization.

Confusion Matrix Analysis: The model performed perfectly on clear classes like 'hour' and 'week'. The most confusion occurred with the 'other' class, highlighting the difficulty of classifying ambiguous, less-defined sentences.

Metric	Result
Final Test Accuracy	**93.1%**
Best Validation Accuracy	**94.6%**

Key Takeaways and Next Steps

Conclusion: This project successfully demonstrates that a Bidirectional LSTM network, enhanced with GloVe embeddings and systematic hyperparameter tuning, can classify complex salary sentences with high accuracy (93.1%).

Future Work:

- **Address Class Imbalance:** Use techniques like data augmentation or class weighting to improve performance on underrepresented classes (like 'biweekly' and 'task').
- **Advanced Models:** Experiment with more advanced architectures, such as adding an **Attention Mechanism** to the BiLSTM or fine-tuning a **Transformer model like BERT**.
- **Expand Feature Engineering:** Incorporate additional signals, such as explicitly flagging whether a sentence contains numerical digits, to help the model distinguish certain classes more easily.