Reflection and Learning from the Project

This project has been a deeply insightful and rewarding experience. At first glance, I **underestimated the complexity of this multiclass classification task.** Initially, it seemed relatively straightforward—but as I began working through the problem, I quickly realized how challenging it truly is, especially due to the unique characteristics of the data such as short, Romanized village names, class imbalance, and ambiguous labels.

One of the most important lessons I learned is the **critical role of encoding and tokenization** in deep learning tasks. Selecting the right tokenization strategy—whether character-level, n-gram, or phonetic—was not just a technical decision but one that fundamentally influenced model performance. I came to appreciate how the wrong choice of encoding could harm model learning, while the right one could significantly enhance generalization.

From a data perspective, I also learned the value of **statistical validation before modelling**. Performing a chi-square test to assess the relationship between village names and their corresponding regions helped confirm the feasibility of the classification task. This **gave me greater confidence in moving forward with model development**, and I now see how important it is to understand data relationships before diving into modelling.

Another key takeaway was that although **trial-and-error modelling** remains a useful approach, it should be grounded in a solid understanding of both the **data context** and the **model architecture choices**. I intentionally avoided stacking many deep layers, recognizing that even the shallow layers were struggling to learn meaningful representations due to the data's sparsity and complexity. This taught me that deeper isn't always better—especially when foundational issues in the data and encoding remain unresolved.

Finally, I found this project intellectually stimulating. It challenged me to think critically about feature representation, encoding, and how deep learning behaves when working with short, non-standardized, and noisy text data. I've gained a stronger understanding of how to approach real-world classification problems, especially those involving low-resource languages or Romanized text.

In summary, this project has strengthened my technical skills, deepened my analytical thinking, and **helped me appreciate the nuance and complexity of natural language classification**. I look forward to building on this experience in future research and applications.

Another key learning point for me was discovering how **Convolutional Neural Networks (CNNs), which I previously associated only with image classification, can be effectively applied to text classification tasks** as well. Before this project, I assumed CNNs were mainly useful for processing spatial features in images. However, through experimentation and research, I learned that CNNs are also powerful for extracting **local patterns and features in text**, especially at the character or n-gram level. For example, in this project, CNNs were used to detect local dependencies between characters or subword units in Romanized village names, which helped the model learn meaningful representations even with short input sequences. This broadened my understanding of deep learning architectures and highlighted how CNNs can generalize beyond their typical image-processing applications.