Sybrid

Proposal

Title of the Thesis	Development of embedding model for local languages (Urdu, Punjabi, Pashto, Sindhi)
Brief project description	Embedding models have revolutionized natural language processing (NLP) by enabling computers to understand and process human languages in a more nuanced manner. However, these models are predominantly fine-tuned for widely spoken languages like English, often neglecting languages such as Urdu and other regional languages. This thesis proposes to address this gap by enhancing embedding models specifically for Urdu and other underrepresented regional languages, thereby improving their applicability in various NLP tasks. The lack of high-quality, large-scale datasets and language-specific fine-tuning hinders the development of accurate and reliable embedding models for these languages. This research aims to investigate and develop methods to create more effective embedding models for Urdu and other regional languages including Pashto, Punjabi and Sindhi. The primary objectives of this thesis are: 1. To explore the limitations of existing embedding models when applied to Urdu and other regional languages. 2. To develop or adapt techniques for enhancing embedding models specifically for these languages. 3. To create a high-quality dataset that captures the linguistic nuances of Urdu and selected regional languages. 4. To evaluate the performance of the enhanced embedding models in various NLP tasks, such as text classification.
Resources required for the project	Hardware: Computer with GPU Software Tools: Python
Area of Specialization	Natural Language Processing
Duration	Final Year Project (9 – 12 months)
Contact Person	Muhammad Murtaza Khan
Designation	Chief Innovation Officer
Email	murtaza.khan@sybrid.com