<u>Unit 8 - Research Proposal Outline:</u>

Project Title:

“Efficient Implementation of Advanced Machine Learning Techniques for Context-Rich Sentiment Analysis”

Significance/Contribution to the discipline/Research Problem:

Sentiment analysis is significant because it enables the extraction of valuable insights from vast amounts of text data, facilitating informed decision-making in fields like marketing, customer service, and social media monitoring (Lai et al., 2023). Adding to the field by improving efficiency of adoption and performance of the used algorithms.

Research Question:

How can advanced machine learning techniques, particularly transformer-based models, improve the accuracy, contextual understanding, and domain adaptability of sentiment analysis systems?

Aims and Objectives:

- Aim: Implement and assess transformer models to improve multi-domain sentiment analysis within six months.
- Objectives:
  1. Review recent literature on ML for sentiment analysis.
  2. Build a SA system using fine-tuned transformers (RoBERTa/DeBERTa).

3. Apply domain adaptation to boost performance across at least two domains (e.g., movies vs. products).
4. Benchmark against classical ML (e.g., SVM) and deep learning models (e.g., LSTM).
5. Deliver a deployable SA pipeline and formal report.

Literature Review:

Sentiment Analysis (SA) still struggles with complex language, such as irony, sarcasm, and domain-specific terms (Helal et al., 2024; Bagate and Suguna, 2022). This work tackles context-awareness and bias issues by leveraging models like BERT (Devlin et al., 2019) and its successors (RoBERTa, Liu et al., 2019; DeBERTa, He et al., 2021) to improve cross-domain performance.

Key gaps addressed include:

- Efficient model implementation within tight project deadlines without compromising accuracy.
- Applying domain adaptation to boost performance on niche or underrepresented datasets (Rostami et al., 2023; Badr, Wanas and Fayek, 2024).

This research focuses on optimising the training and fine-tuning of transformer models for multi-domain SA within a six-month window. The contribution is a streamlined methodology for fast, effective deployment, offering actionable insights for both academia and industry seeking agile NLP solutions. [To be extended]

Methodology/Development strategy/Research Design:

1. Research Approach: A hybrid agile and experimental method for fast model deployment and iterative testing.
2. Phases:
   - Literature Review: Focused review (2021–2025) on transformers in SA.
   - Data Selection: Select two open-source datasets (e.g., IMDb, Amazon reviews) for cross-domain testing.
   - Model Development: Fine-tune RoBERTa or DeBERTa via Hugging Face; apply domain adaptation; optimize with early stopping, learning rate warm-up, and mixed-precision.
   - Benchmark: Train SVM (classical) and LSTM (deep learning) models.
   - Evaluation: Evaluate using Accuracy, F1, Precision, and Recall across domains.

Ethical considerations and risk assessment:

- Data Ethics: Use only public, anonymised datasets.
- Bias Mitigation: Identify and document model biases per ethical standards (Mitchell et al., 2019).
- Risks:
  - Time constraints: Minimised by optimising pre-trained models and two domains.
  - Computational risks: Cloud GPUs will offset hardware limits.
- Ethics: No human subjects; project deemed low-risk.

Created Artifacts:

- Deployable SA pipeline with fine-tuned RoBERTa/DeBERTa for two domains.
- Benchmark report vs. SVM and LSTM.
- Research report detailing methods, experiments, and findings.
- Reusable Jupyter Notebooks or Python package.

Timeline:

Month 1	Literature review, dataset selection, research plan.
Month 2	Data preprocessing, build SVM and LSTM baselines.
Month 3	Fine-tune transformer on primary domain.
Month 4	Domain adaptation on secondary dataset.
Month 5	Model evaluation and performance analysis.
Month 6	Finalise artefacts, complete report, submit.

List of References:

Badr, H., Wanas, N. and Fayek, M. (2024) 'Unsupervised Domain Adaptation via Weighted Sequential Discriminative Feature Learning for Sentiment Analysis', *Applied Sciences* 14(1). Available at: https://www.mdpi.com/2076-3417/14/1/406

Bagate, R. A. and Suguna, R. (2022) 'Sarcasm Detection with and without #Sarcasm: Data Science Approach', *International Journal of Information Science and Management* 20(4), pp. 1-15. Available at: https://ijism.isc.ac/article_698419_e19a565e765e53c527107990b9c63ce6.pdf

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv preprint*. Available at: https://arxiv.org/abs/1810.04805#:~:text=Unlike%20recent%20language%20representation%20models%2C%20BERT%20is%20designed,both%20left%20and%20right%20context%20in%20all%20layers.

He, P., Liu, X. Gao, J. and Chen, W. (2021) 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention', *arXiv preprint*. Available at: https://arxiv.org/abs/2006.03654

Helal, N. A., Hassan, A., Badr, N. L. and Afify, Y. M. (2024) ‘A contextual-based approach for sarcasm detection’, *Scientific Reports* 14. Available at: https://www.nature.com/articles/s41598-024-65217-8

Lai, S., Hu, X., Xu, H., Ren, Z. and Liu, Z. (2023) ‘Multimodal sentiment analysis: A survey’, *Displays* 80. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0141938223001968

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’, *arXiv preprint*. Available at: https://www.semanticscholar.org/paper/RoBERTa%3A-A-Robustly-Optimized-BERT-Pretraining-Liu-Ott/077f8329a7b6fa3b7c877a57b81eb6c18b5f87de

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019) ‘Model Cards for Model Reporting’, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp.220-229. Available at: https://dl.acm.org/doi/10.1145/3287560.3287596

Rostami, M., Bose, D., Narayanan, S. and Galstyan, A. (2023) ‘Domain Adaptation for Sentiment Analysis Using Robust Internal Representations’, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11484-11498. Available at: https://aclanthology.org/2023.findings-emnlp.769/