Winter School for Women 2024

"Natural Language Processing and Large Language Models"

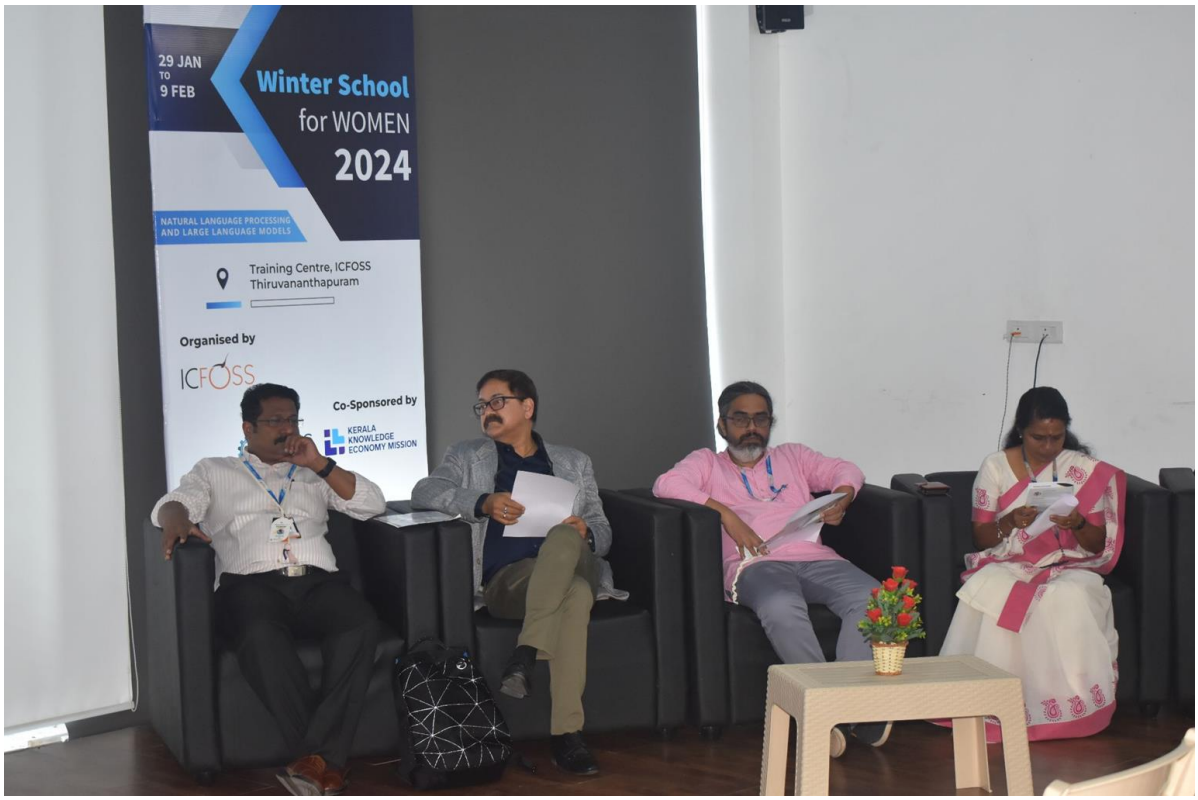
29th January to 9th February 2024

The Winter School for Women 2024 (Fifth edition), focused on "Natural Language Processing (NLP) & Large Language Models (LLM)", was organised by the International Centre for Free and Open Source Solutions (ICFOSS) held from **29th January to 9th February 2024** at ICFOSS. The program aimed to enhance awareness about Free and Open Source Software (FOSS) and address gender issues in technology. Targeting post-graduate students, research scholars, faculty members, and industry professionals, the Winter School featured project-oriented sessions that combined practical concepts with theoretical knowledge. This approach provided participants with a comprehensive understanding of NLP and LLM, fostering a collaborative and inclusive learning environment.

Inaugural Ceremony

The 12-day residential program commenced on 29th January 2024 with an inaugural ceremony at the International Centre for Free and Open Source Solutions (ICFOSS). Dr. Rajeev R R, Head of e-Governance & Development at ICFOSS, warmly welcomed the attendees. The inaugural address was delivered by Dr. Girish Nath Jha, Chairman, CSTT, Ministry of Education, Government of India. Dr. Sunil TT, Director of ICFOSS, delivered the presidential address, emphasising the program's importance. Ms. Chithra MS, Secretary & Registrar at ICFOSS, felicitated the event. Ms. Hasiya Noohu, Programme Coordinator at ICFOSS, extended a heartfelt

vote of thanks, expressing gratitude to all contributors and participants. The primary objective of the program is to enhance participants knowledge and skills in Natural Language Processing (NLP) and Generative Models, providing a comprehensive understanding of cutting-edge NLP technologies and methodologies.

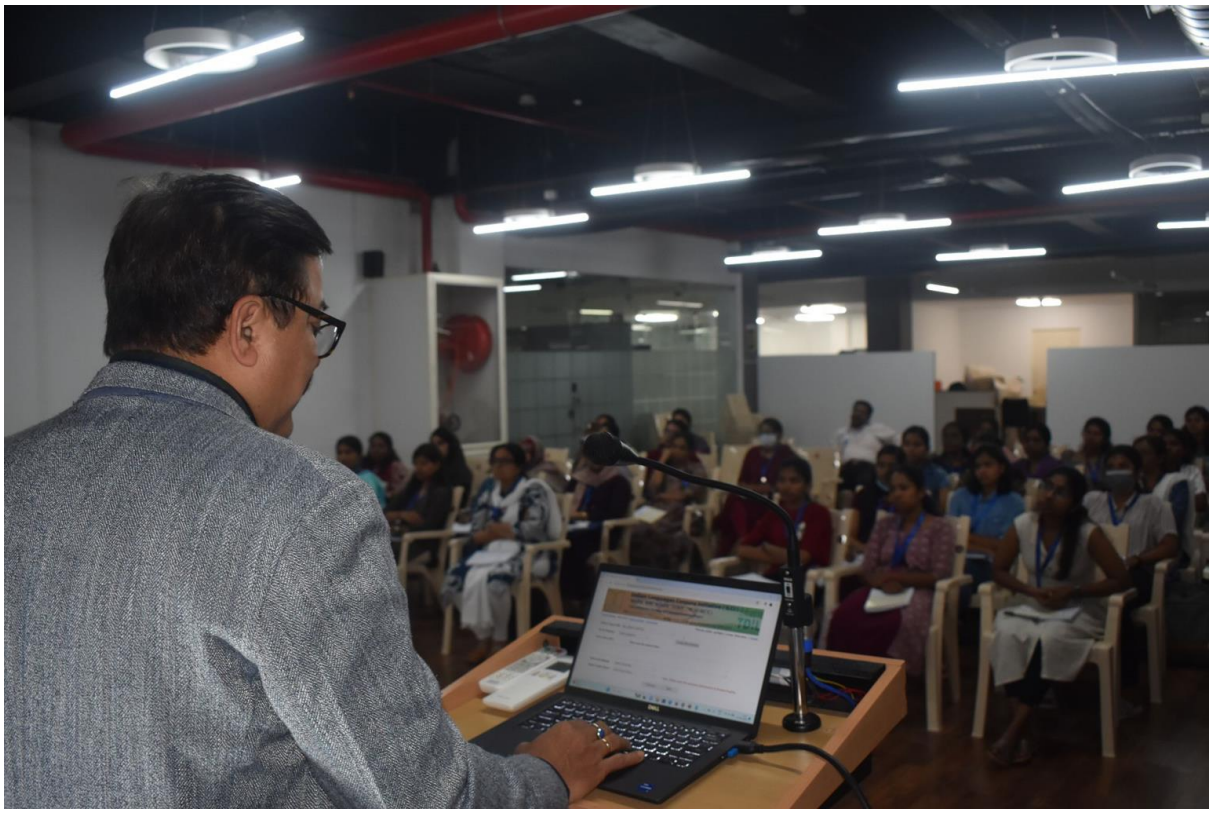


Day 1

Following the inauguration, Dr. Girish Nath Jha delivered a captivating keynote address, emphasising crucial concepts within Natural Language Processing. He began by discussing the significant contributions of key linguists like Ferdinand de Saussure, Leonard Bloomfield, and Noam Chomsky. He mentioned Saussure who is known for his work in structuralism and semiotics, highlighting the importance of the signifier and signified in language and Bloomfield, on the other hand, focused on the descriptive approach to linguistics, emphasising observable linguistic phenomena. Chomsky revolutionised linguistics with his theory of

generative grammar and the concept of a universal grammar underlying all human languages.

Moving on, he explained the different levels of language, starting with phonetics, which deals with the physical aspects of speech sounds, and phonology, which examines how sounds function within a particular language system. Morphology was discussed as the study of word formation and structure, while syntax involves the arrangement of words to create meaningful sentences. Semantics was described as the study of meaning in language, pragmatics as the study of how context influences meaning, and discourse as the analysis of language use in specific contexts.



He also touched upon various types of linguistic enquiry, including descriptive linguistics, which aims to describe and analyse languages as they are spoken, and theoretical linguistics, which seeks to develop theories and models to explain language phenomena. Additionally, he discussed applied

linguistics, which uses linguistic theories to address real-world issues, and sociolinguistics, which examines the relationship between language and society etc

Towards the end, he discussed the challenges and opportunities presented by big data in linguistics. Big data in linguistics refers to the massive amounts of language data that can be analysed to gain insights into language patterns, usage, and evolution. However, issues such as data privacy, data quality, and the need for sophisticated analytical tools were highlighted as key concerns in leveraging big data for linguistic research.

He also discussed the National Manuscript Mission (NMM) in India, which was initiated in 2003 to survey, document, preserve, and disseminate the vast knowledge contained in manuscripts. However, the NMM has encountered various challenges, including the sheer number of manuscripts, their diverse languages and scripts, insufficient awareness and expertise, inadequate funding, and issues related to storage and preservation conditions.

To overcome these challenges, integrating technology into manuscript preservation is crucial. Technologies such as high-resolution imaging, optical character recognition (OCR), and advanced preservation methods like 3D digitization or blockchain storage can significantly aid in digitising and preserving manuscripts for future generations.

Later he explained resource creation in complex linguistic societies which necessitates a comprehensive approach, involving the development of linguistic tools and resources which are tailored to the unique needs of diverse languages and dialects. This includes creating lexicons, grammatical

resources, and corpora for linguistic analysis. Furthermore, tools for speech recognition, machine translation, and natural language processing (NLP) need to be adapted or developed to suit these languages.

He emphasised that Part-of-speech (POS) tag structure is crucial for understanding the grammatical structure of a language and POS tagging assigns a specific tag to each word in a text, indicating its grammatical category and role in the sentence. The structure of POS tags varies across languages, reflecting their unique grammatical features. Domain-specific processing in NLP involves tailoring NLP techniques and tools to specific domains or fields, such as medicine, law or finance. This requires developing specialised language models, ontologies, and corpora for the domain. Domain-specific processing enhances the accuracy and relevance of NLP applications in specialised fields. Annotating data is a crucial step in creating annotated datasets for training machine learning models in NLP. Annotation involves labelling data with linguistic information, such as POS tags, named entities, or syntactic structures. This annotated data is used to train models for various NLP tasks, such as machine translation, text summarization, and sentiment analysis.

He explained that smart computer technologies, such as artificial intelligence and machine learning, require large quantities of annotated data to achieve high performance. Annotated data is used to train models to understand and generate human language. However, annotating data manually can be time-consuming and expensive. Automated or semi-automated annotation tools can help accelerate the annotation process and improve the quality of annotated datasets.

He concluded by stating that ongoing research and collaboration are essential for furthering our understanding of language and its role in society. By addressing these challenges and embracing new technologies, linguistics can continue to illuminate the intricacies of human language and contribute to the development of smart computer technologies that enhance our lives.

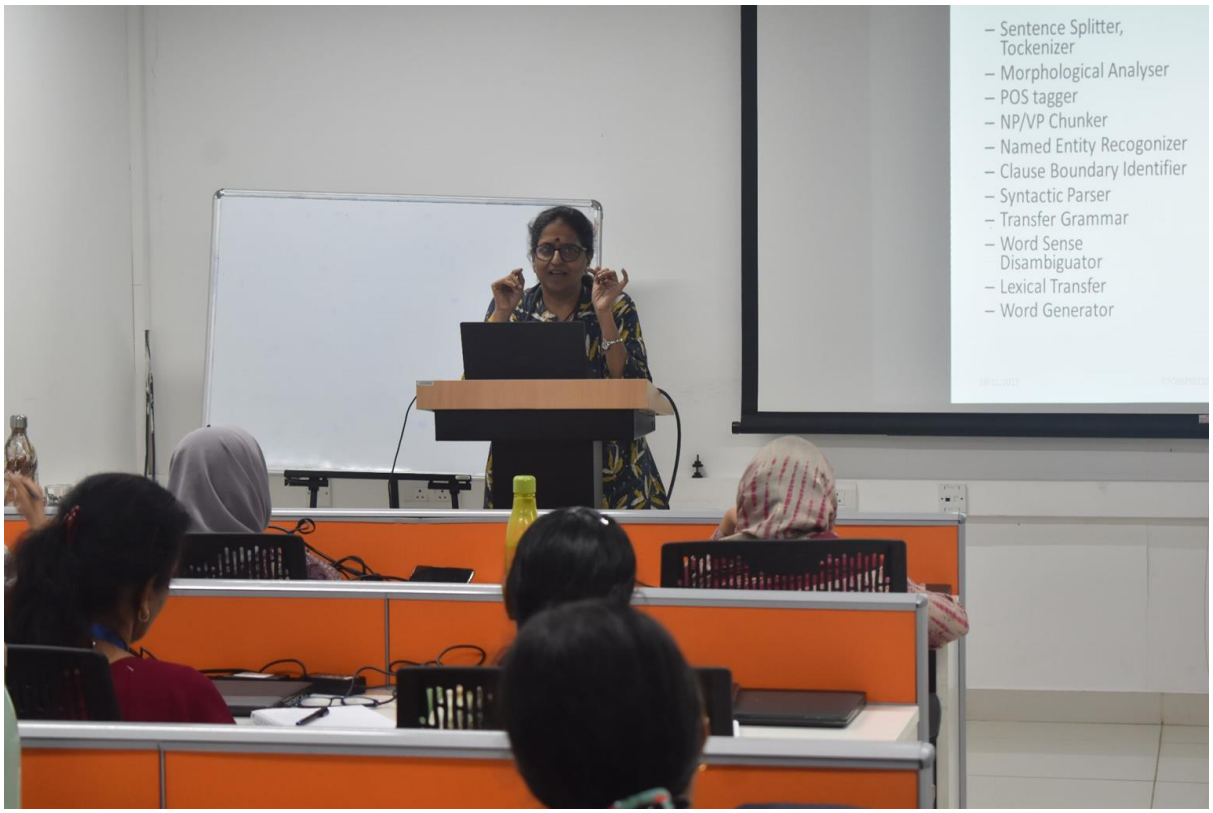
Afternoon Session

Resource Person - Dr Sobha lalitha Devi
Topic - Linguistic Aspects of NLP
Time - 2 pm to 5 pm

Dr. Sobha Lalitha Devi's session on the Linguistic Aspects of Natural Language Processing (NLP) focused on the intricate relationship between language and technology. She began by defining NLP as a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. Central to NLP is the concept of "text," which refers to any form of written or spoken language that computers analyse. She emphasised the importance of linguistic cohesion, which involves the grammatical and lexical relationships that contribute to the overall coherence of a text.

The session then explored various tasks involved in NLP. For instance, a "sentence splitter" identifies and separates sentences within a text, while a "tokenizer" breaks down the text into individual tokens, such as words or punctuation marks. Additionally, a "morphological analyzer" examines the structure of words to determine their grammatical properties, and a "POS tagger" assigns grammatical tags to words based on their parts of speech.

These tasks are fundamental in enabling computers to analyse and understand the structure of human language.



She also discussed the processing of a text, which involves several stages such as preprocessing, deep analysis, and shallow analysis. Preprocessing includes tasks like cleaning and formatting raw text for analysis, while deep analysis involves in-depth linguistic analysis, including syntactic and semantic analysis. Shallow analysis, on the other hand, focuses on surface-level linguistic features.

Furthermore, the session covered computational requirements for NLP tasks, highlighting the need for significant computational resources, especially for tasks requiring deep linguistic analysis. The complexity of NLP tasks and the size of the text being analysed can impact the computational requirements.

In conclusion, Dr. Sobha Lalitha Devi's session provided a detailed overview of the linguistic aspects of NLP, showcasing how computational techniques and linguistic knowledge are integrated to enable computers to interact with human language effectively.

Day 2

Resource person - Dr Umesh P

Topic - Mathematical foundation for Deep learning

Time - 10 am to 5 pm

Dr. Umesh P's session provided a comprehensive view on the mathematical foundation for deep learning, offering attendees a detailed overview of key concepts essential for understanding and applying deep learning algorithms in practice.

The session began with an overview of machine learning, highlighting its various types, including supervised, unsupervised, and reinforcement learning. Supervised learning involves learning from labelled training data, where each example is paired with the correct output, aiming to learn a mapping from inputs to outputs. Unsupervised learning deals with learning from unlabeled data, seeking to find patterns or structure in the data without explicit guidance. Reinforcement learning centres on learning how to make sequences of decisions based on feedback in the form of rewards or punishments.

Linear algebra's role in machine learning was explained, emphasising its significance in understanding and manipulating high-dimensional data. Vectors are used to represent data points, while matrices are used to

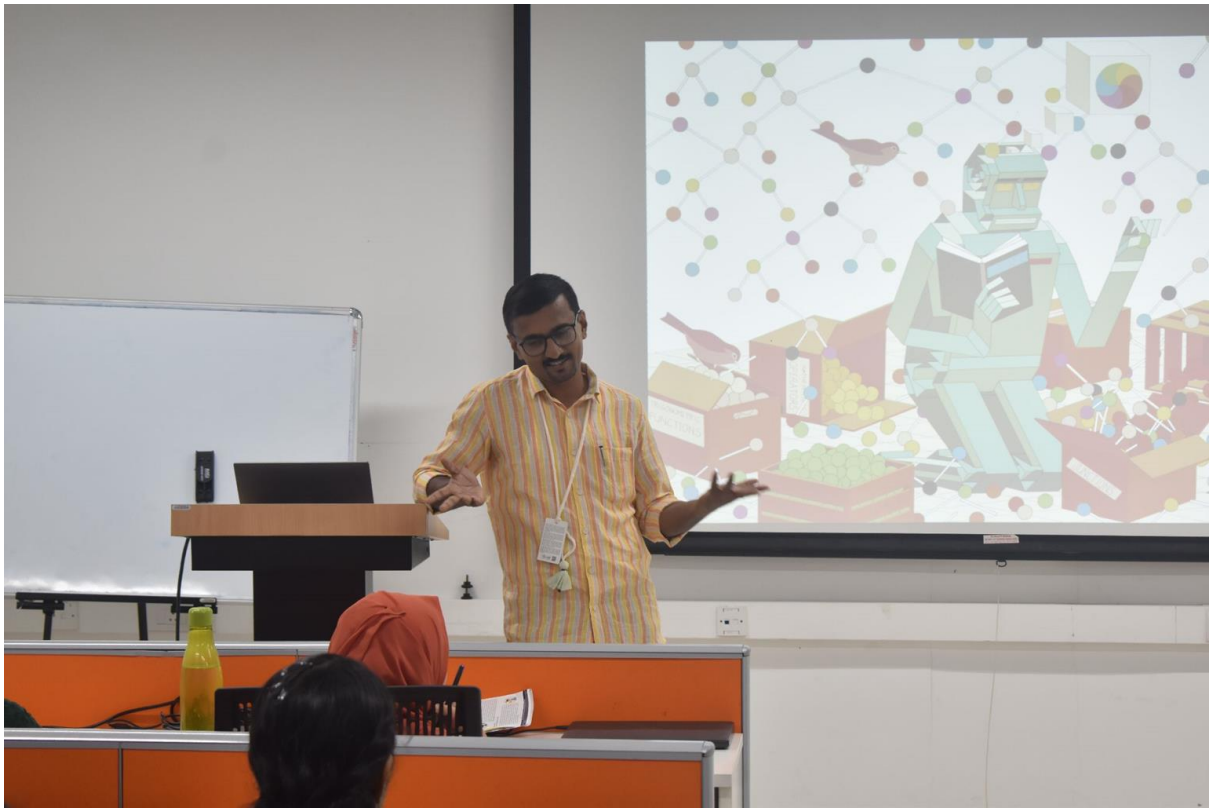
represent datasets or transformations. Operations such as matrix multiplication and transpose were underscored for their importance in various machine learning algorithms.

Calculus was discussed as essential for optimization, which involves finding the best parameters for a model to minimise a loss function. Derivatives and gradients were explained as crucial tools for updating parameters in machine learning models, with the Jacobian matrix and Hessian matrix being instrumental in computing higher-order derivatives.

Probability and statistics were highlighted as crucial for modelling uncertainty in data and making probabilistic predictions. Concepts such as mean, median, mode, covariance, variance, and correlation coefficient were explained as fundamental measures for describing and analysing data distributions. Probability distributions, conditional probability, and probability mass function were explained as essential for modelling uncertainty.

The session also gave an introduction to neural networks, the building blocks of deep learning. Linear regression and logistic regression were discussed as foundational models, with activation functions introducing non-linearity to neural networks. Gradient descent was detailed as an optimization algorithm used to minimise the loss function of a neural network, with regularisation techniques being employed to prevent overfitting. Classification was discussed as a fundamental task in machine learning, with learning algorithm selection being emphasised as crucial for the success of a machine learning project. Model performance assessment metrics such as accuracy, precision, recall, and F1 score were also

discussed.



The session also covered the feedforward and backpropagation operations in neural networks, providing a detailed explanation of these fundamental processes.

Feedforward operation involves the propagation of input data through the neural network, passing through multiple layers of neurons and applying activation functions to produce an output. Each neuron in the network computes a weighted sum of its inputs, adds a bias term, and applies an activation function to produce an output.

Backpropagation operation, on the other hand, is used to update the weights and biases of the neural network based on the error between the predicted output and the actual output. It involves calculating the gradient of the loss function with respect to the weights and biases of the network, and then

using this gradient to update the weights and biases using an optimization algorithm such as gradient descent.

During the session, a worksheet on neural networks was provided, which included concepts such as bias, learning rate, cost function, loss function, and the difference between loss and cost functions.

In conclusion, Dr. Umesh P's session provided a comprehensive understanding of the mathematical foundation for deep learning, equipping attendees with the knowledge necessary to apply deep learning algorithms effectively in real-world scenarios.

Day 3

Morning Session

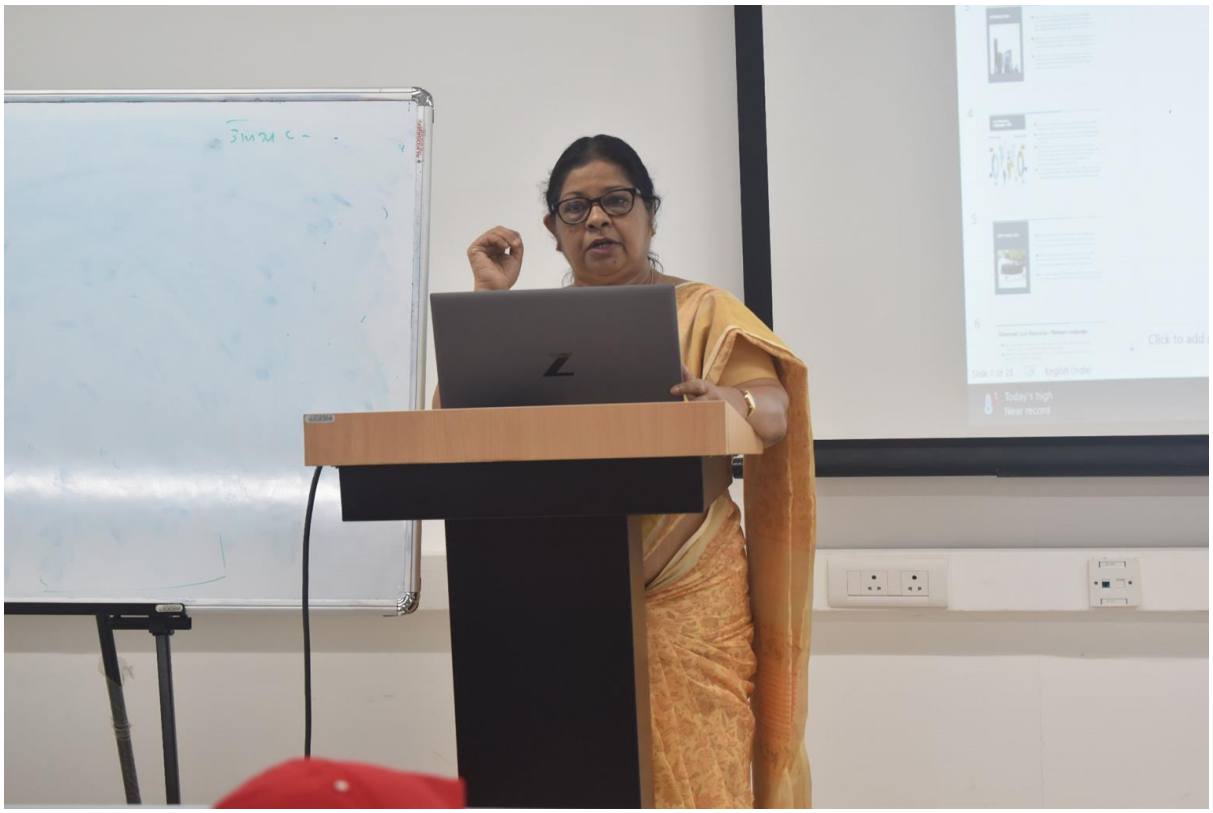
Resource Person - Dr Elizabeth Sherly

Topic - Automatic Speech Recognition (ASR) and Machine Translation (MT) systems for Low Resource Languages (LRLs)

Time - 10 am to 11 pm

Dr. Elizabeth Sherly's session on Automatic Speech Recognition (ASR) and Machine Translation (MT) systems for low resource languages (LRLs) provided a thorough analysis of the challenges and strategies involved in developing effective systems for these languages. The session began with an overview of the unique characteristics of LRLs, emphasising their limited availability of linguistic resources such as annotated data, dictionaries, and language models. Dr. Sherly highlighted the importance of ASR and MT for LRLs, noting their potential to facilitate communication, preserve cultural

heritage, and promote linguistic diversity.



Using Malasar as an example of an extremely low resource language, Dr. Sherly illustrated the challenges faced in developing ASR and MT systems for such languages. She emphasised the need for innovative approaches to address these challenges, including the use of large language models (LLMs) and transfer learning techniques. The session outlined a proposed system architecture for ASR and MT for LRLs, focusing on the use of transformer-based models.

The architecture of the transformer-based ASR system was explained in detail, highlighting its ability to handle the complex linguistic structures of LRLs. Similarly, the architecture of the transformer-based MT model was discussed, with an emphasis on its adaptability to different language pairs and its effectiveness in handling LRLs. She explained why transfer models are used in ASR and MT for LRLs, citing their ability to utilise pre-trained

models and adapt to the specific characteristics of LRLs.

The session also focused on the attention-based encoder-decoder models, explaining their importance in capturing the context and structure of LRLs. She then emphasised the need for corpus creation in developing ASR and MT systems for LRLs, highlighting the challenges involved in collecting and annotating data for these languages. Results from experiments conducted using the proposed system were presented, demonstrating the performance of the system in terms of accuracy and efficiency.

In conclusion, Dr. Sherly reiterated the importance of preserving LRLs for their cultural, historical, and linguistic value. She emphasised the role of technology, particularly ASR and MT systems, in facilitating the preservation and revitalization of LRLs. The session concluded with a call for continued research and collaboration in this area to ensure the continued development of effective language technologies for LRLs.

Afternoon session

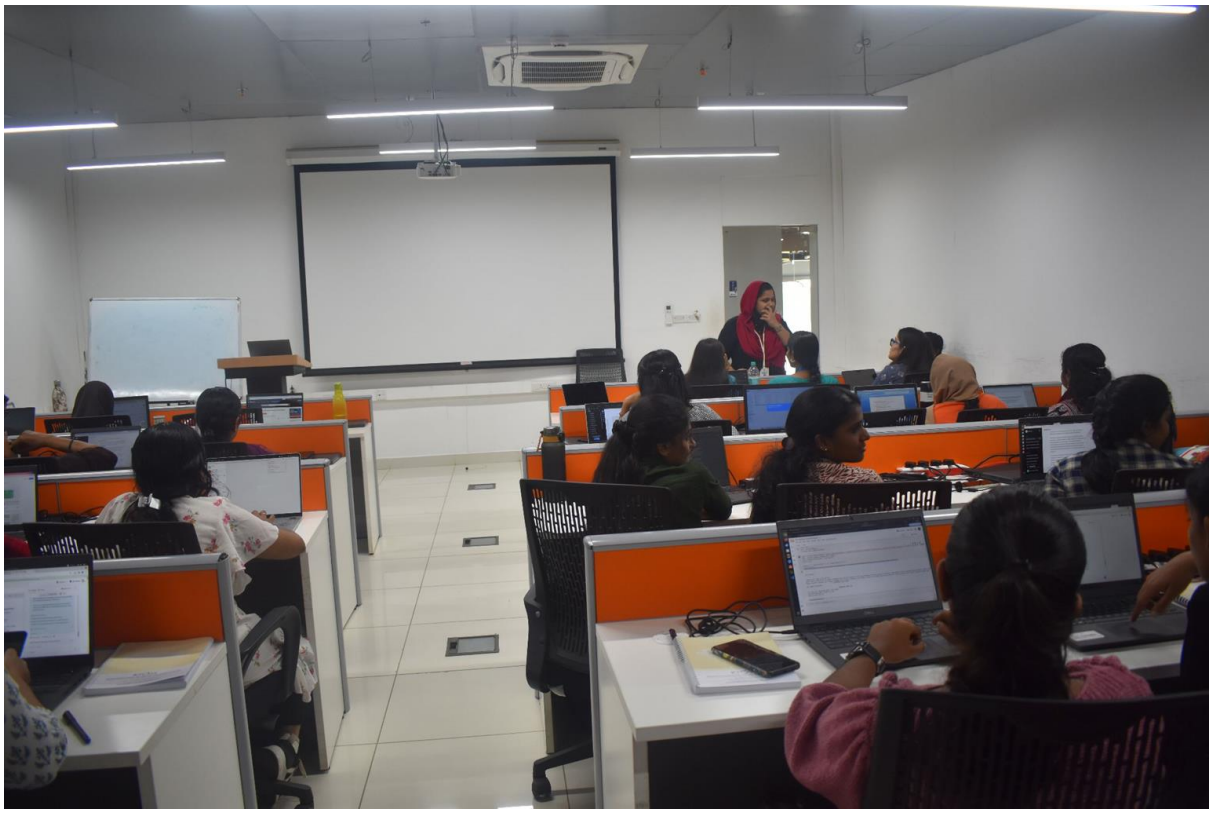
Resource Person - Ms. Meharuniza Nazeem

Topic - Machine Learning in NLP

Time - 2pm to 3pm

Ms. Meharuniza Nazeem's session on Machine Learning through Natural Language Processing (NLP) provided a detailed exploration of how machine learning techniques are applied in various aspects of NLP. She began by discussing the different types of machine learning, including supervised, unsupervised, and deep learning, and their applications in NLP tasks such as text classification, sentiment analysis, and named entity

recognition. She highlighted the importance of feature engineering in machine learning for NLP, where linguistic features such as word embeddings and syntactic structures are used to train models effectively.



The session also covered the challenges of working with large-scale datasets in NLP and the need for scalable machine learning algorithms. She discussed the role of neural networks in NLP, including convolutional neural networks (CNNs) for text classification and recurrent neural networks (RNNs) for sequence modelling. She also highlighted the significance of attention mechanisms in improving the performance of machine learning models in NLP tasks such as machine translation and text summarization.

Moreover, she addressed the ethical considerations in using machine learning in NLP, emphasising the need for fairness, transparency, and accountability in AI systems. She discussed the importance of bias detection

and mitigation strategies to ensure that machine learning models are not perpetuating or amplifying biases present in the data.

In conclusion, Ms. Meharuniza Nazeem's session provided a comprehensive overview of machine learning through NLP, showcasing its potential to revolutionise language understanding and interaction. Her insights into the latest advancements and challenges in the field highlighted the need for continued research and innovation to harness the full power of machine learning in NLP.

Day 4

Morning Session

Resource Person - Ms. Meharuniza Nazeem
Topic - Word Embedding(Word2vec, FastText, Glove)
Time - 10 am to 1 pm

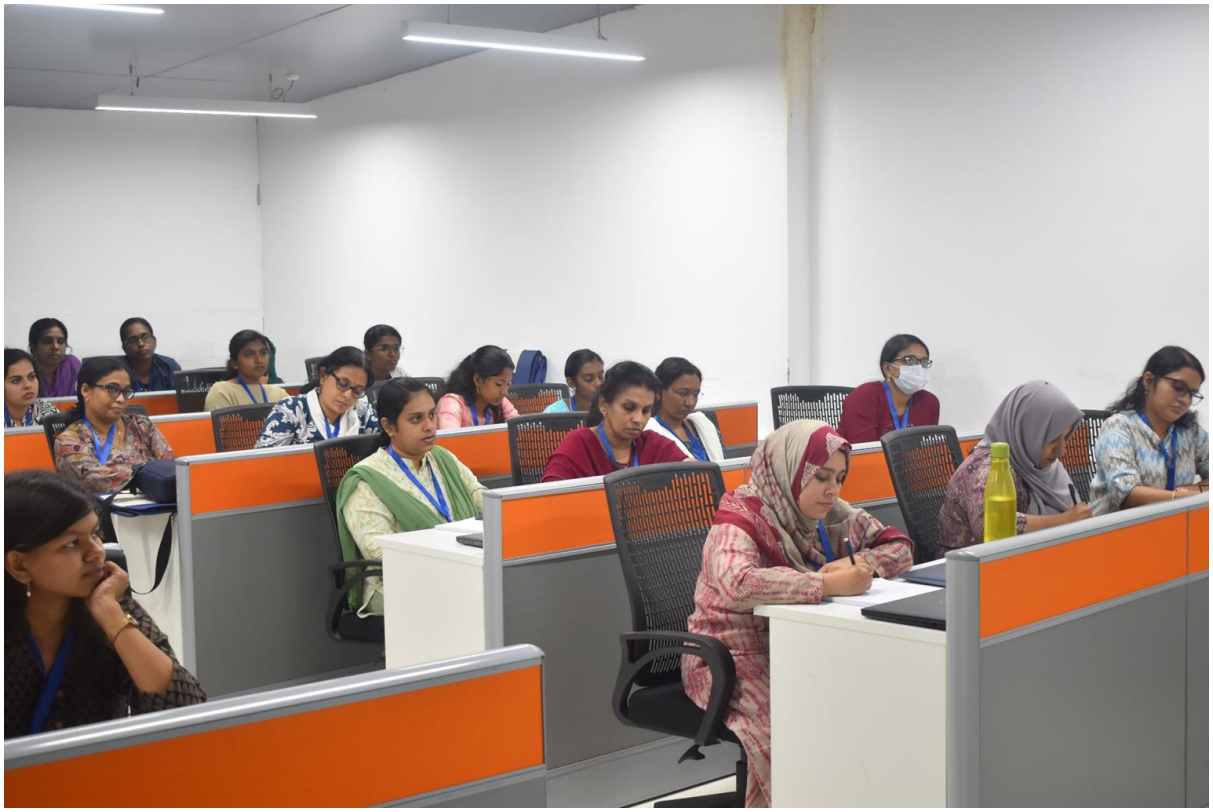
Meharuniza Nazeem's session on word embedding covered three popular techniques: Word2Vec, FastText, GloVe and One-Hot Encoding. Word embedding is a technique used in natural language processing (NLP) to represent words as vectors in a continuous vector space. These word vectors capture semantic relationships between words, enabling NLP models to better understand and process language.

Word2Vec is a neural network-based model that learns word embeddings by predicting the surrounding words in a sentence (skip-gram model) or predicting a word based on its context (continuous bag of words model). The resulting word vectors capture semantic relationships, such as similarity

and analogy, between words. Word2Vec has been widely used in various NLP tasks, including sentiment analysis, machine translation, and named entity recognition, due to its effectiveness in capturing word semantics.

FastText is an extension of Word2Vec that introduces subword information into word embeddings. Instead of representing each word as a single vector, FastText represents words as the sum of their character n-grams. This allows FastText to capture morphological information and handle out-of-vocabulary words more effectively than Word2Vec. FastText has been particularly useful for tasks involving morphologically rich languages or domains with limited training data.

GloVe is another popular word embedding technique that learns word vectors by factoring a matrix of word co-occurrence statistics. Unlike Word2Vec and FastText, which are based on neural networks, GloVe is based on matrix factorization and directly optimises word vectors to capture global word-word co-occurrence statistics. This results in word embeddings that are effective at capturing global semantic relationships between words.



One-hot encoding is a technique to represent categorical data, such as words, as binary vectors. Each word is represented by a vector where all elements are 0 except for one element, which is 1 at the index corresponding to the word's position in the vocabulary. While simple, one-hot encoding does not capture semantic relationships between words and results in high-dimensional sparse vectors.

Word embedding techniques like Word2Vec, FastText, and GloVe are widely used in NLP for tasks such as sentiment analysis, machine translation, and named entity recognition. These techniques enable machine learning models to better understand and process natural language by representing words in a continuous vector space.

Meharuniza Nazeem's session provided a comprehensive overview of

Word2Vec, FastText, and GloVe, highlighting their strengths and applications in NLP. These word embedding techniques have significantly advanced the field of NLP by enabling models to effectively capture semantic relationships between words and improve performance on various NLP tasks. Understanding these techniques is crucial for researchers and practitioners in the field of NLP to develop more accurate and efficient language processing systems.

Afternoon session

Resource Person - Dr. Rajeev R R
Topic - Introduction to NLP
Time - 2 pm to 4pm

Dr. Rajeev R R's session on Introduction to Natural Language Processing (NLP) provided a comprehensive overview of the field, highlighting its significance and applications in today's world. The session aimed to familiarise participants with the basic concepts and techniques used in NLP. He began by defining NLP as a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. He explained the importance of NLP in various real-world applications, such as chatbots, language translation, sentiment analysis, and information extraction.

The session covered several basic techniques used in NLP, including tokenization, which involves breaking text into individual words or tokens; part-of-speech tagging, which assigns grammatical tags to words; and

syntactic parsing, which analyses the grammatical structure of sentences. He also discussed the wide range of applications of NLP in today's world, including machine translation, speech recognition, text summarization, and sentiment analysis. He emphasised how NLP has revolutionised the way we interact with technology, enabling more natural and intuitive communication. The session concluded with a discussion on the challenges facing NLP, such as ambiguity in language, understanding context, and handling different languages and dialects. Dr. Rajeev RR also highlighted the future directions of NLP, including the integration of NLP with other technologies such as machine learning and deep learning to further enhance its capabilities.

The session provided a comprehensive overview of the field, covering its basic concepts, techniques, applications, and future directions. The session was insightful and informative, providing participants with a solid foundation in NLP and its potential impact on society.

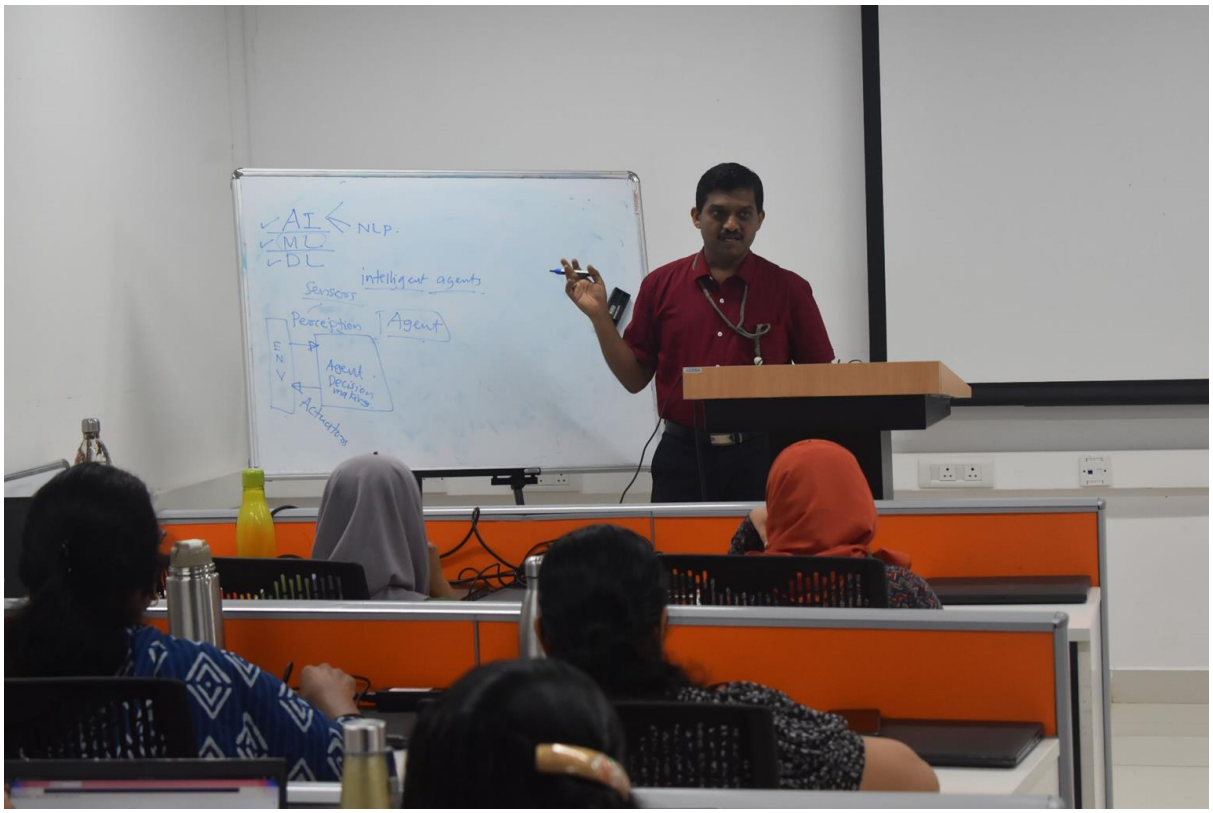
Day 5

Resource Person - Dr Gopakumar G
Topic - Introduction to Neural Networks, RNN, LSTM
Time - 10 am to 5 pm

Dr. Gopakumar G's session on the introduction to Neural Networks, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks provided a comprehensive overview of these fundamental concepts in deep learning. The session began with an introduction to Neural Networks, highlighting their ability to learn complex

patterns in data through interconnected layers of neurons. The concept of Multilayer Neural Networks was discussed, demonstrating how multiple layers allow for more sophisticated modelling of data.

The session then focused on the feedforward operation and classification in Neural Networks, explaining how input data is processed through the network to produce an output, which is then used for classification or prediction tasks. The Backpropagation Algorithm was detailed as a method for updating the weights of the network based on the error between the predicted and actual output, allowing the network to learn from its mistakes and improve its performance.



Two modes of operation were discussed: feedforward and learning. In feedforward mode, the network processes input data to produce an output, while in learning mode, the network adjusts its weights based on the error in the output to improve its performance. The session also covered network

learning, including learning curves, which depict the network's performance over time, and error surfaces, which represent the relationship between the network's weights and its error.

Backpropagation was explained as a form of feature mapping, where the network learns to map input features to output predictions. Practical techniques for improving backpropagation were discussed, including the choice of activation function, scaling input and target values, training with noise, and manufacturing data to augment the training set.

Other factors influencing backpropagation performance, such as the number of hidden units, initialising weights, learning rates, momentum, weight decay, and criteria function, were also covered. The session highlighted the importance of choosing the right training method (on-line, stochastic, or batch training) and knowing when to stop training to avoid overfitting. The number of hidden layers was discussed as a factor influencing the network's ability to learn complex patterns.

In addition to Neural Networks, the session covered Recurrent Neural Networks (RNNs), which are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. The structure of RNNs was explained, illustrating how the hidden state is updated at each time step using an activation function. The session also discussed the different categories of sequence modelling, including one-to-one, one-to-many, many-to-one, and many-to-many models, showcasing the versatility of RNNs in various tasks.

The power of RNNs in representation was emphasised, highlighting their ability to capture long-range dependencies in sequential data. Training

RNNs was explained, focusing on the Backpropagation through Time (BPTT) algorithm, which extends backpropagation to sequential data by unfolding the network over time. The problem of Vanishing gradients in RNNs was addressed, and Truncated BPTT was introduced as a technique to mitigate this issue by limiting the number of time steps considered during training.

Long Short-Term Memory (LSTM) networks were introduced as a variant of RNNs designed to better capture long-range dependencies. The forward pass of LSTM was explained, detailing how it handles input, forget, and output gates to control the flow of information through the network. Training LSTM was discussed, highlighting the importance of initialising weights and choosing appropriate hyperparameters.

Gated Recurrent Unit (GRU) was introduced as another variant of RNNs, which simplifies the architecture of LSTM while achieving comparable performance. Bidirectional RNNs were discussed as a way to improve the representation of sequential data by processing it in both forward and backward directions.

Practical applications of RNNs were explored, including machine translation, where RNNs are used to translate text from one language to another, and generating image descriptions, where RNNs generate textual descriptions of images. The computational graph of RNNs was illustrated, showcasing how data flows through the network during training and inference. Sequence-to-sequence models were discussed as a general framework for tasks that involve mapping input sequences to output sequences, such as machine translation.

Overall, Dr. Gopakumar G's session provided a comprehensive understanding of Neural Networks, RNNs, and LSTMs, equipping attendees with the knowledge necessary to apply these concepts to a wide range of tasks in deep learning.

Day 6

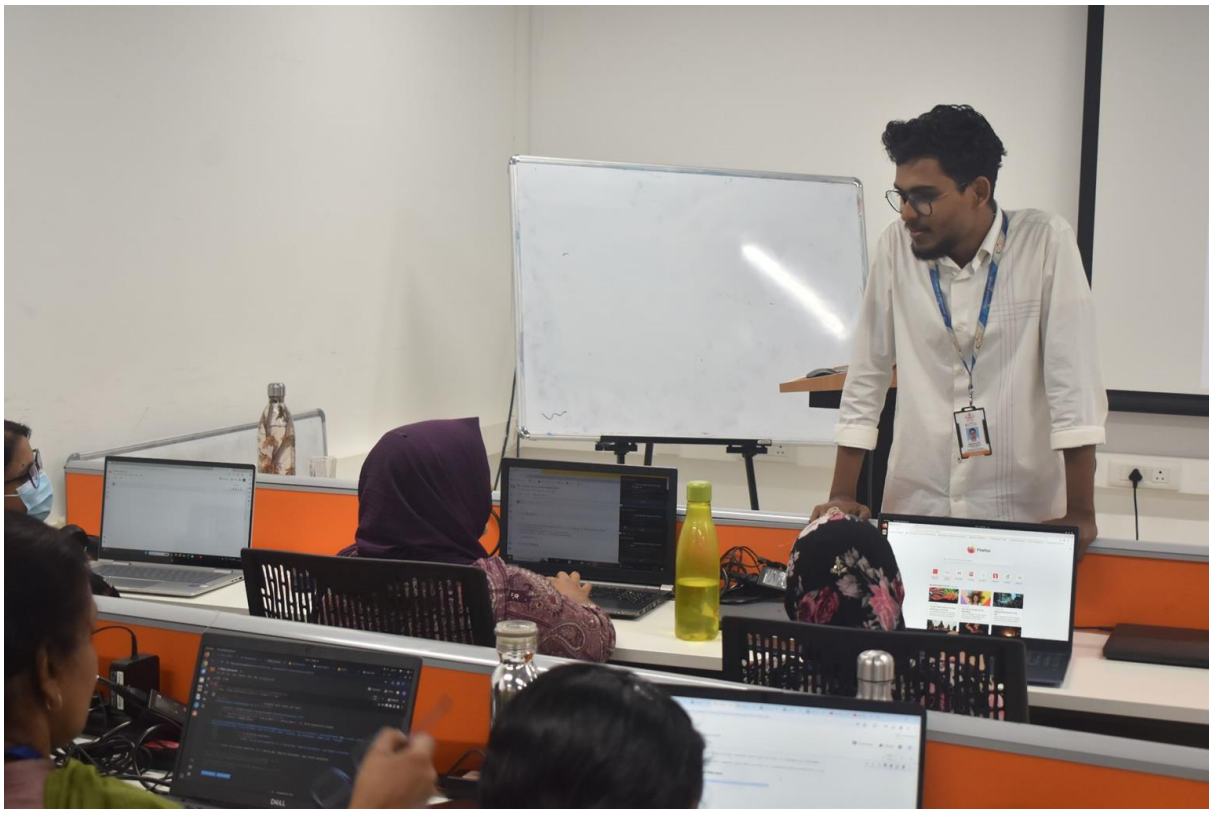
Resource Person - Mr Sabeerali KP
Topic - Deep Learning Frameworks
Time - 10 am to 5 pm

The session on Day 6, conducted by Mr. Sabeerali KP, focused on Deep Learning Frameworks, with a detailed discussion on major frameworks like TensorFlow and PyTorch. Participants were engaged in a hands-on session, where they actively built models using these cutting-edge tools.

The morning session specifically covered TensorFlow, an open-source deep learning framework developed by Google in 2015. TensorFlow is renowned for its extensive documentation, training support, scalable production and deployment options, multiple abstraction levels, and compatibility with various platforms, including Android. It is a symbolic maths library ideal for neural networks and excels in dataflow programming for diverse tasks.

During the hands-on segment, participants were guided through the process of creating a model using the FashionMNIST dataset. The session started with importing the necessary libraries and datasets, followed by data preprocessing, model building, training, and evaluation. Participants learned

how to define a neural network architecture using TensorFlow's high-level Keras API, compile the model with an optimizer and loss function, and train it using the training data. They also explored techniques for model evaluation and learned how to make predictions on new data. Overall, the hands-on session provided participants with practical experience in using TensorFlow to build and train deep learning models.



The afternoon session focused on PyTorch, a machine learning framework based on the Torch library, originally developed by Meta AI and now part of the Linux Foundation umbrella. PyTorch is widely used for applications such as computer vision and natural language processing. It is known for its dynamic computational graph, which allows for more flexible and intuitive model building compared to static graph frameworks like TensorFlow. PyTorch is free and open-source software released under the modified BSD licence.

During the session, participants built the same model using the FashionMNIST dataset in PyTorch to gain a deeper understanding of the differences between PyTorch and TensorFlow. They imported the required libraries and datasets using specific commands, similar to the process in TensorFlow. By comparing the two frameworks in the context of the same task, participants were able to appreciate the unique features and advantages of each framework, enabling them to make informed decisions in choosing the right framework for their future projects.

Day 7

Resource Person - Mr. Sabeerali KP

Topic - Transformers

Time - 10 am to 5pm

The objective of this session was to provide participants with an in-depth understanding of Transformers, a revolutionary deep learning architecture that has transformed the field of Natural Language Processing (NLP). The session aimed to elucidate the architecture, working principles, and applications of Transformers, along with recent advancements and future directions.

The session began with an introduction to Transformers, a deep learning architecture introduced in the paper "Attention is All You Need" by Vaswani et al. in 2017. Transformers have gained widespread popularity in NLP tasks due to their ability to model long-range dependencies efficiently and their parallelizable nature, enabling faster training and inference compared to recurrent neural networks (RNNs) and convolutional neural

networks (CNNs).

Participants were introduced to the key components of Transformers, including:

1. Self-Attention Mechanism: Transformers rely on self-attention mechanisms to weigh the importance of different words in a sequence when generating representations. This mechanism enables the model to capture global dependencies and contextual information effectively.
2. Multi-Head Attention: To enhance representational capacity, Transformers employ multi-head attention mechanisms, allowing the model to focus on different parts of the input sequence simultaneously.
3. Positional Encoding: Since Transformers lack recurrence and convolution, they require positional information to understand the order of words in a sequence. Positional encodings are added to the input embeddings to provide this information.
4. Feedforward Neural Networks: Transformers include feedforward neural networks to process the representations generated by the self-attention mechanism and produce the final output.

The session covered various applications of Transformers across NLP tasks, including:

1. Language Translation: Transformers, particularly the Transformer model introduced in the original paper, have been highly successful in machine translation tasks, achieving state-of-the-art performance on benchmark datasets such as WMT.

2. Question Answering: Transformers have been applied to question answering tasks, where they excel at processing and understanding contextual information to generate accurate answers.
3. Text Summarization: Transformers are effective in text summarization tasks, where they can generate concise summaries of longer documents by focusing on important information.
4. Named Entity Recognition (NER): Transformers have been applied to NER tasks, where they can identify and classify named entities such as names of people, organisations, or locations.

Participants were briefed on recent advancements in Transformer architecture, including:

1. BERT (Bidirectional Encoder Representations from Transformers): BERT, introduced by Devlin et al. in 2018, is a pre-trained Transformer model that has achieved significant improvements in various NLP tasks by pre-training on large corpora of text data.
2. GPT (Generative Pre-trained Transformer): GPT models, developed by OpenAI, are autoregressive Transformer models capable of generating coherent and contextually relevant text.
3. XLNet: XLNet, introduced by Yang et al. in 2019, is a generalised autoregressive pre-training method that outperforms previous Transformer-based models by leveraging permutation-based training objectives.

In conclusion, the session provided participants with a comprehensive understanding of Transformers, including their architecture, working principles, applications, recent advancements, and future directions. By

grasping the fundamentals of Transformers, participants are better equipped to leverage this powerful deep learning architecture in various NLP tasks and contribute to advancements in the field.

Day 8

Resource Person - Mr Navaneeth S and Mr Arun A

Topic - Bert, Hugging face

Time - 10 am to 5pm

The objective of this session was to provide participants with a comprehensive understanding of BERT (Bidirectional Encoder Representations from Transformers) and the Hugging Face library. The session aimed to elucidate the architecture, working principles, and applications of BERT, along with practical demonstrations of using the Hugging Face library for NLP tasks.

The session began with an introduction to BERT, a state-of-the-art pre-trained language representation model introduced by Jacob Devlin and his colleagues at Google AI in 2018. BERT has revolutionised the field of Natural Language Processing (NLP) by utilising the Transformer architecture and pre-training on large corpora of text data in an unsupervised manner.

Participants were introduced to the key features of BERT, including:

1. Bidirectional Context: Unlike previous language models that process text in a left-to-right or right-to-left manner, BERT utilises bidirectional context by considering both left and right context

simultaneously. This enables BERT to capture rich semantic information and context-dependent representations.

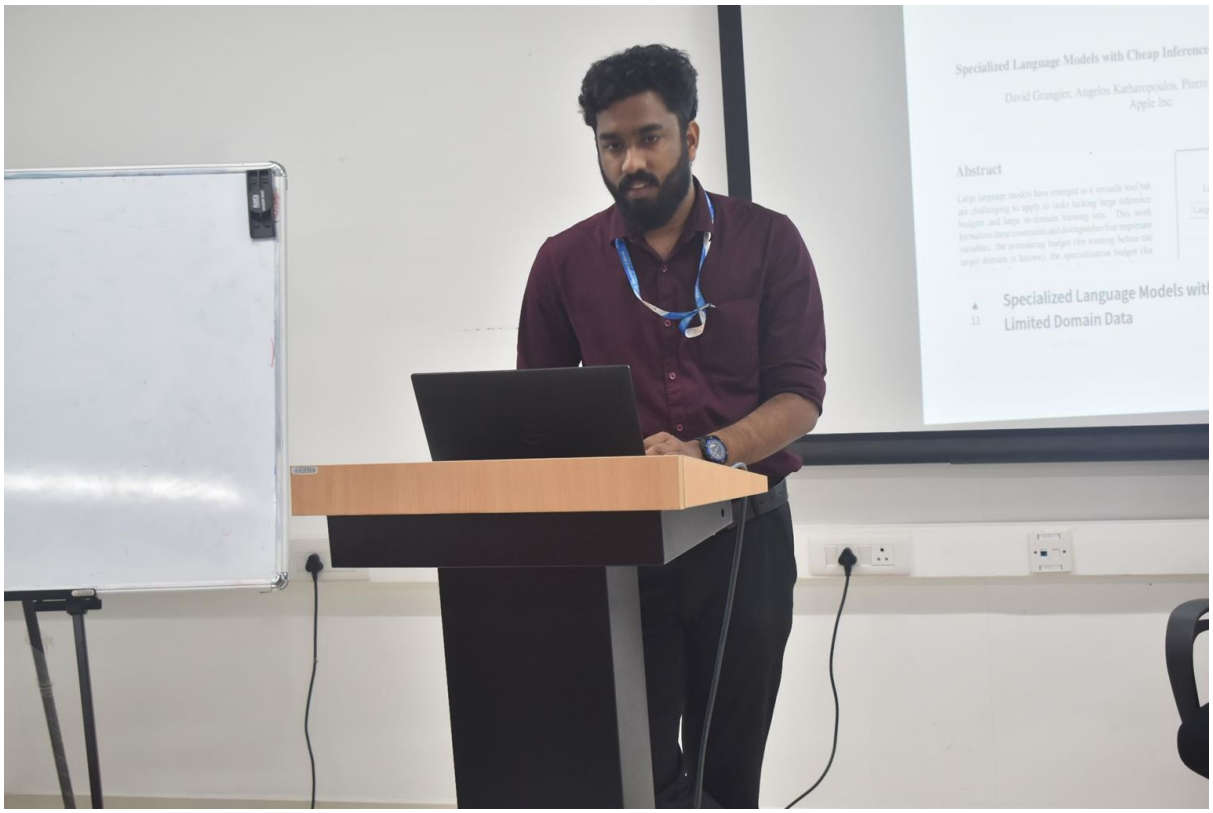
2. Transformer Architecture: BERT is built upon the Transformer architecture, comprising self-attention mechanisms and feedforward neural networks. This architecture allows BERT to model long-range dependencies efficiently and capture hierarchical representations of text.
3. Pre-training and Fine-tuning: BERT is pre-trained on large text corpora using unsupervised learning objectives such as masked language modelling and next sentence prediction. The pre-trained BERT model can then be fine-tuned on specific downstream tasks with task-specific labelled data to achieve state-of-the-art performance.

The session also covered the Hugging Face library, a popular open-source library for natural language processing tasks. Hugging Face provides a wide range of pre-trained models, including BERT, GPT, and many others, along with easy-to-use interfaces for model loading, fine-tuning, and inference.

Participants were provided with practical demonstrations of using BERT and the Hugging Face library for various NLP tasks, including:

1. Text Classification: Using pre-trained BERT models for text classification tasks, such as sentiment analysis or topic classification.
2. Named Entity Recognition (NER): Fine-tuning BERT for NER tasks to identify and classify named entities in text, such as names of people, organisations, or locations.
3. Question Answering: Using BERT for question answering tasks,

where the model is tasked with providing answers to questions based on given contexts.



The session also covered advanced features of the Hugging Face library, including:

1. Pipeline API: Hugging Face provides a pipeline API that allows users to perform various NLP tasks such as text generation, translation, and summarization with pre-trained models in a few lines of code.
2. Model Hub: Hugging Face's Model Hub provides a repository of pre-trained models and community-contributed models that can be easily accessed and used for different tasks.
3. Tokenizer: Hugging Face provides tokenizers for various pre-trained models, allowing users to tokenize text data consistently across different models and frameworks.

In conclusion, the session provided participants with a comprehensive understanding of BERT and the Hugging Face library, including their architecture, features, applications, and practical demonstrations. By grasping the fundamentals of BERT and Hugging Face, participants are better equipped to utilise these powerful tools for various NLP tasks and contribute to advancements in the field.

Day 9

Morning Session

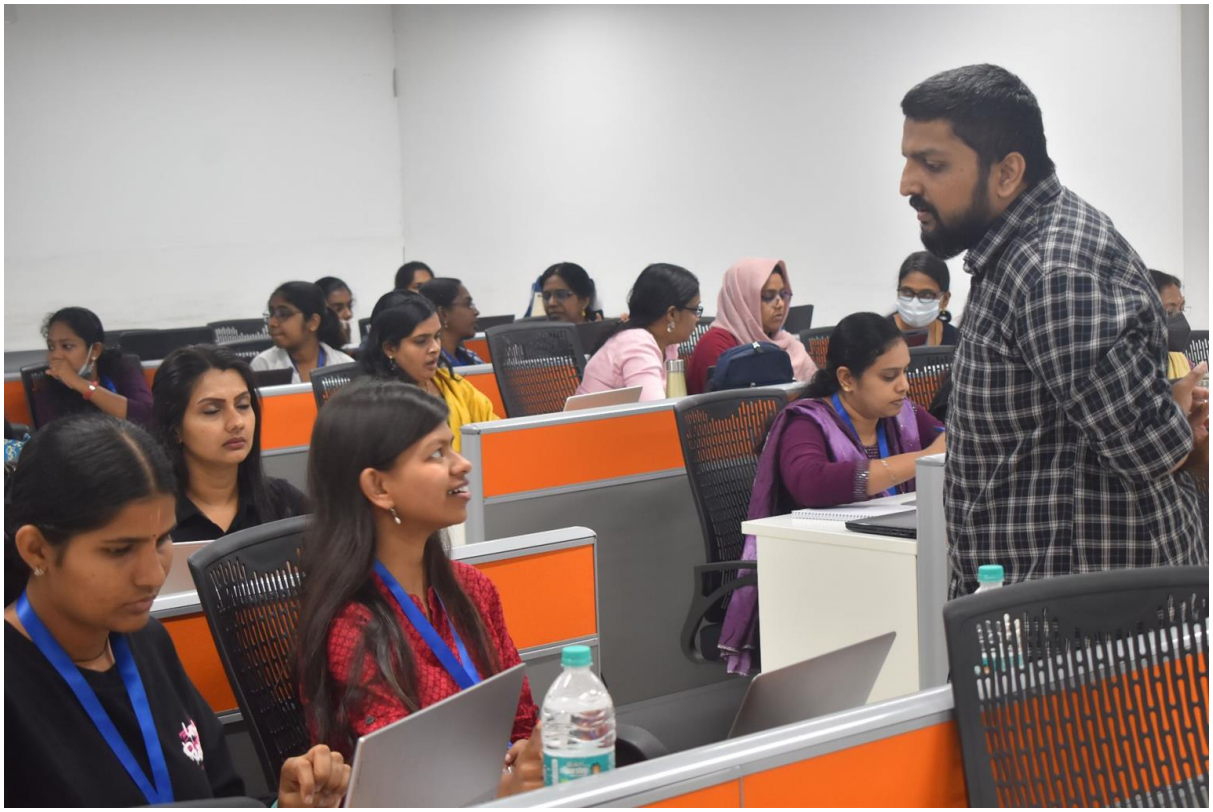
Resource Person - Mr Navaneeth S and Ms. Alaka Krishnan R U

Topic - NanoGPT and Long Chain

Time - 10 am to 1 pm

The objective of this session was to provide participants with an in-depth understanding of NanoGPT, a lightweight and efficient variant of the Generative Pre-trained Transformer (GPT) model developed by OpenAI. The session aimed to elucidate the architecture, capabilities, and potential applications of NanoGPT, as well as its significance in democratising access to powerful language models.

The session commenced with an introduction to NanoGPT, a scaled-down version of the GPT model designed for resource-constrained environments such as mobile devices, edge computing devices, and IoT devices. NanoGPT retains the core architecture and capabilities of GPT while significantly reducing the model size and computational requirements, making it suitable for deployment on devices with limited memory and processing power.



Participants were introduced to the key features of NanoGPT, including:

1. Lightweight Architecture: NanoGPT utilises a compact architecture with fewer parameters compared to its larger counterparts, enabling efficient inference on resource-constrained devices.
2. Efficient Inference: NanoGPT is optimised for inference speed and memory footprint, allowing it to generate text responses quickly and without excessive resource consumption.
3. Scalability: Despite its reduced size, NanoGPT retains scalability and can be fine-tuned on domain-specific data to adapt to specific tasks and applications.
4. Language Understanding: NanoGPT can generate coherent and contextually relevant text responses, making it suitable for applications such as chatbots, virtual assistants, and text generation tasks.

The session covered various potential applications of NanoGPT across different domains, including:

1. Conversational AI: NanoGPT can be deployed in chatbots, virtual assistants, and conversational agents to provide natural and engaging interactions with users.
2. Text Generation: NanoGPT can generate creative and contextually relevant text in applications such as content creation, storytelling, and dialogue generation.
3. Language Translation: NanoGPT can be used for real-time language translation on mobile devices, enabling offline translation capabilities without relying on cloud-based services.
4. Personalization: NanoGPT can be fine-tuned on user-specific data to provide personalised recommendations, responses, and experiences in applications such as content curation and recommendation systems.

The session also addressed challenges and considerations associated with deploying NanoGPT on resource-constrained devices, including:

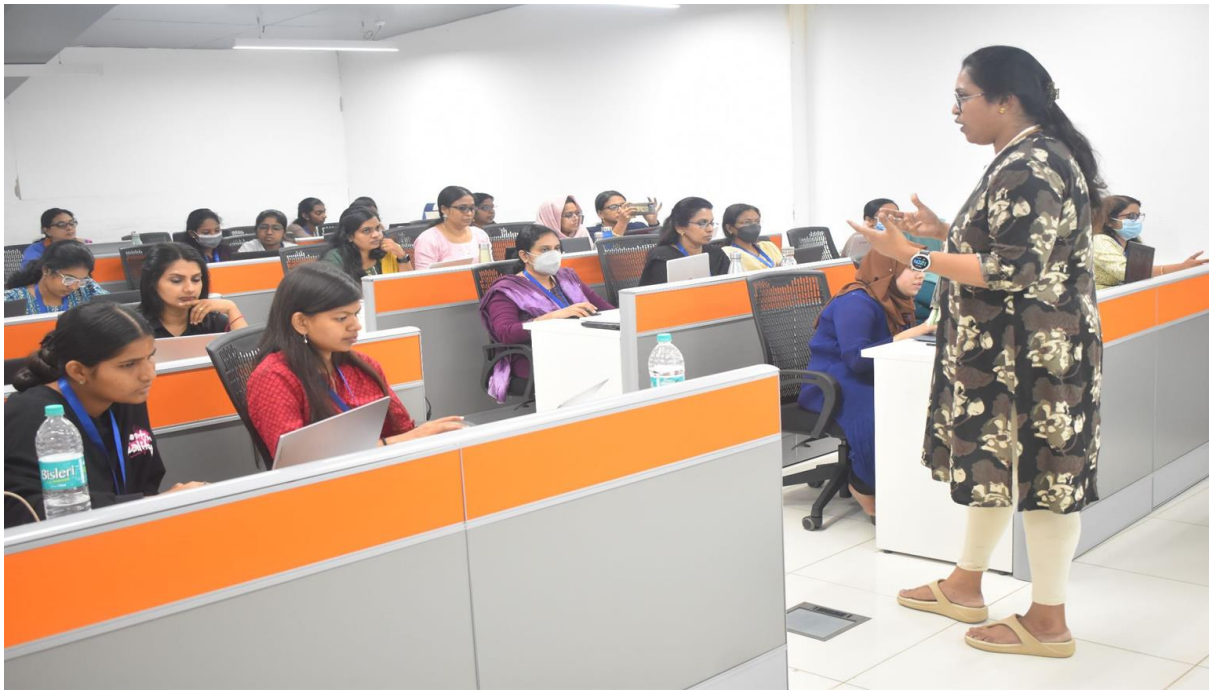
1. Model Size vs. Performance Trade-off: Balancing model size with performance and accuracy is crucial when deploying NanoGPT on devices with limited resources.
2. Optimization Techniques: Various optimization techniques such as quantization, pruning, and knowledge distillation can be employed to further reduce the size and computational requirements of NanoGPT.
3. Privacy and Security: Ensuring privacy and security of user data when deploying NanoGPT on edge devices is essential, requiring robust encryption and data protection measures.

The session provided participants with a comprehensive understanding of NanoGPT, including its architecture, features, potential applications, and challenges. By grasping the fundamentals of NanoGPT, participants got a better idea to utilise this lightweight and efficient language model for various applications on resource-constrained devices, contributing to advancements in edge computing and democratising access to AI technologies.

Langchain

The objective of this session was to provide participants with a comprehensive understanding of LangChain, a novel blockchain-based platform for decentralised language translation services. The session aimed to elucidate the architecture, principles, functionalities, and potential applications of LangChain, as well as its significance in addressing challenges in the language translation industry.

The session commenced with an introduction to LangChain, a pioneering platform that leverages blockchain technology to facilitate decentralised language translation services. LangChain aims to disrupt the traditional centralised model of language translation by providing a decentralised marketplace where translators and users can interact directly, ensuring transparency, security, and efficiency.



Participants were introduced to the key features of LangChain, including:

1. Decentralised Marketplace: LangChain operates a decentralised marketplace where translators and users can connect directly without intermediaries, enabling peer-to-peer transactions and eliminating the need for centralised translation agencies.
2. Blockchain Technology: LangChain leverages blockchain technology to record transactions, verify the authenticity of translations, and ensure data integrity and immutability.
3. Smart Contracts: LangChain utilises smart contracts to automate various aspects of the translation process, including payment, delivery, and dispute resolution, reducing reliance on third-party intermediaries.
4. Token Economy: LangChain incorporates a token economy powered by its native cryptocurrency, enabling incentivization, rewards, and micropayments within the platform.

The session covered the functionality of LangChain, including:

1. Translator Registration: Translators can register on LangChain and create profiles showcasing their skills, experience, and language pairs.
2. Translation Requests: Users can submit translation requests on LangChain, specifying the source language, target language, and desired quality level.
3. Smart Contract Execution: Smart contracts automatically match translation requests with suitable translators based on their profiles and preferences, initiate payment transactions, and facilitate delivery and acceptance of translations.
4. Quality Assurance: LangChain employs reputation systems, crowdsourced validation, and community governance mechanisms to ensure the quality and accuracy of translations.
5. Dispute Resolution: In the event of disputes or disagreements, LangChain provides mechanisms for arbitration and dispute resolution through smart contracts and decentralised governance.

The session discussed various potential applications of LangChain across different domains, including:

1. Globalisation: LangChain can facilitate cross-border communication, collaboration, and commerce by providing on-demand language translation services to individuals, businesses, and organisations worldwide.
2. Content Localization: LangChain can help content creators and publishers localise their content for global audiences, including websites, mobile apps, marketing materials, and multimedia content.

3. Legal and Regulatory Compliance: LangChain can assist legal and regulatory professionals in translating legal documents, contracts, and compliance materials accurately and securely.
4. Education and Research: LangChain can support multilingual education, academic research, and knowledge dissemination by providing access to translation services for academic publications, conference proceedings, and educational resources.

The session also addressed challenges and considerations associated with implementing LangChain, including:

1. Scalability: Ensuring scalability and performance of the LangChain platform to handle a large volume of translation requests and transactions efficiently.
2. Privacy and Confidentiality: Protecting the privacy and confidentiality of user data and sensitive information during the translation process, particularly in industries such as healthcare, finance, and legal.
3. Regulatory Compliance: Navigating regulatory requirements and legal considerations related to language translation services, including licensing, certification, and compliance with industry standards and regulations.

The session provided participants with a comprehensive understanding of LangChain, including its architecture, features, functionality, potential applications, and challenges. By grasping the fundamentals of LangChain, participants are able to appreciate its significance in revolutionising the language translation industry and explore opportunities for utilising

decentralised language translation services in their respective domains.

Afternoon Session

Resource Person - Dr Sunil TT

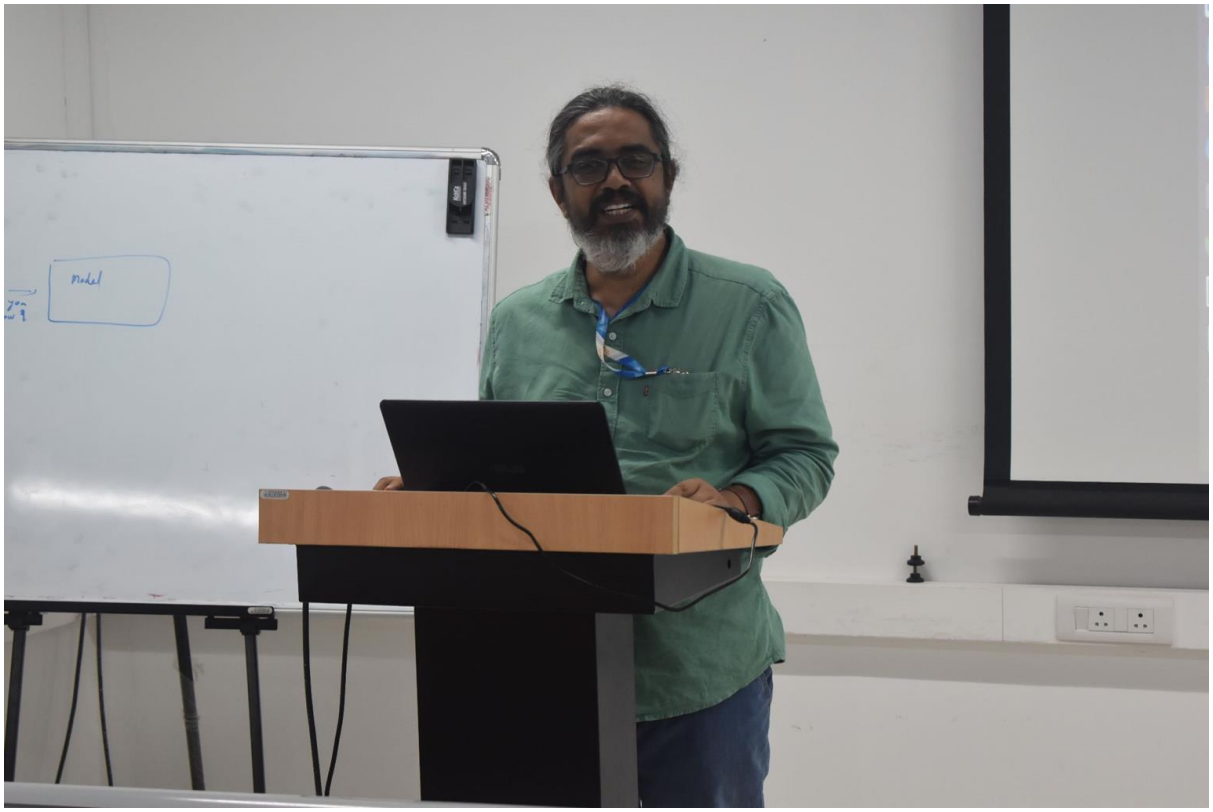
Topic - Prompt Engineering

Time - 2pm to 4pm

Dr. Sunil TT took a session on prompt engineering, focusing on the importance of writing better prompts for achieving better results in natural language processing (NLP) tasks. The session aimed to enhance participants' understanding of the art of crafting effective prompts.

The session began with an overview of the role of prompts in guiding NLP models and influencing their output. He emphasised that well-written prompts can significantly impact the performance of NLP models, leading to more accurate and relevant results. He highlighted the need for prompts to be clear, concise, and tailored to the specific task or dataset.

He provided practical tips and strategies for writing better prompts. He stressed the importance of using specific and unambiguous language, avoiding ambiguity and open-ended questions that could confuse the model. He also discussed the significance of providing context and background information in prompts to help the model better understand the task at hand.



Throughout the session, he engaged participants in interactive discussions and activities to illustrate the principles of prompt engineering. Participants were encouraged to analyse and refine prompts for various NLP tasks, gaining hands-on experience in crafting effective prompts for better results.

In conclusion, the session on prompt engineering provided participants with valuable insights and practical techniques for writing better prompts in NLP. By focusing on the art of prompt design, participants gained a deeper understanding of how to improve the performance of NLP models through thoughtful and strategic prompt engineering.

Day 10 and 11

Resource Person - Dr Premjith B

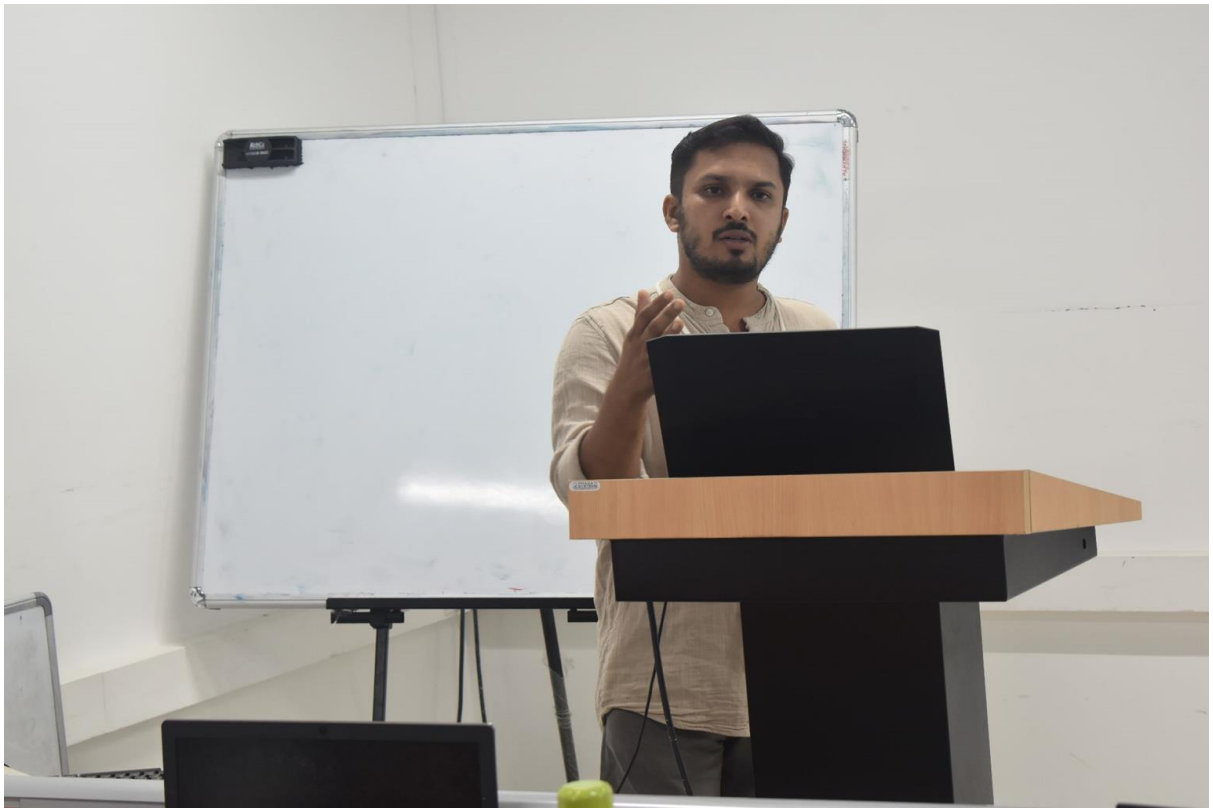
Topic - Large Language Models

Time - 10am - 5pm

Dr. Premjith B's session on Large Language Models (LLMs) provided a comprehensive overview of the evolution and applications of these models in natural language processing (NLP). The session began with a discussion on the revolution sparked by LLMs, which started with autocomplete suggestions and has since evolved into sophisticated language generation capabilities.

He explained the concept of probability and its role in language modelling. LLMs use probability to predict the likelihood of a word or phrase occurring in a given context. This is based on the idea that certain words are more likely to follow others based on the structure and patterns of natural language. He emphasised the chain rule of probability as a fundamental principle in understanding language models. The chain rule of probability is a fundamental concept in language modelling that states the probability of a sequence of events (words in this case) is the product of the probabilities of each event given the previous events in the sequence. LLMs are widely used in various applications, including machine translation, text generation, sentiment analysis, and speech recognition. They have also been applied in healthcare, finance, and other industries for tasks such as data analysis and decision-making.

He discussed the algorithms used for LLMs, including Hidden Markov Models (HMM), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Gated Recurrent Units (GRU), and the Transformer architecture.



He then explained various applications of LLMs including machine translation, text summarization, sentiment analysis, and speech recognition. They are also used in chatbots, virtual assistants, and other natural language processing tasks. Participants learned about the steps involved in fine-tuning LLMs, including preprocessing, pretraining, and fine tuning. The typical life cycle of an LLM involves preprocessing the data, pretraining the model, fine-tuning it on a specific task, and deploying it for use in real-world applications. Preprocessing involves cleaning and formatting the data for training. Pretraining involves training the model on a large dataset to learn general language patterns. Fine-tuning involves further training the model on a specific task or dataset to improve its performance.

Dr. Premjith explained the three major ways to use LLMs: prompting, encoding, and fine-tuning. Prompting involves providing the model with a prompt or starting phrase to generate text. Encoding involves using the

model to encode input text into a fixed-size vector representation. Fine-tuning involves further training the model on a specific task or dataset to improve its performance. Standard prompting involves providing the model with a single prompt to generate text. Chain of thought prompting involves providing the model with multiple prompts to guide the generation of a longer piece of text.

The session also covered prompt engineering, answer engineering, multi-prompt learning, and other prompt-based strategies to enhance the performance of LLMs. Prompt engineering enables practitioners to leverage pre-trained language models efficiently and adapt them to a wide range of downstream tasks with minimal data and computational resources. Participants were introduced to the key concepts of prompt engineering, including:

1. Prompt Design: Designing effective prompts or instructions tailored to the target task or domain is crucial for successful prompt engineering. Prompts should provide relevant context and cues to guide the model's generation or prediction process.
2. Prompt Tuning: Fine-tuning pre-trained language models using task-specific prompts and examples to adapt them to specific tasks or datasets. Prompt tuning involves optimising prompt parameters, such as length, complexity, and wording, to achieve optimal performance.
3. Zero-Shot and Few-Shot Learning: Prompt engineering enables zero-shot and few-shot learning, where models are trained to perform tasks without task-specific training data by providing relevant prompts or examples during fine-tuning.
4. Transfer Learning: Prompt engineering leverages transfer learning

principles, where knowledge learned from pre-trained language models is transferred to downstream tasks by fine-tuning with task-specific prompts.

The session covered various methodologies and techniques used in prompt engineering, including:

1. Prompt Template Design: Designing template-based prompts with placeholders for task-specific inputs or keywords, enabling flexible customization for different tasks.
2. Data Augmentation: Generating additional training examples by augmenting existing data with task-specific prompts or paraphrases, enhancing model robustness and generalisation.
3. Prompt Selection: Selecting appropriate prompts or instructions based on task requirements, domain knowledge, and model capabilities to guide the model's behaviour effectively.
4. Prompt Tuning Strategies: Optimising prompt parameters and configurations through iterative experimentation and validation to improve model performance and convergence.

He discussed various applications of prompt engineering across different NLP tasks and domains, including:

1. Text Classification: Using task-specific prompts to fine-tune language models for text classification tasks such as sentiment analysis, topic classification, and spam detection.
2. Text Generation: Designing prompts to guide language models in generating contextually relevant and coherent text for tasks such as text summarization, dialogue generation, and content creation.

3. Question Answering: Crafting prompts to extract relevant information and generate accurate answers to questions in question answering tasks, including factoid and open-domain QA.
4. Named Entity Recognition (NER): Using prompts to guide language models in identifying and classifying named entities in text, such as names of people, organisations, or locations.

The session also addressed challenges and considerations associated with prompt engineering, including:

1. Prompt Bias: Ensuring prompts are unbiased and representative of the target task or domain to avoid introducing unintended biases or prejudices into model predictions.
2. Prompt Overfitting: Preventing prompt overfitting, where models memorise task-specific prompts rather than learning generalizable patterns from data, by designing diverse and informative prompts.
3. Interpretability: Balancing model performance with interpretability by designing prompts that provide insights into model behaviour and decision-making processes, facilitating model understanding and trust.

He then explained answer engineering which involves designing prompts to elicit specific responses from the model, such as in question answering tasks. Multi-prompt learning involves training the model on multiple prompts simultaneously to improve its performance on a range of tasks. Prompt-based strategies involve using prompts to guide the model's behaviour and improve its performance on specific tasks.

He concluded by stating that LLMs have revolutionised NLP, enabling more

efficient and accurate language processing. Understanding the concepts, algorithms, and applications of LLMs is crucial for researchers and practitioners in the field. As LLMs continue to evolve, they hold immense potential for further advancements in NLP and related fields.

Day 12

Day 12 of the winter school commenced with the project presentations by the participants, which were evaluated by Dr. Anoop VS, Research Officer at the Digital University of Kerala. The project "Machine Translation (English to Malayalam)," guided by Ms. Alaka Krishnan and Athira A S, secured first place. The project "Sentiment Analysis," guided by Ms. Anitha Tilak and Desmi Davis, secured second place. The afternoon was dedicated to the valedictory ceremony, where students received certificates and gifts for their achievements and participation throughout the program.

Valedictory

The Winter School on "Natural Language Processing and Large Language Models," spanning 12 days concluded with a Valedictory Ceremony. In the morning, students presented their projects, which were evaluated by Dr. Anoop VS (Research Officer, Digital University of Kerala). The valedictory ceremony was presided over by Dr. Rajeev RR (Head e-Governance & Development, ICFOSS). The Participants shared their valuable feedback, which is instrumental in enhancing and improving the winter school for future participants. Ms. Chithra M S (Secretary & Registrar, ICFOSS) distributed certificates to the participants for their active involvement throughout the program.

The winners of the project were presented with gifts, with the project "Machine Translation (Eng to Mal)" guided by Ms. Alaka Krishnan and Athira A S securing the first place. The project "Sentiment Analysis," guided by Ms. Anitha Tilak and Desmi Davis, secured second place. Additionally, speakers, project guides, and assistants who contributed to the program's success were duly acknowledged.