Summer School-on "Natural Language Processing Applications"

Report-May-2023

Day 1

Morning session:

The Inaugural ceremony of Summer School on "Natural Language Processing Applications" commenced at 10 AM at ICFOSS. Dr Rajeev R.R(Head, e-Governance & Development, ICFOSS) gave a warm welcome speech addressing the dignitaries present on the dais and participants from various colleges of Kerala. Dr.Rajeev mentioned that the participants of this summer school consists mostly of students and researchers. He concluded his welcome speech citing that he expects this program to ignite a burning passion in researchers to continue their research in NLP.

During his Presidential address, Dr.T.K.Manoj Kumar,Director ICFOSS, stated that the applications of NLP applications in day-to-day life is increasing. He cited the influence of AI language model chatGPT.

A mesmerising Inaugural address was given by Ms. Anu Kumari IAS, Director, Kerala State IT Mission. She began by mentioning the immense impact of chatGPT and how chatGPT has improved in a short span of time. During her speech she insisted on two aspects. The first was that since technology is developing at a faster pace, we have to keep on learning new skills in order to increase our job aspects. She also insisted on making use of expansive and flexible online modes of learning. The second aspect that she specified is that the local language is dying. She also asserted never to let go of your language. She didn’t forget to appreciate the people, especially the student community of Kerala, who value their local language more than any other language.

Dr. Swaran Lata, Advisor to Meity, Department of Electronics and IT delivered the Keynote address on NLP Applications. She started her session explaining the basics of the field from the definition of NLP, NLP and Machine Learning, NLP and AI. She stated that, according to a study conducted by Association of Computing Machinery(ACM) to find out 5 areas for future researchers, the only futuristic area is Machine Learning. She suggested to the participants of Summer School that NLP with Machine learning should be their goal for research.

She explained about Natural Language Understanding(NLU), Dialogue Management and Natural Language Generation. She made reference to automatic speech recognition(ASR), Text- To-Speech(TTS). She touched on the topics Named Entity Recognition(NER), Parts of Speech Tagging(POS), Text Categorization, Syntactic Parsing, Conference Resolution and Machine Translation in NLP; as well as Relation Extraction, Semantic Parsing, Question

Answering(QA), Sentiment Analysis, Summarization, Dialogue Agents, Paraphrase & Natural Language Inference in NLU. She explained various elements of NLP like Machine translation, information retrieval, Sentiment Analysis, Information extraction and question answering systems.

Dr.Swaran Lata instantiated Q&A systems for domain specific areas. She explained Cross-Lingual IR approaches and steps for reducing CLIR into monolingual tasks(from query processing to snippets delivery). She detailed about Large Language Models like chatGPT during her speech. She discussed the typical NLP pipeline that flows through Sentence Segmentation, Word Tokenization, Stemming, Lemmatization,Stop-word Analysis, Dependency Parsing, Parts-of-Speech Tagging and Named Entity Recognition. She told the participants that they will be taught to how to do NLP using this pipeline in the coming sessions.

She briefed about NLP approaches with various Machine Translation Systems like Rule based MT, Corpus based MT and Hybrid MT. She also added the neural machine translation where words are transcribed in to vectors, each with a unique magnitude and direction, in a process of encoding and decoding. Next she explained the Translation Project management aspects like specifications phase, production phase and post project view. She cited that Effective Communication between the requester and the project manager is imperative to the success of a project. She also cited the quote ‘Know Safety, no pain’ in translation Project Management.

Dr.Swaran Lata also discussed the various limitations of Neural machine translation like Accuracy, Expertise, Format, Lexical Robustness (context based),Black Box approach, Requirement of Regular Training on new data and Computational complexity -being some in her list. She added that to leverage AI in NLP, proper expertise is required. Model optimization is also another requirement in NLP.

She mentioned various MT efforts of Indian languages including Siva, Shakthi,MANTRA,English Kannada MTS, Sampark. In her session she elicited the deep learning, ensemble models,Unsupervised machine learning and Supervised / Predictive machine learning and she added that supervised models is better in most cases. Then she analysed the evolution of large language models and how it led to creating personalised chat bots that can help businesses.

Some AI tools in use prompt engineering and artificial general intelligence like hugging chat and Agent GPT were discussed. Various research fields like Text processing tools, Cross Lingual Search, Speech to Speech translation were explained to the participants. Dr. Lata also explained the speech to speech pipeline stages like speech processing using Automatic speech recognition. She concluded her session by mentioning the importance of AI and why we can't pause AI, even though various tech giants like Elon Musk and Steve Wosniak calls to pause any further development posing profound risks to society and humanity.

Smt Chithra M.S. Secretary to ICFOSS, felicitated the event by wishing success to all participants in the summer school programme. Smt. Mehrunisa Nazeem, Research Associate, ICFOSS offered the vote of thanks by thanking all invitees and participants for gracing the occasion by their solemn presence. She also thanked IT Mission for their invaluable cooperation for organising this programme. The inaugural ceremony concluded by 11.45 AM.

After a short tea break, the summer school session started with 30 participants and speakers of various sessions introducing themselves.

Afternoon Session

Resource Person	**Dr. RAJEEV R. R, Head (e-Governance & Development,ICFOSS)**
Topic	**Introduction to Natural Language Processing (NLP)**

The session started at 2:30 pm. Dr. Rajeev R.R. started the session by mentioning that understanding of the research problem and language features is essential before conducting a research in NLP. He then discussed the difficulties in creating a chat gpt-like AI model for malayalam language because of the structure, ambiguity and context of Malayalam language. He also discussed the peculiarities of the family of languages, stating Dravidian languages and concluded that the transliteration is difficult even among different languages of the same family.

Dr.Rajeev handled an interactive session on the “significance of NLP” mentioning first language acquisition, environment dependency and its relation to an individual’s thinking ability and made the session more interesting by citing the story of Birbal. He then discussed the significance of malayalam language computing specifying the agglutinative nature in word formation and difficult character formats. He quoted the works of R.E. Asher, Herman Gundert, A.R. Rajaraja Varma and Dr.Ravishankar while mentioning the scarcity of works in Malayalam syntax and semantics.

The session included the description of definitions of basic terms -natural language, language technology, natural language processing and computational linguistics. Dr. Rajeev pointed out Rule-based NLP and Unstructured NLP citing examples and explained Human aided Machine Translation, Hybrid Machine translation and prediction based on language models like Hidden Markov Models. He made the audience ponder over the challenge of how to perform natural language processing without compromising the ‘naturalness’. He referenced the ASCII code for English and unicode for malayalam language during the lecture.

The session began after the evening tea-break at 4.30 p.m. with the topic ‘where does NLP fit in the CS taxonomy’. The session had a discussion on topics - relation of language technology with multimedia & multimodality technologies, knowledge technologies, speech technologies and text technology. Dr. Rajeev briefed the topics- computational linguistics versus natural language processing and the relation of computational linguistics to other disciplines like Machine learning, Human Computer Interaction, Information Retrieval, Theory of Computation, Psychology, Philosophy of Language, Linguistics, Electrical Engineering (Optical Character Recognition) etc. . He explained about the linguistic levels of analysis pertaining to speech, written languages and gestures. He explained from top to bottom about phonology, morphology, syntax and semantics of written language. He minutely detailed the ‘morphophonemic change’ and ‘morphosyntactic change’ of malayalam language. The intriguing session ended at 5.20 p.m. with a discussion on morphological analyzer, number markers and gender neutrality of words in malayalam language with Dr. Rajeev referencing his publication work on morphological analyzer.

Day 2

Morning Session

Resource person: Ms Sruthi Sara Moses, Intern at ICFOSS

Topic: Linguistic Aspects of NLP

With a view of creating a basic idea of Linguistics, Ms. Sruthi started the session at 10 am. She discussed the three aspects of Linguistics: Phonology, Morphology and Syntax and explained each aspect along with how it is incorporated in Natural Language Processing. The Phonetics/ Phonological aspect of linguistics is concerned only if the text origin is a speech. In the case of Morphology, she clearly described the types of morphemes and challenges faced while doing Morphological analysis which are the inflectional,derivational and agglutinative nature of language. After making a clear understanding of linguistic aspects, she further discussed the NLTK library. Then she explained the text pre-processing steps, which are Tokenization, Text removal, Stemming, lemmatization and Parts of speech tagging(pos tagging). She further described the next aspect of linguistics which is syntax and went on explaining syntactic analysis along with the difference between Chunking and Parsing. She defined chunking, which is used to identify parts of speech and short phrases present in a given sentence. And parsing,which is the process of analysing the strings of symbols in natural language confirming to the rules of formal grammar. The class clearly dealt with types of parsers and emphasised on Regexp parser,which is the most commonly used parsing technique.

The theoretical session came to an end around 11: 45, after the tea break class resumed for hands on session.

Afternoon Session

Resource person: Dr. Prajeesha A K and Mr. Selvaraj R

Topic :Word Sense Disambiguation by Selvaraj and Hands on session by Dr. Prajeesha A K

Afternoon session started with Mr.Selvaraj dealing with the topic Word Sense Disambiguation. This session dealt with yet another aspect of Linguistics which is Semantics. After creating a basic idea of semantics that is the study of meaning of words and sentences, he explained what ambiguity in a sentence means. He discussed the three types of ambiguity; Lexical ambiguity, Structural ambiguity and Referential ambiguity. The session got interesting when each type was explained with examples and each of the participants were told to write examples of their own. Some of them even shared their examples with the whole class. Then the session entered to the main topic, ‘Word Sense Disambiguation’ is a key enabling technology that automatically chooses the intended sense of a word in context. The computational identification of the correct sense of a word in a context is known as Word Sense Disambiguation (WSD). It’s a machine learning technique which is used to teach a machine what ambiguity is. The WSD approaches can be grouped into two main categories: methods based on supervised machine learning (supervised methods) and knowledge- based methods (unsupervised methods).The supervised approach is based on trained sense annotated corpus to build classifiers. Initially, annotated corpus is required to build a classifier. Supervised machine learning algorithms are “supervised” in that they receive labelled training data and go through a training process. Later, he described the four main approaches to supervised WSD and went on to unsupervised approach.The Unsupervised approach is based on unannotated corpora. It is based on clustering of words.

Day 3
Morning Session
Resource person: Dr. Umesh P.
Topic: Introduction to ML Linear Algebra and Review - Probability, and Statistics, or Data Analysis

The session started at 10.am. He has started from the very beginning of Mathematics behind ML concepts. Different areas of Mathematics that are used in Data science and ML are calculus, probability, statistics, and linear algebra. Linear algebra deals with lines and all problems that can be solved with lines.

Different distance measures in different problems. Euclidean, Manhattan, Minkowski, Hamming distance etc. Hamming distance is mainly used in NLP to find the word distance.

Matrix is the key factor behind the data transformations. The transformation output will purely depend on the values of the matrix used. Determinant of the matrix is the measure by how much a matrix can be scaled or it determines the value of transformation. Determinant 1 means no change.

Eigenvalues are a concept in linear algebra that are used to describe the behaviour of linear transformations. Specifically, given a square matrix A, an eigenvalue λ is a scalar value such that there exists a nonzero vector x (called an eigenvector) that satisfies the equation $Ax = \lambda x$. Eigen vectors are those values that can be used as characteristic of some vector.

GeoGebra is a free and open-source software that combines geometry, algebra, and calculus in one easy-to-use package. It is available on multiple platforms, including desktop, tablet, and mobile devices.

Singular value decomposition(SVD): SVD is a powerful technique that is widely used in a variety of applications, including image compression, data analysis, and machine learning. It can be used to reduce or increase the dimensionality of data, to identify patterns in data, and to solve linear equations.

After explaining all the relevant theoretical aspects, he did the hands-on session using Google Colab regarding Linear Algebra problems. It includes how to find Eigenvalues, Eigen vectors, Eigen decomposition, SVD and practical application of SVD. He clearly explained how an image can be reconstructed using SVD and gave an example. At the last, described Principal component Analysis(PCA) for dimensionality reduction. The PCA can be done from a covariance matrix by selecting highest values of Eigenvalues with their Eigenvectors.

Afternoon Session

Resource Person : Dr. Umesh. P.

Calculus

Note : Use GeoGEBRA for Graphical Representation

Explained to draw a graph of f(x) = x^2 and f(x) = x^3 using GeoGEBRA.

Derivative and Slope

The slope of a curve is very important in calculus. Slope of a curve is not constant and can vary from point to point.The slope is defined as the derivative of the curve at that point. The next topic was *Derivative and Gradient* in which he explained about the basic equation that describes the definition of gradient and multiplication of a scalar and vector constants. Then explained *Higherorder Derivatives* (Second Order Derivatives) that is an important concept in calculus.

Taylor Series

It allows us to approximate the behaviour of a function.

Note : Derivative of e^x is e^x itself, because the slope of e^x is always e^x.

Partial Derivatives

It is used to compute the derivative of a multivariate function.

Normalization

Normalization involves choosing appropriate scaling factors. Explained to perform normalisation on constant values.

Next, he took the *Hessian Matrix* in which he explained how to calculate the second order derivatives of partial derivatives. Finally took optimization of objective function.

Probability and Statistics: In statistics he took a session on summary statistics that includes Mean, Median, Mode, Variance, Covariance, Correlation Coefficient, Standard Deviation, Univariate, Bivariate, and Trivariate.

Gave a brief information about *conditional probability* and explained a problem on it.

Finally he took a session on **Machine Learning** in which he covered ML Model Life cycle, Linear Regression, Loss Function, Cost Function, Recall in Linear Regression, Logistic Regression, Classification (Linear Classifier).

Day 4
Morning Session
Resource person: Dr. Shabina Bhaskar 11th May 2023

Resource Person	**Dr. Shabina Bhaskar, Mrs.Saleena T S**
Topic	**Data Analysis using Neural Networks**
Time	**10.00am -1.00pm**

This session deals with NLP basics starts with showing a NLP flowchart and further followed by NLP techniques.Speech and text are different spaces of NLP Show a briefing of text by using NLP techniques.Syntatic parsing is a stage i that include stemming, Lemmatization and stop word removal.

Text representation.
Text represent as numerical form different method Bag of words,TF-IDF,N-grams. In N grams probability of next coming word. Further discussed about NLP tasks such as Chat-bots,Machine translation.

Modern NLP
Modern NLP uses word embeddings. Showed different NLTK like libraries like pytorch-NLP,CoreNLP. Now discuss about corpus creation in NLP: A corpus is collection of authentic text or audio organised into datasets,corpus can be made from newspaper,articles etc.

Application of NLP
Text classification,Language modelling
Text classification refers as categories document into different
eg:Newsarticle into sports,email spam classification,sentiment analysis etc

Image captioning
Task of generating a textual description for a given image

Coco dataset,Flickr8k
Speech recognition
Timit acoustic-phonetic continuous speech corpus
Vox forge open source data for speech recognition
https://nlpforhackers.io/corpora/ :Referance more corpus in NLP
Further discussed about webscraping by telling its use case like Twitter sentiment data.
To check whether a website allow webscrapping by looking at robots.txt.
Ways to gather webscraping data is by using Api and by scrapping libraries.
Next further discuss about pypdf2 library for pdf reading and NLTK library operations.
After that showed how to webscrap youtube comments by digitalmethods.net web based tools and discussed current scenario of twitter like websites provide api for data request as pay subscription.Showed demo of apify.com to retrieve data of websites which includes a pay as you use function.

The rest of the session has been handled by Mrs.Saleena T.S regarding the Text Analytics tools, especially TextHero. Before starting the main session she tell about difference between traditional system and machine learning system and has covered all the steps involved in an NLP project. The pre-processing has been explained in detail with steps involved in it, data cleaning, data integration, data transformation, data reduction (dimensionality reduction) and data discretization. After that explained the difference between Text Analysis and Text Analytics. Then a hands-on session explaining how NLTK can be used on a .csv file and how different preprocessing like stop word removal, missing data handling, stemming, data cleaning, punctuation digits removal and vectorization using TF IDF can be done on the data. Then as the last topic, TextHero has introduced and done a hands-on session on that also. All the pre-processing techniques mentioned above can be done very easily in this without using many lines of code. That library contains all the functions that can be used to do the above mentioned works. Apart from that it can visualise the processed data using its own function without using other libraries like ‘matplotlib’.

Afternoon Session

Resource Person	**Alaka Krishnan R U**
Topic	**Machine learning in NLP**
Time	**2.00pm -5.00pm**

Started session with purpose of language as a wonderful way of communication,and concept of human learn from past experience in relation with artificial intelligence , in mimics human behaviour,how AI,ML,deep learning related.Explained about relation between machine learning,Artificial intelligence and deep learning in a descriptive way.Next discussed regarding NLP,NLU AND NLG and how are they related Natural language understanding helps machine to understand data,meaning of data processed accordingly,solves understanding context,semantics of data Natural language generation and application of

natural language processing include email filtering,chatbots,autocomplete in search engines,Language translation etc. Levels of natural language processing Phonology,Morphology,Lexical,Syntactic,Semantic,Discourse,Pragmatic
Phonology:interpreting speech sounds Morphology:interpreting componential nature of words

Steps in making a model
Preprocessing L2-Bag of words,TF-IDF,Unigram,Bigram
Text preprocessing-Gensim,Word2vec
Natural language processing techniques
Bag of words,TF-IDF,Tokenization,stop words removal,stemming,lemmatization,Topic Modeling-Consider data and use linear regression like algorithms,word embeddings.
Describe definition of word standalone ,carries a meaning
Next define Morphemes with examples
Bag of Words:Technique of preprocessing text by converting it into a number/vector format
Built vocabulary
TF-IDF
One of fundamental tasks in natural language processing
Term frequency*inverse document frequency
tf-idf(t,d) = tf(t,d)*log(N/(df+1)
Bag of words and TF-IDF problems
Both do not store semantic information
Word embeddings:Word2vec is common method of generating word embedding has variety of application text similarity,recommendation systems
N-gram models:
Is a statistical language model works based on probability of occurrence of next word.
Next discussed machine learning basics and algorthims in a brief way.Discussed about different machine learning algorithms focused more on unsupervised learning algorithms like Support vector machines, Decision tree and Random forest.

Day 5

Morning Session

Resource Person	**Dr. Gopakumar G**
Topic	**Data Analysis using Neural Networks**
Time	**10.00am -1.00pm**

Started the session with the importance of Data,Types of data, Entities and relations and Graph.

Discussed basics of Machine learning which include importance of features,classification and regression,overfitting and underfitting, Polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression. It is a linear model with some modification in order to increase the accuracy. Where degree is 0=it is a constant,1=a straight line ,2- it is a parabola, 3-cubic….nth dimension=10 is a hyperplane.

Statistical Learning

Statistical learning theory deals with the statistical inference problem of finding a predictive function based on data.

Parametric and structured models

Eg;-how to calculate the rent of a house from features,to fit the model by training data.

β1*size+ β2*Furnished β3*location….+ β0

Understand the important features to predict the outcome.

Spline- Larger degree curve.How to find the Model accuracy,MSE ,Residual Sum of Squares (RSS) value should be less,**Indicator Function** ,Hypothesis testing

Assessing the accuracy of the coefficient Estimates

Bayes theorem for classification

Bayes' Theorem calculates the conditional probability of an event, based on the values of specific related known probabilities.

Cross-validation and the Bootstrap and concluded Training versus test-Set Performance to explain overfitting and underfitting .

K-fold cross-validation (widely used approach for estimating test error).

Cross-validation is **a resampling method** that uses different portions of the data to test and train a model on different iterations.

Afternoon Session

Resource Person	**Dr. Gopakumar G(contin.)**
Topic	**Neural network**
Time	**2.00pm-5.00pm**

Architecture of Neural network

A neural network works by taking in a set of inputs, which are then processed through a series of hidden layers to produce an output. Each neuron in the network receives inputs from the previous layer and applies a mathematical function to them before passing them on to the next layer.

Perceptron, weight, bias, activation function , the problem that happens when classifying data in a hyperplane and the role of neural networks in such problems.

Input Layer- Original features

The perceptron consists of a set of input values, each of which is multiplied by a weight and then summed together. This sum is then passed through an **activation function,** which produces the output of the perceptron. The output can be either 0 or 1, depending on whether the activation function threshold is crossed or not.

Loss Function

Loss is the difference between actual and predicted output that represents the error. States local minima and Global minima. Gradient Descent and showcasing the landscape of errors.
Layers of neural networks input,hidden,output layers are here and finalise the number of hidden layers depending on trial and error or cross validation method for making a model.How to minimise the loss value by updating the weights in backpropagation.
Epoch - Epoch is completed when the neural network has seen and processed all training examples at once which completed both forward and backward propagation ones.
During Network Learning, training Error J(W)=1/2sum(tk-zk) from first to the all output c.
Back Propagation learning rule which is based on gradient descent.
Given detailed description of weight update or learning rule for the hidden to output weights.
Importance of **Activation Functions** and categories.
Detailed explanation in all mathematical aspects for data analysis and neural networks.

Day 6
Morning Session

Resource Person	**Dr. K Satheesh Kumar**
Topic	**Automatic Knowledge Discovery**
Time	**10.00am -1.00pm**

Today’s session is handled by Dr. K Satheesh Kumar, Kerala University and the topic is Automatic Knowledge Discovery.

Previously research work was published as printed form and nowadays all the materials are available in computer readable form. Today, he has discussed the information extraction from text format. Human beings can understand the information from paragraphs but computers can't. Two methods he explained.

1. Word to vector

2. Knowledge graph

This method can be applied in any area like computer science, mathematics, social science etc. There is one research paper entitled “PaperRobot: Incremental Draft Generation of Scientific Ideas” discussing how to write a research paper by a machine.

Large Language Models (LLM)-Ask anything it will write. Why is it called Natural Language-same sentence can be expressed in different ways.

Ethical issues are one important criteria in this type of research. Now google announced before any conversation the machine has to tell that I am a robot. Ethics associated with AI is now an important research area. GPT ZERO is a model introduced by a student discussed about How to identify human writing or Machine writing. Complex sentences and simple sentences are mixed in human writing. But text uniformity is the same in LLM models.

By the emergence of AI, automated knowledge discovery became conceivable. DENDRAL(1960) a system analysing mass spectrometry data to discover molecular structures is introduced at the initial stage of AI. Mycin(1970) is used to diagnose bacterial infection. All these systems come under expert systems. An expert system needs knowledge, reasoner and a user interface. Two types of expert systems -Rule based, logic based expert systems, case-based, Model based, fuzzy based and Bayesian reasoner. BACON another system for rediscovering scientific laws. LISP, in full list processing, a computer programming language developed about 1960 by John McCarthy.

Three types of proving method

Empirical proof-Experimental proof

Mathematics proof

Statistical proof- hypothesis based proof

-Automated Mathematician (1977) is developed using LISP programming.

Big data and semantics was introduced in 2000 and the development of the semantic wed has led to the new approach of knowledge representation and reasoning. Semantic web is a vision for linking data across web pages, applications and files. First convert to Resource document framework (RDF). The basic concept behind this is every sentence expressed in subject-predicate-object order. Web ontology language (WEL) is based on First order logic. Example All men are mortal is represented by For every x man(x) implies mortal (x).

Word2vec

word2vec was created , patented and published in 2013 by google.We are converting words to vectors because if we vectorized words then we can add two words. The two papers “Efficient estimation of word representations in vector space and Distributed representations of words and phrases and their compositionality” discuss word2vec. Material informatics. Word2vec is a technique for NLP published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonym words or suggest additional words for a partial sentence.

Knowledge graph

Introduced by Google in 2012. The heart of the knowledge graph is a knowledge model: a collection of interlinked descriptions of concepts, entities, relationships and events. Knowledge graphs put data in context via linking and semantic metadata and this way provide a framework for data integration, unification, analytics and sharing. Drug Repurposing- one medicine can be used for other purposes- can be done with the help of a knowledge graph.

Applications of knowledge graph

1. Patent database

2. Cybersecurity

3. Multilingual knowledge graph

4. Agricultural knowledge graph.

5. Design and implementation of curriculum systems based on knowledge graphs.

Category theory: Category theory is a general theory of mathematical structures and their relations that was introduced by Samuel Eilenberg and Saunders Mac Lane in the middle of the 20th century in their foundational work on algebraic topology. Road network: an example of category theory.

Ontology Graph (Olog) :An ontology is a description of data structure–of classes, properties, and relationships in a domain of knowledge. It is meant to serve as a basis for instances of knowledge graphs, ensuring data consistency and understanding of the data model.

After noon session

Resource Person	**Dr. K Satheesh Kumar**
Topic	**Automatic Knowledge Discovery (contd)**
Time	**2.00pm -1.00pm**

This session starts with word embeddings which converts words in vector space One way to do this is by one hot representation.
Co occurrence matrix in this discuss about a matrix approach in showing occurrence of words in sentences.
Co-occurance with svd(singular value decomposition) Its expensive and not scalable With neural network
Explains about a basic neural network which is feedforward neural network,it can be used for word embeddings
Word2vec,cbow,skip-gram

CBOW continuous bag of words ,connect words in a particular manner

Skip-gram another method of word embedding

Introduced glove python library and following that introduced web articles to refer NLP basics based on these libraries. For reference showed some blogs from towards data science about glove library ,also discussed an implementation of glove from github.

Discussed spaccy library for building knowledge graph.Knowledge graph is representation using relation of words .

Showed how to use google scholar in taking more research content in word2vec.Introduced knowledge engineering,BERT like terms to students.The session end with how to do fine tune skills to research and further studies in NLP.The vote of thanks of session done by Mr.Rajeev by giving memento .

Day 7

Morning Session

Resource Person	**Ms.Dhanya L.K.**
Topic	**Deep Learning in NLP**
Time	**10.00am -1.00pm**

This session starts with Chat generative pre-trained transformer, then discussed about NLP - Blending of cs and linguistics and about two types of NLP - NLU and NLG Deep learning Vs Machine learning : point out about image processing, recommendation system and then how to develop a NLP System : Dataset , preprocessing dataset, word embedding (Vector), Model creation (Mention about Explainable AI, Tools : LIME , SHARP). Resources of NLP(Dataset) : TDIL, Tree bank ,NLTK, Penn university.

Structured data/Unstructured data
To handle unstructured dataset : SMOT Algorithm Take f1 Score.
Detailing the theorem Cohen's Kappa-Statistical measure that is used to measure the reliability of two raters for rating to identify how frequently the raters are in agreement.
Shown an example of Hate Speech Recognition.

Word Embedding
For representing words input to the encoder.It is a vector representation of words.
How to input text data to the neural network and how it works till to reach the output.
Describe the NN , starting with simple neural network(ANN) how it converges to deep neural networks by showing Multiple hidden layers architecture. Explain weights, perceptron,output,activation function,Loss etc.

Objective of the deep learning algorithm is to minimise the loss value.Loss is the difference between actual and predicted value in each epoch. How Forward Propagation and Backward propagation make it Loss function lesser.

It should be important to add an Embedding Layer in multilayers NN.
TF-IDF- Term Frequency Inverse Document Frequency
It is an ML algorithm based on a statistical measure of find the representation of a word

Word2Vec
Consider sentence by sentence
CBOW and Skip Gram are the two methods used in this

GloVe is another kind **Embedding mechanism.**

FastText

OOV(out of vocabulary words) which do not occur while training the data are not present in the model's vocabulary ,i.e; can find the word embeddings that are not present at the time of training.Word2Vec and Glove is differentiate from FastText in the same.

BERT

ChatGpt works on Bert. It is a deeply Bi-directional for word embedding as well as classification.It learns information from both the left and right side of a token's context during the training phase.Multilingual Bert is good for dravidian language and Glove and word2vec is not that much good.

Problems:

Sequence labelling-POS Tagging,NER, **Classification**:-Sentiment Analysis,**Sequence Transformation:-** Machine Translation,Question Answering,Summarization

RNN

Feed forward neural network with no loops.Dependency between the words in the text while making predictions.

$$\mathbf{ht}{=}f\left(Xt-ht-1\right)$$

Input,output,hidden,context layers are there in RNN.Vanishing Gradient and Exploding Gradient Problem are the two problems in RNN.

Vanishing gradient problems-They start to forget about the previous data they have seen. So to resolve it, we need to use Long memory,using the ReLu function(rectified Linear unit)Which takes the value max(0,x) which leads to the Exploding **Gradient problem.**

LSTM(Long Short term Memory)

Special ANN which needs memory.First step to decide how much past that should remember. Forget Gate Layer is the most important one.In Step2 decide how much this unit adds to the current state.Input gate layer working on it.Finally Decide which part of the current cell state contains Output layer.

Transfer Learning

Transfer learning is a machine learning technique where knowledge gained from training a model on one task is applied to another related task. Rather than training a model from scratch on a new task, transfer learning leverages the pre-trained knowledge from a different task to enhance learning and improve performance on the new task.

NLP libraries WordNet,Nltk

Afternoon Session

RESOURCE PERSON	**Dr.RENU S**
TOPIC	**Text-to-Indian Sign Language (ISL) Translation**

Time	**2.00pm-5.00pm**

SIGN LANGUAGE

It is the communication medium for hearing impaired people. Using facial expressions, body movements, and gestures, sign languages provide a three-dimensional representation of thoughts and feelings.

Variants of Sign Language

American Sign Language (ASL), British Sign Language (BSL), Argentinian Sign Language (LSA), Indian Sign Language (ISL)

Sign Translation Research

Two Vertices of Sign Language Translation research:

Sign-to-Text (Video to Text\Speech) - converting a video image or a dynamic video sequence, then finding the appropriate text or speech.

Text- to-Sign - Translating text to its corresponding sign language.

Sign-To-Text

It focuses only on finger movements

1. Glove Based
2. Video\Image Processing

Text-To-Sign

1. Notation
2. Avatar Videos
3. Sigml player

Indian Sign Language

Although it is not widely recognised, Indian Sign Language is the preferred indication language in India.ISL is a complete language with defined grammar rules.

Text-to-Indian Sign Language Translation System

Text Preprocessing

Unlike English and other Indian languages, Indian Sign Language has its own set of grammar rules.

The three basic processes used to translate text into its ISL are phrase reordering, stemming, and elimination.

Elimination

Eliminating unused words, connection words, etc. are performed in this stage.

Never use linking verbs (am, is, are,etc) and gerunds (-ing).

Input text

After Elimination

I am Renu

Me\I Renu

The rose is red

Rose Red

He is a teacher

He teacher

We are friends

We friends

Stemming

ISL always uses root words and stemming means converting words into their root form.

Input text

After Stemming

Ramu is running

Ramu run

She is singing

She sing

They played Cricket

They play Cricket

She wrote a book

She write book

Phrase Reordering

Phrase reordering is a common format in ISL.

'adjective+noun' combination it will be signed as 'noun+adjective '

Input text-

After Reordering

Red Rose

Rose Red

Beautiful Girl

Girl Beautiful

WH- questions are always at the end.

Phrase Reordering

Input text-

After Reordering

What is your name?

Your name what?

What is the time?

Time what?

'NOT' is always at the end.

Phrase Reordering

Input text

After Reordering

I don't have any children

I children no

I don't know how to cook.

cook know not

ISL text follows ‘subject +object+verb ’ structure.

Preprocessing Challenges

1. Dinner, for instance, is indicated as "Night" and "Food".
2. ISL also makes use of the sentence's context.
3. For example the sentence "He listened to what I said” is signed as “He listens to me”.

Notations System

As a visual-spatial language, sign language is not capable of writing like other spoken languages.A textual representation of sign language has been established by scholars who are passionate about sign translation. Notations were used to describe these representations. The population of hearing-impaired people is unfamiliar with the notation systems.

These notation systems help to represent a 3D language in a written format.
Some of the commonly available notation systems are:

- Bebian Notation,
- Stokoe Notation
- Gloss Notation
- Hamburg Notation System (HamNoSys)
- SignWriting (SW)
- Si5s
- SignFont
- SignScript
- SLIPA

A standard phonetic notation method was required for the text to Indian Sign Language translation system in order to translate a three-dimensional language into a two-dimensional space.

HamNoSys Notation System

In order to translate a text into its equivalent visuals, the Institute of German Sign Language at the University of Hamburg created a notation system in 1980 that incorporates more than 200 symbols.
The HamNo notation has six parts. The first two parts Symmetry operator and Non-manual Features(NMF)are the optional features.

Hand-held ISL expressions can be divided into four.

- hand shapes
- hand orientation
- hand location
- hand movement.

BASIC HAND SHAPES

- hamfist
- hamflathand
- hamfinger2
- hamfinger23
- hamfinger23spread
- hamfinger2345
- hampinch12
- hampinchall
- hampinch12open
- hamcee12
- hamceeall
- Hamceeopen

To adapt and create additional symbols, such as basic hand forms, diacritics, symbols for fingers, and finger bends, the hand shape can be separated into four main categories.All the hand shapes are a combination of these different

Basic Hand shapes- Diacritics

- Hamfingerstraightmod
- hamthumboutmod
- hamthumbacrossmod

- hamthumbopenmod

Basic Hand shapes- Finger bend

- Hamdoublehooked
- Hamfingerbendmod
- Hamfingerhookmod
- Hamdoublebent

Basic Hand shapes- Symbol for Finger

- Hampinky
- Hamthumb
- Hamindexfinger
- Hammiddlefinger
- Hamringfinger

Basic Hand shapes- Finger Parts

- Hamfingerside
- Hamfingertip
- Hamfingernail
- Hamfingerpad
- Hamfingermidjoint
- Hamfingerbase

Direction of Forefinger

The direction of the extended fingers points in the same general direction as the vector that extends from the wrist along the back of the hand and continues in that same general direction.

HamNoSys has 26 possible direction values.

Orientation of Palm

The palm's orientation with respect to the hand's shaft is referred to as palm orientation. The palm's orientation can take up to eight different values for each extended finger direction.For example, if the EFD is forward, palm orientation may be up, down, right, left, and four orientations intermediate to these.

Day 8

Morning Session

RESOURCE PERSON	**Premjith B**
TOPIC	**Text classification and Machine Translation**
Time	**10.00 am - 1.00 am**

Sir started the class by briefing the language models, usage of the language models.eg., chat GPT, google keyboard and all the text prediction models. He familiarised about the n gram models and feedforward networks that are used in the Language Models.

Note : About the Windowing approach.

Explained about the advantages and disadvantages of the language models.Talked about the Hidden Markov Model.

Explained about the Task, Problem, and the Solution of the NLM. Next he discussed the embedding of previous words using the One-hot-Encoding Representation in NLM.Discussed about the Fixed Window problem of feed forward network and its solution.

Discussed about the difference between Convolutional Neural Network (CNN) and Feed Forward Network(FFN). He explained the CNN and FNN for image processing. Briefed about the POS Tagging and prediction of next word using the parts of Speech Tagging. Discussed the Designing Criteria for the Sequence Modeling. Explained about the mathematical concept of word prediction.

He talked about the problems of prediction of future words and finalised that, the solution for this is using the recurrent neural network (Hidden Layer Neural Networks) to tackle the problem of feasibility of prediction of words.

Discussed about the parameters, weight and bias used in Neural Network. Explained about the architecture of the Recurrent Neural Network (RNN) using a flow chart.

Note : In image classification you can use RNN but DNN is better, as it gives more accuracy.

For image captioning models you can make use of one to many architecture of RNN and for sentimental analysis you can use many to one architecture. Discussed the DALLE.2 for image generation.

Discussed about the encoder Decoder Architecture and its use in the Generative Models like Chat GPT, DALLE.2 etc.

LSTM and GRU(Long short term Memory and Gated Recurrent Unit)

Vanishing Gradient and Exploding gradient problem

Lengthy sequences can cause gradients to vanish or explode.So,solution is clips gradient to a small number whenever they explode.Whenever the gradient reaches a certain value which comes down.

Solutions:

ReLU instead of sigmoid function.Derivative for the ReLU is either 0 or1.

RNN can't capture long-term dependencies.But LSTM does the same.

LSTM captures very long term data dependencies to some extend. Computation is much higher than others. GRU is the evolution of the language model.

Bi–RNN/LSTM/GRU

Afternoon Session

RESOURCE PERSON	**Dr. Premjith B**
TOPIC	**Text classification**
Time	**2.00 pm - 5.30.00 pm**

Practical Session handled Sentiment_analysis_using Deep_Learning methods.

Given two datasets for analysing sentiment of the text.

Work the code with RNN,LSTM,Bi-LSTM,GRU

Word embedding is the vectorization method and applies word2Vec instead of embedding layers.

Day 9

Morning Session

Resource Person	**Dr. Premjith**
Topic	**Machine translation**
Time	**10.00am -1.00pm**

Encoder Decoder architecture

- In machine learning we have to give feature or feature extraction have to be done by user while deep learning feature extraction done automatically
- Resnet is used as pretrained model for feature extraction for object detection or image detection
- At decoder level the future should be predicted
- Vectors made from input layer made decode at decode layer.
- Loss happens at decoder layers and transmitted back to input using back propagation.
- Vision transformers can be used for image detection as pretrained deep learning architecture.

Machine translation

- In machine translation input and output are sequence while in image input is sequence
- Bidirection lstm/gru for less resource data otherwise unidirection lstm/gru
- Text summarization is similar in architecture with machine translation
- For contextual translation in nmt attention mechanism used.
- Attention mechanism takes the relevance between the words and gives number
- Different attentions are bahdanaus attention,luongs attention and dot product
- Bahdanaus attention more efficient
- Bahdanaus attention is used for attention scores computing

  Transformer Network

  - Transformer model is designed for NMT
  - Gpt is used at decoder level
  - BERT is used at encoder level.
  - Self attention is used at encoder level
  - Multihead attention is the combination of self attention
  - Head is the parameter
  - Multihead attention dimension is transformed to input dimension
  - Autoregressive models are for predicting output based on trained inputs

BERT

Sentence piece tokenization is used

Afternoon Session

Resource Person	**Dr.Premjith**
Topic	**Machine translation**
Time	**2.00pm-5.30 pm**

Afternoon Session handled with practicality .Explaining translation code which could be seen in tensorflow library,explained the parameters and encoder decoder model in depth.

Day 10

Morning Session

Resource Person	**Sabeerali K P, Swathy A S**
Topic	**OCR**
Time	**10.00am -1.00pm**

Ocr is a combination of patterns and features.In OCR the underlying patterns are recognized and features are the uniqueness in the given characters.

Applications of OCR(Optical character recognition)
In word processing,Legal documentation,Banking,Translation applications OCR is widely used.

Types of OCR
HTR: Handwritten text recognition is a complex than normal OCR

Number plate recognition is a part of OCR.
ICR :Intelligent character recognition
MICR :Magnetic ink character recognition used in Bank passbooks.

OCR working:
First step is Image acquisition.In preprocessing it include binarization ,noise removal,skew correction,Thinning and orientation detection.

Next step is Layout Analysis in which text and non-text recognition include text line extraction ,word segmentation.

Next introduced about open source libraries which include Tesserect developed by Google and keras ocr,paddle ocr,Esy ocr.

Tesseract one of widely used ocr support about 100 languages and highly accurate than rest of them.and discussed about difficulties with document recognition difficulty.

MMOCR(Multi modal OCR)

Limited support to more languages.

Easy OCR
Based on tensorflow deep learning framework ,but like MMOCR it support limited languages. Next Shabeer discussed his work on hand written recognition in Malayalam and its details connecting how a simple OCR works
Mainly discussed on Binarization , Skew correction
Thinning :giving uniform thickness to characters.Followed by usage of deep learning using LSTM and post processing using spell checker.

Moving to next section of coding with Google colab with show how keras ocr which is a deep learning framework works on a written english text image.After giving a proper understanding of showing the implementation code next move to implementation with Easy OCR.
For easy ocr the students tried also with arabic image document also.Next showed an interesting library in OCR which is paddleocr which works for chinese and english much better.
Further discussed usage of streamlit in creating web app and usage of pytesseract.
After Noon session start at 2:15pm with further implementation of streamlit handon session with pdf and jpg files.

Afternoon Session

Resource Person	**Meharniza Nazeem**
Topic	**Explainable AI in NLP**
Time	**2.00pm -5.00pm**

She started with Introduction about Machine Learning, then Explainable AI in NLP , utilising AI to provide understandable and transparent explanations for their decisions and actions. Explained scenario in which explainable AI used for different times such as why a loan is given or not and showed different flow charts related to map of explainability. Then she mentioned the methods of explainable AI, its features and why they are important in AI applications. The map of different approaches is explained. She started to explain algorithms: LIME and SHAP.

LIME(Local interpretable model-agnostic explanations)

Model agnosticism refers to property of LIME to black box model. Local explanations mean that LIME gives explanations that are locally faithful within the surroundings or vicinity of the observation/sample being explained. LIME is currently available for tabular,image,text datasets. Limits to Supervised Machine Learning and Deep Learning(now).

Steps: a) Sampling and obtaining a surrogate dataset
b) Feature Selection from the surrogate dataset.

SHAP(SHapley Additive exPlanations)

Provides useful model explainability using simple plots such as summary and force plots.It is based on a game theoretic approach and explains the output of any machine learning model using visualization tools.

After that it is given research paper on Explainable AI. Briefly explain SHAP theory based on user docs. Hands-on session on the implementation of shap in dataset and introduced how to create an explainable ai project followed by implementation with LIME. Hands-on session on Sentiment Analysis using SHAP.

Day 11

Morning Session

Resource Person	**Dr. Anoop V S**
Topic	**Natural Language Generation**
Time	**10.00am -1.00pm**

The session started with an insight into how “Data is the new oil ” which gives us a preface on the importance of machines to understand and process natural language. Context Surrounding facts regarding data.Data without context mean little.Knowledge and insights gained from the information when used to take proactive decisions can be termed as intelligence in a broader understanding.

Types of data : Structured data : Unstructured Data : Semi Structured Data eg : XML
80 % data available is unstructured and it needs text mining and Natural Language processing to make sense out of it.Natural language generation and Natural Language understanding are subsets of NLP .

NLU
> Interpret the natural language
> Derive meaning
> Identify context
> Draw insights
While **Intent** refers to the goal the customer has in mind when typing in a question or comment,
Entity refers to the modifier – fields, data, or text, the customer uses to describe their requirement while the intent is what they really mean.

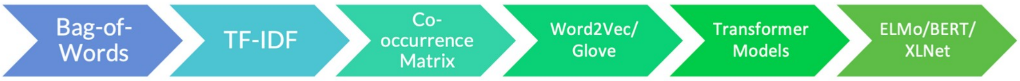

Fig. Evolution of NLP Models

Binary Bag of words
Count vectorizer
TF-IDF - Identifies the importance of a word in corpus.Syntactic similarity
Co occurrence matrix
Distributional Semantics - The words that are having similar meaning or similar context will be coming closer and be clustered . The words can be represented as vector and their similarity can be specified by the cosine between the vectors.
Word2Vec , Glove are such models which use semantic similarity.
This resulted in the transformers model , which is derived from the paper"Attention is all you need".
In traditional RNN , it will check previous 1 or 2 words only. Which can be overcome by using LSTM , but it has computational complexity once there is a large amount of data. Thus a transduction model relying on self attention was recommended .
Domain specific language models

- FinBERT
- ClimateBERT
- LawBERT

Afternoon Session

Resource Person	**Dr.Anoop V S**
Topic	**Natural Language Generation**
Time	**2.00pm-5.00 pm**

An engaging hands-on session was conducted on the art of text generation employing advanced GPT models. Additionally, participants learned the text summarization using the power of GPT models. The session concluded gracefully at 5 p.m.

Day 12

Project Evaluation & Valedictory Function

The project evaluation commenced at 9:30 a.m. and concluded at 11:45 a.m. Subsequently, Dr. Girish Nath Jha delivered a captivating keynote talk, delving into the present landscape of

NLP research and highlighting diverse challenges within the field. The session concluded at 1 p.m. Following a refreshing lunch break, the feedback session resumed, and it was followed by an award ceremony honoring the winners of the project presentation. The distribution of certificates concluded at 4 p.m.

****