Development of Abstractive Text Summarizer

Internship Report

Submitted for the partial fulfilment of the degree of

Bachelor of Technology

In

Mathematics and Computing

Submitted By

Naman Pasari

0101MC201044

UNDER THE SUPERVISION AND GUIDANCE OF

Dr. V P Shinde

Professor & Head of Department

Department of Engineering Mathematics and Computing

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR (M.P.), INDIA

माधव प्रौद्योगिकी एवंगवज्ञान संस्थान, ग्वागियर (म.प्र.), भारत

Deemed to be University

NAAC ACCREDITED WITH A++ GRADE

January - June 2024

DECLARATION BY THE CANDIDATE

I hereby declare that the work entitled **Development of Abstractive Text Summarizer** is my work, conducted under the supervision of Dr. V P Shinde, **Professor & HoD (Dept. Engineering Mathematics and Computing), during** the session Feb - May 2024. The report submitted by me is a record of bonafide work carried out by me.

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Naman Pasari
0901MC201044

Date: 16/05/2024
Place: Gwalior

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief.

Guided By:

Dr. V P Shinde
Professor & HoD
Department of Engineering Mathematics and Computing
MITS, Gwalior

Departmental Project Coordinator

Dr. D K Jain
Professor
Department of Engineering
Mathematics and Computing
MITS, Gwalior

Approved by HoD

Dr. V P Shinde
Professor & HoD
Department of Engineering
Mathematics and Computing
MITS, Gwalior

PLAGIARISM CHECK CERTIFICATE

This is to certify that I/we, a student of B.Tech. in **Engineering Mathematics & Computing** have checked my complete report entitled **Development of Abstractive Text Summarizer** for similarity/plagiarism using the "Turnitin" software available in the institute.

This is to certify that the similarity in my report is found to be 16% which is within the specified limit (20%).

The full plagiarism report along with the summary is enclosed.

Naman

Naman Pasari

0901MC201044

Checked & Approved By:

Dr. J K Muthele

Associate Professor

Department of Engineering

Mathematics and Computing

MITS, Gwalior

CERTIFICATE OF INTERNSHIP

AD Infocom Systems

62, Bajrang Nagar, Manewada Road, Nagpur – 440027
Contact No.: 9860455757 **Email:** adinfocomsystems@gmail.com
Website: www.adinfocomsystems.in

Approved By
All India Council For Technical Edu
AICTE(NEW DELHI)

Ref. No. - ADIS/2024/33

Date - 30/05/2024

Internship Certificate

We are happy to announce that **Naman Pasari** has successfully completed the internship at AD Infocom Systems in the domain of **"Artificial Intelligence and Machine Learning"** from **30/01/2024** to **30/05/2024**.

During internship, you have successfully completed all the modules and project work.

We wish you all the best for your future endeavors.

Thank you.

Prabhakar Dorge
Managing Director
AD Infocom Systems

AD Infocom Systems Seal Nagpur

ABSTRACT

Interned as an AI/ML Intern at AD Infocom Systems, gaining hands-on experience in core concepts (supervised, unsupervised, reinforcement learning) and applying them to projects. Strengthened Python skills with libraries like NumPy, Pandas, and Scikit-learn for data manipulation, model building, and visualization. Developed problem-solving and communication skills by tackling real-world challenges and collaborating with the team. This internship deepened my understanding of AI/ML and equipped me for a career in this field.

Main Project: Abstractive Text Summarization with Pegasus

This section details the development of an abstractive text summarization system using the Pegasus model. The report explores the selection of Pegasus, a transformer pre-trained for summarization, and the SAMSum dataset for training. It also discusses techniques for mitigating challenges with Large Language Models (LLMs) and the modular approach taken for code development and deployment.

Keywords: Artificial Intelligence, Machine Learning, Natural Language Processing, Deep Learning, Python, transformers, encoder-decoder, Pegasus, Frameworks, Abstractive text summarisation.

ACKNOWLEDGEMENT

I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit**, and the Dean Academics, **Dr. Manjaree Pandit** for this project. It is my great pleasure to express sincere gratitude to my mentor and supervisor, **Dr. V P Shinde, Professor & HoD**, Department of Engineering Mathematics & Computing for his expert guidance and constant encouragement. I acknowledge that it is because of his interest that I enjoyed working on this project and express my earnest and heartfelt thanks to him for his time, support, and efforts. I am also thankful to all the faculties of the **Department of Engineering Mathematics & Computing** for their encouragement, who had invested their valuable time in providing their feedback with a lot of useful suggestions. I am highly obliged to all my friends for their encouragement and for helping me at the points where I got stuck. I am deeply indebted to all of them for always helping and inspiriting me.

Naman

Naman Pasari

0901MC201044

CONTENT

Table of Contents

ACRONYMS

For better understanding, this report uses the following acronyms:

AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CBF	Content-Based Filtering
CF	Collaborative Filtering
CNN	Convolutional Neural Network
DL	Deep Learning
F1 score	F1 Score (harmonic mean of precision and recall)
IT	Information Technology
KNN	K-Nearest Neighbours
LLM	Large Language Model
MATLAB	Matrix Laboratory
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
NS2	Network Simulator 2
ROUGE score	Recall-Oriented Understudy for Gisting Evaluation score
RS	Recommendation System
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

AD Infocom Systems, a leading software company headquartered in Nagpur, Maharashtra, India, stands out for its comprehensive IT service offerings. Their expertise encompasses web development, app development, software development, and advanced coding using platforms like MATLAB and NS2. **Beyond traditional software development, AD Infocom Systems is at the forefront of Artificial Intelligence (AI) and Machine Learning (ML) solutions**. The company recognizes the transformative potential of these fields and actively invests in developing solutions that leverage AI and ML algorithms.

In addition to crafting innovative solutions, AD Infocom Systems fosters future talent through its extensive training and internship programs. These programs encompass a wide range of Engineering and Technology disciplines, empowering aspiring professionals to develop their skills and meaningfully contribute to the industry.

AD Infocom Systems embodies a commitment to continuous innovation and deep technical expertise. By consistently delivering custom-tailored solutions that address the evolving needs of their clients, the company has established a strong reputation within the ever-expanding information technology landscape. **Driven by a relentless pursuit of excellence and a pioneering spirit, AD Infocom Systems continues to solidify its position as a leader in the IT industry**.

CHAPTER 2: LITERATURE SURVEY

The convergence of recommender systems (RS) and Natural Language Processing (NLP) techniques has become a cornerstone of personalized user experiences across various domains. This literature survey explores the synergy between these fields, focusing on how NLP empowers the development of more sophisticated and user-centric recommendation systems.

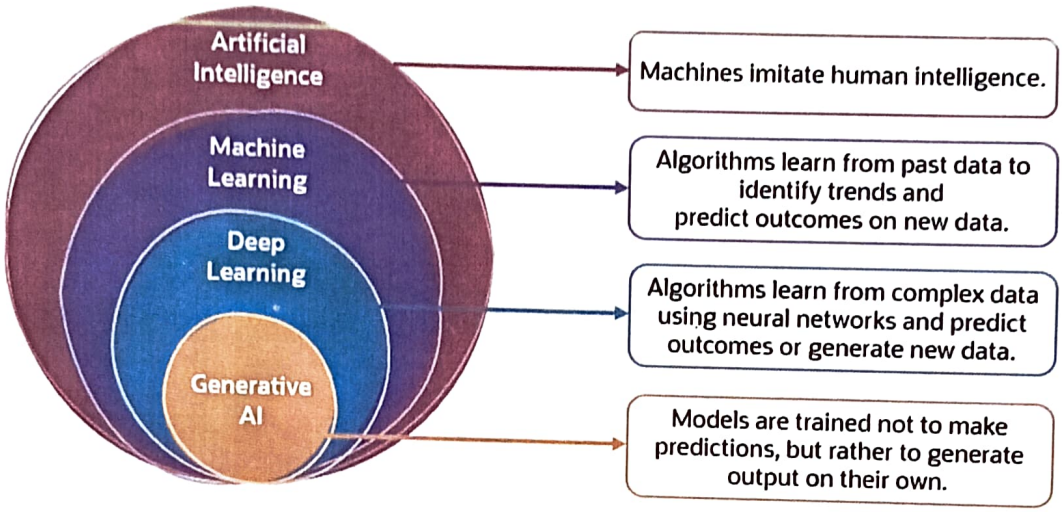

Figure 1: Figuration of AI, ML, DL and Generative AI Courtesy: Oracle University [1]

1. Recommender Systems: A Brief Overview

Recommender systems (RS) aim to predict user preferences and suggest relevant items (products, services, content) based on their past interactions and behaviour. Common RS approaches include:

- **Collaborative Filtering (CF):** This technique identifies users with similar preferences and recommends items enjoyed by similar users.
- **Content-Based Filtering (CBF):** This approach focuses on item attributes and recommends items similar to those the user has interacted with previously.

2. NLP Techniques for Recommendation Systems

NLP empowers RS in several ways:

- **Feature Extraction:** NLP techniques can extract valuable features from textual data associated with users and items. These features, such as sentiment analysis, topic modeling, and named entity recognition, can enhance the recommendation process.
- **Understanding User Reviews and Feedback:** NLP can analyze user reviews, ratings, and social media posts to understand user preferences and opinions. This helps identify emerging trends and personalize recommendations.
- **Natural Language Generation (NLG):** NLG techniques can be used to generate personalized recommendations with clear explanations, improving user transparency and trust.

3. Research Trends and Applications

Several research areas leverage NLP for advanced recommendation systems:

- **Movie/Book Recommendations:** NLP can analyse reviews and plot summaries to recommend content that aligns with user preferences and interests.
- **E-commerce Recommendations:** NLP techniques can analyse product descriptions and user reviews to suggest complementary products or items based on user needs.
- **News Recommendation:** NLP can assess user reading history and extract topics of interest to suggest relevant news articles.
- **Conversational Recommender Systems:** Integrating NLP with chatbots allows for personalized recommendations through interactive conversations.

4. Challenges and Future Directions

Despite the advancements, NLP-based RS face challenges:

- **Data Sparsity and Cold Start Problem:** Limited user data or lack of prior interactions for new users can hinder accurate recommendations.
- **Scalability:** Processing large volumes of textual data requires efficient NLP algorithms and infrastructure.

- **Explainability and Bias:** Ensuring transparency in recommendations generated by complex NLP models is crucial.

Future research directions involve:

- **Incorporating Explainable AI (XAI) techniques:** This can improve user trust and understanding of how NLP models influence recommendations.
- **Leveraging deep learning models:** Deep learning can extract complex patterns from textual data and personalize recommendations further.
- **Combining NLP with other modalities:** Multimodal approaches, integrating visual and audio information with NLP, can create a more holistic understanding of user preferences.

This literature survey provides a foundational understanding of how NLP techniques enhance recommendation systems. By addressing the challenges and exploring future directions, NLP can revolutionize personalized user experiences across diverse applications.

CHAPTER 3: PROBLEM FORMULATION

Abstractive Text Summarization with Pegasus

This project aimed to develop an abstractive text summarization model using the Pegasus model and the SAMSum dataset. Abstractive summarization goes beyond extractive techniques, aiming to capture the key meaning of a text and generate a concise summary that retains the core information while potentially using different words and phrasing.

Transformer Architectures:

Transformer models revolutionized the field of NLP. Several transformer architectures exist, each with its strengths. This project specifically chose the Pegasus model, which falls under the category of encoder-decoder transformers.

- **Encoder-Decoder Transformers:**
  - **Encoder:** The encoder portion of the model processes the input text, capturing the relationships between words and extracting the main ideas.
  - **Decoder:** The decoder utilizes the encoded information to generate a new sequence of words, forming the concise summary.

Architectures at a glance

Task	Encoders	Decoders	Encoder-decoder
Embedding text	Yes	No	No
Abstractive QA	No	Yes	Yes
Extractive QA	Yes	Maybe	Yes
Translation	No	Maybe	Yes
Creative writing	No	Yes	No
Abstractive Summarization	No	Yes	Yes
Extractive Summarization	Yes	Maybe	Yes
Chat	No	Yes	No
Forecasting	No	No	No
Code	No	Yes	Yes

Figure 2: Comparative analysis of LLM architectures. Courtesy of Oracle University [7]

Strengths of Pegasus:

- **Pre-training:** Pegasus benefits from being pre-trained on a massive dataset of news articles. This pre-training allows it to understand complex sentence structures and the flow of information within a text, crucial for generating summaries that capture the essence of the original content.
- **Efficiency:** The google/pegasus-cnn_dailymail version specifically offers a good balance between performance and computational efficiency. This is important, especially if the project aims for real-world deployment where resources might be limited.

SAMSum Dataset - A Closer Look:

While various summarization datasets exist, SAMSum offers several advantages for this project:

- **Abstractive Summaries:** This dataset includes not just the source articles but also human-written abstractive summaries. This allows the model to learn the task of generating summaries that convey the core meaning in new phrasing, rather than simply copying sentences from the source text.
- **Variety:** SAMSum provides a diverse collection of texts, ensuring the model is exposed to different writing styles and topics. This helps the model generalize its summarization capabilities to unseen data.

Choosing the Right Tools:

- **AutoTokenizer:** The Hugging Face AutoTokenizer simplifies the process of tokenizing text data for the Pegasus model. It automatically handles vocabulary building and special token management, ensuring compatibility with the pre-trained model.

CHAPTER 4: PROPOSED METHODOLOGY

A Modular Approach

To achieve a robust and scalable summarization system, a modular approach was adopted throughout the project.

Addressing LLM Challenges in Detail:

The project tackled common challenges associated with Large Language Models (LLMs) like Pegasus through several techniques:

- **Early Stopping - Preventing Overfitting:** Imagine memorizing every detail from a textbook instead of understanding the key concepts. Similarly, overfitting occurs when the model focuses too much on training data specifics and fails to generalize to unseen examples. Early stopping monitors the model's performance on a separate validation set. When the validation performance stops improving or starts declining, training is terminated to prevent overfitting to the training data.
- **Warmup Steps - A Gentle Introduction:** Imagine a student overwhelmed with a sudden influx of new information. Similarly, abruptly applying large learning rate updates can be detrimental to LLM training. Warmup steps gradually increase the learning rate during the initial training phase, allowing the model to adapt to the data before applying bigger updates. This helps the model converge on a better solution faster.
- **Hyperparameter Tuning - Finding the Sweet Spot:** Hyperparameters control various aspects of the training process. Finding the optimal settings for these parameters is crucial for achieving peak model performance. Techniques like grid search or random search were likely employed to explore different combinations of hyperparameters (e.g., epochs, batch size, learning rate) and identify the configuration that yielded the best results on the validation set.

Benefits of Modular Coding:

- **Readability:** Well-defined modules with clear functions improve code comprehension, making it easier for others (or even yourself in the future) to understand the project's logic.
- **Maintainability:** Modular code allows for easier modifications and bug fixes. By isolating functionality within modules, changes can be made without affecting unrelated sections of the code.
- **Reusability:** Reusable modules can be incorporated into future projects, saving development time and effort.

Logging and Pipelines - Streamlining the Process:

- **Logging:** Imagine training a model without any record of its performance. Logging mechanisms track essential information during training, such as loss values, metrics (e.g., ROUGE score for summarization), and hyperparameter configurations. This data provides valuable insights into the training process and allows for analysis and troubleshooting if needed.
- **Pipelines:** Building reusable pipelines for data pre-processing, training, and evaluation stages can significantly improve workflow efficiency. These pipelines automate repetitive tasks, ensuring consistency and reducing the risk of errors.

This project demonstrates the effectiveness of using Pegasus for abstractive text summarization and showcases a well-structured approach encompassing data selection, model choice, hyperparameter tuning, modular coding practices, and deployment strategies. The use of open-source tools from Hugging Face and frameworks like PyTorch and FastAPI further demonstrates the project's efficiency and potential real-world applications.

CHAPTER 5: RESULT & DISCUSSION

Foundational Knowledge:

- **Grasp core concepts of AI and ML:** This includes understanding the difference between supervised, unsupervised, and reinforcement learning. Learners gain a solid foundation in areas like neural networks, decision trees, and various algorithms such as Support Vector Machines (SVM) [1], Random Forest [2], and K-Nearest Neighbours (KNN) [3].
- **Explore data science principles:** The internship delves into data preparation, cleaning, and feature engineering techniques. Learners explore various data analysis libraries like pandas and understand the importance of data visualization for exploratory analysis.
- **Comprehend machine learning evaluation metrics:** The program emphasizes the importance of evaluating model performance using metrics like accuracy, precision, recall, F1-score [4], and AUC-ROC curve [5].

Technical Skills:

- **Proficiency in Python programming:** Learners gain a strong grasp of Python syntax, data structures, and libraries like NumPy, Pandas, Matplotlib, and Scikit-learn. These tools become essential for data manipulation, model building, and visualization.
- **Experience with deep learning frameworks:** The internship may involve working with frameworks like TensorFlow or PyTorch. Learners gain hands-on experience building and training deep learning models for various tasks like image recognition, natural language processing, and time series forecasting.
- **Exposure to cloud-based platforms:** Interns may explore deploying trained models on cloud platforms like Microsoft Azure. This provides experience with cloud infrastructure and scalability.

Problem-Solving and Soft Skills:

- **Develop critical thinking and problem-solving skills:** The internship presents real-world AI/ML challenges, requiring interns to analyse data, identify patterns, and design appropriate machine learning solutions.
- **Enhance communication skills:** Interns learn to effectively communicate technical concepts to both technical and non-technical audiences. This may involve writing reports, presenting findings, and collaborating with team members.
- **Embrace continuous learning:** The dynamic nature of AI/ML requires continuous learning and adaptation. Interns develop the ability to learn new tools, frameworks, and algorithms as the field evolves.

Career Development:

- **Gain a competitive edge in the job market:** By mastering AI and ML techniques, interns become highly sought-after candidates for various industries leveraging AI/ML.
- **Build a strong portfolio:** The internship provides opportunities to work on real-world projects, allowing interns to showcase their skills and knowledge to potential employers.
- **Network with professionals:** Interns gain valuable connections with mentors and colleagues within the AI/ML field, potentially leading to future career opportunities.

This internship provides a comprehensive learning experience, equipping aspiring professionals with the skills and knowledge to thrive in the ever-growing field of Artificial Intelligence and Machine Learning.

CHAPTER 6: CONCLUSION

My internship at **AD Infocom Systems** as an AI/ML intern proved to be an invaluable experience that significantly deepened my understanding of Artificial Intelligence and Machine Learning. Throughout this internship, I had the opportunity to:

- **Gain practical experience with various AI/ML algorithms:** This included working with algorithms like e.g., Support Vector Machines (SVM) [1], Random Forest [2], K-Nearest Neighbours (KNN) [3] etc. for tasks such as data preprocessing and gaining helpful insights from data for analysis.
- **Develop proficiency in Python programming and essential libraries:** I honed my skills in Python and gained a strong grasp of libraries like NumPy, Pandas, Matplotlib, and Scikit-learn, which are crucial for data manipulation, model building, and visualization in the AI/ML field.
- **Explore deep learning frameworks:** The internship provided the opportunity to work with deep learning frameworks like TensorFlow or PyTorch, allowing me to build and train deep learning models for tasks like image recognition, natural language processing etc.
- **Gain exposure to cloud platforms:** This experience exposed me to deploying trained models on cloud platforms like Microsoft Azure, fostering an understanding of cloud infrastructure and scalability.
- **Apply problem-solving skills to real-world AI/ML challenges:** By analyzing data, identifying patterns, and designing appropriate solutions, I further developed my critical thinking and problem-solving abilities.
- **Enhance communication skills:** Throughout the internship, I honed my ability to communicate technical concepts effectively to both technical and non-technical audiences through reports, presentations, and collaboration with team members.

This internship has instilled in me a strong foundation in AI/ML and a passion for continuous learning in this rapidly evolving field. I am confident that the skills and knowledge I gained will be instrumental in my future endeavours within the exciting world of AI/ML.

I would like to express my sincere gratitude to Mr. Prabhakar Dorge for their guidance and mentorship throughout the internship. I would also like to thank the entire team at AD Infocom Systems for providing me with this invaluable learning opportunity.

REFERENCES

[1] Oracle University. [Online]. Available: https://blogs.oracle.com/oracleuniversity/post/announcing-oci-ai-foundations-certification.

[2] Oracle University. [Online]. Available: https://education.oracle.com/.

[3] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[4] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.

[5] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html#sklearn.impute.KNNImputer.

[6] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score.

[7] [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html#sklearn.impute.KNNImputer.

TURNITIN PLAGIARISM REPORT

Similarity Report

PAPER NAME

plag_check.pdf

WORD COUNT

2664 Words

CHARACTER COUNT

16972 Characters

PAGE COUNT

15 Pages

FILE SIZE

393.8KB

SUBMISSION DATE

May 16, 2024 11:44 PM GMT+5:30

REPORT DATE

May 16, 2024 11:45 PM GMT+5:30

● 16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 14% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material

Summary

MPRs

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV Bhopal)
NAAC Accredited with A++ Grade

MONTHLY PROGRESS REPORT (MPR) FROM INDUSTRY/COMPANY MENTOR

Name of Student	Naman Pasari		Enrollment No.	0901MC201044	
Department	Mathematics & Computing		Sem.	8 (session: Jan-June 24)	
Industry/Organization	Ad Infocom Systems		Date/Duration	01/02/24 - 01/03/24	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work		✓			
Learning capacity/Knowledge up gradation				✓	
Performance/Quality of work				✓	
Behaviour/Discipline/Team work			✓		
Sincerity/Hard work			✓		
Comment on nature of work done/Area/Topic	He has completed all the tasks on time, being a fresher he has quite good knowledge of machine learning and currently is under training period and delivers on the project he is assigned.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry/Company Mentor	Mr. Prabhakar Dorge				
Signature of Industry/Company Mentor					

Receiving Date	02/03/24	**Name of Faculty Mentor**	Dr. Vikas Shinde	**Sign**	

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV Bhopal)
NAAC Accredited with A++ Grade

MONTHLY PROGRESS REPORT (MPR) FROM INDUSTRY/COMPANY MENTOR

Name of Student	Naman Pasari		Enrollment No.	0901MC201044	
Department	Mathematics & Computing		Sem.	8 (session: Jan-June 24)	
Industry/Organization	Ad Infocom Systems		Date/Duration	01/03/24 - 01/04/24	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work		✓			
Learning capacity/Knowledge up gradation			✓		
Performance/Quality of work			✓		
Behaviour/Discipline/Team work			✓		
Sincerity/Hard work			✓		
Comment on nature of work done/Area/Topic	He has completed all the tasks on time.				
OVERALL GRADE (Any one)	**GOOD**				
Name of Industry/Company Mentor	Mr. Prabhakar Dorge				
Signature of Industry/Company Mentor					

Receiving Date	**05/04/24**	**Name of Faculty Mentor**	Dr. Vikas Shinde	**Sign**	

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV Bhopal)
NAAC Accredited with A++ Grade

MONTHLY PROGRESS REPORT (MPR) FROM INDUSTRY/COMPANY MENTOR

Name of Student	Naman Pasari		Enrollment No.	0901MC201044	
Department	Mathematics & Computing		Sem.	8 (session: Jan-June 24)	
Industry/Organization	Ad Infocom Systems		Date/Duration	01/04/24 - 01/05/24	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work			✓		
Learning capacity/Knowledge up gradation		✓			
Performance/Quality of work				✓	
Behaviour/Discipline/Team work				✓	
Sincerity/Hard work				✓	
Comment on nature of work done/Area/Topic	He has acquired descent machine learning practices and handles given projects in a structured manner.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry/Company Mentor	Mr. Prabhakar Dorge				
Signature of Industry/Company Mentor					

Receiving Date	**04/05/24**	**Name of Faculty Mentor**	Dr. Vikas Shinde	**Sign**	