



VAIBHAV JAIN

✉ vaibhav.jain174@gmail.com [in linkedin](#) [github](#) [Portfolio](#)

I am a M.Sc. Data Engineering and Analytics student at the Technical University of Munich majoring in Natural Language Processing. I have 1 year of Professional experience and 2 years of Academic experience in working in the field of data science. I want to further expand my practical data science and ML/NLP engineering knowledge by working on real-world research problems. Please visit my portfolio to know more about my project work.

EXPERIENCE

Siemens

Apr 2022 – Present

Working Student in NLP

Munich, Germany

- Working as a student assistant in the Data Analytics and Artificial Intelligence (DAI) team at the Siemens HQ.
- Implementing NLP and machine learning based research prototypes that solve business problems for other internal units (Python, ML, NLP)
- Working on the problem of domain adaptation of Large language models under low resource constraints.(Semantic analysis, Hypothesis Testing)
- Contributed in building a python library to easily build and share domain adapted language models for Siemens industrial/technical domains(Python, Git, NLP)
- Cloud deployment of domain adapted models for ease of access by other teams with Siemens(AWS)

EDUCATION

Technical University of Munich

Oct 2020 – Jul 2023

M.Sc Data Engineering and Analytics

Munich, Germany

Birla Institute of Technology, Mesra

Aug 2016 – May 2020

B.E Computer Science

Ranchi, Jharkhand

PROJECTS

Generating Knowledge Graph from PDF's using self supervised learning.(Steering Lab-Horváth)

- **Task:-** Building an end-to-end pipeline to generate Knowledge graphs from PDF's from German Environment Agency (Umweltbundesamt).
- Building a rule based approach for entity extraction and combining it with neural NER to extract all possible entities from the text.
- Trained a transformer model to extract relationships using self supervised learning.
- Used Kmeans clustering approach to cluster sentences containing entities to build a training data for relationship classification.
- **Key contribution:** The approach does not require labelled data for relationship classification. It uses clustering to build data on its own using self supervision.

Movies and Songs Recommendation system

- **Task:** Implement different recommendation algorithms for songs and movies recommendation
- popularity based recommendation
- Content based filtering using movies genres for movies recommendation
- User and Item based collaborative filtering using movie genres and co-occurrence matrix
- Pearson's correlation to find similar users and items

MCQ AND T/F Question generation using Transformers

- **Task:-** Generate back the MCQ and T/F questions from the given correct answer and paragraph from which correct answer was extracted.
- Fine tune the T5 transformer model using SQuAd Data set to generate questions from one word answer, context (word sense) and text from which question to be generated from.
- Adapt BERT to perform Word sense disambiguation using positive negative context-gloss pair.
- Generate wrong choices for MCQ using co-hyponyms of the correct answer in the WORDNET.
- Generate False statements for T/F type questions by removing ending verb phrase or noun phrase from the sentence and completing the sentence using by wrong verb/noun phrases generated using GPT3.

End to End implementation of a simple ML use case

- **Task:** To understand the complete life cycle of a data science project
- Predict the maths marks of a student based on the student's attributes like gender, previous marks, age, classes attended etc
- EDA and data transformation
- Pipeline that tests multiple regression models and picks the best based on R2 score
- Grid search for hyper parameter tuning integrated with the model pipeline
- Deployment and prediction using flask web application

TECHNICAL SKILLS

Programming Languages	Python, C, SQL, R
Frameworks	Pandas, NumPy, Keras, Tensorflow, Pytorch, HuggingFace, NLTK, SpaCy, Scikit-Learn, Seaborn
Visualization tools	Tableau, Excel, Tensorboard
Version Control	Git

- Predictive and descriptive analytics experience.
- Professional experience in NLP frameworks.
- Experience in low resource model training.

PUBLICATION

Image and Video colorization system

Jan 2020

International Journal for Research in Applied Science and Engineering Technology (IJRASET)

PAPER LINK