PROJECT 1

Domain: Tourism/Social Media
Techniques: NLP, Machine Learning
Title: Fake Review Prediction

Overview and Problem Statement:

One of the fields in which disinformation is seen often is hotel reviews, both for positive and for negative reviews. There may be an interest to spread positive or negative fake news about hotels to gain unfair competitive advantage or as (unethical) means for profitability. The main idea of this project is to reduce the impact of disinformation by building a machine learning model to predict fake reviews pertaining to the hotel industry. This will also involve using computational stylometry based analysis to detect the specific stylistic features that differentiate fake from real (truthful) reviews.

Dataset:

The Deceptive Opinion Spam corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels and contains 400 truthful positive reviews from TripAdvisor and 400 deceptive positive reviews from Mechanical Turk in addition to 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp and 400 deceptive negative reviews from Mechanical Turk. Each dataset consists of 20 reviews for each of the 20 most popular Chicago hotels. In total there are 1600 reviews, and the task is to classify the real vs fake hotel reviews using a classical machine learning or deep learning algorithm. Dataset source: https://myleott.com/op-spam.html

Specific Challenges:

1. Detection of author specific writing stylistic features to differentiate truthful vs fake reviews
2. Obtaining a good accuracy on new reviews from other websites

References:

1. Dataset link: https://cdn.iisc.talentsprint.com/CDS/Datasets/Deceptive_Opinion_Spam_Corpus_v1.4.zip

2. Pascucci et al., (2020): https://aclanthology.org/2020.stoc-1.6.pdf.

3. Kennedy et al., (2020). Fact or factitious? Contextualized opinion spam detection. *arXiv preprint arXiv:2010.15296.*

PROJECT 2

Domain: Plant Pathology
Techniques: Deep Learning
Title: Image-based plant disease identification

Overview and Problem Statement:

Crop losses due to diseases are a major threat to food security every year, across countries. Conventionally, plant diseases were detected through a visual examination of the affected plants by plant pathology experts. This was often possible only after major damage had already occurred, so treatments were of limited or no use. Recently, access to smartphone based image capturing has highly increased amongst farmers and agriculturists. This has led to the successful adoption of plant disease diagnostic applications based on deep learning techniques. This is of immense value in the field of agriculture and an excellent tool for faster identification and treatment of crop diseases. It holds key importance in preventing crop based food and economic losses. The goal of this project is to build a convolutional neural network or to use transfer learning and develop a plant disease identification tool.

Specific Challenges:

1. Extending the applicability of the plant disease identification tool for farmers through the use of Indian regional languages [Optional challenge]

DataSet:

The PlantVillage dataset consists of 54303 healthy and unhealthy leaf images divided into 38 categories by species and disease. Dataset link:
https://data.mendeley.com/datasets/tywbtsjrjv/1

References:

1. Dataset link: https://data.mendeley.com/datasets/tywbtsjrjv/1
2. Dataset source details: https://www.tensorflow.org/datasets/catalog/plant_village
3. Liu and Wang (2021):
   https://plantmethods.biomedcentral.com/articles/10.1186/s13007-021-00722-9
4. Mohanty et al., (2016):
   https://www.frontiersin.org/articles/10.3389/fpls.2016.01419/full

PROJECT 3

Domain: Food / Wellness industry
Techniques: CNN based Image segmentation
Title: Food Image Segmentation

Overview and Problem Statement:

Worldwide, obesity has nearly tripled since 1975. In 2016, more than 1.9 billion adults, 18 years and older, were overweight (WHO sources). In such a situation, documenting dietary caloric intake is crucial to manage weight loss. Food image segmentation is a critical and indispensable task for developing health-related applications such as automated estimation of food calories and nutrients as a means for dietary monitoring. One of the challenges in this area is the improvement of accuracy in dietary assessment by food image analysis. However, how to derive the food information (e.g., food type and portion size) from food images effectively is a challenging task and an open research problem. In this project, participants are expected to make a model that can segment the food components present in an input food image and build an application that can predict the food class and the food portions from it.

Specific Challenges:

1. The complex appearance of food makes it difficult to localize and recognize ingredients in food images, e.g. the components may overlap with one another in the same image, and the identical ingredient may appear distinctly in different food images.

Data Description:

Food recognition dataset (Ciocca et. al., 2017) is one of the few publicly available, pixel segmented datasets on food. It contains 1,027 images of food on trays, with 73 classes of food and 3,616 labelled instances of food. The tray images have been manually segmented using carefully drawn polygonal boundaries.

Application:

It has many applications such as calorie estimation, diet management, industrial compliance and customer satisfaction.

References:

1. Ciocca et al., 2017: Food Recognition: A New Dataset, Experiments, and Results
2. Yang, 2020: Food Image Segmentation with fast.ai
3. Liu et al., 2016: DeepFood: Deep Learning-based Food Image Recognition for Computer-aided Dietary Assessment

PROJECT 4

Domain: Financial Services
Techniques: NLP, Machine Learning, Large Language Model (LLM)
Title: Personalized Financial Advisor using Large Language Model (LLM)

Overview and Problem Statement:

The field of finance can be complex and overwhelming for individuals seeking personalized financial advice. In order to make informed decisions regarding investments, retirement planning, budgeting, and financial products, individuals often require guidance from financial experts. The aim of this project is to develop an Intelligent Financial Advisor powered by a Large Language Model (LLM) to provide personalized financial advice and guidance to individuals. By leveraging NLP and machine learning techniques, the Intelligent Financial Advisor will assist users in making informed financial decisions and achieving their financial goals.

Specific Challenges:

1.Training and fine-tuning a Large Language Model (LLM) on financial datasets

2.Ensuring accurate understanding and response generation for financial queries.

Data Description:

Dataset: Alpha Vantage Financial Market Data Source: Alpha Vantage (https://www.alphavantage.co/). Alpha Vantage is a provider of financial market data APIs that offer real-time and historical data for various financial instruments such as stock prices, technical indicators, sector performance, exchange rates, and more. This data should be utilized to train and fine-tune the Large Language Model (LLM). Real-time financial data and market news and trends should be incorporated to provide current and accurate financial advice.

Applications:

1. Improved financial decision-making and goal achievement for individuals.
2. Offering personalized recommendations for investment portfolios, retirement plans, insurance policies, and other financial products based on individual goals and risk tolerance.

References:

1. Yang, Liu and Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *ArXiv, abs/2306.06031*.

PROJECT 5

Domain: Education
Techniques: Siamese Text Similarity Model, NLP, Deep Learning
Title: Automated Answer Validation for Science Question Answering using Siamese Text Similarity Model

Overview and Problem Statement:

In the domain of scientific question answering, validating answers and providing accurate feedback is critical for effective learning. The goal of this capstone project is to develop an automated answer validation system using a Siamese text similarity model. The system will compare student responses with the correct answer and distractors to determine the level of correctness and provide appropriate feedback. The automated answer validation system for science question answering will benefit educators and students in science-related subjects. It will streamline the assessment process, reduce manual effort, and ensure consistent evaluations, leading to improved learning outcomes in the science domain.

Specific Challenges:

1. Implementing a robust Siamese text similarity model for answer validation in the science domain.
2. Preprocessing and structuring the dataset from the provided SCIQ dataset to create meaningful pairs of student responses and correct answers.
3. Handling varying lengths of text data during model training and inference.
4. Addressing potential semantic ambiguities and domain-specific challenges in science questions and responses.

Data Description:

The SciQ dataset contains 13,679 crowdsourced science exam questions about Physics, Chemistry and Biology, among others. The questions are in multiple-choice format with 4 answer options each. For the majority of the questions, an additional paragraph with supporting evidence for the correct answer is provided. (string).

- answer1: One of the candidate answers for the question (string).
- answer2: One of the candidate answers for the question (string).
- answer3: One of the candidate answers for the question (string).
- answer4: One of the candidate answers for the question (string).
- correct_answer: The correct answer for the question (string).
- support: The supporting text for the question (string).

Methodology:

- Combine the question and supporting text to create meaningful context for answer validation in the science domain. Structure the data to form pairs of student responses and correct answers, along with corresponding labels (0 for incorrect and 1 for correct).
- Implement a Siamese neural network architecture using NLP libraries and deep learning frameworks. Train the model on the pairs of student responses and correct answers to learn the underlying text representations
- Develop an interface for educators to input student responses and query the Siamese text similarity model for validation in the science domain.
- Fine-tune the model and update the answer validation system based on the evaluation results.

References:

1. SCIQ (Science Question Answering) dataset from Kaggle:
https://www.kaggle.com/datasets/thedevastator/sciq-a-dataset-for-science-question-answering
2. Siamese network for image and text similarity:
https://medium.com/@prabhnoor0212/siamese-network-keras-31a3a8f37d04

PROJECT 6

Domain: Interior Design
Techniques: Stable Diffusion, Transformers
Title: AI-Driven Interior Design Using Stable Diffusion and Transformers

Problem Statement:

AI-driven innovation in creative industries such as design offers a very good potential. Be it architectural design, landscape design or interior design, the possibilities are endless. Such generated designs have the potential to drive rapid growth and profits in the interior design industry. The goal of this project is to generate realistic and novel interior room designs by leveraging Stable Diffusion and Transformer models, trained on the IKEA Interior Design Dataset. These generated designs can revolutionize the interior design industry by providing rapid, high-quality design options.

Specific Challenges:

1. Distinguishing a good design from a poor one or developing an evaluation metric is a complicated problem because good design is 'subjective'.
2. Ensuring the stable and reliable training of Stable Diffusion models to avoid issues like generating nonsensical designs.
3. Integrating Transformer models to effectively understand and apply design principles and context.

Dataset Description:

The dataset was collected from IKEA.com website for the purpose of building Style Search Engine (note: only for non-commercial use). It consists of: 2193 object (product) photos, 298 context (room scene) photos in which those objects appear, text descriptions for products, ground truth information on which items appear in which rooms.

Applications:

1. Web/app based designing tool for designers and real estate market - similar to Heavenly, Planner 5D
2. Automated interior design suggestions for enhancing user experience and creativity.

References:

1. Image to Image using Artificial Intelligence — Automating Room Interior Design, David Oluyale:
https://medium.com/@oluyaled/image-to-image-using-artificial-intelligence-automating-room-interior-design-5351bdfdd4bf

2. Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation: https://www.researchgate.net/publication/378154158_Integrating_aesthetics_and_efficiency_AI-driven_diffusion_models_for_visually_pleasing_interior_design_generation
3. Tautkute et al., 2017, ACICS, IKEA: Interior Design Dataset
4. Denoising Diffusion Probabilistic Models, Ho et al., 2020: https://arxiv.org/abs/2006.11239
5. Attention is All You Need, Vaswani et al., 2017: https://arxiv.org/abs/1706.03762

PROJECT 7

Domain: Market analytics and segmentation

Techniques: Data Engineering, ML, Google Data Studio

Title: Big data analytics and clustering of Farmers markets

Overview and Problem Statement:

Farmer's markets have emerged as an alternative and an improvement to grocery stores in the US. The Farmers' markets are a great place where people can consume nutritious fresh food while helping small family farms. Here, using a huge Farmer Market dataset, business insights in products and sales will be obtained through Big Data Analytics and market segmentation.

Specific Challenges:

1. Find important market insights through Big Data Analytics.
   Eg. Use Hadoop MapReduce to find the US state which has the lowest number of Farmer markets. Find the top 5 states that have the maximum number of products available. Assign ratings 1-5 for all the states based on the availability of products
2. Provide analytics to investors on states where new Farmer markets could be setup and prove to be profitable with a specific set of grocery products. Integrate other geographic/ state wise datasets eg. population/weather/average income if required to make better predictions
3. Clustering/market segmentation of the Farmer Markets to understand their distribution in terms of products, regions and sale

Data description:

The Farmers Market dataset is from the US data.gov site. This USDA National Farmers Market Directory data contains information on all registered farmers' markets in the US. Each row contains the information on the farmers market's name along with its city, county, types of food products sold, and the accepted forms of payment.

Application:

1. Obtaining business intelligence for investments in the Food retail industry
2. Obtaining competitive intelligence of understanding a company's industry and industry rivals so the company can make better business decisions

References:

1. USDA Farmers Markets:
https://www.ams.usda.gov/local-food-directories/farmersmarkets
2. Dataset source: https://data.world/johnsnowlabs/us-farmers-market-locationFarmer
3. Data Visualization: https://www.susielu.com/data-viz/farmers-markets

Project 8

Domain: Supply chain – Food Delivery
Techniques: ML Algorithms
Title: Driver Demand Prediction for Optimal Food Delivery Charges

Overview and Problem Statement

The food delivery industry relies heavily on the efficiency of its delivery processes, where timely and accurate deliveries significantly impact customer satisfaction and overall experience. This project focuses on developing an effective predictive model that accurately estimates the time it takes for orders to reach customers. Delivery time is influenced by various factors, including the delivery person's age, ratings, geographic coordinates, and time-related variables. The successful implementation of this food delivery time prediction model will enable business to optimize their operational processes, leading to more accurate delivery time estimates and an improved overall delivery experience for customers. The aim is also to predict the demand for delivery drivers in specific regions and times, by analyzing order requests, driver activity, and related parameters, thereby optimizing delivery charges, ensuring consistency and minimizing customer drop-offs.

Dataset:

The Kaggle Food demand forecasting dataset consists of the features 'ID', 'Delivery_person_ID', 'Delivery_person_Age', 'Delivery_person_Ratings', 'Restaurant_latitude', 'Restaurant_longitude', 'Delivery_location_latitude','Delivery_location_longitude', 'Order_Date', 'Time_Orderd','Time_Order_picked', 'Weatherconditions', 'Road_traffic_density','Vehicle_condition', 'Type_of_order', 'Type_of_vehicle','multiple_deliveries', 'Festival', 'City', 'Time_taken(min)'

Challenges:

- Handling potential data inconsistencies, such as missing data or outliers.
- Predicting sudden surges in order demand, which may not always follow historical patterns.
- Geographical feature engineering
- Optional: Real-time driver demand prediction without compromising the accuracy or speed of the system.

Additional Reading:
Urban distribution of short-term food delivery demand

Project 9

Domain: Human Resources and Technology
Techniques: Large Language Model (LLM) Fine-Tuning, Natural Language Processing (NLP), Text-to-Speech, Video Generation
Title: Automated Video Creation from Resumes Using GPT

Overview and Problem Statement: Creating a concise and engaging video summary of a resume can significantly enhance the recruitment process. Traditional methods of resume screening are time-consuming and may overlook key aspects of a candidate's profile. This project addresses the need for an automated solution to convert text-based resumes into 1-minute video summaries, effectively showcasing the candidate's skills, experiences, and achievements. The project involves fine-tuning a GPT model to generate a script from a resume and utilizing text-to-speech and video generation technologies to create the video.

Dataset: The primary dataset used for this project will be the Resume Dataset from Kaggle. This dataset contains a diverse collection of resumes, covering various industries and roles. The resumes are annotated to highlight key sections such as contact information, summary, skills, work experience, education, and achievements. This comprehensive dataset will be instrumental in training the model to accurately interpret and summarize resume content.

Specific Challenges:

- Ensuring the model accurately interprets the context and importance of various resume sections to create a coherent and relevant script.
- Tailoring the video script to reflect the unique strengths and experiences of each candidate.
- Generating human-like and engaging scripts that effectively summarize the resume content.
- Integrating text-to-speech technology to produce natural-sounding narration and combining it with relevant visual elements to create a professional video.

References:

1. Text-to-Video Generative AI Models: The Definitive List:
https://aibusiness.com/nlp/ai-video-generation-the-supreme-list

2. Everything to Know About OpenAI’s New Text-to-Video Generator, Sora:
https://www.scientificamerican.com/article/sora-openai-text-video-generator/

PROJECT 10

Domain: E-commerce and Fashion
Techniques: Deep Learning (Stable Diffusion, Transformer)
Title: Fashion Compatibility Prediction

Overview and Problem Statement:

The fashion domain is a very important and lucrative application of computer vision. According to a recent study by Statista, the fashion industry's worth was estimated to be $1.5 trillion in 2020 and it keeps growing, representing a huge market for garment companies, designers, and e-commerce entities. Fashion image retrieval and fashion image attribute learning have been the two main areas of study in this domain. The goal of this project is to compose or predict fashion outfits automatically, working to address challenges in compatibility and aesthetics using advanced deep learning models like Stable Diffusion and Transformers.

Specific Challenges:

1. Learn compatibility relationships among fashion items to facilitate effective fashion recommendation using a Transformer model. Transformers are highly effective in capturing contextual relationships in sequential data.
2. Utilize Stable Diffusion for generating high-quality, realistic images of fashion outfits that follow compatibility criteria.
3. Note that this project is compute intensive. This might require the use of AWS Sagemaker free/purchased account or equivalent high-performance computing resources.

Data Description:

Polyvore is a popular fashion website, where users create and upload outfit data. These fashion outfits contain rich multimodal information like images and descriptions of fashion items, number of likes of the outfit, hash tags of the outfit, etc. Researchers have utilized this information for various fashion tasks. Therefore, here, a curated part of the Polyvore dataset (made available for research purpose through the ACM MM'17 paper "Learning Fashion Compatibility with Bidirectional LSTMs" [paper] [code]) will be used. It contains 164,379 items (each item contains a pair - product image and a corresponding text description). The average number of fashion items in an outfit is 6.5. The fashion-compatibility-prediction.txt contains ~7,000 outfits, where 4,000 are incompatible and 3,000 are compatible. In each line the first number indicates the compatibility (1 is compatible, 0 is not) followed by a sequence of fashion items consisting the outfit.

Applications:

1. Outfit match recommendation
2. Accessories recommendation
3. Realistic generation of fashion outfits

4. Personalized fashion recommendations

References:

1. Polyvore dataset: https://github.com/xthan/polyvore-dataset
2. High-Resolution Image Synthesis with Latent Diffusion Models, Rombach et al., 2021: https://arxiv.org/abs/2112.10752
3. Virtual Fashion Designer: https://www.e2enetworks.com/blog/fine-tuning-stable-diffusion-to-create-a-virtual-fashion-designer-for-customers

PROJECT 11

Domain: Automotives & Insurance
Techniques: Deep Learning (Transfer Learning, Mask R-CNN)
Title: Automated car exterior damage assessment

Overview and Problem Statement:

Automated detection of car exterior damages and subsequent estimation of damage severity is of immense value for car dealers and insurance providers as it eliminates the manual process of damage assessment. This project is focussed on the detection of car damage and estimating its severity by using computer vision based Mask R-CNN technique. Mask R-CNN is an instance segmentation model that allows us to segment individual objects within a scene, regardless of whether they are of the same type.

Specific Challenges:

1. Data collection and annotation: Collect more car damage images under different weather conditions and different levels of illumination, enhance the data, improve the edge-contour enhancement of images, and make the masking of the damaged areas of the car more accurate.
2. Precise object detection
3. Robustness
4. Compute intensive. This might require the use of AWS Sagemaker free/purchased account.

Data Description:

1. Use web scraping from google images for downloading damaged and undamaged car images. Also include the COCO car damage dataset.
2. Annotate the images based on damage severity

Application:

1. Automated car damage assessment tool
2. Automated insurance claim prediction tool

References:

1. Sourish Dey, TowardsDataScience: Detecting car exterior damage
2. Priya Dwivedi, AnalyticsVidhya: Mask R-CNN model for detecting car damage
3. Zhang et al., 2020: https://ieeexplore.ieee.org/document/8950115

PROJECT 12

Domain: R & D
Techniques: Large Language Model (LLM) Fine Tuning, Computer Vision, NLP
Title: Gemini-SCICAP: Enhancing Scientific Figure Captioning with a Language Model

Overview and Problem Statement:

Scientific figures often contain crucial information, and providing accurate captions is essential for better comprehension. Existing generic captioning models may not capture the specialized terminology and context found in scientific literature. This project addresses the need for a dedicated model for scientific image captioning. This project involves fine tuning the Gemini Large Language Model (LLM) to generate accurate and contextually relevant captions for scientific figures. A model capable of understanding and describing complex scientific visuals will be created, combining the power of NLP with computer vision

Dataset

SCICAP (Scientific Captioning) is a large-scale image captioning dataset containing real-world scientific figures and captions. The dataset is constructed using over two million images from more than 290,000 papers collected and released by arXiv. It covers a wide range of scientific domains, making it a comprehensive resource for training and evaluating the model.

Specific Challenges

- Scientific Terminology: Adapting the language model to understand and generate captions with domain-specific scientific terminology.
- Complex Visuals: Handling intricate scientific visuals that may include charts, graphs, and diagrams.
- Contextual Understanding: Ensuring the model captures the context of the scientific content to provide informative and coherent captions.
- Multi-Modal Learning: Integrating both text and visual information for effective image captioning.

References

1. Gemini: A Family of Highly Capable Multimodal Models
2. Gemini: An Overview of Multimodal Use Cases
3. SCICAP Dataset: A Large-Scale Scientific Image Captioning Benchmark

PROJECT 13

Domain: Music Industry
Techniques: Association Rule Mining, Collaborative Filtering
Title: Market Basket Analysis for Personalized Music Recommendations

Overview and Problem Statement:

Music platforms with a large user bases often invest more in advanced recommender systems to cater to diverse user preferences. Competition among music streaming services drives innovation in recommender systems. Companies strive to differentiate themselves by offering more accurate and personalized recommendations, which can impact the overall market value. This project aims to create personalized music recommendations by employing Market Basket Analysis on the LFM-2b dataset. Using Data Mining, Market Basket Analysis and Recommender System Techniques, participants will utilize methodologies to uncover user behaviour patterns, fostering a deeper understanding of individual music preferences.

Dataset:

The LFM-2b dataset (http://www.cp.jku.at/datasets/LFM-2b/) contains the listening records of over 120,000 users of the music platform Last.fm. These users provide a total of more than two billion individual listening events that span a time range of over 15 years, from February 2005 until March 2020. These listening events refer to a total of 50 million distinct tracks of 5 million distinct artists. Beside the common metadata (i. e., artist and track name), LFM-2b contains additional information both regarding the users and items. This includes the demographic information of users, namely country, gender, and age, and the fine-grained genre and style of items together with the vector embeddings of their lyrics.

Specific Challenges

- Identifying frequent itemsets: Discovering commonly co-occurring music tracks in user listening histories and generating Association Rules to extract meaningful associations to understand music preferences.
- Utilizing user behavior patterns for personalized music recommendations through Collaborative filtering
- Optional: Apply matrix factorization techniques like Singular Value Decomposition (SVD) for latent factor modeling.

Reference:

LFM-2b Dataset Publication:
https://humrec.github.io/publication/schedl-chiir-2022/schedl-chiir-2022.pdf

PROJECT 14

Domain: Real Estate

Techniques: EDA, ML, DL, Time Series Analytics

Title: Market House price trends and prediction for metropolitan cities of India

Overview and Problem Statement:

Accurately predicting house prices is a challenging task. Potential home buyers assess multiple factors such as the size of the property, the location, road density, vicinity to offices, schools, stores, parks, restaurants, hospitals, airport, public transport etc. The goal of this project is to perform exploratory data analysis, time series analysis and predictive modelling for house prices in the metropolitan cities of India: Mumbai, Delhi, Bengaluru, Kolkata, Hyderabad and Chennai. This will involve deriving house price analytical insights through EDA and improving the accuracy over the current benchmarks using feature engineering (including geocoding) and model tuning.

Specific Challenges:

1. Feature engineering and model tuning
2. High predictive accuracy

Data Description:

1. For capturing house pricing trends over the years, National Housing Bank (NHB) data comprising of city wise housing price indices is available. It tracks changes in housing prices, measured against a base period and provides quarterly data on housing prices up to a neighbourhood level.

2. For predictive modelling, the dataset to be used is ‘The housing prices in Metropolitan areas of India’ (Kaggle). This dataset comprises of web-scraped data (~4000-7000 records per city) including the collection of prices of new and resale houses located in the metropolitan areas of India and the amenities provided for each house.

Applications:

Web based tool or mobile application for house price estimation, comparison and urban residential amenities assessment in metropolitan cities. Useful for individual home buyers and real estate businesses.

References:

1. Dataset: National Housing Bank - Residex
2. Dataset: Housing prices in Metropolitan areas of India
3. Sarkar and Purohit, Reuters, 2022. India house price rises pick up pace
4. Thakur and Satish, IRJET, 2021. Bangalore House price prediction

PROJECT 15

Domain: Sports
Techniques: Statistics, Probability, Machine Learning, Data Visualization, Big Data
Title: Exploratory and predictive data analytics on the Indian Premier League (IPL) dataset with LLM Integration

Overview and Problem Statement:

The Indian Premier League (IPL) is a professional Twenty20 cricket league, contested by eight teams based out of eight different Indian cities. The league was founded by the Board of Control for Cricket in India (BCCI) in 2007. IPL is one of the most lucrative and powerful cricket tournaments for BCCI as well as the IPL players. The huge data collected from IPL matches over 10 years presents a vast opportunity for data analytics and insights from predictive studies. The goal of this project is to obtain statistical and predictive insights from the IPL dataset. This project will involve the use of statistical, probabilistic, machine learning, big data and data visualization tools to derive valuable information and predictive insights by integrating an LLM interface to allow users to interact with the data. A data dashboard of the analytics should also be created. The project should focus on integrating a Large Language Model (LLM) interface with the data analytics dashboard. This interface will allow users to ask natural language questions and receive statistical insights, making the data analysis more accessible and intuitive. This LLM-driven approach will not only facilitate traditional data analysis but also enhance user interaction, enabling more profound and dynamic insights into the IPL data.

Data Description:

The IPL Data till 2017 dataset includes detailed ball-by-ball data of all IPL matches up to the 2017 season. It consists of five files with 97 columns (attributes) in total:

- Ball_by_Ball.csv
- Match.csv
- Player_match.csv
- Player.csv
- Team.csv

By using the IPL dataset and incorporating LLM for a conversational interface, this project aims to create a powerful, user-friendly tool. The project ensures that the solution remains robust, data-driven, and user-centric, aligning with the evolving needs of the sports analytics field.

Applications:

1. Providing a statistical and data-driven basis for team formulation.

2. Developing in-game strategies.
3. Making real-time outcome predictions.
4. Gaining insights into cross-team player strengths and weaknesses.

References:

1. Data.world: IPL Data till 2017
2. Srinivasan, 2018: IPL and Big Data Analytics: A match made in heaven?
3. Kargal, 2021: IPL 2021: How data analytics is changing cricket

Project 16

Domain: Healthcare and Predictive Analytics
Techniques: Machine Learning, Predictive Modeling, Data Preparation, Model Validation, Model Deployment, Real-time Data Integration
Title: Hospital Emergency Prediction using ETS Data

Overview and Problem Statement:

The objective of this project is to develop and implement a predictive model to anticipate emergent medical situations at the point of patient admission to a hospital. This predictive model will leverage historical patient data, particularly the Emergency Triage Score (ETS) dataset, to probabilistically anticipate instances where a significant influx of patients is likely to occur. This project seeks to harness data-driven insights to enhance the anticipatory and preparatory capacities of healthcare facilities in responding to emergent medical scenarios, thereby improving overall healthcare service efficiency and patient care outcomes. The goal of this project is to predict hospital emergencies at the time of patient admission. This predictive model will help healthcare providers allocate resources efficiently and prepare for potential surges in patient volume.

Dataset:

The project will utilize the Hospital Triage and Patient History Data from Kaggle: https://www.kaggle.com/datasets/maalona/hospital-triage-and-patient-history-data, specifically the ETS (Emergency Triage Score) dataset. This dataset contains valuable information about patient demographics, medical history, and triage scores, which can be used to predict the likelihood of an emergency upon admission.

Methodology:

1. Data Preparation: Prepare and clean the data for analysis.
2. Model Development: Create a predictive model using the cleaned data.
3. Model Validation: Validate the model using a separate test dataset.
4. Model Deployment: Deploy the model in a hospital setting for real-time predictions.
5. Model Monitoring and Updating: Continuously check and improve the model's performance.
6. Documentation and Reporting: Keep records and share findings with others.

Challenges:

1. Model Generalization: Creating a model that works in different hospitals.
2. Data Privacy: Handling patient data while adhering to privacy laws.

3. Model Interpretability: Making the model's decisions understandable.
4. Real-time Data: Integrating real-time data for timely predictions.
5. Model Monitoring: Continuously checking and adapting the model.
6. Resource Constraints: Working within budget limitations.

Significance:

This project aims to use machine learning to enhance emergency prediction in hospital admissions. By developing an accurate and reliable predictive model, the project seeks to provide healthcare providers with a valuable tool for optimizing resource allocation and improving patient care in emergency situations.

Reference:

1. Vântu A, Vasilescu A, Băicoianu A. Medical emergency department triage data processing using a machine-learning solution. Heliyon. 2023 Jul 22;9(8):e18402. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10412878/

Project 17

Domain: Customer Service and Conversational AI
Techniques: Natural Language Processing (NLP), Large Language Models (LLMs), Sentiment Analysis, Intent Recognition, Topic Modeling
Title: Customer Conversational Intelligence Platform Powered by an LLM Agent

Overview and Problem Statement:

This project aims to develop a state-of-the-art Customer Conversational Intelligence Platform powered by a Large Language Model (LLM) agent. The LLM's advanced language understanding will drive the analysis of customer interactions across diverse channels (chatbots, call centers, email, social media). The platform will extract actionable insights from this data, enabling businesses to optimize customer service processes and significantly enhance the overall customer experience.

Datasets:

Name: Relational Strategies in Customer Interactions (RSiCS)

- Description: This dataset contains a corpus for improving the quality and relational abilities of Intelligent Virtual Agents.
- Link: Link to the dataset

Name: 3K Conversations Dataset for ChatBot from Kaggle

- Description: The dataset includes various types of conversations such as casual or formal discussions, interviews, customer service interactions, and social media conversations.
- Link: Link to the dataset

Name: Customer Support on Twitter Dataset from Kaggle

- Description: This is a large corpus of tweets and replies that can aid in natural language understanding and conversational models.
- Link: Link to the dataset

Challenges:

1. Data Collection: Gather customer conversations from diverse sources like voice calls, chat transcripts, emails, and social media interactions.
2. Use LLM-Agent for:
    a. Sentiment Analysis – accurate detection of customer emotions (positive, negative, neutral) and granular sentiment categories (frustration, satisfaction, inquiry, etc.) throughout conversations.

b. Intent Recognition - understanding the underlying purpose behind customers' queries, enabling tailored responses and resolutions
c. Topic Modeling - discovering recurring themes and patterns within conversations, highlighting trending issues, feedback topics, and potential areas for improvement.
d. Agent Performance Evaluation - Analyzing agent interactions to provide constructive feedback, identifying training needs, and recognizing exceptional service.
e. LLM-Driven Real-time Recommendations - Empowering agents with suggestions for next-best actions or responses during active conversations, optimizing outcomes.

Methodology:

Select GPT2/GPT3, fine-tune the LLM agent extensively on a large dataset of customer conversations annotated for sentiment, intent, topics, etc. Develop ML algorithms to support the LLM agent. The primary focus will be on the LLM's ability to perform sentiment analysis, intent recognition, topic modeling, and agent performance assessment. Utilize platforms like SageMaker or equivalent to automate the ML workflow.

Example:
Customer: Hello, I ordered a laptop from your website, and it's been a week, but I haven't received it yet. Can you help me track my order?

Platform Analysis:

Categorization: Inquiry about order tracking.
Sentiment Analysis: Neutral sentiment.
Resolution Status: Unresolved.
Support Agent: Hi there! I apologize for the delay in your order. Could you please provide me with your order number? I'll check the status for you.

Customer: My order number is 123456789.

Platform Analysis:

Categorization: Providing order information.
Sentiment Analysis: Neutral sentiment.
Resolution Status: In progress.

Support Agent: Thank you for providing the order number. Let me check that for you. [Platform sends a real-time request to the order tracking system]

Platform Analysis:

Real-time Analysis: The platform receives updated order tracking information. The laptop is currently in transit and is expected to arrive in two days.
Support Agent: Good news! Your laptop is on its way and should be delivered within the next two days. Here's your tracking number: ABC123XYZ. You can use this number to monitor its progress.

Customer: Thank you for the information. I appreciate your help.

Platform Analysis:

Sentiment Analysis: Positive sentiment.
Resolution Status: Resolved.
Support Agent: You're welcome! If you have any more questions or need further assistance, feel free to ask. Have a great day!

Significance:

As companies accumulate immense volumes of customer interaction data, the ability to unlock meaningful insights and streamline customer service processes becomes a competitive advantage. The envisioned platform, with its real-time analysis capabilities, has the potential to revolutionize customer service, ultimately translating into greater customer satisfaction, increased operational efficiency, and a strengthened market position for businesses.

Reference:

1. Conversational Health Agents: A Personalized LLM-Powered Agent Framework, Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, Ramesh Jain: https://arxiv.org/html/2310.02374v4
2. Building a Conversational AI Agent with Long-Term Memory Using LangChain and Milvus, Zilliz:
https://medium.com/@zilliz_learn/building-a-conversational-ai-agent-with-long-term-memory-using-langchain-and-milvus-0c4120ad7426

Project 18

Domain: Search Engine Optimization, Generative AI, LLMs
Techniques: Generative pre-trained transformers (GPTs)
Title: Automated SEO using ChatGPT

Overview and Problem Statement:

The goal of this project is to create an automated Search Engine Optimization (SEO) tool using ChatGPT, an AI-based chatbot system. The tool will use natural language processing (NLP) and machine learning (ML) algorithms to analyze website content, identify SEO issues, and provide recommendations for improvement. The tool will help website owners and SEO professionals to optimize their website's content and improve search engine rankings more efficiently and effectively.

Methodology:

1. Design a chatbot interface that can interact with website owners and SEO professionals and collect website information such as website URL, keywords, and target audience.
2. Train the ChatGPT model with SEO-specific data and vocabulary, including industry keywords, search engine algorithms, and best practices for on-page and off-page optimization. *Note: Dataset for this project is not readily available and should be collected/generated*
3. Implement an NLP algorithm to analyze website content, including meta tags, headings, images, and links, and identify SEO issues such as duplicate content, missing tags, and broken links.
4. Develop an ML algorithm that can learn from the SEO analysis data and provide personalized recommendations for optimization based on the website's content and target audience.
5. Integrate the automated SEO tool with popular SEO platforms and tools, such as Google Analytics and Google Search Console, to provide more comprehensive data and insights.

Project Deliverables:

1. An automated SEO tool that uses ChatGPT to analyze website content and provide recommendations for optimization.
2. A user-friendly chatbot interface that can interact with website owners and SEO professionals and collect website information.

Project 19

Domain: Healthcare
Techniques: Natural Language Processing, Generative AI
Title: Fine-Tuning the IndicTrans NMT Model for Healthcare Conversations

Overview and Problem Statement:

This project aims to enhance the performance of the IndicTrans Neural Machine Translation (NMT) model specifically for healthcare conversations between patients and hospital staff (doctors/nurses). The current model has very poor accuracy on medical terminology translations from Indian languages to English. The objective is to fine-tune the model using a dataset of healthcare conversations to improve its accuracy and fluency in translating medical terminology and patient queries from Indian languages to English. By fine-tuning the model on domain-specific data, we aim to optimize its performance for real-world healthcare communication scenarios.

Dataset:

Custom data is required for this project, wherein the patient's statement in a non-English Indian language (eg. Telugu) and its corresponding English translation are generated. These conversations may cover various medical scenarios, including patient inquiries, symptom descriptions, treatment discussions, and medical advice exchanges. [See a related (paid) dataset comprised of a healthcare discussion in Telugu here: https://www.futurebeeai.com/dataset/speech-dataset/healthcare-call-center-conversation-telugu-india#contact-form]. Other possibilities are to create synthetic medical conversation dataset (Telugu (or another Indian languae) and English translation) using an LLM. For an example, see Fig. 1.

Methodology:

The IndicTrans NMT model, a pre-trained machine translation model tailored for Indian languages, will be selected as the base model for fine-tuning.

Dataset Preparation:

The healthcare conversation dataset will be preprocessed and formatted to prepare it for fine-tuning. This involves cleaning the data, tokenizing the text, and organizing it into a suitable format for training.

Fine-Tuning:

The pre-trained IndicTrans NMT model will be fine-tuned on the healthcare conversation dataset.

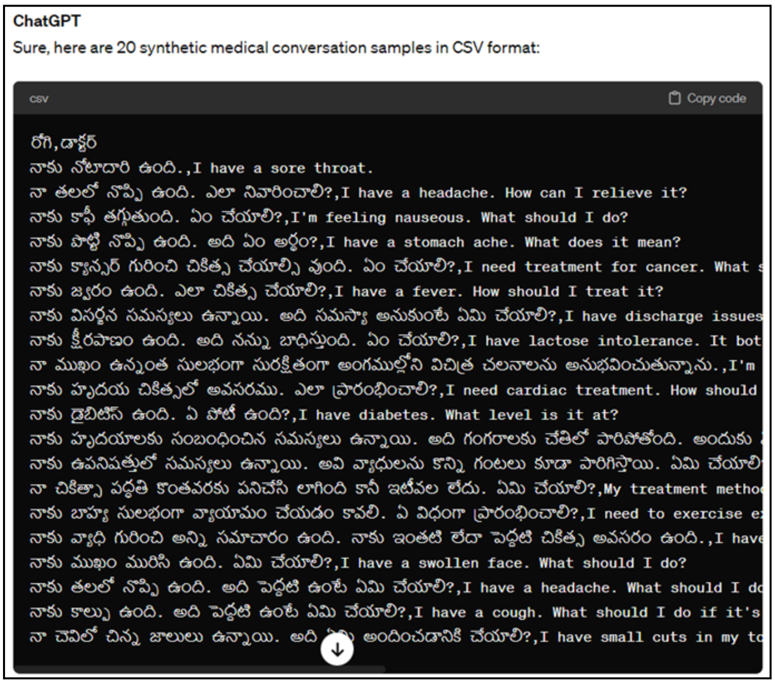
Evaluation:

The fine-tuned model will be evaluated on a separate validation set to assess its performance in translating healthcare conversations accurately and fluently. Evaluation metrics such as BLEU score, accuracy, and fluency will be used to measure the model's effectiveness.

Model Deployment:

Once the fine-tuning process is complete and the model demonstrates satisfactory performance on the validation set, it will be deployed for use as a Whatsapp based app in healthcare communication scenarios.

By fine-tuning the IndicTrans NMT model on healthcare conversation data, this project aims to address the specific challenges and nuances of medical communication, thereby improving the accessibility and quality of healthcare services for Indian language speakers



ChatGPT

Sure, here are 20 synthetic medical conversation samples in CSV format:

```csv
రోగి,డాక్టర్
నాకు నోటాదారి ఉంది.,I have a sore throat.
నా తలలో నొప్పి ఉంది. ఎలా నివారించాలి?,I have a headache. How can I relieve it?
నాకు కాఫీ తగ్గుతుంది. ఏం చేయాలి?,I'm feeling nauseous. What should I do?
నాకు పొట్టి నొప్పి ఉంది. అది ఏం అర్థం?,I have a stomach ache. What does it mean?
నాకు క్యాన్సర్ గురించి చికిత్స చేయాల్సి వుంది. ఏం చేయాలి?,I need treatment for cancer. What s
నాకు జ్వరం ఉంది. ఎలా చికిత్స చేయాలి?,I have a fever. How should I treat it?
నాకు విసర్జన సమస్యలు ఉన్నాయి. అది సమస్యా అనుకుంటే ఏమి చేయాలి?,I have discharge issues
నాకు క్షీరపాణం ఉంది. అది నన్ను బాధిస్తుంది. ఏం చేయాలి?,I have lactose intolerance. It bot
నా ముఖం ఉన్నంత సులభంగా సురక్షితంగా అంగముల్లోని విచిత్ర చలనాలను అనుభవించుతున్నాను.,I'm
నాకు హృదయ చికిత్సలో అవసరము. ఎలా ప్రారంభించాలి?,I need cardiac treatment. How should
నాకు డైబిటిస్ ఉంది. ఏ పోటీ ఉంది?,I have diabetes. What level is it at?
నాకు హృదయాలకు సంబంధించిన సమస్యలు ఉన్నాయి. అది గంగరాలకు చేతిలో పారిపోతోంది. అందుకు
నాకు ఉపనిషత్తులో సమస్యలు ఉన్నాయి. అవి వ్యాధులను కొన్ని గంటలు కూడా పారిగిస్తాయి. ఏమి చేయాలి
నా చికిత్సా పద్ధతి కొంతవరకు పనిచేసి లాగింది కానీ ఇటీవల లేదు. ఏమి చేయాలి?,My treatment metho
నాకు బాహ్య సులభంగా వ్యాయామం చేయడం కావలి. ఏ విధంగా ప్రారంభించాలి?,I need to exercise e
నాకు వ్యాధి గురించి అన్ని సమాచారం ఉంది. నాకు ఇంతటి లేదా పెద్దటి చికిత్స అవసరం ఉంది.,I have
నాకు ముఖం మురిసి ఉంది. ఏమి చేయాలి?,I have a swollen face. What should I do?
నాకు తలలో నొప్పి ఉంది. అది పెద్దటి ఉంటే ఏమి చేయాలి?,I have a headache. What should I do
నాకు కాల్పు ఉంది. అది పెద్దటి ఉంటే ఏమి చేయాలి?,I have a cough. What should I do if it's
నా చెవిలో చిన్న జాలులు ఉన్నాయి. అది [illegible] అందించడానికి చేయాలి?,I have small cuts in my to
```

Fig. 1 Example of synthetic data creation using ChatGPT (Telugu->English medical conversation samples)

Project 20

Domain: Legal Studies
Techniques: Natural Language Processing, Generative AI
Title: AI Patent Advisor: Leveraging Large Language Models for Patent Analysis and Technology Transfer Facilitation

Overview:

With the exponential growth of patent databases, extracting valuable insights and facilitating technology transfer has become increasingly challenging. This project aims to develop an AI-powered advisor that can analyze patents, provide comprehensive summaries, and recommend potential commercial applications and licensing opportunities. By fine-tuning the model on the AI-Growth-Lab patents and claims dataset, sourced from various patent repositories, the AI Patent Advisor will empower inventors, businesses, and legal professionals to navigate the complex landscape of patent law and innovation.

Dataset:

The AI-Growth-Lab patents and claims dataset (https://huggingface.co/datasets/AI-Growth-Lab/patents_claims_1.5m_traim_test) will serve as the foundation for training the AI Patent Advisor. This dataset comprises a vast collection of patents and their associated claims, covering diverse technical fields and industries. With over 1.5 million patent documents, the dataset provides a rich source of information for training and fine-tuning the LLM model. Each patent document includes detailed descriptions, claims, and metadata, enabling comprehensive analysis and understanding of patented technologies.

Methodology:

- Data Collection and Preprocessing: Gather AI-Growth-Lab patents and claims, preprocess to clean noise, standardize formatting, and tokenize text
- Fine-tuning LLM: Utilize state-of-the-art LLM architecture, fine-tune on AI-Growth-Lab dataset for patent-specific language adaptation.
- Analyze patents, extract key concepts, and generate concise summaries for efficient knowledge extraction.
- (Optional) Implement semantic matching algorithms to map patented inventions to potential commercial applications and licensing opportunities.
- Design a user-friendly interface enabling users to input patents, explore summaries, and receive technology transfer recommendations.

- (Optional) Testing and Validation: Assess AI Patent Advisor's performance, including accuracy of summaries, relevance of transfer recommendations, and overall usability.
- (Optional) Deployment and Maintenance: Deploy the AI Patent Advisor, ensuring scalability and reliability, with protocols for regular maintenance and updates to align with evolving patent language and industry trends.

Challenges:

- Handling the complexity and variability of patent language and terminology; Ensuring the accuracy and relevance of patent summaries and technology transfer recommendations

Significance:

The AI Patent Advisor project holds immense potential to transform the landscape of patent analysis and technology transfer facilitation: It will enable inventors, businesses, and legal professionals to efficiently analyze patent documents and extract valuable insights; bridge the gap between patented inventions and commercial applications by identifying potential licensing opportunities and strategic partnerships Overall it will promoter innovation by accelerating the pace of innovation by facilitating knowledge dissemination and collaboration within the patent ecosystem.