Section 1 : Topic Submission Form

This form should be submitted by the mentioned deadline.

Name: Sanjay Saini

Student Number: 1096513

Course: LJMU Masters in Machine Learning and AI July 2023

Fill your topic/s below.

Project Title/Area 1:

Hybrid Model of Bidirectional Long-Short Term Memory and CNN for Multivariate Time Series Classification for ecommerce sales Forecasting

Dataset: https://www.kaggle.com/code/ryanholbrook/hybrid-models/data

The dataset contains the sales for the thousands of product families sold at Favorita stores located in Ecuador. The training data includes dates, store and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models.

Description:

By fusing the strengths of Bidirectional Long-Short Term Memory (Bi-LSTM) and Convolutional Neural Networks (CNN), this hybrid model aims to capture both temporal sequences and intricate patterns in multivariate time series data. The bidirectional nature of LSTM ensures the model comprehends past and future data points, enhancing its predictive accuracy. Meanwhile, the CNN component extracts spatial features and detects patterns from the time series data.

When applied to ecommerce, this model can provide accurate sales forecasts, enabling businesses to optimize inventory management, marketing strategies, and overall operational efficiency.

It will address the growing need for accurate ecommerce sales forecasting in an increasingly digitalized economy. By enhancing prediction accuracy, businesses can optimize supply

chain operations, reduce waste, and ensure product availability, leading to improved customer satisfaction. Furthermore, accurate sales forecasts will help small and medium enterprises (SMEs) compete effectively, fostering economic growth and job creation. In a broader context, as ecommerce becomes a significant part of global commerce.

Implementation approach:

- In this project will use the dataset of ecommerce sales, including multivariate time series data.
- Cleaning of the data to handle missing values, outliers, and any inconsistencies. Normalize or standardize the data.
- Design the hybrid model by integrating Bi-LSTM, CNN layers, number of layers, nodes, and other hyperparameters based on preliminary tests.
- Divide training, validation, and test sets.
- Train the hybrid model using the training set. Monitor the model's performance on the validation set to prevent overfitting and adjust hyperparameters.
- Evaluate its performance and use metrics MAE, RMSE, and others.
- Compare the model's performance with traditional forecasting models (like ARIMA) and standalone deep learning models (like LSTM or CNN alone).

Project Title/Area 2:

Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction and sentiment analysis from news feeds

Dataset: https://www.bseindia.com/Indices/IndexArchiveData.html

This is the official website of BSE (formerly Bombay Stock Exchange) where we can download historical data related to various indices

News Dataset for Sentiments: https://www.alphavantage.co/documentation/#intelligence

This API returns live and historical market news & sentiment data from a large & growing selection of premier news outlets around the world, covering stocks, cryptocurrencies, forex, and a wide range of topics such as fiscal policy, mergers & acquisitions, IPOs, etc. This API, combined with our core stock API, fundamental data, and technical indicator APIs, can provide you with a 360-degree view of the financial market and the broader economy

Description:

This project will excavate the development and evaluation of a hybrid predictive model that combines the strengths of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and the XGBoost algorithm for stock market forecasting. The

primary objective is to harness the power of deep learning and ensemble learning to predict stock prices and analyse sentiments from news feeds.

By integrating attention mechanisms with CNN, LSTM, and XGBoost, this project aims to create a more accurate and robust stock prediction model. The inclusion of sentiment analysis further enhances the model's ability to factor in the impact of news on stock prices. This project is at the intersection of deep learning, ensemble learning, and natural language processing, aiming to provide more accurate stock predictions by considering both historical stock data and the sentiment derived from related news articles.

The project will help in many ways to the society like Informed Investment Decisions, Financial Stability, Enhanced Trust in Financial Markets, Empowerment of Small Investors, Economic Growth, Sentiment-Driven Insights, Transparency & Accountability and Educational Value. The utilization of advanced predictive models in stock forecasting, combined with sentiment analysis, can lead to a more informed, stable, and inclusive financial ecosystem. This not only benefits individual investors but also contributes to the broader economic well-being of society.

Implementation approach:

- This project will use the dataset of the BSE index (Sensex) for past years from bseindia.com and also use the news data related to the BSE index from alphavantage.co for the same period.
- Cleaning of the data to handle missing values, outliers, and any inconsistencies. Normalize or standardize the data.
- Use NLP techniques to analyse the sentiments of new articles and convert them into numerical data (Positive, Neutral & Negative)
- Design the hybrid model by CNN, LSTM, and attention mechanisms. Integrate the XGBoost algorithm to refine and enhance the predictions from the deep learning model.
- Divide training, validation, and test sets.
- Train the hybrid model using the training set. Monitor the model's performance on the validation set to prevent overfitting and adjust/tune hyperparameters.
- Evaluate its performance and gauge its accuracy, precision, recall, and other relevant metrics in stock prediction.
- Compare the model's performance against traditional stock prediction models or benchmarks to determine its advantages and efficacy.
- Refine the model based on insights gained, potentially incorporating additional data, or adjusting the model architecture.

Project Title/Area 3:

Analyzing the Impact of Data Distribution on Knowledge Base Question Answering Systems

Dataset: https://dki-lab.github.io/GrailQA/

Strongly Generalizable Question Answering (GrailQA) is a new large-scale, high-quality dataset for question answering on knowledge bases (KBQA) on Freebase with 64,331 questions annotated with both answers and corresponding logical forms in different syntax (i.e., SPARQL, S-expression, etc.). It can be used to test three levels of generalization in KBQA: i.i.d., compositional, and zero-shot.

Description:

In the AI and NLP realm, Knowledge Base Question Answering (KBQA) systems are designed to provide precise answers to user queries by leveraging structured knowledge bases. The project aims to investigate how the distribution of data affects the performance and accuracy of KBQA systems. Data distribution refers to the way data is organized, spread, and accessed within the knowledge base. Factors such as data sparsity, imbalance, and granularity can influence the system's ability to understand and respond to queries accurately.

The project can revolutionize information retrieval, making it more precise and user-friendly. Enhanced systems can aid in better decision-making, reduce misinformation, and promote accessibility. This research can drive economic growth, improve healthcare, and foster educational advancements. Overall, it holds the potential to benefit various sectors of society, elevating the quality of life.

Implementation approach:

- Define clear research objectives and hypotheses.
- Collect diverse datasets varying in distribution, sparsity, and granularity.
- Preprocess data to ensure quality and consistency.
- Choose a standard KBQA system and define evaluation metrics.
- Establish a performance baseline using a balanced dataset.
- Systematically alter data distribution and test the KBQA system.
- Use statistical tools to analyze results and determine significance.
- Propose strategies to optimize KBQA performance based on findings.

Fill in this section if a member of staff has agreed to be your supervisor:

Member of Staff:_____

If you have found a supervisor, then you and the member of staff who agreed to supervise your project should sign below.

Sanjay Saini	**Shubham Gupta**
_____	_____
Student Signature	Supervisor Signature
10-Oct-2023	**10-Oct-2023**
Date	Date

Section 2 : Topic Selection Research

Table 1 : Topic 1

Title	Link to the Paper	Understanding of the Dataset	Understanding the Methodology Used	Dataset Link
Time-series forecasting of seasonal items sales using machine learning – A comparative analysis	https://www.sciencedirect.com/science/article/pii/S26670968220000027?ref=pdf_download&fr=RR-2&rr=81266970e8b48adc	Superstore sales dataset is used, which contains sales information of furniture, technology goods, and office supplies from 2014 to the end of 2017. contains nearly 10,000 data points and 21 features. It displays seasonality in its sales pattern and does not contain any missing values.	Stacked LSTM method is found to be the best performing algorithm for furniture sales prediction, followed by Prophet and CNN. Stacked LSTM method is found to be the best performing algorithm for furniture sales prediction	https://community.tableau.com/s/global-search/%40uri#q=Superstore%20sales&t=All
Prediction of Soybean Price Trend via a Synthesis Method With Multistage Model	https://www.semanticscholar.org/paper/6255aeeee5416b63fb351369780b658a059eef5b	1. data set consists of soybean prices and is clustered into four patterns using TICC. 2. Uses a multivariate time series data set composed of four variables: soybean purchase price, corn market price, soybean futures price, and soybean oil futures price	1. uses Toeplitz inverse covariance-based clustering (TICC) to cluster the soybean prices. 2. Used long short-term memory (LSTM) to forecast the prices and multivariate long short-term memory (MLSTM) to classify the risk levels.	Not provided
A Deep Stacked Bidirectional LSTM (SBiLSTM) Model for Petroleum Production Forecasting	https://www.sciencedirect.com/science/article/pii/S187705092300248X	1. It uses two oilfield production time series datasets, one from the Huabei-China oilfield and another from the Cambay Rift Basin, India oil-fields. 2. both datasets is used to evaluate the performance of the proposed Stacked Bi-LSTM (SBiLSTM) model in petroleum production forecasting.	1. The empirical results shows proposed Stacked Bi-LSTM (SBiLSTM) model outperforms other standard approaches, including RNNs, multi-layer RNNs, DGRU, and DLSTM models. 2. SBiLSTM model able to acquire long and short-range interdependent features of univariate time series data without requiring large memory. 3. It highlights that the proposed DLSTM model, which is a part of the SBiLSTM architecture, performs better than other approaches based on different	https://www.ndrdgh.gov.in/NDR/?page_id=629 https://www.globaldata.com/data-insights/listing/search/?q[]=1st%20block%20of

			measurement criteria. 4.The prime contribution of It is the introduction of the SBiLSTM architecture, which is an adaptation of the traditional deep LSTM model, and its superior performance in petroleum production forecasting	%20the%20Huabei-China%20oilfield
Diabetes Prediction Using Bi-directional Long Short-Term Memory	https://link.springer.com/article/10.1007/s42979-023-01831-z	Aims to reduce diabetic patients' pain and improve their lives by developing an expert system for diabetes mellitus classification using a bi-directional long short-term memory (BLSTM) model. The proposed methodology achieves better accuracy, sensitivity, specificity, and F1 score compared to earlier studies in classifying diabetic patients.	BLSTM model achieved better accuracy, sensitivity, specificity, and F1 score compared to earlier studies in classifying diabetic patients	https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
Bidirectional LSTM-RNN-based hybrid deep learning frameworks for univariate time series classification	https://link.springer.com/article/10.1007/s11227-020-03560-z	1. Datasets are already z-normalized, with zero mean and a unit of standard deviation. 2. The number of classes ranges from 2 to 60. and Sequence length varies between 24 and 2,709 observations. 3. The datasets include various types of collected sources such as Image, Sensor, Motion, Spectro, Device, ElectroCardioGram (ECG), and Simulated.	After evaluating the performance over 85 datasets from different domains, the proposed models consistently outperform existing state-of-the-art techniques.	https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018/

Table 2 : Topic 2

Title	Link to the Paper	Understanding of the Dataset	Understanding the Methodology Used	Dataset Link
Attention-based CNN-LSTM and XGBoost	https://arxiv.org/pdf/2204.02623v2.pdf	data used in the study comes from the open and free public dataset in Tushare, which provides basic market data of	The hybrid model combines the time series model ARIMA, Convolutional Neural Networks with Attention mechanism, Long	Not mentioned

hybrid model for stock prediction		stocks including opening price, closing price, highest price, lowest price, trading volume, and trading amount. The ARIMA-processing sequence and residual sequence are also included as features	Short-Term Memory network, and XGBoost regressor	
A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model	https://www.mdpi.com/2227-7390/11/3/590	daily transaction data of the Shanghai Composite Index (000001) is used as the dataset, covering the period from 3 July 1997 to 24 January 2022. The training set contains 4761 trading days data, while the validation and testing set contains 1190 trading days data	proposed multi-parameter forecasting model using LSTM and deep learning outperforms existing methods in predicting stock prices. Improves the accuracy of close price and high price forecasts, providing investors with valuable information for making profitable decisions.	https://www.statista.com/statistics/410726/sse-composite-index-performance/
Predicting Stock Market time-series data using CNN-LSTM Neural Network model	https://arxiv.org/pdf/2305.14378v1.pdf	stock market datasets, including Shanghai A-Share composite index, Sinopec, TOPIX (Japan), NASDAQ, NSE, BOVA11, BBDC4, ITUB4, CIEL3, PETR4, and NYSE	CNN-LSTM model achieves high accuracy even with real-time stock market data.	Multiple datasets from Kaggle, finance APIs like Yahoo Finance and Alpha Vantage, and also used Google Sheets
A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News	https://arxiv.org/pdf/2007.12620v1.pdf	S&P 500 Index 1. Data used in the research includes adjusted closing stock prices and news sentiment compound scores. 2,. News data is obtained from prominent financial news organizations such as CNBC, Reuters, WSJ, and Fortune. 3. Stock data is the S&P 500 Index, which represents the performance of the 500 largest	Blending ensemble deep learning model outperforms the best existing prediction model.	Stock data used is the S&P 500 Index, data range from December 2017 up to the end of June 2018

		publicly traded companies in the US		
A Hybrid Model of Bidirectional Long-Short Term Memory and CNN for Multivariate Time Series Classification of Remote Sensing Data	https://www.semanticscholar.org/reader/90591939f8a40b915677913ec296952b2fe181fe	data used in this study is multivariate time series data of Landsat 8 satellite images, specifically for land cover classification.	The proposed Conv-BiLSTM model demonstrates superior performance compared to other classifiers such as BiLSTM, CNN, and RF in terms of precision, recall, F-score, and classification accuracy.	https://sites.google.com/site/dinoienco/tiselac-time-series-land-cover-classification-challenge

Table 3 : Topic 3

Title	Link to the Paper	Understanding of the Dataset	Understanding the Methodology Used	Dataset Link
A Flexible and Efficient Framework for Knowledge Base Question Answering	https://aclanthology.org/2021.acl-demo.39.pdf	GrailQA dataset, which is used for evaluation. GrailQA dataset is used for evaluation and includes examples with semantically equivalent and inconsistent schema items	checker module in ReTraCk significantly improves performance, as removing it leads to a drop in F1 points on GrailQA and WebQSP datasets. ReTraCk demonstrates high efficiency, processing queries in an average of 1.62 seconds per query	https://dki-lab.github.io/GrailQA/
A Sequential Flow Control Framework for Multi-hop Knowledge Base Question Answering	https://aclanthology.org/2022.emnlp-main.578.pdf	two datasets for evaluation: MetaQA and WebQSP . MetaQA is a large-scale dataset of multi-hop KBQA with over 400k questions generated using templates and has up to 3 hops. It includes a knowledge graph from the movie domain with 43k entities, 9 predicates, and 135k triples . WebQSP is a subset of WebQuestions and consists of 4,737 questions based on Freebase. It contains both training and test sets, with questions that can be solved under 1 or 2 hops of reasoning. The KB for WebQSP is pruned to include only mentioned relations and within 2-hop triples of mentioned entities	GFC framework achieves new state-of-the-art performance on WebQSP and is effective even when the knowledge base is incomplete	https://paperswithcode.com/dataset/metaqa https://www.microsoft.com/en-us/download/details.aspx?id=52763

Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases	https://dl.acm.org/doi/10.1145/3442381.3449992	https://dki-lab.github.io/GrailQA/ GRAILQA dataset with 64,331 questions, diverse characteristics, and high quality	presents a systematic study on three levels of generalization for KBQA and emphasizes their importance for practical KBQA systems. It also highlights the effectiveness of pre-trained contextual embeddings in generalization. Fine-grained analyses are provided, suggesting potential areas for improvement	https://dki-lab.github.io/GrailQA/
Case-Based Reasoning for Natural Language Queries over Knowledge Bases	https://arxiv.org/pdf/2104.08762.pdf	The paper mentions using KBQA datasets that contain complex questions, such as the COMPLEXWEBQUESTIONS dataset	CBR-KBQA outperforms the current state of the art by 11% on accuracy on the COMPLEXWEBQUESTIONS dataset	http://docs.deeppavlov.ai/en/master/features/models/kbqa.html https://allenai.org/data/complexwebquestions
Semantic Parsing for Knowledge Graph Question Answering with Large Language Models	https://2023.eswc-conferences.org/wp-content/uploads/2023/05/paper_Banerjee_2023_Semantic.pdf	Primarily focus on datasets with more than 5,000 questions, including LC-QuAD 2.0. T5 and BART, the two most popular T2TLMs, are tested and fine-tuned on selected datasets	Best metric for evaluating semantic parsing to logical forms remains an open question. Execution-based metrics such as F1-score can be used to evaluate grounded queries. Existing semantic parsing and KGQA systems are used for comparison	https://paperswithcode.com/dataset/lc-quad-2-0