**Section 1: Topic Submission Form**

This form should be submitted by the mentioned deadline.

Name: **\_\_\_\_\_\_\_\_HARISH BEKANAHALLI NANJUNDAPPA\_\_\_\_\_\_\_\_**

Student Number: **1080274\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Course:  **Masters in ML and AI**             \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Fill your topic/s below.**

Project Title/Area 1: Large Language Models for Efficient and Effective Text Summarization

Dataset:  CNN DailyMail Dataset from HuggingFace portal

<https://huggingface.co/datasets/cnn_dailymail>

Description: Leveraging pretrained language models has emerged as a transformative approach for text summarization. This research aims to explore and enhance text summarization techniques using state-of-the-art pretrained language models, including BERT, GPT, T5, and Pegasus.

Study aims to explore extractive, abstractive and hybrid approach for text summarization.

Subjected to time and cost constraints few techniques and language models will be selected for exploration and fine tuning the models to generate concise, contextually relevant summaries.

News data from CNN or BBC will be used as primary data source for exploration. Later, if time permits this experiment can be extended to other datasets, domains, multi-modal content, low resource languages. ROUGE-\* metrics will be used as evaluation metrics for quantifying the quality, coherence, and informativeness of generated summaries.

Project Title/Area 2: Enhancing Stock Market Predictions with Deep Neural Networks: A Time Series Analysis

Dataset: National Stock Exchange (NSE) India Stock Data (1990-2021)

<https://www.kaggle.com/datasets/stoicstatic/india-stock-data-nse-1990-2020>

Description: This research aims to leverage deep learning techniques for a comprehensive and advanced analysis of the stock market. By employing deep learning architectures, the research aims to enhance the accuracy, efficiency, and predictive power of stock market models. Some of the DL techniques could involve models like RNN, LSTM, BiLSTM, GRU, FB Prophet, Google TFT. Initial research will be focused on univariate analysis and predicting few selected companies stock prices or overall market stock index for a short to medium term horizon. Later, if time permits would like to expand the research on multivariate analysis, attention mechanism, Hybrid models and predictions on other stock commodities like gold, crude oil, etc. Also, it would be interesting to evaluate the efficacy of prediction model across different stock markets like Nasdaq, FTSE, NYSE and EuroNext.

Project Title/Area 3: Deep Learning Approaches for Cyberbully Detection in Social Media.

Dataset: Social media platforms Twitter, Wikipedia Talk pages and YouTube dataset <https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset>

Description: By harnessing the capabilities of deep learning and natural language processing, this research seeks to address the pressing issue of cyberbullying on social media platforms. The proposed methodologies and experiments aim to enhance the accuracy, efficiency, and real-time capabilities of cyberbullying detection models, fostering a positive and secure online experience for users.

Initially, Twitter data will be used as primary datasource for this research. Intention is to explore SOTA DL, NLP, Transformers and Hybrid models. Later, if time permits would like to explore these models on other social media data like Youtube and Facebook comments.

**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.

HARISH BEKANAHALLI NANJUNDAPPA       MOUNIKA MARREDDY\_

Student Signature                                                                         Supervisor Signature

29-12-2023                                                                             29-12-2023

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** |
| A Comparative Study on Transformer-based News Summarization | <https://ieeexplore.ieee.org/document/10099798> | BBC News dataset  Columns: Title, PubDate, Guid, Link, Description  Size = 9.02 MB  Rows = 25,962 | Analysis of news data using transformer-based models for extractive and abstractive summarization  Following models are explored in this paper:   * BART, * T5, * PEGASUS,   ROGUE-1, ROGUE-2, ROGUE-L metrics were used for evaluation. | <https://www.kaggle.com/datasets/gpreda/bbc-news> |
| Automated News Summarization Using Transformers | <https://arxiv.org/abs/2108.01064> | BBC News dataset  Columns: Title, PubDate, Guid, Link and Description.  Size: 9.02 MB  Rows: 25,962 | Analysis of news data using transformer-based models for extractive and abstractive summarization  Following models are explored in this paper:   * BERT, * GPT-2, * XL Net, * BART, * T5   ROGUE-1, ROGUE-2, ROGUE-L metrics were used for evaluation. | <https://www.kaggle.com/datasets/gpreda/bbc-news> |
| Text Summarization using Transformer Model | <https://ieeexplore.ieee.org/document/10062698> | UCI drug review dataset  Columns: Unique Id, Drug Name, Condition, Review, Rating, Date, Useful Count  Size:110.63 MB  Rows: 232,000  BBC News Dataset  Columns: Title, PubDate, Guid, Link and Description.  Size: 9.02 MB  Rows: 25,962 | Study proposes abstractive text summarization method based on the Text-to- Text Transfer Transformer (T5) model.  This model was fine-tuned and evaluated on BBC news dataset.  ROUGE-1, ROUGE-2, ROUGE-L metrics are used for evaluation | Dataset 1 link:  <https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018>  Dataset 2 link:  <https://www.kaggle.com/datasets/gpreda/bbc-news> |
| A Hybrid Extractive-Abstractive Framework with Pre & Post-Processing Techniques to Enhance Text Summarization | <https://ieeexplore.ieee.org/document/10275584> | Samsum dataset  Columns: Id, Dialogue, Summary  Size: 2.94 MB  Rows: 16,369 | The TextRank algorithm is used first for extractive summarization followed by T5-small pre trained model for abstractive summarization. Finally, Levenstein algorithm is used to find the matching sentence that will replace the missing sentence in the previous model output.  ROUGE-1, ROUGE-2, ROUGE-L metrics are used for evaluation | <https://huggingface.co/datasets/samsum> |
| LSTM Encoder Decoder Based Text Highlight Abstraction Method Using Summaries Extracted by PageRank | <https://ieeexplore.ieee.org/document/10078652> | BBC News Dataset  Columns: Title, PubDate, Guid, Link and Description.  Size: 9.02 MB  Rows: 25,962  CNN Daily News dataset  Size: 1.76 GB  Rows: 935,913 | Hybrid model for text summarization. Initially PageRanker model performs extractive summarization then Encoder-Decoder (LSTM) model performs abstractive summarization.  ROUGE-1, ROUGE-2, ROUGE-L metrics are used for evaluation | Dataset 1 link:  <https://www.kaggle.com/datasets/gpreda/bbc-news>  Dataset 2 link:  <https://huggingface.co/datasets/cnn_dailymail> |

**Table 2: Topic 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Link to the Paper** | **Understanding of the Dataset** | **Understanding the Methodology Used** | **Dataset Link** |
| Stock Market Analysis and Prediction for Nifty50 using LSTM Deep Learning Approach | <https://ieeexplore.ieee.org/abstract/document/9754148> | 10-year historical data from the National Stock Exchange (NSE) for NIFTY 50 equities for the period 10-Dec-2011 to 10-Dec-2021  Columns: Date, Symbol, Series, Prev Close, Open, High, Low, Last, Close  Size: 54.36 MB | The DL based LSTM Algorithm is proposed in this study.  RMSE, MSE, MAE, MAPE metrics used for evaluation | <https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data> |
| Stock Market Prediction Using LSTM Recurrent Neural Network | <https://www.sciencedirect.com/science/article/pii/S1877050920304865> | New York Stock Exchange (NYSE)  GOOGL and NIKE stocks price data extracted from yahoo finance website.  Columns: Date, Open, High, Low, Close, Adj Close, Volume | RNN based LSTM model is proposed to predict future values for Google and Nike company stock.  Metrics like processing time / sec for different epochs and Loss is considered for evaluation | Dataset 1 link:  <https://finance.yahoo.com/quote/GOOG/history/>  (Date Filter - 8/19/2004 to 12/19/2019)  Dataset 2 link:  <https://finance.yahoo.com/quote/NKE/history?p=NKE>  (Date Filter - 1/4/2010 to 12/19/2019) |
| Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models | <https://arxiv.org/abs/2009.10819v1> | NIFTY 50 index values of the National Stock Exchange (NSE) of India, during the period December 29, 2014, till July 31, 2020  Columns: Date, Symbol, Series, Prev Close, Open, High, Low, Last, Close  Size: 54.36 MB | LSTM based univariate model is used to predict the stock price.  Grid search technique is used to optimise the LSTM model.  Other 8 ML based Regression models are used to compare the performance.  RMSE used for evaluation | Dataset 1 link:  <https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data>  Dataset 2 link:  <https://www.niftyindices.com/reports/historical-data> |
| A novel ensemble deep learning model for stock prediction based on stock prices and news | <https://link.springer.com/article/10.1007/s41060-021-00279-9> | S&P 500 Index dataset  US Financial News from following:   * Bloomberg.com, * CNBC.com, * reuters.com, * wsj.com, * fortune.com   Columns: date, open, high, low, close, volume, Name  Size: 29.58 MB | Ensemble learning method to combine two RNN networks, followed by a fully connected neural network.  VADER model for sentiment rating  RMSE, F1, Recall metrics for evaluation | Dataset 1 link:  <https://www.kaggle.com/datasets/camnugent/sandp500>  Dataset 2 link:  <https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles> |
| A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing | https://arxiv.org/abs/1912.07700v1 | National Stock Exchange (NSE) NIFTY 50 index daily values are collected from Jan 2015 till Dec 2017 for training various ML and DL Models  Columns: Date, Symbol, Series, Prev Close, Open, High, Low, Last, Close  Size: 54.36 MB | LSTM - based deep learning network for predicting the closing price of the NIFTY stock.  Twitter sentiment analysis model is added to augment the LSTM model to correlate the public sentiment of stock price with market sentiment.  To compare the LSTM model performance 7 other ML models are trained for sentiment analysis and Regression.  MAPE and correlation metrics are used for evaluation | https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data |
| Deep Stock Predictions | <https://arxiv.org/pdf/2006.04992v1.pdf> | Nasdaq Stock Exchange prices for following automobile manufacturers:   * Ford * GM * Toyota * Tesla   Columns: Date, Low, Open, Volume, High, Close, Adjusted Close  Size: 10.23 GB | BiLSTM with Attention model is proposed to predict the next day stock price.  Arima model is proposed to compare the results with the BiLSTM model.  MSE metrics is used for evaluation | <https://www.kaggle.com/datasets/paultimothymooney/stock-market-data> |

**Table 3: Topic 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Link to the Paper** | **Understanding of the Dataset** | **Understanding the Methodology Used** | **Dataset Link** |
| Cyberbullying detection solutions based on deep learning architectures | <https://link.springer.com/article/10.1007/s00530-020-00701-5> | The data is from different social media platforms like Twitter, Wikipedia Talk pages and YouTube. The data contain text and labelled as bullying or not. The data contains different types of cyber-bullying like hate speech, aggression, insults and toxicity.  Dataset 1 Size: 179.8 MB  Dataset 2 Size: 7.17 MB | In this article following 4 DL models are used BiLSTM, GRU, LSTM and RNN.  Confusion Matrix, Precision, Recall, Accuracy, F1 score are used for evaluation.  Word cloud visualization is used for displaying the most frequently used words | Dataset 1 link:  <https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset>  Dataset 2 link:  <https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification> |
| Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques | <https://www.mdpi.com/2079-9292/10/22/2810> | Two datasets are used in this paper:  1. Wikipedia Talk dataset  Columns: rev\_id, worker\_id, quoting\_attack, recipient\_attack, third\_party\_attack, other\_attack, attack  Size: 105 MB  2. Jigsaw toxic data set  Columns: id, comment\_text, toxic, severe\_toxic, obscene, threat, insult, identity\_hate  Size: 55.18 MB | This paper proposes a novel neural network framework with parameter optimization.  Several ML/ DL models are used for comparison.  XG Boost, Naïve Bayes, SVM, LR  CNN, LSTM, GRU, BiLSTM, BiGRU, CNN-BiLSTM, Att- BiLSTM  F1 Score used for evaluation | Dataset 1 link:  <https://www.kaggle.com/datasets/jigsaw-team/wikipedia-talk-labels-personal-attacks?select=attack_annotations.csv>  Dataset 2 link:  <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data?select=train.csv.zip> |
| Detecting Hate Speech with GPT-3 \* | [[2103.12407] Detecting Hate Speech with GPT-3 (arxiv.org)](https://arxiv.org/abs/2103.12407) | ETHOS: onlinE haTe speecH detectiOn dataSet. This repository contains a dataset for hate speech detection on social media platforms, called Ethos.  Rows: 1431 | GPT-3 model is to identify sexist and racist text passages with zero-, one-, and few-shot learning.  Research shows that with zero- and one-shot learning, GPT-3 can reach an average accuracy between 55 per cent and 67 per cent.  With few-shot learning, the model’s accuracy can be as high as 85 per cent.  Accuracy, Precision, Recall, F1 Score is used for evaluation | <https://huggingface.co/datasets/ethos> |
| Cyberbullying Detection: An Ensemble Based Machine Learning Approach | <https://ieeexplore.ieee.org/abstract/document/9388499> | Benchmarked Twitter dataset taken from previous research paper. Dataset file contains tweet message and class representing labelled as offensive or non-offensive.  Size: 2.43 MB | In this article single and double ensemble-based voting model is used for classification.  Four ML models like MNB, LR, DT, LSVC and three Ensemble strategies like GBoost, ADB and Bagging are explored.  Accuracy, F1 Score and AUC metrics are used for evaluation | <https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data> |
| Enhancing Cyber Bullying Detection Using Convolutional Neural Network | <https://ieeexplore.ieee.org/document/10276007> | Sample dataset from Twitter contains actual tweet message and class that labels the cyber bullying type.  The data contains different types of cyber-bullying harassment based on religion, age and others.  Size:7.17 MB  Rows: 46,017 | CNN based Deep learning model is used to identify the cyberbullying in the Twitter message.  NLP techniques are used to prepare the dataset for training.  Twitter API is used to fetch new tweets.  Precision, Accuracy, F1 Score metrics are used for evaluation | <https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification> |