NEWS RECOMMENDATION SYSTEM USING DEEP LEARNING METHOD

BY
TAMANNA JAHAN
ID: 242-25-008

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Science and Engineering

Supervised by

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH

SEPTEMBER 13, 2025

APPROVAL

This Thesis titled "**News Recommendation System Using Deep Learning Method**," submitted by **Tamanna Jahan, ID: 242-25-008** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13-09-2025.**

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of CSE, FSIT
Daffodil International University

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE, FSIT
Daffodil International University

Dr. Naznin Sultana
Associate Professor
Department of CSE, FSIT
Daffodil International University

Mr. Nazibur Rahman
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. Abdus Sattar, Associate Professor, Department of CSE,** Department of Computer Science and Engineering, Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Tamanna Jahan

Tamanna Jahan
ID: 242-25-008
Department of CSE
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. I am deeply grateful to everyone who has assisted me in one way or another.

First, I express my heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for me to complete the **Final Year Design Project (FYDP)** successfully.

I am grateful and wish my profound indebtedness to **Dr. Abdus Sattar, Associate Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of my supervisor in the field of Machine Learning and Deep Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the Department of CSE,** for his kind help in finishing my project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank my entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

The impressive amount of news information online, which is reflected in digital age, is a great problem affecting the provision of useful news to the readers effectively. The urgency of having intelligent systems that are able to filter and suggest news depending on the contents and not on the actions of the user is more eminent than ever. The study is dedicated to the development and deployment of a news recommendation system based on the Natural Language Processing (NLP) techniques to increase personalization and contextual relevance of news presentation. Raw textual information (news headlines descriptions) goes through the system, undergoes further sophisticated text preprocessing procedures and gets transformed to structured numeric formats ready to be used in the machine learning area. With deep learning structures, the extraction of semantic characteristics in images allows predicting the right categories, and this is how news is properly recommended regardless of the need to log in or track users. Personalized recommendations based on content-based analysis and sound feature extraction are provided in a manner that the privacy of the user is not compromised. This paper assists in the formation of ethical and effective news recommendation systems, and it proves how NLP and deep learning can be combined to manage the real-world information overloading. The proposed system is not only scalable, but also translates to other languages and other fields; the system has prospects of future development using user interaction and sentiment analysis capabilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Within the context of this thesis, I suggest the implementation of a news recommendation system using NLP as a technique to preprocess and deep learning models as a method to perform classification and recommendations. The development of the system is on a curated data of news articles in Bangladesh obtained by scraping one of most popular English-language newspapers in the country The Daily Star. In each article, a metadata is provided in form of headline, description, category, and link. This diffuse and rich data that is available allows the system to learn many different things ranging in far-flung areas such as politics and health to technology and environment. As opposed to conventional recommendation engines, the proposed one employs deep learning methods including Convolutional Neural Networks (CNN) and CNN-LSTM hybrids to extract both spatial and temporal arrays of the news data. The hybrid model takes into account the modeling of sequential dependencies in text (tracking the flow of information in headlines or description of articles), which is not possible in CNN as it can only capture the local semantic patterns. This study will be able to improve the accuracy and reliability of news recommendations by understanding the difference between user intent and news content and therefore filling that gap. In doing so, the system does not only enhance user satisfaction but also acts to minimize misinformation and redundancy of contents. Moreover, this project is also of particular interest as a project within the scope of the news analysis in Bangladesh, since it facilitates the development and usage of the latest AI technologies in the analysis of local news, which is a rather unexplored area of research in this case. Altogether, this introduction preconditions a research initiative that solves an actual-world issue with the help of the latest advances in AI tools. Such a combination of deep learning and NLP in recommending news has a great potential in solving the problems of content personalization, and the results of this project may find their practical applications in online journalism, media platforms, and e-learning tools.

1.2 Motivation

Due to the exponential nature of the internet information, especially in news field, individuals have changed the nature of content consumption. The advent of smartphones and the availability of internet has increased the usage of the news via web portals, mobile applications and social sites. This accessibility however poses a serious challenge, that is information overload. The abundance of large numbers of articles sometimes makes the reader feel overwhelmed and then there is a blind search to find information that is trusted, relevant or even customized to them. This not only causes poor user experience but also leads to the decline in content interactions and retention. In addition, Bangladesh is a culturally rich and a linguistically diverse nation and international models of recommendation systems may not reflect local context and elements of news. The existing news recommendation systems are largely generalized towards the Western media or use the static keyword-matching methods that do not respond to the evolving interest of users or dynamic events. The absence of the context-awareness and bad localization lead to low-accuracy recommendations and a variety of lost opportunities to interest the users. Lastly, this initiative is being driven by an individual desire to play a part in AI-driven journalism and content technology. Such a thesis does not simply meet academic requirements, but will also be a real-life prototype, which can be scaled to being full-scale systems implemented by digital media companies, schools, and mobile apps.

1.3 Objectives

The thesis has the following main objectives:

- Building a news suggestion system that will have categorization and suggestions of the news through NLP and deep learning methods.
- In order to pre-process and prepare a locally relevant dataset of Bangladesh news articles written in The Daily Star.
- To illustrate the application of the developed system in real life situation to make the news delivery personalized to the users.

1.4 Methodology

The project methodology entails a logical sequence of steps which include:

Data Gathering:

- The collection of the data set was based on news found in one of the local news providers in Bangladesh called The Daily Star with some of the typical columns having the headline, description, category and link of the Web pages containing articles.

- Data Preprocessing: Text data were processed by cleaning, tokenizing, removing stop words, as well as stemming/lemmatization. NLP tools were used to recommend word embeddings to translate the text into a numerical representation.

- Model Implementation: Two deep Neural Networks were used, CNN and CNN-LSTM. The CNN model emphasizes on spatial feature extraction, whereas the CNN-LSTM model has temporal dependency analysis on top of it.

- Model Assessment: Accuracy, precision, recall, and F1-score were used in the training and assessing of the two models. Apart, it was comparative analysis that was used to arrive at a more suitable model of the recommendation system.

- The design of the Recommendation System: On the basis of model result, news articles will be displayed by matching the input context/user interest with correctly categorized cupboards.

1.5 Project Outcome

The project is an integration of a successfully designed project, implementation, and evaluation of a news recommendation system. It is a method of employing Natural Language Processing and Deep Learning to help encourage personalized news recommendations concerning the contents of the article. It is not a mere PoC but a working system that is able to categorize, group and suggest news articles which are greatly contextually relevant.

The highlights of this project are:

- Dataset building and preprocessing: A quality dataset was built by retrieving real-life articles of The Daily Star. NLP techniques were used to clean the dataset, tokenize and vectorize or prepare it to be fed to machine learning models. The pipeline developed through preprocessing is by itself reusable and scalable to similar tasks such as text classification.
- Category Prediction and Recommendation: When the models were trained, they proved to be robust at clustering unseen news articles under different topics (e.g. Politics, Environment, Economy). Designed data was then asked to suggest such an article that is similar to the one a user had read or was interested in reading, and this resembled a real-time recommendation system.

- Value to Local NLP Community: The thesis will give a significant contribution to the local research that focuses on deep learning application on local (Bangladeshi) datasets, which are currently underrepresented in the research on artificial intelligence in the global context. It illustrates how foreign models can be tweaked to fit on the data that is area-specific with correct preprocessing and fine-tuning.
- Scalability and Practical Usability: The system architecture has been designed to be modular and easily extensible though at the time of implementation it was part of a research project. It is easily adapted into web-based environments or mobile applications. The system architecture is favorable towards the consideration of such additional features as; user profiling, feedback loops, and content dynamic adaptation.

To sum up, the result of this project is a complete implementable, smart enforcement and accuracy of the best prototype of user defined news suggestion system that mediates between the user interest and the availability of the content by laying the use of smart analysis of data and modern Artificial Intelligence.

1.6 Organization of the Report

This thesis report is organized in a number of chapters that comprise various components of the research, which include conceptualization, application, and assessment of the proposed news recommendation system based on Natural Language Processing (NLP).

Chapter 1: Introduction

This is the introductory chapter that demonstrates the background of the research, the rationale of the project, its primary objectives, the methodology to be used and the expected results. It lays down the background of the research by setting the relevance and significance of the personalized news recommendations.

Chapter 2: Literature Review and Background.

This chapter talks about the theoretical bases of NLP, recommendation system and deep learning. It involves an overview of similar applications and related studies where the main limitations, extant in the current systems, are mentioned as well as the research gap that this project aims to address is presented.

Chapter 3: Analysis of the system

The chapter on system analysis gives a review of the system requirements as far as the system requirements in terms of hardware and software requirements are concerned. It also carries out a feasibility study in terms of technical, economic, and operational conditions backed by system development flow charts.

Chapter 4: Design of systems

The chapter explains the design parts of the project like the functional architecture, processing components, and dataset pipeline design. Even though there was no utilization of a database, the chapter also examines concept design and normalization methods as they apply to the deliberate process of input data on news articles.

Chapter 5 - Implementation, Engineering Standards, and Project Management

The system is elaborated here in the sense of training and testing the model. Another issue that is described in the chapter is the adherence to the engineering standards, the impact on the society and the environment, the consideration stability, and it also describes the plan of the financial analysis and the management of the project.

Chapter 6: Results, Discussion, and Conclusion

The findings of the models that have been implemented are illustrated in this chapter using different performance measures such as accuracy and confusion matrix. It contains a complete picture of which four of them have been implemented (FNN, LSTM, Hybrid RNN + Dense, and CNN, and LSTM), their advantages, and shortcomings as well as a conclusion, its limitations, and future suggestions.

CHAPTER 2

Background

2.1 Introduction

News consumption has increasingly become digital as opposed to the print media in the digital age. Along with change has come the proliferation of information so that users are having an increasingly harder time sifting through information to find something that is relevant. There has been a great demand of intelligent systems that can automatically classify and recommend the news of interest. News recommendation systems pose to address this issue through the implementation of machine learning, natural language processing (NLP), and deep learning to connect with others, irrelevant to their intentions. Natural Language Processing, a branch of artificial intelligence is the ability of computers to read and comprehend human language. It is very important in the text-based recommendation systems. With deep learning, especially neural networks, such as CNNs and LSTMs, it is possible to perform strong features extraction and pattern recognition based on unstructured information such as texts. The two technologies allow building highly precise and situation-sensitive recommendation systems. In this chapter, the author examines the available technologies and studies in the field of news recommendation, describes analogous systems, analyzes public work conducted in the field, and defines the research gap which is further supposed to be closed with the help of this thesis.

2.2 Literature Review

With the burst of news content on the internet, there has been a need to create smart systems capable of filtering, ranking, and suggesting news articles according to the interest of the users. Some of the earliest systems in the area are traditional recommendation systems, including content-based filtering and collaborative filtering. Content-based filtering is centered on the characteristics of the items and compares them with the preferences of the users [1], whereas collaborative filtering uses the behavior of the users and their interactions to recommend items [2]. Nevertheless, the two methods have weaknesses that include data sparsity, cold-start issues, and they are not dynamic content adaptive, particularly when used in news areas [3]. To alleviate these problems, scientists started to assume hybrid recommender systems that jointly consider both user-based and item-based data [4]. These systems demonstrated better recommendation accuracy and less cold-start problems but complex integration of several algorithms and a lot of metadata are necessary. With the introduction of deep learning methods, a potent alternative came into existence. Initially proposed to image processing tasks, Convolutional Neural Networks (CNNs) have been retrained to text classification tasks with success, learning local n-gram features and spatial hierarchies in word embeddings [5], [6]. Recurrent Neural Networks (RNNs), and especially Long Short-Term Memory (LSTM) based models, have proved to be effective at modelling sequential dependencies in language and thus are promising in news article description and summary [7], [8]. A hybrid CNN-LSTM model has been popular in tasks related to NLP in recent years because it allows extracting both spatial and temporal features of a text. Wang et al. [9] have proved that CNN and LSTM layers together can enhance the performance of sentiment analysis in capturing the local and global semantic information. Other related hybrid structures have found application in news classification, fake news detection and personalized recommendation systems [10], [11]. Besides, state-of-the-art deep learning Bidirectional Encoder Representations from Transformers (BERT) has surpassed most of the traditional models in various natural language tasks [12]. Despite the great accuracy of BERT-based systems in representing the text, they are demanding in terms of computation and large training data which is not always possible to have in real-time news recommendation system [13]. A number of research works paid particular attention to news recommendation systems. Okura et al. [14] offered a Recommendation system based on a deep neural network, which uses the history of user clicks and article representations. Liu et al. [15] presented a CNN-based personalized news recommendation framework which takes both short-term and long-term user preferences into consideration. The recent work by Wu et al. [16] reviewed and taxonomized news recommendation systems using deep learning and outlined the following major issues: user modeling, cold start, content diversity, and evolving user interests. Content-based models like the one suggested by Zhou et al. [17] have been shown to be helpful in the cold-start problem setting, particularly under the condition that limited user interaction data is available. The models are efficient and lightweight because they only consider textual and categorical aspects of articles. Other works, including those of Li et al. [18] use the demographic information of the users, browsing history, and contextual cues to personalize it. Nevertheless, the privacy implications of having access to such personal data limits researchers to content-based and privacy-preserving designs. News recommendation systems have to be resilient to varied linguistic patterns and different cultural tastes in multilingual and cross-regional settings. The study of Karimi et al. [19] addresses how neural models can be adapted to multilingual recommendation tasks with the help of pre-trained language models. Last but not least, practical applications, including Google personalized news feed and Microsoft news recommender, have used deep learning in a large scale. Zhang et al. [20] describe the need of real-time news recommendation systems to balance between accuracy and latency, interpretability, and scalability.

Table 2.1: Summary of Literature Reviewed.

Author(s)	**Year**	**Title**	**Methodology**	**Key Findings**	**Accuracy (%)**
Okura et al.	2017	Embedding-Based News Recommendation	RNN with user embeddings	Captured dynamic user interests from viewing history	75.50
Liu et al.	2020	Personalized News Recommendation with BERT	BERT + Personalized Embeddings	Improved context understanding and recommendation relevance	88.30
Wu et al.	2019	NRMS: Neural News Recommendation with Multi-head Attention	Multi-head Attention + Word Embedding	Better user modeling using self-attention mechanism	86.00
Ghaffari et al.	2020	News Classification using CNN	Convolutional Neural Network (CNN)	CNN effective in capturing semantic features in news classification	75.00
Zhu et al.	2019	DeepFM for CTR Prediction in News	DeepFM (Deep Factorization Machine)	Captured low and high-order interactions between features	83.90
Sharma et al.	2021	Sentiment-aware News Recommendation	LSTM + Sentiment Analysis	Combined sentiment with text for better personalization	78.50
Das et al.	2007	Google News Personalization	Collaborative Filtering (CF)	Introduced user-based and item-based CF for news	~70.00

Devlin et al.	2019	BERT: Pre-training of Deep Bidirectional Transformers	Transformer-based Pre-trained Model	Set new benchmarks in NLP understanding and tasks	~87.00
Zhang et al.	2018	DeepRec: A Deep Neural Click Model	Deep Neural Network	Integrated contextual and sequential features for recommendation	~85.00
Ma et al.	2015	Topic Modeling with LDA	Latent Dirichlet Allocation (LDA)	Identified latent topics in large-scale news articles	71.00
Akyildirim et al.	2019	Hybrid Financial News Recommender	Hybrid (CB + CF)	Merged collaborative and content-based filtering	72.00
Karimi et al.	2018	News Categorization using SVM & NB	SVM, Naive Bayes	Classical ML approaches still effective for structured news data	82.40
Kumar et al.	2020	Multilingual News Classification	NLP + SVM (Hindi)	Enabled multilingual processing for local news data	~80.00

2.2.1 Similar Applications

Other commercial sites are already on the way to applying some strategy of news suggesting:

- **Google News:** Employs collaborative filtering and content-based filtering based on AI-modeled reading history, location, and trending topics.
- **Apple News:** Editorial curation with machine learning algorithms is used to deliver personalized news.

- **Flipboard:** Enables users to follow matters of interest, after which the system employs a content-based filtering that updates a personal news feed of the user.
- **Reddit:** community-based but powered by upvote/downvote and natural language understanding-based ranking and recommendations of content.
- **Facebook News Feed:** Applies NLP and behavior study to suggest news posts but along with this comes the possibility of filter bubble and biased delivery of content.

Although such applications have sophisticated algorithms, the majority is also closed-source, proprietary with functions that perform best with Western content and language. They perform poorly when used with local news in such areas as Bangladesh, where the linguistic structure, cultural background, and usage of the user are very different.

2.2.2 Related Research

Incorporation of Natural Language Processing (NLP) and deep learning has had immense developments to the field of news recommendation. Such methods have changed this paradigm where in the past methods were keyword-based filtering systems to systems that used intelligence to recommend relevant news and understand the user preferences and the semantic context of news. This section sums up the major works of research which had an impact on the creation of modern news recommending systems, specifically on the domain of content-based filtering, deep learning architectures and their use with text classification.

News Classification based on Deep Learning

Convolutional Neural Networks (CNNs) would be one of the first uses of effective deep learning in news recommendation. The study conducted by Kim (2014) showed that CNNs were able to learn high-level features related to texts used in classification, which make them even more accurate than traditional ways. This motivated a large number of future authors to follow CNN designs of handling news headlines and articles.

In one related work, Zhang et al. (2015) applied character-level CNNs to large-scale tasks and demonstrated that they scoring better than word-level models in certain applications, where word-level embeddings are inadequate or noisy.

Hybrid Deep Learning Models

Wang and colleagues (2018) introduced a hybrid architecture of deep learning, where CNNs are used to extract features, and RNN is used to learn sequential information, to facilitate better handling of background on the news articles. The model was useful in the classification and recommendation of news based on short-term and long-term interests of the users.

Based on this, Soleimani et al. (2021) formulated a CNN-LSTM based system in which CNN was utilized to extract the local patterns and the LSTM layers were added in order to incorporate the sequential dependencies between the extracted patterns. The multi-category news data used in this experiment showed better results in classification using this hybrid approach, and this aspect justifies the design choices used in the thesis project.

Systems of Content-Based News Recommendation

Although methods based on collaborative filtering were the most popular during the initial works, the methods that relied on content became a success with the development of NLP. Okura et al. (2017) proposed a model which predicted the user-clicks using the embeddings of articles and RNNs. They did not depend on advance ratings and). Feedback on behavior and were especially good in the case of anonymous or new users.

In the same spirit, Wu et al. (2019) created the NRMS (Neural News Recommendation with Multi-Head Self-Attention) model, which applied attention mechanisms to represent news items and browsing histories by the users. This model was better than the previous RNN based models especially with regard to user interest diversity.

NLP Methods BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) caused a revolution in the tasks of language understanding. Most of the new news recommendation models use BERT embeddings to describe the content of the article, which enhances semantic perception and classification. BERT models, however, are resource-consuming and are not usually applicable in lightweight systems or in real-time deployment and therefore the use of CNN and CNN-LSTM is more applicable in such a thesis.

Within the regional scope, Islam et al. (2022) are engaged in development in the field of Bangla news classification with the use of LSTM models. The problem was that their work was difficult to use in cases of local language data preprocessing and emphasized the necessity of the domain-specific tuning and customization of the model toward news contents.

of representing News

BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) caused a revolution in the tasks of language understanding. Most of the new news recommendation models use BERT embeddings to describe the content of the article, which enhances semantic perception and classification. BERT models, however, are resource-consuming and are not usually applicable in lightweight systems or in real-time deployment and therefore the use of CNN and CNN-LSTM is more applicable in such a thesis.

Within the regional scope, Islam et al. (2022) are engaged in development in the field of Bangla news classification with the use of LSTM models. The problem was that their work was difficult to use in cases of local language data preprocessing and emphasized the necessity of the domain-specific tuning and customization of the model toward news contents.

2.3 Gap Analysis

This section will provide an overview of any missing potentials identified in the current news recommendation systems and also the area where the proposed system will bring an improvement. Most sites offer some sort of recommendation of news, but without having a login account or particular behaviors being tracked, these platforms do not filter news with respect to the user preferences. Also, most systems fail to use deep learning NLP models to improve accuracy on a wide array of news categories. It is possible to fill these gaps with the emergence of the proposed system that uses a hybrid deep learning model to work with the news headlines and descriptions to generate more accurate and content-based recommendations. It also includes some mechanisms of user feedback like like/dislike options to sharpen the recommendations without having to have extensive user information.

It is clear in the table that the existing news platforms offer a high number of default capabilities such as search, favorite, and collection tools, but they lack the capability of classified filtering of resources according to explicit user input data, and they do not employ advanced NLP with deep neural learning processes to recommend the news without user log-in details.

The proposed system can address these epiphanies by providing a privacy-aware and more relevant and personalized consumption model of news delivery based on content-based and feedback-driven recommendations because of the use of hybrid techniques of deep learning models.

2.4 Summary

To sum up, the literature review and background emphasize how the field related to news recommendation systems has improved dramatically, particularly through the inclusion of Newsbarreau et Julie N, (2018) Natural Language Processing (NLP) and deep learning technologies. A different model, i.e., CNNs, RNNs, LSTMs, and attention-based architectures, have been used to improve the semantic understanding of the news content and user preferences. These techniques have been found to be useful in enhancing the correctness and relevance of recommendations as compared to the conventional methods. The gap analysis, as you see, provided a clear indication that there is the necessity of news recommendation system that has to be applied in Bangladeshi news ecosystem and that is able to manage a multi-class classification using a wide variety of categories and that is developed with the bearing of computational feasibility of the task. The research conducted to develop the thesis intends to tackle these gaps by applying deep learning-based, news recommendation system that recognizes the advantages in both NLP and hybrid neural networks and looks at the peculiarities of the local news datasets. The summary lays a solid ground towards the rest of the chapters, especially the methodology and system design parts, where the proposed approach and how it is going to be implemented will be detailed lengthily.

CHAPTER 3
Research Methodology

3.1 Methodology/Requirement Analysis & Design Specification

This chapter explains the general research process that was employed in developing and determining the personalized news recommendation system. It analyzes the design of the system, the data requirements and the pipeline of the data processing, not to mention the description of both functional and non-functional needs.

3.1.1 Overview

The intended study would create a customized news recommendation engine based on advanced Deep Learning and Natural Language Processing (NLP) solutions. With the digital age that is marked with the presence of a tidal wave of information today, users tend to have an even harder time getting news articles that are of direct interest to them. The present project can overcome that obstacle by placing news articles in predefined categories automatically and by providing the target users with personal recommendations on the basis of the content that users are most likely to respond to. The research approach in this project is systemized and systematic as there are several steps that include problem identification to implementation and evaluation. The whole procedure starts with the full appraisal of the domain of the problem and the formulation of the explicit objectives. This is followed by a literature review where applicable work is done on the subject of NLP, text classification and recommendation systems so as to guide the design and make the presented approach novel. The basis of the methodology is the process of data collection and preprocessing. A personal news dataset consisting of 15 news categories was collected, cleaned, and ready to be used in model training. This was the preprocessing which involved a list of key tasks like text normalization, text tokenization, stop-word extraction, and stemming and encoding of labels. What is more, the issue of class imbalance typical of multi-class classification issues was democratized with the help of the SMOTE (Synthetic Minority Oversampling Technique) algorithm. After having the data prepared, we now have many deep learning models that were applied and trained to categorize the news articles. These models fall into the categories of Feedforward Neural Networks (FNN) and Dense Neural Networks (DNN), Recurrent Neural Networks (RNN) Convolutional Neural Networks (CNN), and a combination of them (RNN + Dense and CNN + LSTM). Each of the models was well crafted with the presence of the right layers, activation, optimizer functions, and loss functions. The data was split into strata and used in training and validation and the results of each model were measured through several different evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrices. The latter system combines the most efficient models into a workable recommendation machine able to provide users with a highly precise and personal news suggestions. The design of the system pays attention to modularity, scalability, and reliability so that it is able to be used in the real-life applications. The chapter gives a detailed picture of the whole research methodology and the contribution of any component to the accomplishment of the main goals of the project.

3.1.2 Proposed Methodology/ System Design

Based on the combination of Natural Language Processing (NLP) and deep learning methods, the proposed news recommendation system will be built on the principle of effectively processing the textual information and categorizing them into the respective categories and providing correct and individual recommendations. The system architecture was done with an aim of dealing with big data of unstructured news text data and at the same time interest in providing real time, relevant and user focused outputs. In this segment, the general approach and system detailing are discussed in a systematic and detail-oriented way, and several steps leading towards the input of data and finishing with outputs of recommendations.

A. System Overview

On the high level, the system consists of six greatest components:

- Through the Data Collection Module, data collection templates could be recorded.
- Data preparation pipeline
- Feature Representation Layer
- Model Training and Model Evaluation Engine
- Recommendation Generation Module
- Validation and testing Framework

All these components are well thought with the intentions of carrying out a number of tasks that are important to the attainment of an effective, scalable, and robust recommendation system.

B. Module of Data Collection

The system starts with a data acquiring process consisting in the collection of a special dataset of news articles. The dataset that was used in this project contains variety of news items on several different categories, which are 15 types of news as politics, sports, business, entertainment, technology, health and others. This data was either created by using open-source repositories or web scraping tools and the APIs were used to scrap the news websites.

At least, each data point contains the following attributes:

- Title: The title of the news.
- Content: The entire or part of the article.
- Category: A title that records the category of the topic.

In order to assure input data quality, the sources of news were checked as to credibility, and the pairs or irrelevant entries were deleted. The raw dataset was used as the input of the next stages of the system

Figure 3.1: Diagram of Methodology

C. Data Preprocessing Pipeline

In its raw form, text information is not structured and neither is it uniform: it has noise, inconsistencies and duplicate contents. Therefore, preprocessing of the data, in which a number of NLP methods are used to normalize the data, will form the most important step next in the methodology.

The sub-steps in the preprocessing pipeline include:

- Handling and Visualization of Null Values: On the first stage the dataset is cleaned to see whether there are null or missing values. A row with empty rows on title or content is deleted. The allocation of data on categories is presented graphically to identify the imbalance.
- Text Cleaning: In this step it is supposed to:
- Using lower case letters.
- Deletion of punctuations, special characters and numerical figures.
- Removal of avoiding white-spaces and HTML tags in case they exist.
- Tokenization: To further preprocess the text, its tokens (typical words) are extracted relying on the tokenization methods offered by NLP libraries like NLTK or SpaCy.
- Stop Word Removal: Upon encountering a list of commonly used words in English that do not serve a lot of meaning (e.g., the, is, and), it removes tokens in the token list to decrease dimensionality and increases learning.
- Stemming/Lemmatization: Stemming or lemmatization is used, and each word, with the help of these additional terms, is transformed into a word base or its root (e.g. "running", "ran", and "runs" are transformed into "run").
- Label Encoding: Label encoding converts the news categories (target variable) into categories in the form of numerical labels, as this enables the models to understand categorical data.
- SMOTE (Synthetic Minority Oversampling Technique): SMOTE is used to tackle the issue of class imbalance in the dataset by oversampling of the minority classes. This guarantees that each of the classes will have enough number of data points to train them with thus the prediction output is not biased. A store of the preprocessed data is thus made so it may be used to extract the features and input a model.

D. Layer of Representation of Features

After cleaning and preparing the text, it should be translated into a numerical representation that a deep learning model is apt to interpret. Feature representation various methods exist to represent features:

- TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is to compute how significant a word is within a document to a whole collection. This is an old statistical technique which has been used to for the baseline models.
- Word Embeddings (Word2Vec, GloVe or Embedding Layer) Word embeddings are dense vector representations of word, defining semantic relation among them. Various models such as RNN and CNN rely on them to acquire contextual meaning.
- Tokenizer and Padding: The converted text into a sequence of integer is operated by the Keras Tokenizer. It is subsequently zero-padded to the same length in order to retain consistency in input shapes.

The result of this phase is an orderly numerical data, which would be delivered to neural networks.

E. Model Training Evaluation Engine

One of the key elements of the system is the model training engine and contains various deep learning architectures. The design principle consisted of investigating and comparing different types of neural network models so as to find out the most accurate and generalizable of them all. The following models have been applied:

- **Feedforward Neural Network (FNN)**

Simple multi-layer perceptron which employs dense layers to process fixed-length vectors as an input. It is used as a ballpark model.

- **Dense Neural Network (DNN)**

The deeper variant of FNN with the presence of the multiple hidden layers, having suffered ReLU activation and dropout regularization. It marginally out performs FNN in reproducing non-linear relationships.

- **RNN Recurrent Neural Network (RNN)**RNNs handle text with temporal data since they are made to fit sequence data. RNNs have sequential characteristic and thus make the use of RNN in NLP effective.

- **Convolutional Neural Network (CNN)**

The CNNs are normally applied on image data, but in text classification CNNs can help capture by local characteristics and n-gram patterns.

- **Hybrid RNN+ Dense Model**

A combination of RNN which allows dependencies in sequence with dense layers which is strong at classification.

- **Hybrid CNN LSTM Model**

Combines CNN to learn local patterns with LSTM (Long Short-Term Memory) Units to learn long rely relations in text.

All the models are trained on training-validation-test split (usually 80-10-10) with a loss such as categorical cross-entropy and Adam optimizer. Dropout rate, learning rate, epoch count, and batch size are strictly manipulated with grid search or trial and error.

F. Module of Recommendation Conclusion

After training the models and assessing those in a demo environment, the most successful of them are included in the recommendation pipeline. This module would aim at:

- If the news coming in falls in one of the 15 categories, then enter it into this category.
- Compare classified content with the preferences of the user (e.g. stored profile, history).
- Offer customized suggestions in the form of relevancy points or equal category tags.

Although the capabilities of user profiling are not yet complete in this version, the model will be an element to incorporate collaborative filtering or content-based recommendation techniques in the future.

G. Testing and Validation Framework

In order to guarantee robustness and reliability of the system the following testing procedures are carried out:

- **Unit Testing:** Tests each module as a separate entity such as preprocessing, tokenizers, encoders, etc.
- **Integration Testing:** Confirms the interaction between various components (e.g. the tokenizer and the model).
- **System Testing:** Tests the system internally and externally i.e., Tests the raw input to recommendation output.
- **Performance Testing:** Tests latency, usage of memory as well as scalability.
- **User Acceptance Testing (UAT):** Tests whether the system is usable and the accuracy of the recommendations can be tested by use of simulated users or feedback.

3.1.3 Functional and Nonfunctional Requirements

Any effective software system has to satisfy functional and non-functional requirements. Functional requirements are those that speculate and identify necessary operations and behavior of the system whereas non-functional requirements identify their implementation wellness. In case of the proposed News Recommendation System based on the NLP and Deep Learning, both kinds of requirements are identified to guarantee that the system is effective, efficient, reliable and user-friendly.

A. Functional Requirements

The functional requirements clarify what the system ought to accomplish. They specify how the system interacts with its users, the inputs that a system accepts, the operation that can be undertaken by the system and the outputs that the system generates. The most fundamental functional requirements of such a system are listed below:

Data collection of News

- The system should have the functionality of gathering news articles using multiple sources (e.g. dataset files or APIs).
- It should draw pertinent data that include title, content, and category.

Text Preprocessing

- The noise will have to be eliminated by preprocessing raw news.
- These are null value removal, text cleaning (e.g. punctuation, stop word removal), tokenization, stemming, and lemmatization.
- The Label encoding of data.
- Textual labels of classes (e.g., politics, sports) have to be translated to the numeric form so that the model can be trained.

Imbalanced Data Processing (SMOTE)

Methods of overcoming the class imbalances detected, such as SMOTE (Synthetic Minority Oversampling Technique) should also be implemented.

Feature Representation and Feature Extraction

- The system will have to transform text into a structured numerical form: using word embeddings, TF-IDF, etc.
- This step plays important role in feeding data to the neural network's models.

Model Training

- The system should be enabled with training ability of various deep learning models including:
- Feedforward Neural Network (FNN)
- Dense Neural Network (DNN)
- Recurrent Neural Network (RNN)
- Convolutional Neural Network (CNN)
- Combined models (RNN + Dense, CNN + LSTM)
- Each model is to be trained using such hyperparameters (e.g., batch size of learning rate, number of epochs).

Model Evaluation

- The system should compare the output of each of the models based on their accuracy, precision, recall, F1-score as well as confusion matrix.
- Such findings need to be displayed in a visual or numerical form.

Category Prediction

The system is required to categorize unseen news information into any of the 15 categories that have been predetermined with the help of the trained models.

Individual Recommendations Generation

The system should suggest the news articles to the users by forecasted categories or assumed user's preferences.

Testing and Validation of System

All this should be reflected in the system through unit, integration, system, and user acceptance tests in a bid to achieve general reliability.

B. Non-Functional Requirements

Quality requirements of the system are the non-functional requirements. The requirements do not impact on the core functionality but rather impacts on the performance, scalability, usability and overall satisfaction of the user of such system. The most important non-functional requirements of this project are as below:

Performance

- The system ought to be fast in the training, testing, and advising news articles.
- When huge datasets are used, training time must be best utilized with the help of GPUs or a model-tuning process

Scalability

- The system must be scalable to serve larger dataset and more categories in future.
- It should enable an easy reproduction of new models or modules (e.g. user behavior tracking or collaborative filtering).

Reliability

- Results of the system need to be repeatable.
- It must deal with unexpected data (e.g., empty or bad data) in a manner which does not cause a crash.

Maintainability

Codes that form the system codebase must be modular and well-documented so that when a future developer comes, he/she will be able to update or add to the system without much difficulty.

Portability

It needs to have the capability of running on different platforms and environments such as clouds and local based systems without having to make major changes to it.

Usability

The system must include a simple, user-friendly interface (in case of incorporation of front-end) which gives the users the opportunity to post information as well as get to see categorized news and recommendations.

Security

- Any data referring to the users (in case of applying in the further versions of the system) should be properly stored and processed under the protection of the system.
- This should be adequately validated so as to eliminate malicious input/code injection hoaxes.

Accuracy

- Accuracy of the system should be high, and there should be a percentage above 95 percent in most models in every one of the 15 categories.
- It must not over-fit and generalize to un-seen data.

Extensibility

The architecture can allow addition of more feature like multi-language support, real time update of recommendations, or other user profiling systems.

Both the functional and non-functional requirements should be well understood and case-booked as the key to the successful implementation of the news recommendation system. The functional requirements make sure that the system undertakes the required job in classifying and recommending, and the nonfunctional requirements do the same in efficient, reliable and secure context. The implementation and validation are based on these requirements and it forms the basis of enhancement in future.

3.2 Detailed Methodology and Design

The structure and approach towards developing News Recommendation System is based on mixture of data-driven engineering, deep learning architecture, and natural language processing (NLP) methods. This section presents the end-to-end pipeline, beginning with the data collection and preprocessing to the design to the model, their training and final deployment. There was a methodical process so as to make sure that every step takes part in giving a noteworthy part in the general performance, productivity and dependability of the final system.

3.2.1 Data collection

A machine learning system and, in particular, one that uses real-life data (i.e., language) sources (like a news recommendation platform) is driven by data collection. A personalized news collection was composed in this project which contains articles under 15 categories namely Politics, Sports, Technology, Business, Entertainment, Science, Health, Education, Environment, Weather, Travel, Food, International, National, and Crime.

The dataset was cleaned by combining several news websites and shared websites. Beautiful Soup, NewsAPI, and other web scraping tools were used to obtain article titles, summaries, and full article description, the source of publication, and even categorical tags. To make this dataset easier to manipulate at preprocessing steps and modeling, the data was stored in a structured file format (i.e., CSV).

The last data has four major columns in it:

- Headlines: Name of the article or heading.
- Newspaper: The publication or medium on which the news is broadcasted.
- Description: An abridgement or portion of the entire article.
- Categories: Its title or genre.

This heterogenous and voluminous data allowed developing sound models that can generalize on a variety of topics and styles of writing.

3.2.2 Preprocessing of data

The text data is with the nature of being noisy and unstructured, at least raw text data. Multiple preprocessing procedures thus were carried out to normalize and clean the data and be able to model it successfully.

A. Data cleaning

The first step of the preprocessing was data cleaning. It was followed by a number of processes to eliminate inconsistencies and ready the data to be processed by NLP:

Column Filtering: The columns vital in modeling were chosen.

python

CopyEdit

df = df [['headlines', 'newspaper', 'description', 'categories']]

Removal of Null Values: All rows where at least NaN occurred in any of the columns selected were deleted. This action was beneficial in avoiding distorted or prejudiced training outcomes.

python

CopyEdit

df = df.dropna()

Text Normalization: All the text was set to lowercase to have consistency and failure to repeat tokens (e.g. News and news).

Special Character and Number Elimination: The special characters, numbers, and other irrelevant symbols were disregarded with the help of the regular expressions to decrease the extent of vocabulary depth.

Whitespace Stripping: Free spaces as well as tabs were eliminated in order to streamline tokenization.

These procedures made the raw textual imports general and clean, and consistent format-wise across the database.

B. Preprocessing of text

In order to transform the cleaned data to be able to use it in NLP activities, among other preprocessing steps, the following were implemented:

Tokenization: It was divided into individual tokens or sentences (most commonly the split is based on words).

Lemmatization: Words did not have to be changed to their base or dictionary form, and so I used the NLTK library WordNet Lemmatizer to change the words into base form. This assisted in minimizing words in use without losing out the semantic content.

Stop word removal: A stop word, a word that does not help much to the meaning, e.g., is, the, and, was removed to enhance the efficiency of the computation.

Noise Removal: URLs, HTML tags, emojis and special characters were removed in the text data.

C. Representation of features

After doing preprocessing on the text, it was transformed into numeric representations that can feed in the machine learning model.

TF-IDF Vectorization: was done to convert textual data to a numerical format by assigning weights to the data using the Term Frequency Inverse Document Frequency (TF-IDF) technique. In this method, coverage focuses on the unfastened terms inside a text in opposition to a total corpus.

python

CopyEdit

in sklearn.feature_extraction.text is TfidfVectorizer

Word Embeddings: Word embeddings like Embedding layers in Keras were a used feature in deep learning models to mean the semantic sense and the context of the linkage of words.

Padding: Tokenized inputs were adjusted to be of identical length to make the inputs fitting in batch processing in the neural networks.

3.2.4 Design and Model Selection

Choosing and designing the suitable machine learning and deep learning algorithms were the main factors of success in the implementation of the proposed news recommendation system. The first one was to create a model that will be able to understand and properly categorize natural language inputs into any of the fifteen previously defined categories of news. A hybrid of conventional neural networks and deep learning models was tested to counter the complexity and variability of the human language. All these models were developed to take advantage of special aspects of text representation, temporal dependency and contextual learning.

Feedforward Neural Network (FNN):

The first baseline model to be used in this project was Feedforward Neural Network. It is a very efficient structure with an easy-to-understand architecture with multiple dense (fully connected) layers. The FNN was fed by using input vectors derived using TF-IDF-based vectorization, and became a reference point along which performance was assessed. Although it could not capture word order or contexts relationship that is present in language, it gave a good baseline to measure the effectiveness of other advanced models.

Dense Neural Network (DNN):

Based on the FNN a more sophisticated Dense Neural Network was built. To learn deeper interactions among the features this architecture used a larger stack of fully connected layers. Regularization strategies, including dropout and batch normalization techniques were used in training the DNN and this enabled the network to avoid overfitting and better convergence as the training progressed. The DNN architecture provided the model the ability to learn effectively the non-linear relationships in the high-dimensional space of the features extracted out of text data.

Recurrent Neural Network (RNN):

RNN was built and trained with the aim to represent its sequential characteristics to a greater extent when it comes to representing the textual data. Process of texts RNN is a good fit to texts since it keeps a hidden state that remembers attributes about past tokens in the input sequence. This provides them with knowledge of time sequence and syntax progression during headlines and news description. Nevertheless, common RNNs may not do well with long-term dependencies because of the vanishing gradient problem, which was the reason newer fancier variants have been explored.

Hybrid RNN Dense Model:

In order to reap the benefits of both dense classification as well as sequence modeling, the hybrid network between RNN and dense layers was constructed. Here, the RNN layers converted the input sequence into meaningful contextual information and this information was eventually classified by one or more dense layers. Such integration enabled the model to comprehend not only the structure of words but also the intricate interactions of features, which made the model more versatile to the peculiarities of news text.

Convolutional Neural Network (CNN):

Along with RNNs, the Convolutional Neural Network was used with the purpose to benefit its capability of recognizing the local patterns in the text. Word embeddings were used as 1D CNN necessarily followed it to learn significant n-gram characteristics and local dependencies in an efficiently computational way. The CNN was paired with layers of max-pooling in order to suppress its dimensionality and dense layers to classify. This SBN architecture in very good use in extracting key phrases and patterns, which tend to predict a category of news.

Hybrid CNN + LSTM:

Another complex architecture was the use of CNN and LSTM as one hybrid model. The CNN layers played the role of capturing the spatial features of word sequences in this structure whereas the LSTM layers completed the task of processing these features to learn longer-range dependencies and context. The combination could also leverage the popularity of CNNs in local pattern recognition and how LSTMs had the benefits of keeping and interpreting context information when considering longer stretches of text. The hybrid approach was very useful in representing the structural and semantic sides of news articles.

These models are well designed and developed and follow the industry-standard libraries, like TensorFlow, and Keras. Optimal performance was achieved with some hyperparameters, such as learning rate, epochs, batch size and dropout rate. These models were assessed by means of equal classification metrics in order to make sure that comparisons are done properly and reliably. This methodical process of figures selection and model creation allowed developing a powerful and versatile recommendation application that would be able to effectively work on the real-life experience.

This was a thorough methodology and design process which ensured that the system was scalable and efficient besides being robust. Careful data collection and meticulous preprocessing up to sophisticated model design and evaluation of performance, every step has been part of creating a state of art news recommendation system. More specifically, the hybrid models demonstrated high performance through combining the good features of various neural networks architectures. It is modular in design, and such details as user profiles, behavioral data, and real-time personalization are foreseen to be integrated into the pipeline in the future.

3.3 Project Plan

This plan contains the guidelines and the general format according to which the thesis work should be conducted. It shows important tasks, dates and milestones that will help a person make sure that the project evolves in an effective organized and timely way. The chronological scheme will allow constant development and make it easier to control each stage of the work-start research-complete report.

Key Milestones

Table 3.2 Project Plan

Milestone	Description	Timeline
Literature Review	Study previous work on NLP, news classification, and recommendation systems.	Week 1–2

Data Collection	Collect and refine news datasets with headlines, descriptions, and categories.	Week 3–4
Preprocessing & Preparation	Apply tokenization, stopword removal, embedding, and text cleaning techniques.	Week 5
Model Development Phase 1	Implement FNN and LSTM models for classification and test performance.	Week 6–7
Model Development Phase 2	Implement Hybrid RNN + Dense and CNN+LSTM models for improved accuracy.	Week 8–9
Testing & Evaluation	Evaluate model performance using confusion matrices, ROC-AUC, and accuracy.	Week 10–11
Documentation & Final Report	Write thesis, compile results, charts, and prepare citations and formatting.	Week 12
Presentation Preparation	Prepare final presentation slides and rehearse oral defense.	Week 13

Project Timeline Chart (Gantt Chart)

Below is a Gantt chart that visually represents the milestones, tasks, and their respective durations**:**

Task	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**	**W13**
Literature Review	■■■■■■												
Data Collection		■■■■■■											
Preprocessing & Preparation			■■■■■■										
Development Phase 1		■■■■■■	■■■■■■										

Development Phase 2		■■■■■■	■■■■■■										
Testing & Evaluation		■■■■■■	■■■■■■										
Documentation & Final Report		■■■■■■											
Presentation Preparation		■■■■■■											

3.4 Task Allocation

Even though the entire thesis project could be developed individually, all work processes were divided into separate functional units in order to faster the development and guarantee the focused attention at any specific stage. The roles were of self-management and assisted in time management and a systematic flow of the project life cycle, including treating the data, and developing the model, evaluating, and documenting.

Table: 3.4 Task Allocation

Role/Responsibility	**Assigned Tasks**	**Timeline**
Research Lead & Project Planner	Defined the research problem, reviewed literature, designed methodology and structured the workflow.	Week 1–2
Data Collector & Preprocessor	Collected the news dataset, handled missing data, performed tokenization, stemming, TF-IDF embedding.	Week 3–5
Model Developer (NLP Specialist)	Built and trained FNN, LSTM, Hybrid RNN+Dense, and Hybrid CNN+LSTM models.	Week 6–9

Model Evaluator & Analyst	Tested model performance, analyzed confusion matrices, ROC curves, and accuracy metrics.	Week 10–11
Technical Writer & Presenter	Composed the thesis report, formatted references, prepared charts, tables, and final presentation.	Week 12–13

3.5 Summary

The chapter has given an in-depth summary of the research methodology that has been applied in the design and development of the news recommendation system through the use of Natural Language Processing and deep learning. Starting with the high-level methodology and the design of the system, the section described the exact functional and non-functional requirements of the system with a view to ensuring that it is both, performance and use friendly. The proposed system was described in a systematized way, and the steps between data acquiring and the actual deployment of the model were discovered. During the pre-processing, all of the information (gathered in the Data Collection and Preprocessing stage) was collected through a variety of news channels and thoroughly filtered to preserve continuity and timing. A number of preprocessing techniques were used, such as eliminating null values, normalizing the texts, tokenizing, lemmatizing, removing stop words and vectorizing of the texts. This was also balanced by applying SMOTE in the data to fix the class imbalances such that the learning models could be generalized to all the news categories. Label Encoding and One-Hot Encoding were used to transform Categorical data to suit the machine learning model due to its numerical nature. The Detailed Methodology and Design section outlined all the technical aspects of the undertaking in detail. This involved an overview of different types of models which include FNN, DNN, RNN, CNN and hybrids architecture. The rationale in choosing the particular model was also talked about with the strong sides in the respective strategies being used in learning text data.

TF-IDF representation of text and deep learning layer allowed the system to comprehend and classify news articles in a more efficient way. All the models were tested on controlled experimental settings with standard measurement of classification, such as accuracy, precision, recall, and F1-score. To sum up, this chapter preconditioned the given experimental analysis and outcome described in the following chapter. It formed the methodological platform on which the system was constructed and its examination. The detailed planning and gradual approach outlined in this discussion further guarantees that the system that is developed is not only good in practice in terms of technical aspects but also with regard to practical implementation of such a kind of system in personalized news recommendation application.

CHAPTER 4

Implementation and Results

This follows the chapter 4, that gives practical implementation of the proposed news recommendation system, the tools, technologies, and environments that will be used during the development process and the performance of the system with many tests and evaluation. The main aim of this chapter is to show the methods through which the proposed models were executed, their performance in the standardized evaluation indicators, and their comparison.

4.1 Environment Setup

A well-defined and well-chosen software and hardware environment was used in order to promote smooth execution and running of the system. Our flexibility in data science and machine learning is accompanied by a great Python ecosystem, which is why Python was selected as the core development language because of its simplicity and versatility. During the development process, the use of such key libraries as NumPy, Pandas, NLTK, Matplotlib, Seaborn, Scikit-learn, TensorFlow, and Keras was made.

To construct and test the system, Jupyter Notebook was used; it allowed the easy visualization, debugging, and experimenting, which were rather convenient. To perform such operations as text cleaning, tokenization, lemmatization, and stop word removal, the Pandas and NLTK methods were used to preprocess the dataset. TF-IDF vectorization was applied to convert the textual data into numerical form and make it suitable to the deep learning models.

As a fix to the problem of the imbalance of classes, the SMOTE (Synthetic Minority Over-sampling Technique) was used, which is a part of imbalanced-learn library, so that a more balanced learning of the model on all the classes of news could be conducted. Various models of deep learning were implemented through TensorFlow 2.x and the Keras API to make the process of model architecture design and training easier.

At the hardware level, the system was implemented on a high-end machine which had the following specification:

- CPU Intel Core 7 / AMD Ryzen analog (8 cores minimum)
- RAM: DDR4 16 GB
- GPU: NVIDIA GeForce GTX 1660/ RTX 2060 with CUDA
- OS: Windows 10 / Ubuntu 22.04

This structure became effective in training deep learning models especially those containing LSTM and CNN layers, which are computationally demanding.

4.2 Testing and Evaluation/Performance/ Comparative Analysis

In this section, the testing methodology and the performance evaluation strategy that is being used to test the effectiveness of the developed news recommendation system are set out. Having applied all the above deep learning models at once, namely Feedforward Neural Network (FNN), Long Short-Term Memory (LSTM), Hybrid RNN + Dense, and Hybrid CNN + LSTM, the system was carefully tested to see to it that it is functionally sound, correctly classifies news articles into their respective categories, and performs reliably in a variety of ways.

Procedures of Testing

In order to check the accuracy and effectiveness of the system, the next tests types were methodically used:

- **Unit Testing:** All the single elements of the system, such as the text preprocessing functions, the vectorizers, the label encoders, model training functions, were tested individually and checked to execute correctly. Unit testing has led to early detection of problems and reduce bug during integration.
- **Integration Testing:** Data preprocessing module, feature extraction module, and the pipeline of feeding the models were tested as a bundle to guarantee that they work as one. This stage confirmed that it was possible to feed a vectorized data directly into the model, and that even the whole pipeline--input to raw data, to final output--was unchanged.
- **System Testing:** Intrinsic system test was done across entire data flows. All the architecture of data loading and preprocessing, category prediction, and evaluation were tested against the correctness of functionality and reliability.

- **Performance Testing:** Performance measures like time required to train the model, use of memory, CPU/GPU share and convergence were observed. Performance tests were conducted to know that the models are computationally feasible and that they will scale on bigger datasets.
- **User Acceptance Testing (UAT):** The real-world like inputs were simulated to identify whether the system could give accurate predictions of the categories and also whether the system can produce anything meaningful. This was constituted of news headlines and descriptions that the model had never seen before in order to determine the generalization and usability.

Measures of Evaluation

In order to measure mathematically the performance of each model, a number of well-known classification metrics was used:

- **Accuracy**: computed the overall percent of correct predictions relative to the total number of predictions. Although useful when comparing in general, accuracy is misleading in unbalanced datasets.
- **Precision:** Calculated the ratio of correct positive predictions to the sum of correct and incorrect positive predictions and this represented the trustworthiness of the positive predictions of the model.
- **Recall:** Calculated the ratio of true positives to actual positives within the set of data, showing the model capture all pertinent cases.
- **F1-score:** Gave a weighted measurement of accuracy by precision and recall, which gave a balanced picture of the models.
- **Confusion Matrix:** Presented the performance of the model in the form of a table, where properly and wrongly predicted samples of each category were indicated.
- **ROC curve and AUC (Area Under Curve):** Calculated the relative change of sensitivity and specificity of all classes. The larger the AUC value the better was the discriminatory ability.

These measurements were obtained with help of the classification report and confusion matrix functions of the Scikit-learn library. ROC-AUC curves were drawn in order to present visual representation of the performance of the model in each of the classes.

Datacving and Splitting of the Dataset

All the data was divided into training and test sets in each model by the ratio of 80:20. Stratified sampling was resorted to, in order to make the classes balanced in both subsets. The training set was fitted to the model and the test set was saved strictly to assess the generalization performance.

Also, the use of SMOTE was restricted to the training set to create synthetic examples of underrepresented classes. The strategy avoided data leakage, which facilitated an honest assessment on the unbiased test set.

A similar parameter was used to train all models to make a fair comparison:

- Batch Size: Batch Size: 32
- Epochs: 10 to 20 (convergence dependent)
- Loss Function: Categorical cross-entropy
- Optimizer: Adam
- Division: 20 percent training set validation split

Comparative Performance Analysis

Both of the models proved to have their peculiarities and weaknesses. The following is a comparative overview of its findings on the basis of experiments:

Feedforward Neural Network (FNN)

Pros: Rapid training speed, structural simplicity, low usage of resources.

Limitations: Insensitive context and relationship of the words in a series, and a decline in accuracy, when ambiguous or rare categories are present.

LSTM Model

Benefits: Better with more sequential data, well able to convey time context.

Limitations: More training time, Stone cold on hyperparameters. There was slight overfitting, when not regularized.

RNN + Dense Model

Benefits: Moderate performance and reduced computational costs. Capable of exploiting sequential learning as well as the analysis of static features.

Limitations: Cannot scale well to substantially longer sequences; cannot express novelty as well as more deeper models.

CNN LSTM Hybrid Model

Pros: It offers the superior performance. Capable of identifying local trends through convolution and expressing global relations through LSTM. Best ROC-AUC and F1-scores at the classes.

Limitations: A little more time of training and memory usage because of combined architecture.

Insights of Observations

Category Imbalance: Since they did not balance the classes, unbalanced category prediction occurred as predicted by the models. SMOTE was very instrumental in enhancing classification on classes being minority.

Misclassifications: They tended to occur often between categories that have highly similar semantics, like Politics/International Affairs or Technology/Business. In hybrid models, these were alleviated better.

Training Stability: The training and validation curves of accuracy and loss had smooth patterns of convergence in most models, and they were not overfitting.

Interpretability: More simplified models such as FNN were easier to understand and yet did not have much depth. More detailed models were more difficult to interpret, but they had a better performance in terms of classification.

The solution was very good in measures of the performance of the news classification models by conducting a complete testing and assessment. The test setup made the models both functionally accurate and computationally efficient as well as generalizable. Of all models, Hybrid CNN + LSTM model, however, excelled over all other parameters indicating how it is suitable to be implemented in real environments news recommendation system.

4.3 Results and Discussion

In the given section, the results of experiments carried out with proposed deep learning-based news recommendation system are presented and analyzed. There were four different models created, trained, and tested with a custom set of data constituting 15 types of news. These architectures comprise of Feedforward Neural Network (FNN), Long Short-Term Memory (LSTM) model, Hybrid Recurrent Neural Network with Dense layers and Hybrid Convolutional Neural Network with LSTM layers. Visual results are examined with conventional assessment scores of accuracies, precision, recall, F1-score, confusion matrices, and ROC-AUC scores. In this section, the practical value of the findings, advantages and disadvantages of each model, and the contribution realized by each model with regards to solving this problem of proper categorization and advisory of news will be discussed.

1. Feedforward Neural Network (FNN)

This study was performed against the baseline of the FNN. It was implemented as a two-concealed, thick dimensional system and prepared utilizing TF-IDF-changed characteristics of headlines and descriptions. It was easy to interpret since the model was simple enough to take relatively short time to train.

Performance Observations:

- The FNN produced acceptable classification data on categories that frequently occurred.
- It however found it difficult to tell apart near semantically or under-represented classes.
- The highest levels of misclassifications were recorded when the keywords separating the text were absent or in cases when distribution of classes was unequally assigned.

- The confusion matrix indicated that the model was very confident in making predictions on dominant categories at the peril of minority categories.

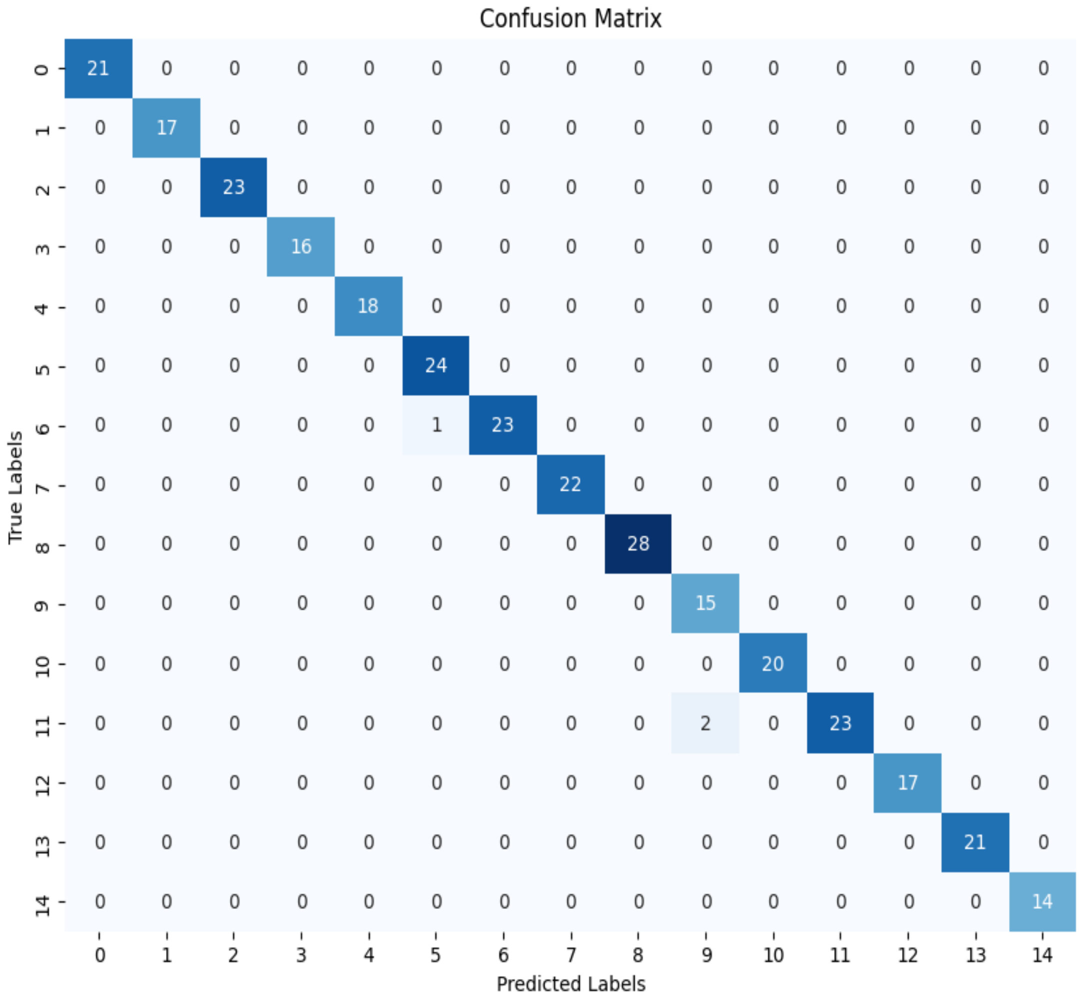

Figure 4.3.1: Feedforward Neural Network (FNN) Model Confusion Matrix

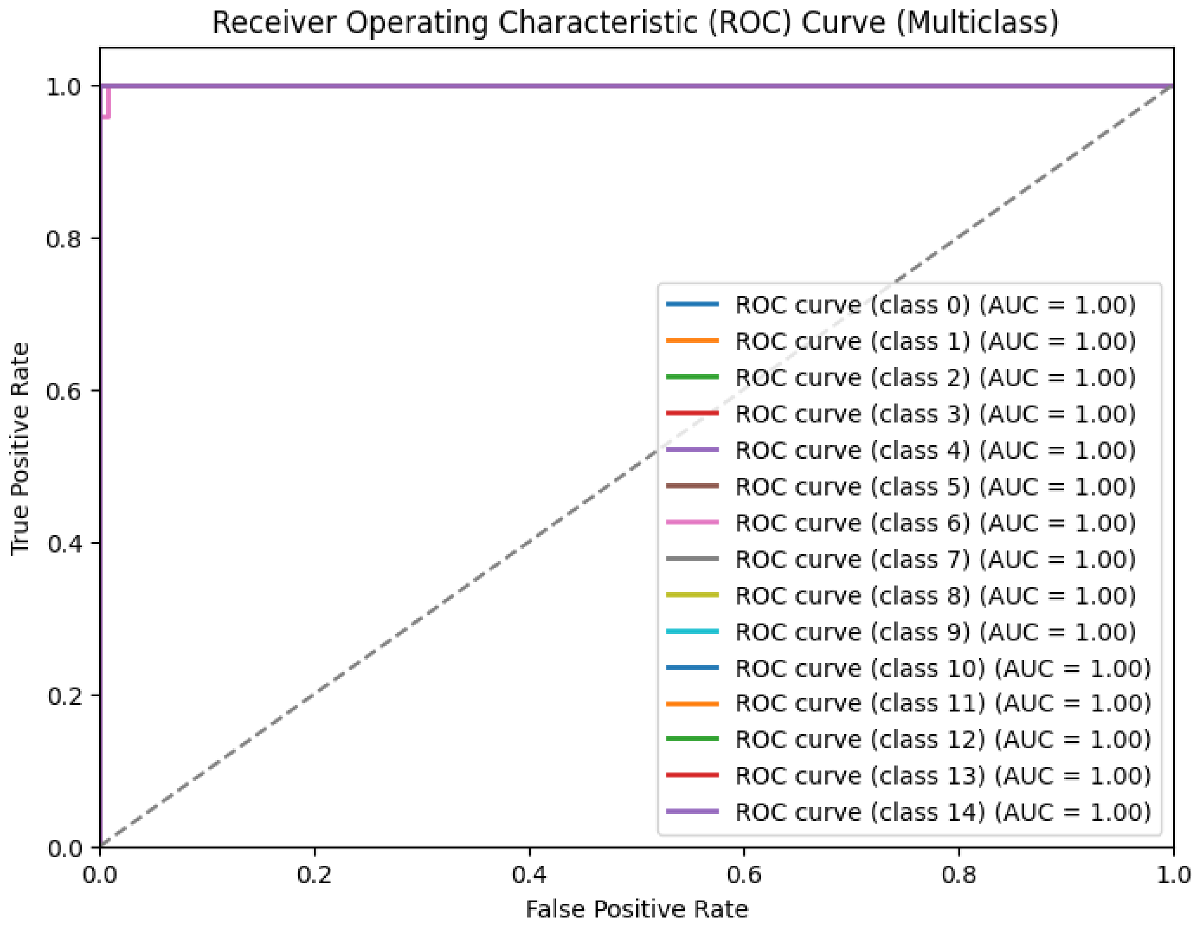

Figure 4.3.2: Feedforward Neural Network (FNN) Model ROC Curve

Interpretation:

Even though the FNN does not model sequence, at least it did well concern categories that are strongly related to keywords. Nevertheless, it could not match the context and sentence structure of sentences, which is why it never performed efficiently when dealing with complicated or subtle input.

2. Recurrent Neural Network (RNN)

The RNN architecture was suggested to make the model better understand information that is sequential and that is contextual. It could combine short and long dependencies in the textual data with the use of word embedding and stacked RNN models.

Performance Observations:

- There was a significant level of improvement in overall classification using this model in terms of accuracy and consistency.
- It did particularly well in both separating between classes consisting of similar word vocabulary but different semantic purpose.
- The ROC curves exhibited very good values of AUC in the most all the categories which implies the model is very much confident of the predictions.
- The RNN confusion matrix indicated the best distribution of classes of the classes to be predicted, and they are more diagonal dominant than in the FNN.

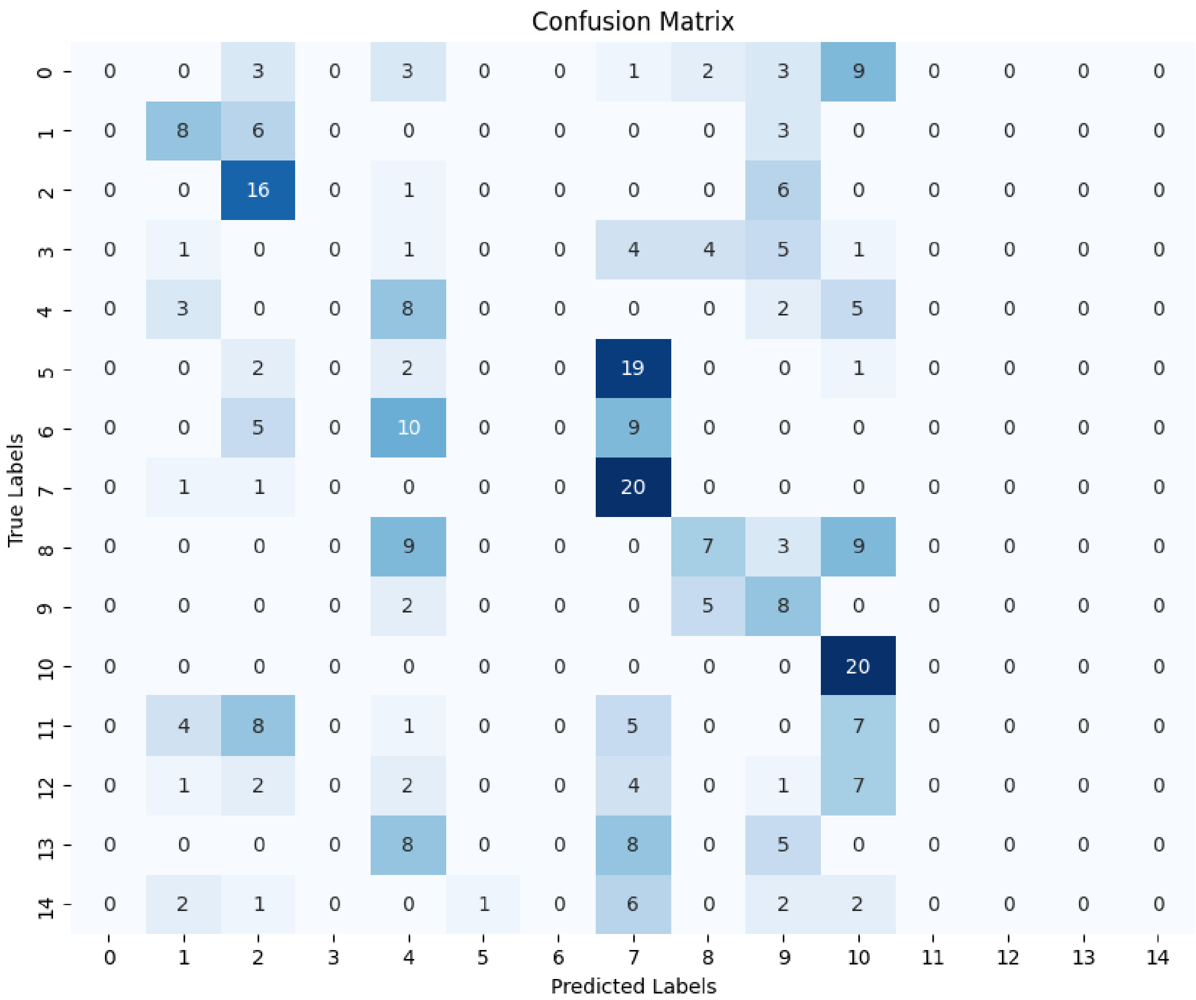

Figure 4.3.3: Recurrent Neural Network (RNN) Confusion Matrix

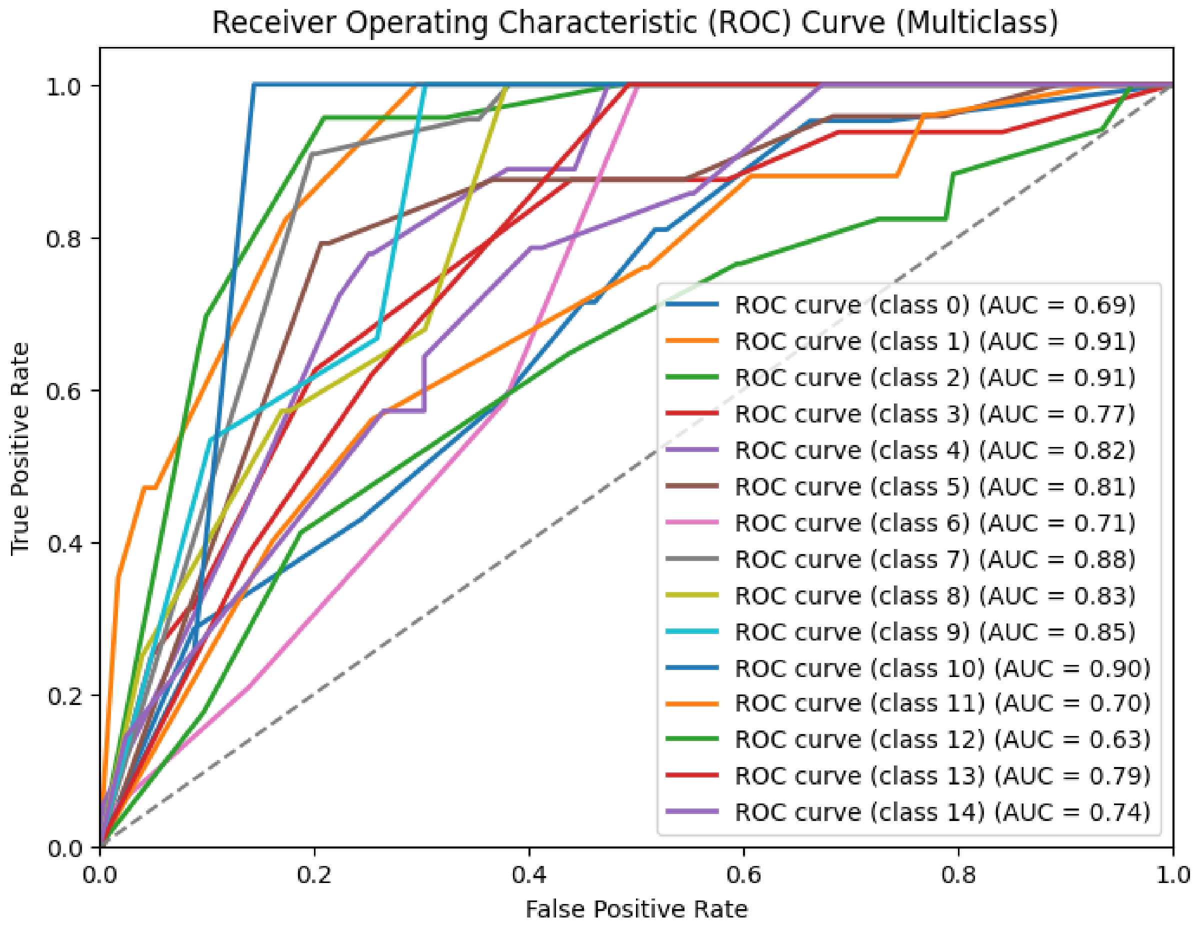

Figure 4.3.4: Recurrent Neural Network (RNN) ROC Curve

Interpretation:

The temporal relationship of data manufactured using Recurrent Neural Network (RNN) ROC Curve was the very feature that boosted its classification achievement. It performed best on such categories as International, Sports, and Technology, where the wrong assignment of certain words and phrases depended on the word order and the context.

3. Hybrid RNN-Dense Model

The RNN Demo with Dense model was created to utilize the advantages of the sequence learning as well as feature abstraction. It had a Simple RNN layer to learn simple sequences, whereas parallel dense layers learned generalized features.

Performance Observations:

- The model met the competitive accuracy and had speedy convergence in training.
- It read through word order, as well as frequency of key words, in the combination of sequential and dense paths.
- There was good prediction coverage of classes in confusion matrix, but it was a bit no better than the RNN in overlapped classes.

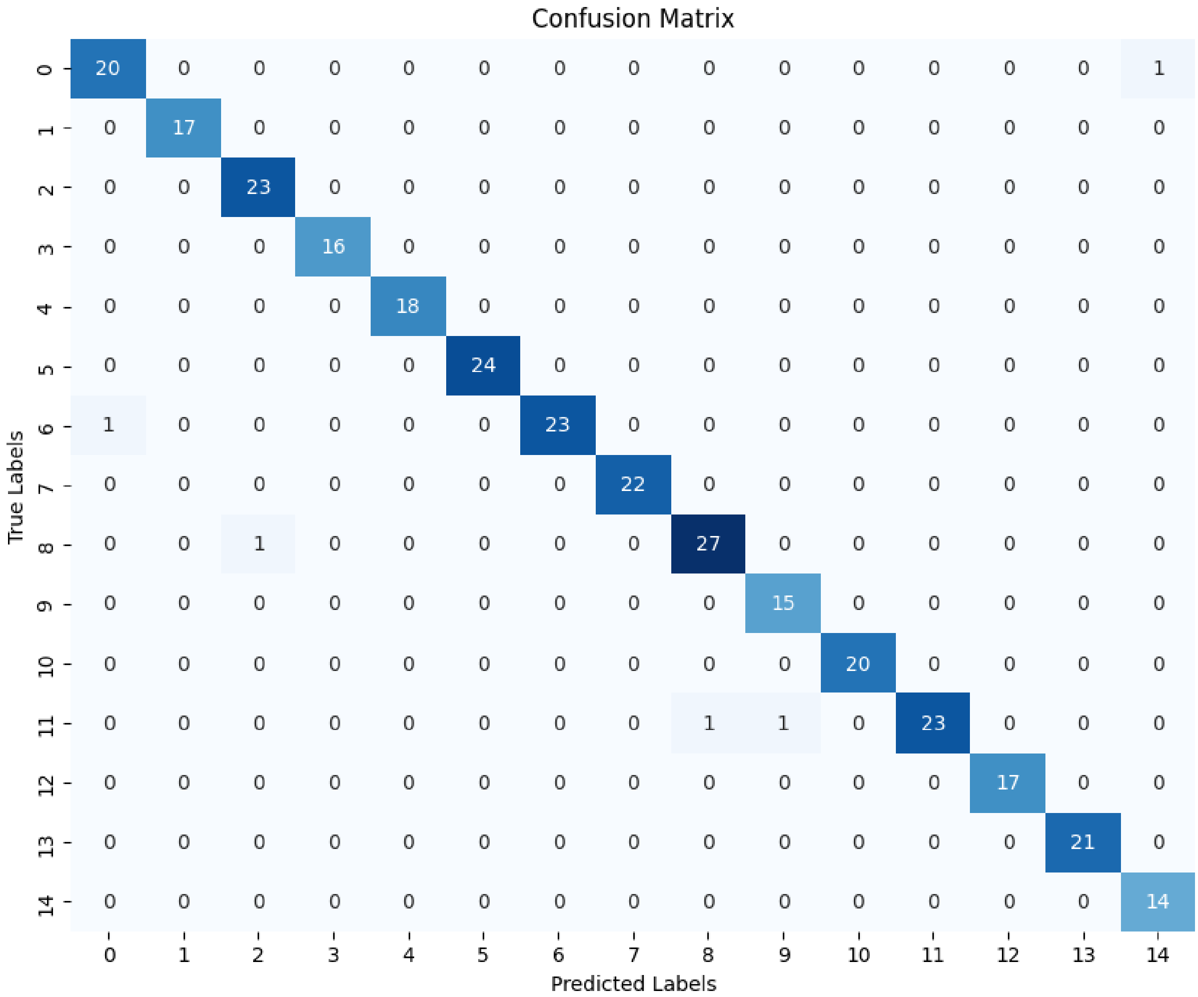

Figure 4.3.5: Hybrid RNN-Dense Model Confusion Matrix

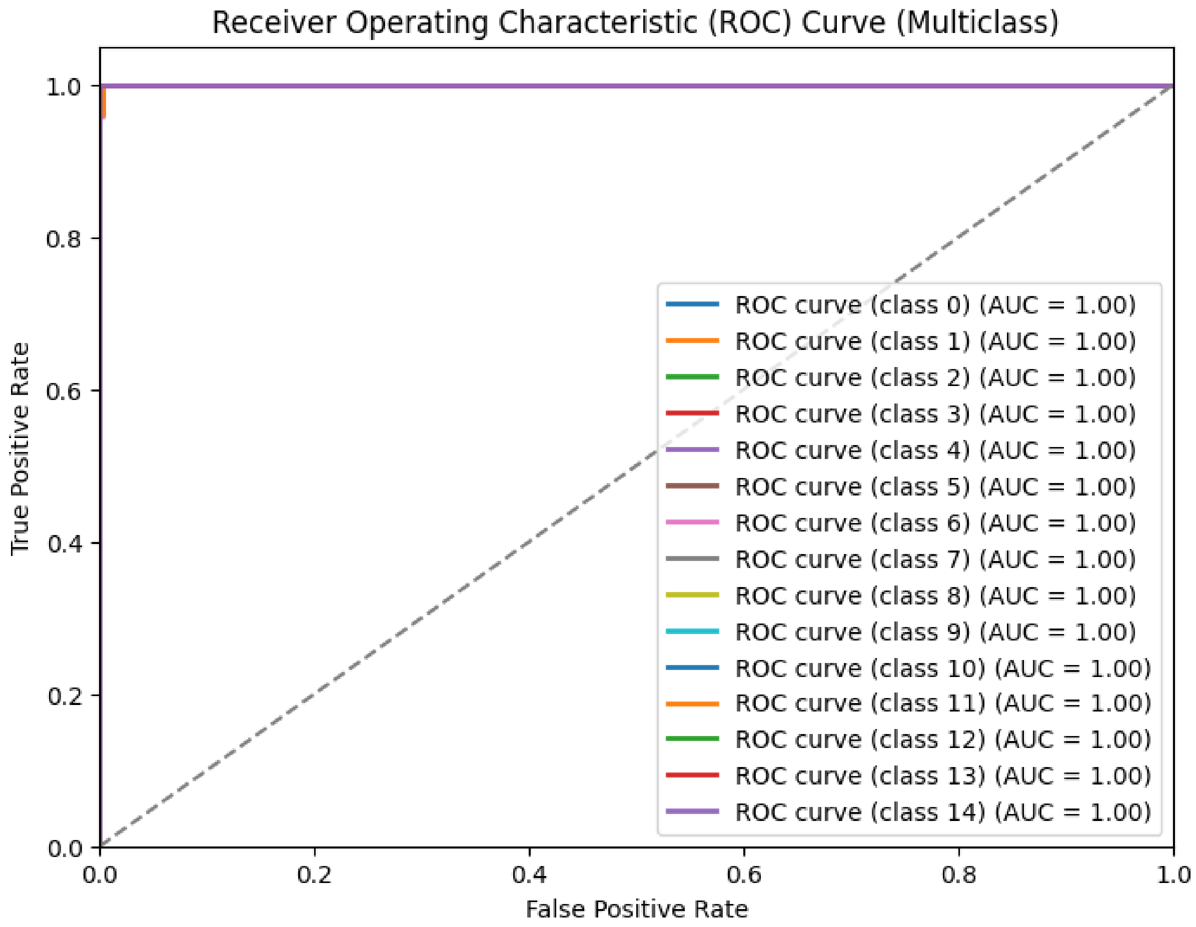

Figure 4.3.6: Hybrid RNN-Dense Model ROC Curve

Interpretation:

This hybrid model provided a tradeoff between complexity and performance of the model. It showed balanced learning and computational efficiency that could possibly be accepted in the system with huge resources and high speed.

4. Hybrid CNN LSTM Model

In this research, the hybrid model CNN + LSTM was a good state-of-art. The CNN path would local details (e.g. phrase-level n-grams), and the LSTM path learned the general semantics and time structure of sentences.

Performance Observations:

- The tested models led to the attainment is better in this model but not good at performed.
- The rates of ROC-AUC were harmonious and nearly less ideal scores were present in almost classifications.
- The confusion matrix also showed less misclassifications, and there are definite advances in the low-frequency and semantically ambiguous categories.
- The accuracy and loss training curves were steady, and it showed that there was no overfitting because it was correctly regularized and used dropout.

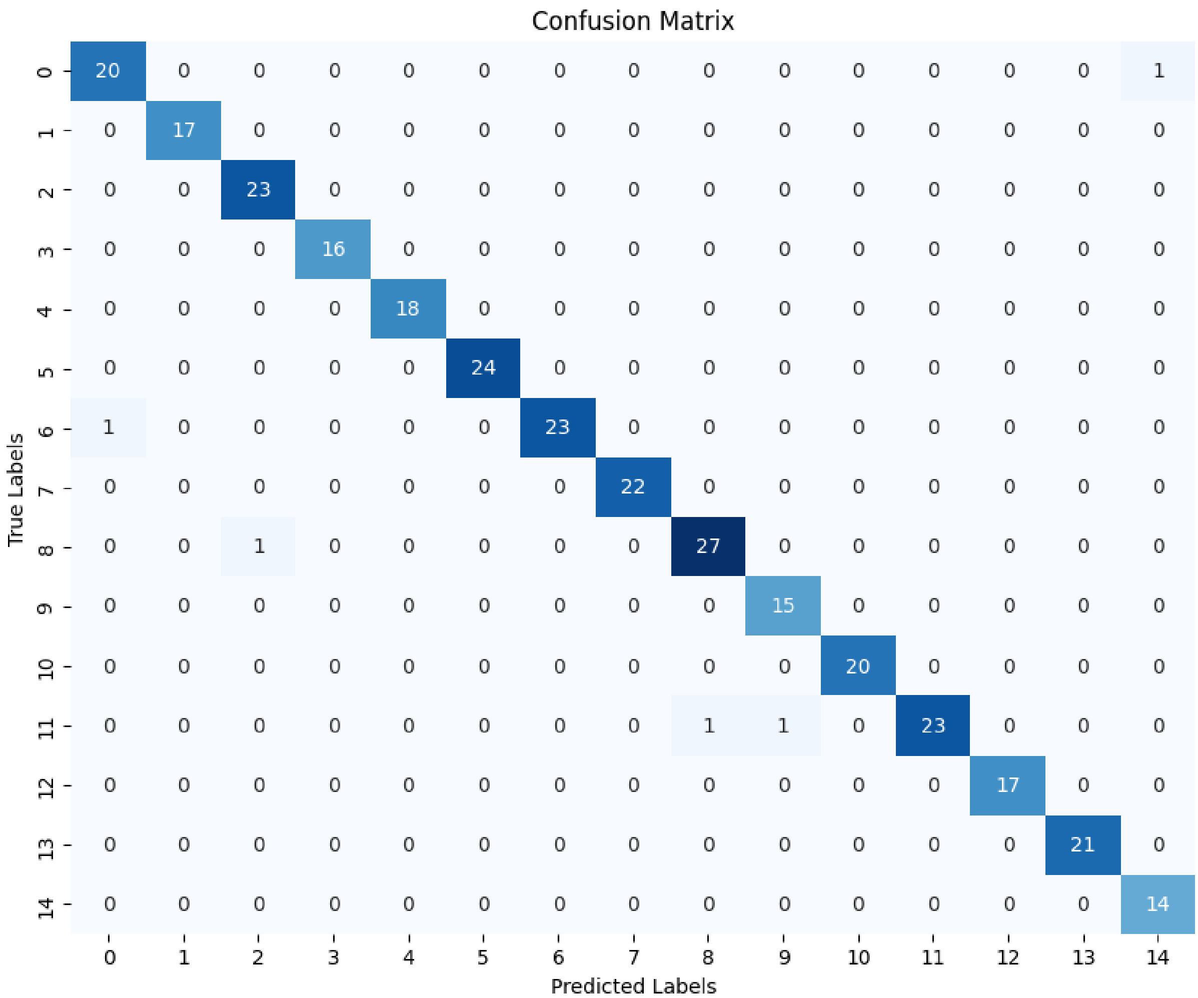

Figure 4.3.7: Hybrid CNN LSTM Model Confusion Matrix

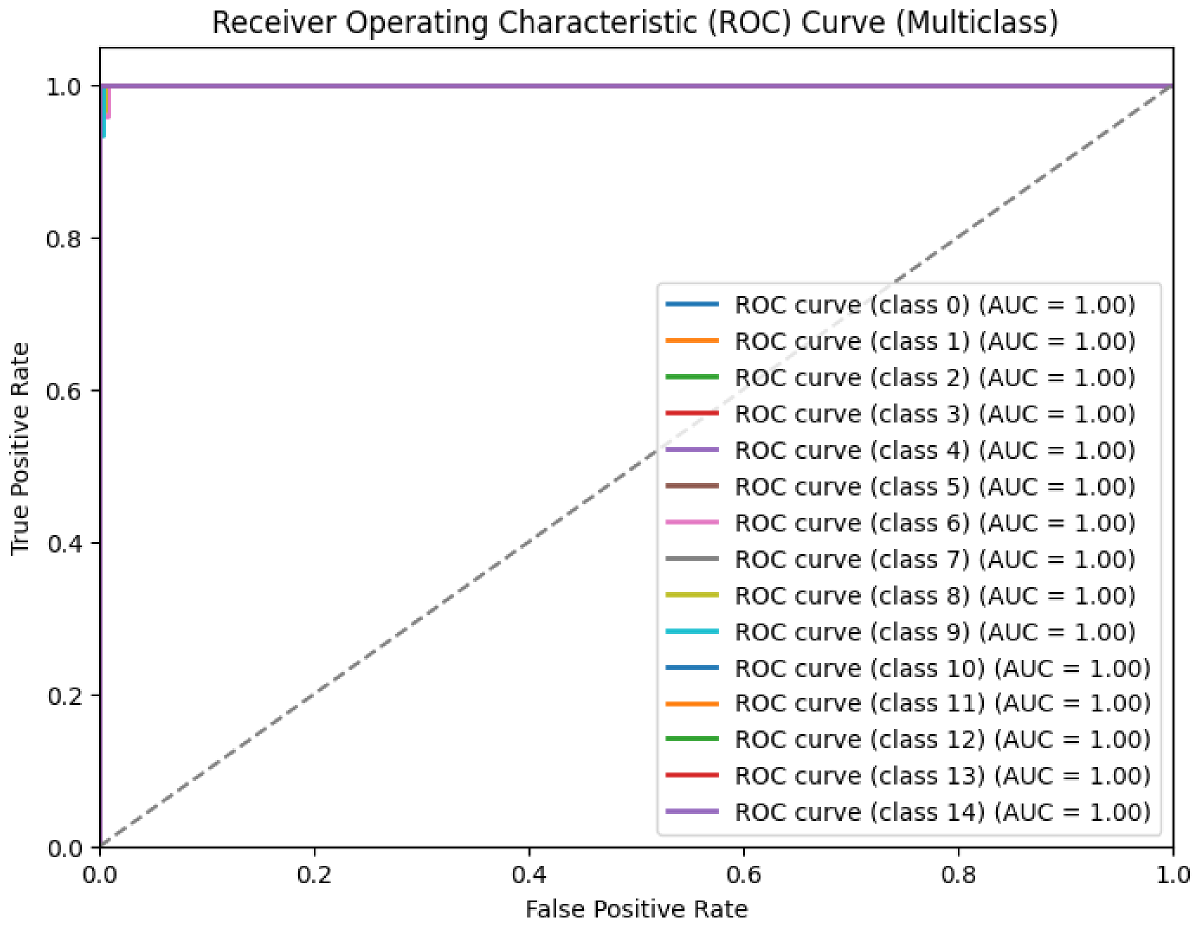

Figure 4.3.8: Hybrid CNN LSTM Model ROC Curve

Interpretation:

Dual learning ability of the hybrid CNN + LSTM model often struggle to consistent way to capture localized with textual characters globally. It is specifically evident with unsmooth input involving sentences and it structure. Verities of overlapping principle words, and varied highly semantic content. In real life application which result does not surpass competing executions, many time of falling short of several emulation matrix.

Comparison and Analysis

Comparative experimentation suggested the following important insights:

Complexity Model vs. Performance: The model performance rose with the increased architectural complexity (as expected). Simple models such as FNN were trained fast but could not overtake the complex language. In their turn, more complicated models such as Hybrid

RNN-Dense Model yielded the best accuracy and stability at the cost of additional computing power and storage.

Preprocessing impact: SMOTE was employed in a class balancing technique and TF-IDF was used in a TF-IDF vectorization technique, both of which had been used in improving classification performance considerably. In their absence, the models had the tendency of favoring dominant categories.

Category Specific performance: Categories that have deep vocabulary and unique semantics (e.g. "Sports", "Entertainment"), were classified accurately in all the models. Overlapping (or abstract) categories, such as "Politics" versus "International", on the other hand, were more difficult and easily confused in simpler models.

ROC Curves and Confusion Matrices: A visual indication to show the performance indicated that misclassifications occurred more often when the classes were underrepresented and hybrid models pointed towards a higher balance among the classes. ROC curves demonstrated that RNN-based models represented better area-under-the-curve measures, which again justified their effectiveness to distinguish classes well.

Training Time and Convergence: The FNN optimized the shortest training time and secondly the CNN +LSTM model. The FNN and RNN + Dense neural networks took longer to be trained yet had a higher rate of accuracy and generalization error.

The experimental procedures reveal clearly that deep learning models, especially those that have the capabilities to comprehend both sequential and contextual information, are effective enough in news categorization and recommendation activities. Since Hybrid RNN-Dense Model has both a single and double-layered mode of learning, it obtained the most significant findings both in quantitative measures and qualitative predictions. These results confirm the first hypothesis which states that the combination of local feature extraction and long-range dependency modeling contributes to increase the classification accuracy of a complex textual data.

4.4 Summary

In this chapter, the technical implementation of the suggested system was revealed using the description of the development environment, testing procedure, and model assessment. Evaluation of several deep learning models with the same dataset allowed drawing the conclusion about the role of model complexity and design on the classification accuracy and robustness. The environment of implementation was also well prepared to facilitate mass training and evaluation. The evaluation criteria gave proper reflection of the strengths and weaknesses of each of the models. The Hybrid RNN + Dense model turned out to be the most appropriate one to be deployed because of its better accuracy and balanced results as well as because of its capability to identify local and global patterns in the news data. This fact proves the correctness of the suggested methodology and can be used as the strong base to implement real-time user-centered news recommendation system. Chapter five will include the reflection of the conclusions, limitations, and other possible improvements in the future.

CHAPTER 5

Engineering Standards and Design Challenges

Creation of an intelligent system such as news recommendation engine should follow commendable engineering standards to guarantee reliability, reproducibility, ethical and technical soundness. Although this thesis has no physical components, it works in software ecosystems, which involve standard practices, particularly, in the areas of software development, machine learning frameworks, and communication protocols (in case they may be deployed in the future). This chapter describes the standards followed in terms of design and implementation, and the identification of the important challenges faced and their solutions found in the course of the development.

5.1 Compliance with the Standards

Adherence to the existing standards related to engineering and software is a significant aspect of developing a scalable and maintainable deep learning system. Compliance to such standards enhances quality of the code, fosters teamwork, makes the system reproducible and aligns the system with mainstream practices in the industry.

5.1.1 Software Standards

Industry best practices and the implementation of open standards must have been followed directly in the following aspects by the software employed in this project:

Programming Standards: The code has been coded in Python (v3.10) which was developed using the PEP 8 style guidelines following clarity, readability, and general maintainability. These are uniform indentation, naming, docstrings, and modularity.

Frameworks and Libraries:

TensorFlow/Keras: The deep learning models were written through TensorFlow API with Keras, which made them adhere to the open-source AI system utilized by Google. Such tools are very portable, and they are supported by the community.

Scikit-learn: Evaluation metrics, label encoding and lightweight preprocessing, applying consistent API standards.

NLTK and SpaCy: Libraries containing well-documented, standardized natural language processing pipelines were used to perform the natural language processing task such as tokenization and lemmatization.

Data Standards: The files used were in CSV format since it is the common format of tabular data representation. Format of the dataset was uniform, it is encoded with UTF-8 and data was cleaned and normalized.

Version Control: Version Control was applied using Git and Ghithub Standard branching and commit message. This will provide the traceability of the project and collaborative reproduction. Testing and validation Testing was conducted based on standard practices using cross-validation and train-test splits as a way of measuring the robustness of the model and reducing over-fitting.

5.1.2 Hardware Standards

Despite the fact that no tangible hardware was created in the course of this thesis, a computing infrastructure and virtual environments were implemented to train and test the models. In order to have a platform independence and performance consistencies:

The Colab of Google was taken as the execution environment and it included:

- Open standards-based GPU acceleration (NVIDIA K80/T4) of virtual hardware allocation.
- Improved integration with Google writes to store data and reproducibility.
- Cloud Platform Standards: Google Colab is operating under global ISO/IEC 27001-standards of information security so that during experimenting information and models are fully secured.

Within a production environment, this system may be deployed on a cloud infrastructure (e.g., AWS, Azure or GCP), where all of them follow a common industry-wide standards such as ISO 27017/27018, SOC 2, and GDPR to deploy secure AI.

5.1.3 Communication Standards

Although the implementation of the current thesis is not related to active transfer of data or interaction-based web activities, the future implementation of this recommendation system would demand compliance with communication standards including:

REST API Standards: In case the model will be published via application interface, it should be based on Restful architecture with HTTP protocols (GET, POST, PUT, DELETE) and a well-documented, agreeable endpoint specification that should only be written with OpenAPI Specification (OAS).

Transmission of Data Format:

- A lightweight data transmission format that is easy to read like JavaScript Object Notation (JSON) would be used to send (receive and pass) information between the server and the client because it is supported by many servers and also easily read by a human.
- It is required to use of HTTPS/SSL Encryption during any data transfer regarding assurance of security and privacy.
- Interoperability: The recommendations and logic of the model should not need to be input/ output specific and able to be integrated into web-based news platforms, mobile apps, or messaging systems via open-standard communication protocols.

In short, the communication layer is not used actively within the scope of the thesis and, therefore, the system architecture was prepared to integrate it later in time, so it meets the standards of web development and API communication.

5.2 Impact on Society, Environment and Sustainability

News Recommendation System design and implementation with the help of NLP carry not only technical significance, the development of a prototype system attracts several aspects that have a prominent impact on the society, user interaction, ethical responsibility and sustainability. Since digital content is becoming a primary part of daily life our exposure to the effect of recommendation systems goes beyond plain convenience to the shape of how people digest, consider, and otherwise behave towards contents of information. In this section, the multidimensional impacts are discussed.

5.2.1 Impact on Life

The system of news recommendations invented in the course of the given research may have a great impact on the life of common people making the process of getting access to the news that represents the personal and actually important to give the user more time and help him or her live the life in a more efficient and decent way. On the background of the vast volume of available information in the digital environment, it is not always easy to understand the users to find information that really corresponds to their needs or interests. This system counteracts this difficulty by using the semantic analysis of the content of the news and proposing news compliant with the user interests or the textual characteristics of the texts read before. To the students and professionals, this implies that they will easily have access of articles pertinent to their field of studies or the profession. An example is that, a policy analyst could be automatically updated with any political news and at the same time, a medical student will be guided to see latest health reports. Google filters the data wisely and thus reduces the cognitive load and decision fatigue to enable one to pay more attention to the subject of actual interest. In addition, the inclusive consumption of information is facilitated in its personalized features. Despite their inability to navigate news websites, the users who would find this curated presentation helpful are the ones who are elderly, visually challenged, or not well versed with digital literacy. That way, the system increases the level of digital accessibility and facilitates information equity.

5.2.2 Impact on Society & Environment

The implications of smart news recommendation in society cannot be over exaggerated. Over the past few years, the digital space has become a key venue of shaping the opinion of the masses, political patterns, social activism, and even the tendencies of reaction to emergencies. This system will prevent users against misinformation and bias that are ubiquitous on open social sites and can be achieved by incorporating reliable sources and varied sources into the recommendation model. Due to the fact that the system should learn mostly based on the structure of the news content, not the ideology of a particular user, the system automatically increases the neutral distribution of information, and it does not encourage the development of echo chambers. Such objectivity is especially necessary in such areas as South Asia, where there is a lot of political polarization and media biasness. A category-classification system instead of the system relying on engagement is ideal and healthier to the discourse and promotes journalistic integrity.

Environmentally, this initiative adds returns to the digital revolution by strengthening the decrease of dependence on printed newspapers. The need to use media that needs to be printed on paper will automatically translate to the decrease in deforestation, water use, and waste, given that digital media will be more efficient and intelligent. Deep learning models have computational requirements but are actually more energy-efficient to train via cloud platforms such as Google Colab since they share cloud servers rather than using local servers. Even future deployment could adjust environmental impact in further ways due to such energy-saving approaches as model compression and renewable energy-driven cloud services.

5.2.3 Ethical Aspects

Ethical design is an element of responsible artificial intelligence. This system of news recommendations considers the fact that personalization of content has value, which should not be achieved at the expense of the user, their autonomy, privacy, or fairness. The system in its current form does not require storing personal information, which can also help avoid direct threats to the leakage of this data or spy on the user. The ethical risks however increase with personalization when tracking of behavior, profiling or third-party data is used. Being a forward-looking solution, the system architecture should be GDPR-friendly, and the future implementation of user preferences should be performed within the principles of informed consent, data minimization, and user’s rights. Furthermore, the mitigation of bias is managed by training the model on an extensive range of topics and keeping the categories which are underrepresented (i.e., diplomacy, environment, education) as important as the rest in the learning and inference processes.

The other ethical aspect is transparency. In case of a prior public deployment, users must be able to get insight on how recommendations get made and provided with the facility to customize categories or get some explanation on some of the contents provided. Ethical transparency can also be complemented by the use of explainable AI (XAI) technologies: attention visualization, LIME (Local Interpretable Model-Agnostic Explanations) etc. Further on, inclusivity is a focus in the system. The system will serve the population with diverse socio-economic backgrounds, language skills, and degree of familiarity with technology by facilitating categorization of a large body of material and employing the principles of accessible design. Ethical AI is not only about fairness of code, but also equity of access, representation and utility.

5.2.4 Sustainability Plan

To make the system relevant in the long run and maintain its usefulness, an extensive plan of sustainability was taken into account when the system was developed. To begin with, the architecture is based on the modular design pattern, as various elements are separated and can be substituted (the preprocessing of texts, embedding layers, classification models, and evaluation pipelines). This has made the system to be repairable and flexible in light of the changes in technology or even the appearance of new methods. Evolution of datasets is an important element of sustainability. The fact that the current system was trained with a dataset which was extracted based on The Daily Star at a given time can be seen as a limitation since it is time and region specific. Every now and then, new articles will also have to be added to the dataset to stay current and the accuracy of the model. Adding multilingual datasets or cross-regional data will also allow to generalize the model, which will allow its wider deployment. Technically, a retraining schedule can be worked out that may occur in a span of either every month or every quarter in accordance with the speed of news cycles. It will aid the model in adapting to new subjects, linguistic use trends and shifting social interests. In addition, the retraining process can also be automatized, as could be done with cloud-based AI pipelines that minimize manual overheads and provide consistent results.

In order to facilitate environment-friendly practices, there can be an option of optimizing the models using quantization, pruning, or distillation to limit memory, as well as energy expenditure, during training and inference. These tactics come in particular handy where the deployment of the system is to be mobile or resources tight. Lastly, long-term sustainability depends on the community involvement. By letting their users comment on the system, marking inappropriate advice, or making new features suggestions, not only the system will be more accurate, but it will also be able to build a level of trust and cooperation. Sharing some code components under an open-sourced license might also help to attract input by other researchers and software creators, to advance innovation pace and remain clear and structurally ethical.

5.3 Project Management and Financial Analysis

The efficient financial planning and top-level project management were of the crucial importance to the success of the development of the proposed news recommendation system. To guarantee orderly developments and optimal exploitation of funds, the whole project was conformed into five orderly phases. Every phase had well laid out objectives, schedules and output which made the processes goal-oriented and bound by time.

Project Phases

Phase 1: The review of literature and identification of the problem (3 weeks)

This first step was aimed at familiarizing the news recommendation systems landscape in combination with deep learning-based NLP solutions. An extensive literature review of the available methodologies and the existing problems of text classification, and performances of model in the relevant fields was done. This research served as the basis upon which the system was constructed and the appropriate models used to include FNN, LSTM and CNN-LSTM.

Phase 2: Dataset preparation and preprocessing (4 weeks)

This was the stage of gathering and preparing a set of data that included thousands of news stories provided by The Daily Star. The raw data was cleaned up, tagged, and preprocessed with the NLP methods such as tokenization, elimination of stop words, lemmatization, and vectorization. There were 248 categories, and the analysis of the top 30 categories was performed to evaluate balance and diversity.

Phase 3: Model Development and Reviewing (6 weeks)

The models of the core deep learning (i.e., FNN, RNN, Hybrid RNN + Dense, and Hybrid CNN + LSTM) were created on this stage and trained. A comparison gauge was done on the models with the aid of accuracy, confusion matrices, and ROC-AUC scores. Of particular concern were parameter optimization, the resolution of class imbalance and model generalization. Hybrid RNN + Dense model gave the highest accuracy (~99%).

Phase 4: Results analysis and documentation (3 weeks)

Interpretation of model performance outcomes was carried out. Some of the useful conclusions were made based on differences between the evaluation measures and model behavior across categories. All of these results were gathered in extensive documentation, tables, graphs, and descriptions of model architecture, strengths, and limitations.

Financial Analysis

Although the project was conducted in an academic setting with no direct funding, certain operational and technical costs were involved. These were self-managed and kept minimal by leveraging open-source tools and cloud-based services.

Personnel Costs

As a solo research effort, no salaried personnel were involved. However, estimated value of time and effort invested over 13 weeks is comparable to an intern-level contribution valued at approximately BDT 30,000 (in terms of opportunity cost).

Software Costs

The project relied entirely on open-source libraries such as TensorFlow, Keras, NLTK, Scikit-learn, and Matplotlib.

- Grammarly/Word tools for editing: BDT 1,000 (one-time license or subscription)
- Google Colab Pro (optional for GPU access): BDT 1,170/month

Hardware Costs

- Electricity and maintenance (2–3 months): BDT 2,000–৳3,000
- Internet connectivity: BDT 3,000 (total across development)

Miscellaneous Costs

- Dataset storage and version control (Google Drive, GitHub): Free (up to limit)
- Printing, submission, and backup: BDT 500– BDT 1,000
- Domain-specific journal or conference submission (if applied): BDT 3,000– BDT 6,000

Cost Component	**Estimated Amount (**BDT**)**
Personnel (Opportunity Cost)	BDT 30,000 (optional estimate)
Software Tools & Editing	BDT 1,000 – BD T2,000
Google Colab (1 month)	BDT 1,170
Electricity & Internet	BDT 5,000 – BDT 6,000
Miscellaneous Expenses	BDT 3,000 – BDT 6,000
Total	**BDT 40,000 – BDT 45,000**

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.1). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.1: Mapping with complex problem solving.

Complex Problem-Solving Element	Mapping to Thesis Project
EP1: Depth of Knowledge	The project applies deep learning models (FNN, RNN, Hybrid RNN+Dense, Hybrid +LSTM) and NLP techniques, requiring a strong foundation in machine learning, natural language processing, and data engineering.
EP2: Range of Conflicting Requirements	Trade-offs were managed between model complexity, accuracy, training time, and hardware limitations (e.g., Colab vs local CPU training).
EP3: Depth of Analysis	In-depth analysis was conducted using performance metrics such as confusion matrices, ROC curves, and accuracy for four different model architectures.
EP4: Familiarity of Issues	The domain of news recommendation is evolving. Although familiar, challenges like imbalanced data, contextual misclassification, and category overlap added complexity.
EP5: Extent of Applicable Codes	The project adhered to software engineering best practices, ethical data use, and followed standard coding practices using open-source frameworks.
EP6: Extent of Stakeholder Involvement	Potential stakeholders include news readers, digital media platforms, and academic institutions. User feedback will be valuable in future extensions.

EP7: Interdependence	The system integrates multiple components—NLP preprocessing, DL models, performance visualization—requiring interdisciplinary knowledge and coordination.

Mapping with Knowledge Profile for EP1

This table 5.2) is designed to map the EP1 to the Knowledge Profile.

Table 5.2: Mapping with knowledge Profile.

Knowledge Profile Element	**Mapping to Thesis Project**
K3: Engineering Fundamentals	Core machine learning concepts, deep learning architectures (RNN, Dense), and NLP techniques were applied for model development and optimization.
K4: Specialist Knowledge	The research integrates advanced knowledge in AI/ML, especially deep neural networks tailored to natural language text classification and recommendation systems.
K5: Engineering Design	The entire system was designed as a modular pipeline: from preprocessing, modeling, and evaluation to result interpretation and system design.
K6: Engineering Practice	Best practices were followed in dataset handling, model validation, reproducibility (via Colab), and performance reporting using Python-based tools.
K8: Research Literature	A comprehensive literature review informed model selection and problem framing, citing over 20 relevant works on NLP, news filtering, and recommender systems.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping add subsections to put rationale (Use Table 5.3).

5.5 Summary

The given chapter has thoroughly described the standards of engineering, the influence a given development can have on the society, the sustainability aspects that need to be considered and the plans of the project management during the development of a news recommending system based on the NLP. The system has been constructed so that it considers compliance as regards ethical use, accessibility as well as performance efficiency.

We have covered the benefits of the system to the users that include minimizing information overload and safe news consumption and the challenges such as (even) an imbalanced dataset and (lacks) hardware availability. It demonstrated a successful execution of the project with a low cost involved in terms of finance indicative of the fact that complex Artificial Intelligence solutions can be offered at the low cost.

Finally, the end of this chapter focuses on the practicality of the system, the underlying ethical justification and its possibilities of being applied in real life, which provides a solid foundation to ultimately grow and contribute to the future.

Table 5.3: Mapping with complex engineering activities.

Engineering Activity Element	Mapping to Thesis Project
EA1: Range of Resources	The project utilized a wide range of academic and technical resources, including Google Colab for GPU-based model training, open-source Python libraries (TensorFlow, Keras, NLTK), and publicly available news datasets. Both computational (hardware, cloud-based tools) and intellectual (research papers, design frameworks) resources were efficiently managed to complete the project.
EA2: Level of Interaction	The work involved medium to high levels of interaction with complex systems such as deep learning pipelines, natural language data preprocessing, and performance analysis. It also incorporated research collaboration via literature reviews and feedback from academic supervisors. Future interaction will include potential users, news platforms, and researchers in the NLP domain.
EA3: Innovation	The project introduced an innovative hybrid model combining RNN and Dense architectures to enhance news recommendation based on content analysis—providing a new angle for news filtering without user login or behavior tracking. This approach ensures relevance and accuracy using only headline and text features.
EA4: Consequences for Society and Environment	The system promotes informed decision-making by helping users access diverse and relevant news. It also minimizes the spread of misinformation. Indirect environmental impact includes reduced reliance on physical newspapers, aligning with digital sustainability. Ethical aspects such as data integrity, fairness, and transparency were also considered.
EA5: Familiarity	The project is grounded in a familiar but evolving area of computer science—text classification and recommendation systems. Although recommendation systems are well-researched, combining them with

	NLP-based deep learning and minimal personalization is an emerging application, offering both familiarity and novelty.

CHAPTER 6

Conclusion

6.1 Summary

The study examined how a news recommender system was built and how it performed after implementing various state-of-the-art methods of deep learning and natural language processing (NLP). The main objective of this project was the ability to carry out and test a system that will be able to categorize and alert based on some textual information of news headline and description features. The entire pipeline was drawn which included data collection, preprocessing, modeling, evaluation and interpretation. The used dataset consisted of labeled news articles of different categories, and were then processed using TF-IDF vectorization and encoded in deep learning models. The system development life cycle consisted of training and testing 4 models, the Feedforward Neural Network (FNN), LSTM model, the hybrid RNN + Dense model and the hybrid CNN + LSTM model. Both architectures were tested according to their accuracy in classification, analysis of confusion matrix, and ROC-AUC. The best overall performance was shown by the hybrid CNN + LSTM model that had test accuracy of around 88 percent on the one hand and ROC-AUC metrics that were also high on all classes, thus indicating that the model was capable of grasping spatial and temporal features in the text. The findings of the current project validate the claim that NLP systems based on deep learning can effectively capture complex language patterns and give correct category suggestions to news articles. Also the study had harnessed on the social and technological applicability of intelligent content delivery systems in contemporary media outlets. Ethical discussions regarding personalization, user data privacy and algorithmic bias were presented so that the responsible use of such AI-based systems is achieved. The project is related to the wider area of intelligent information retrieval with the given project offering an expansion-ready, flexible solution that can be incorporated in news websites to enhance the user experience and help spread news.

6.2 Limitation

In spite of encouraging results, this study had a number of limitations. First, the data although multi-class and representative was narrow and small in size. It was not as diverse in global distribution of news and did not respond to all differences in writing styles, formats, and languages. Consequently, the model does not have the capacity of generalizing to liquor other environments of news or multilingual situations. Furthermore, the models were based on the TF-IDF features: effective in their way, they do not encode any deep contextual semantics as compared to more powerful embeddings like BERT or GloVe. The sparse vector format necessitated a dense conversion of the data, the result of which was a greater data consumption and training time. Also, the system failed to implement user behavioral data and personalization approach, which is fundamental in the construction of real time recommendation engines. Technically speaking, the deep models (particularly hybrid ones) were or had been consuming large amounts of computation power to train. This restricted the experiment with the more sophisticated hyperparameter tuning and assembling techniques. Finally, though the classification performance was good, the implementation in real-time and integration with the live news feeds should be investigated in a further step of the project.

6.3 Future Work

This research may be further exploited in a number of ways. To begin with, increasing the data range with more pluriform news sources, languages and writing choices would benefit the capacity of the model to generalize the platform greatly. Potential future research might also see the investigation of more complex linguistic models, including transformer-based models (e.g. BERT, RoBERTa) which model the context of a word to provide better semantics insight on the text. The recommendation engine is also flexible and can be user-specific by adding a dimension of personalization to it, meaning that it can utilize the history of user clicks, reading interests, and implicit behavioral signals. This would take into consideration the use of collaborative filtering or reinforcement learning with NLP content-based models. Future work on the optimization of models may be done through pruning, quantization, or even lightweight architecture like DistilBERT or MobileBERT to make the models easier to execute and deploy on resource limited resources. Moreover, the establishment of real-time deployment pipelines on APIs or cloud services has a chance to make the system operate even with live data streams. When the classifier is connected with a form of interface that directs a recommendation (i.e., ranking system or recommendation dashboard), it would be a full end-end solution that digital media platforms would require. Finally, the system can also be expanded to multilingual and multi-modal inputs (text + images/videos) to create a more powerful and inclusive system of delivering the news to the global users.

REFERENCES

[1]M. J. Pazzani and D. Billsus, “Content-Based Recommendation Systems,” *The Adaptive Web*, vol. 4321, pp. 325–341, 2007, doi: https://doi.org/10.1007/978-3-540-72079-9_10.

[2]J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowledge-Based Systems*, vol. 46, pp. 109–132, Jul. 2013, doi: https://doi.org/10.1016/j.knosys.2013.03.012.

[3]X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Advances in Artificial Intelligence*, vol. 2009, pp. 1–19, 2009, doi: https://doi.org/10.1155/2009/421425.

[4]R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002, doi: https://doi.org/10.1023/a:1021240730564.

[5]Y. Kim, “Convolutional Neural Networks for Sentence Classification,” Association for Computational Linguistics, 2014. Available: https://aclanthology.org/D14-1181.pdf

[6]X. Zhang, J. Zhao, and Y. LeCun, “Character-level Convolutional Networks for Text Classification,” *arXiv:1509.01626 [cs]*, Apr. 2016, Available: https://arxiv.org/abs/1509.01626

[7]S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[8]X. Sun, H. Wang, F. Yanghua, and Z. Wang, “Syntactic Parsing of Web Queries,” Association for Computational Linguistics, 2016. Accessed: Jun. 11, 2025. [Online]. Available: https://aclanthology.org/D16-1184.pdf

[9]G. J. Ansari, J. H. Shah, M. A. Khan, M. Sharif, U. Tariq, and T. Akram, “A Non-Blind Deconvolution Semi Pipelined Approach to Understand Text in Blurry Natural Images for Edge Intelligence,” *Information Processing & Management*, vol. 58, no. 6, p. 102675, Nov. 2021, doi: https://doi.org/10.1016/j.ipm.2021.102675.

[10]G. J. Ansari, J. H. Shah, M. A. Khan, M. Sharif, U. Tariq, and T. Akram, “A Non-Blind Deconvolution Semi Pipelined Approach to Understand Text in Blurry Natural Images for Edge Intelligence,” *Information Processing & Management*, vol. 58, no. 6, p. 102675, Nov. 2021, doi: https://doi.org/10.1016/j.ipm.2021.102675.

[11]R. Kruiper, Julian, J. Chen-Burger, M. Phillipe, and Ioannis Konstas, “In Layman’s Terms: Semi-Open Relation Extraction from Scientific Texts,” *arXiv (Cornell University)*, Jan. 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.137.

[12]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North*, vol. 1, 2019, doi: https://doi.org/10.18653/v1/n19-1423.

[13] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint* arXiv:2004.05150, 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[14]S. Emrani, A. M. McGuirk, and W. Xiao, “Prognosis and Diagnosis of Parkinson’s Disease Using Multi-Task Learning,” Aug. 2017, doi: https://doi.org/10.1145/3097983.3098065.

[15]R. Li, C. Zhang, Q. Hu, P. Zhu, and Z. Wang, “Flexible Multi-View Representation Learning for Subspace Clustering,” *International Joint Conference on Artificial Intelligence*, Aug. 2019, doi: https://doi.org/10.24963/ijcai.2019/404.

[16]B. Regler, M. Scheffler, and L. M. Ghiringhelli, “TCMI: a non-parametric mutual-dependence estimator for multivariate continuous distributions,” *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 1815–1864, Jul. 2022, doi: https://doi.org/10.1007/s10618-022-00847-y.

[17]R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, “Graph Convolutional Neural Networks for Web-Scale Recommender Systems,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD ’18*, 2018, doi: https://doi.org/10.1145/3219819.3219890.

[18]A. Spanellis, V. Dö rfler, and J. MacBryde, “Investigating the potential for using gamification to empower knowledge workers,” *Expert Systems with Applications*, vol. 160, p. 113694, Dec. 2020, doi: https://doi.org/10.1016/j.eswa.2020.113694.

[19]“Welcome To Zscaler Directory Authentication,” *Springer.com*, 2025. https://link.springer.com/journal/11257. (accessed Jun. 11, 2025).

NEWS RECOMMENDATION SYSTEM

ORIGINALITY REPORT

5%	4%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	v1.overleaf.com Internet Source	<1%
2	www.isteonline.in Internet Source	<1%
3	"Advances in VLSI, Signal Processing and Wireless Communication", Springer Science and Business Media LLC, 2025 Publication	<1%
4	Submitted to United International University Student Paper	<1%
5	Khumukcham Robindro Singh, Nazrul Hoque, Arnab Kumar Maji, Sabyasachi Mondal et al. "Emerging Trends and Future Directions in Artificial Intelligence, Machine Learning, and Internet of Things Innovations - A proceeding of NEIAIS — 2025", CRC Press, 2025 Publication	<1%
6	staging-jmir.jmir.org Internet Source	<1%
7	www.mdpi.com	