National College of Ireland

Detecting misinformation Across Social Media, Healthcare, and Job Posting Websites using Machine Learning and Deep Learning Techniques

MSc Research Project
Data Analytics

Navya Ravichandran
Student ID: 22241990

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Navya Ravichandran
Student ID:	22241990
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Athanasios Staikopoulos
Submission Due Date:	12/12/2024
Project Title:	Detecting misinformation Across Social Media, Healthcare, and Job Posting Websites using Machine Learning and Deep Learning Techniques
Word Count:	8961
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Navya Ravichandran
Date:	28th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting misinformation Across Social Media, Healthcare, and Job Posting Websites using Machine Learning and Deep Learning Techniques

Navya Ravichandran
22241990

Abstract

In recent days there is a rapid growth in technology across various streams like social media, Health care and Job Posting websites. This growth has led to the spread of false information called rumors among the public which have reduced their safety and trust in usage of technology and social websites. To address this, in our paper we will be developing a model that uses both Machine learning and Deep Learning algorithms in detection of fake news among multiple streams like healthcare, job posting websites and social media. Here datasets that have both images and texts are trained and tested against multiple advanced algorithms. A hybrid model is developed as a result to predict the false information with various data collection and Data Preprocessing techniques. Results show the efficiency of hybrid model by comparing its prediction accuracy across the techniques implemented. The implementation of this multi modal data helps to share only reliable online information across social media and other platforms. The model in final helps to promote social safety and trust among the end- users. The study includes combination of ten machine learning and deep learning algorithm across two datasets and CNN has better results comparatively. A detailed comparison is explained further in the implementation section.Here the machine learning data performs well on text data and deep learning algorithms performance is better on image data. Here the real time deployment of the model is a challenging and it will be addressed in future studies.

1 Introduction

1.1 Background

The rapid growth of social media and other digital platforms have made all information available in a single place for end-users. This results in a digital society where the spread of false information across social media, HealthCare and Job Posting websites has become a critical issue where the safety and trust of public is threatened using digital content. This viral false information has resulted in a great question among the public in trusting social media and technology. The false news across healthcare might mentally affect the patients and doctors lead them to danger practices. As thousands of youngsters and the public start their day searching for a job, this type of fake job postings can mitigate their job search practices emotionally. This single phenomenon plays a great role in the mental

and financial loss of every single person in society who uses social media and other digital platforms. A lot of economic consequences are also faced by the government, and they have been trying in several ways to overcome this distress but most of them have not shown efficient results.

To overcome the trust among the public the development of a hybrid model is focused on this research. Advanced Machine and Deep Learning algorithms are implemented and the prediction accuracy among them is compared to produce a robust final model in detection of fake news in digital platforms. Here the complexity of using both texts and images is addressed by using multimodal data which comprises of both images and texts. In this a structured methodology process is undertaken from data collection, data preprocessing, implementation of algorithms and final evaluation of the models using its performance metrics. In both image and text datasets, this aids in the development of a scalable solution using sophisticated machine and deep learning techniques such as support vector machines, K-Nearest Neighbor, Decision Trees,Random Forest, Naive-Bayes, Visual-Bert, Convolutional Neural Network, Recurrent Neural Network, Artificial Neural Network, and LSTM. A robust model that may be utilized to detect misleading information across social media applications is found by comparing evaluation criteria such as accuracy, precision, recall, F1-score, and confusion matrix of all the algorithms.

1.2 Research Motivation

The ultimate motivation in this research is to increase the trust among the common public in usage of social media and other technological websites. As social media usage and applications increase daily the spread of false information pretends to increase where this results a great question among end users in sharing their data across such platforms. There is huge financial and mental loss for both application provider and its end users. In social media most of the fake news is about politics, health care and fake job postings. In this research the goal is to implement both machine learning and deep learning algorithms among both image and text datasets across various domains and a detailed comparison of such techniques is explained. The unique motivation here is to use advanced techniques on various domain multimodal data in order to find a robust model in detection of false and misinformation.

1.3 Research Question and Objectives

The ultimate aim of the proposed research is to answer the below developed research questions:

- **RQ:** What are the finest and most efficient deep learning and machine learning methods for detecting fake news on multimodal datasets from various digital media platforms?

The core objectives of the research is to address the following key points :

- Implementing advanced Machine learning and Deep Learning algorithms to find a suitable model that can be used in detection of false news.

- A comparison study between each algorithm is performed by using their evaluation metrics such as Accuracy, Precision and recall score.

- The model is tested against multimodal datasets, the dataset with texts and images.
- Datasets are collected from three domains i.e. Health care, social media and Job posting websites and their performance all evaluated.

The research topic chosen has more significance and contribution to public safety and trust which can reduce the crimes and disappointments in usage of digital platforms. The study follows KDD Methodology, and a robust model is developed to detect the false rumors' a result a detailed comparison is performed among both machine learning and deep learning algorithms to find a robust model.

1.4 Document Structure

The entire research document has several sections that explains the entire research from scratch.Document starts with abstract and introduction where the aim,objective,motivation and research question is outlined in brief.Next is the related work section in which a detailed study about existing and previous works are carried out.In third section its methodology where the project flow and implementation is explained.In design section project design and algorithms are outlined for better understanding of the user.The fifth section is evaluation section where the results of implementations are compared and explained.The research document is wrapped with conclusion and references used in entire research.

2 Related Work

2.1 Comprehensive Review of Machine Learning Techniques for Fake News Detection.

Dsouza and French (2024) conducted research to overcome the threat faced by society in spread of fake information in their daily usage of social media. A dataset that has real and false news has been taken from kaggle for the study.LSTM is used and as it have shown better results and then adversarial approach using GPT2 and SeqGAN few machine learning algorithms was performed and it have showed less results comparatively than LSTM.The research remained incomplete as the obtained results were bias and it was conducted only on limited data. Researchers concluded that a lot of future improvements need to be implemented further.

This research conducted by Choudhary et al. (2021) includes implementation of machine learning techniques called Naive Bayes, Convolutional Neural Network, LSTM, Neural Network and Support Vector Machines in social media websites. It is a very basic study which has lot of gaps in it in model implementation like there is imbalance so in the result all the implemented algorithms have shown great metrics score which need to examine further by the author to develop a robust model without misdetections in it.

According to this article proposed by Abdullah et al. (2024) three machine learning algorithms Decision Tree, Naive Bayes and SVM is implemented to classify the fake news across social media websites. Real time data was taken from twitter with the help of API access. Covid-19 false data is also included in this research. Overall healthcare and politics are considered here.Hyperparameter tuning is also considered to measure the performance of the model. Finally, the Decision tree algorithm has shown better performance for both text and meta data than hypertuned model.

In the research done by Jouhar et al. (2024) a basic comparison is made between six machine learning algorithms. SOT fake news dataset is used for model implementation.TF-IDF is used for embedding. Logistic regression is initially implemented as base model by author following Decision Tree, Random Forest, Gradient Boosting, XGBoost and Passive Aggressive Classifier was used in which XG-Boost achieved highest performance in detection of fake news from the given data.

Authors Babu et al. (2023) Conducted a study using Machine Learning algorithms with detailed architecture flow to detect the misinformation spread across news articles. Dataset used in this research is self-sourced from the internet and finally two datasets are merged with fake and truth columns in it for further analysis. All the models have performed better but Decision tree outperforms other models with a high metrics score. Overall, the study has shown better results comparatively.

2.2 Deep Learning Methods for Effective Fake News Detection

There was huge false information's spread during Covid-19 all over the world.In this paperWani et al. (2024) have proposed a model to predict toxic and toxic behaviors that have been spread as rumors across social media websites which have leaded to multiple issues across every individual. Here a combination of BERT SVM and BERT RF is used. Here four datasets D1, D2, D3 and E1 were used which were collected for Facebook and twitter for analysis. BERT SVM and BERT RF have shown better results than all other combinations implemented in the study by author.

This research conducted by Al-Tarawneh et al. (2024) includes implementation of both machine learning and deep learning techniques to maintain the information Morality among the public users. TruthSeeker a social media fake and real news collection dataset for the period of (2009 to 2022) is used in the research Various word embedding techniques have been considered in the study. Here SVM have shown better model performance in terms of all evaluation metrics. Deep learning algorithms like Convolutional neural network techniques have been implemented and overall comparison analysis shows SVM as the robust model.

This research proposed by Men and Mariano (2024) is about detecting fake news during Covid-19 using advanced deep learning techniques. "COVID Fake News Dataset" is used in the study. Here Bi-LSTM is chosen as a base model and further BERT and SHAP models have been implemented. Knowledge distillation with BERT is performed in this study to achieve efficiency. Various combinations with Bi-LSTM and BERT are performed and detailed comparative analysis is shown. Results have shown good results in implementation of SHAP and BERT model techniques. Further analysis is mentioned to be implemented in future using real world live data.

Authors Vu et al. (2024) developed a solution to identify fake job predictions across job posting websites and online portals using Deep learning and NLP Techniques.Word2Vec is used to extract text data features. BERT is also used along with Word2Vec, and it is evident that Word2Vec outperforms through all the combinations. After all combinations, finally BERT-NLP2FJD performs a little better than Word2Vec-NLP2FJD in detecting Fake Job Description. This problem statement is really a great thought by author which eliminates fake job frauds among youngsters and job seekers.

In this research done by Mahara et al. (2023) CBM and FBM models are considered for implementation. CNN-LSTM along with CNN-BiLSTM is compared with all machine learning algorithms for further analysis. Feature based models have shown better per-

formance than CBM techniques. AdaBoost-Random Forest have shown a good accuracy than other FBM models in classifying the fake news among Internet.

2.3 Cross-Modal Deep Learning and Machine Learning Approaches and Techniques for Detecting Fake News

In the study proposed by LAHLOU et al. (2024) and embedding system is developed using multimodal data in fake news prediction. Here only the transformer models BERT, ELECTRA, and XLNet have been implemented through the articles collected from the twitter. Few basic Machine learning algorithms also have been tested on the model developed where those models results in less accuracy and validation due to inefficient way of data handling comparing to other transformer based models. BuzzFeed and PolitiFact are the major resources used to build an dataset called Fake news net used in the study. The comparison of models used in the study shows that the BERT model shows better performance than other implemented models. From this study we can incorporate the usage of BERT and domain-specific transformer models in our study which improves the feature extraction and accuracy of the classification model developed.

Agrawal et al. (2024) Proposed an study using multimodal data in detecting fake news using BERT and ResNet110 by performing hyperparameter tuning. Here HTBERT is used for text data and ResNet110 is used for images. Dataset named Fakeddit is used in the research and have shown better results comparing RCNN and other deep learning algorithms. Here even the results are good the model developed is not optimized to predict the results simultaneously and only basic hyperparameter tuning steps have been followed by the author to achieve results.A proper feature fusion should be adopted in the current research from this study.

In the study Dwivedi and Wankhade (2024) have proposed both text and image combined model using multiple deep learning techniques in order to detect false informations.Weibo dataset is used here with pretrained modals and accuracy metrics is compared across Single Text Modal , Single Vision Mode and Multimodal.Among all the techniques KMGCN multimodal outperforms with greater accuracy.Albation analysis is also performed but same results are obtained.The study have acheived its aim by implementing the semantic-enhanced using multiple techniques by trial and error methods.

This research done by Festus Ayetiran and Özgöbek (2024) is proposed to detect fake news and harmful languages in internet.Deep neural network (DNN) is used here. Text,Image,Meme and Audio data formats are used.A survey about the topic is included in the research.More than ten datasets are used for implementation across various combination of deep learning techniques.An comparison of evaluation metrics is performed across the model implementation.Addressing Missing modalities is considered to be done in future implementation.

This author Giachanou et al. (2020) have conducted research in multi-image fake news detection using BERT. VGG-16 model is used to extract the visual features as BERT only can handle text-data. Real world dataset is used for implementation and BERT and Spotfake is used as baseline model. Embedding is also performed a=on image captions. Here multi-image system outperforms better. Semantic techniques are also implemented.

2.4 Key Findings and Gaps From Literature Review

Overall, from the literature finding there are minimal gaps and takeaways considered in the below research. First is the data collection in this research the dataset includes both texts and images from three domains namely Job postings, Healthcare and social media. A detailed comparison between traditional machine learning and advanced deep learning algorithms is performed. Multimodal implementation using deep learning is implemented and has shown better results comparatively. The gaps like usage of imbalanced datasets, Bias and overfitting issues are addressed. Text and image data are treated equally. The main takeaway from this section is that some machine learning and deep learning evaluations and methods can produce predictions that are more accurate. This study will look for patterns using a range of pre-processing methods, feature engineering, and feature selection in fake news detection across three domains combined Job postings, Healthcare and social media.

3 Methodology

The Methodology section includes the research plan and overflow of entire project that is considered. Ultimate motivation of the research is explained from the technical perspective. To create the final robust model, the study uses machine learning and deep learning algorithms on a set of text and image data utilizing the Knowledge Discovery Database technique..KDD process includes data collection, preprocessing, and transformation and final step is evaluation and comparison of performance metrics.

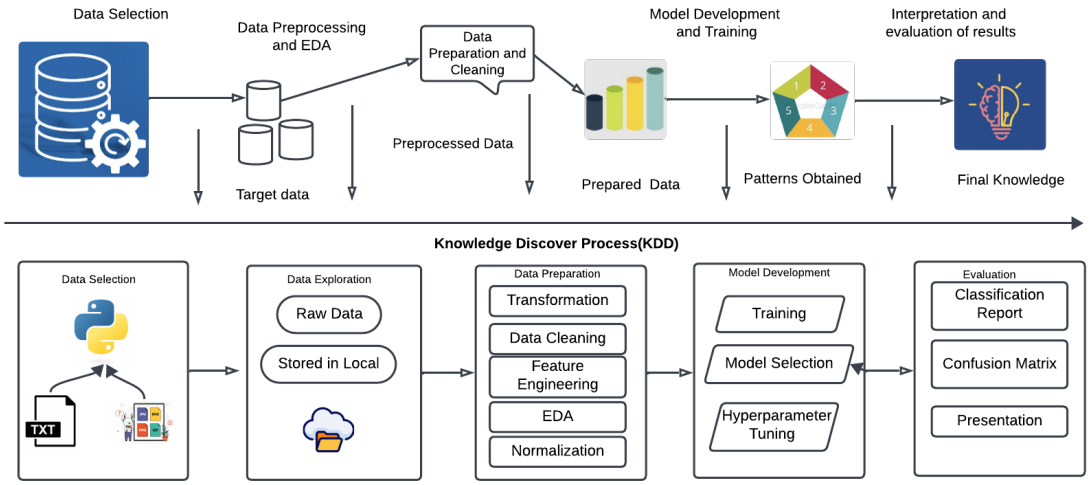


Figure 1: Project Process Flow Diagram

3.1 Data Selection

The initial state in Data mining research is collection of appropriate data by open-source or self-collection methods. Here in Fake news detection three domains have been considered Healthcare, Social media applications and Job posting websites. Considering the domain two distinct data types - text and image - is used for further analysis. The text data is sourced from an open-source website called Kaggle which contains fake and real

job postings with eighteen columns and huge records. Image data is sourced from GitHub repository which has collection of large images from twitter and other social media platforms related to healthcare, politics and some environmental related issues. Both the text and image data are consistently reviewed for further implementations. By using these datasets, the study helps to build a robust model in detection of fake news.

3.2 Data Exploration

After finding the suitable data for implementation the next step is Data exploration to understand the behavior of data and its features used for analysis. In this research two distinct types of data formats - image and text data - are used.

Text data (.csv file): It is taken from kaggle an open - source large community website used for publishing data science projects.The data contains 18 columns and 17880 records of fake and real job information.

Image Data (.jpg file): Image data which has fake and real news about healthcare, politics and other relevant environmental issues sourced from twitter is taken from an open-source GitHub repo. It has around 2000 plus images.

3.3 Data Preparation for Analysis

Once Data understanding is done the next crucial step is Data Preparation/Exploratory data Analysis where the data is inspected and modified in an efficient manner to obtain better outperforming results. In this phase we should ensure that the data is clean and structured for analysis. As both text and image data is used both undergoes distinct process when it comes to Preprocessing and other steps. For example, for text data tokenization,stopwords using wordcloud,stemming and lemmatization is performed. In image data image augmentation, resizing image to required pixels and other basic implementations are performed.

3.3.1 Data Cleaning

Figure 2(a) depicts the columns in the fake job posting dataset. There are 18 columns with unique names here fraudulent column acts as a target column. This column is used for prediction it has two Boolean values 0 - Not Fraud and 1 Fraud. Figure 2(b) are the images present in the image data. In Data cleaning the first step is to remove or replace the null values from the dataset as it may lead to inefficient models. Next is handling duplicate values in our text data i has no duplicates in it.Outliers removal should be considered but for this classification task we have only class imbalance problem. Removing Noise means eliminating special characters should be considered as the text data doesn't have any special characters only stop word removal using wordcloud is implemented. Additionally, Tokenization and Lemantization is also done for the text data.Next,For image data all the images uploaded are only in .jpg format. Image resolution is maintained accordingly, and irrelevant images are removed to reduce the noise of the data. Above data cleaning steps help to develop a robust model.

3.3.2 Handling missing values

As mentioned, the first step in Data cleaning/Data Preprocessing is to remove the null values or replace them with appropriate values. In our case we have null values in a few

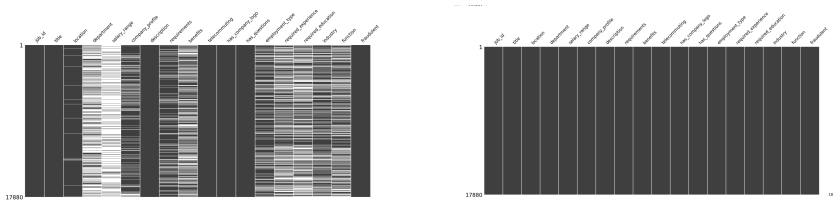
0	job_id
1	title
2	location
3	department
4	salary_range
5	company_profile
6	description
7	requirements
8	benefits
9	telecommuting
10	has_company_logo
11	has_questions
12	employment_type
13	required_experience
14	required_education
15	industry
16	function
17	fraudulent

(a) Attributes in Text Data

(b) Image Data's

Figure 2: Image and Text Data:(a)Columns that are included in image data;(b) Images that are used in implementation?

columns as shown in Figure 3(a). Due to huge missing values here, we have replaced those values accordingly to maintain the weightage of data. Figure 3(b) depicts the columns after replacement of null values. The missing values handling is only performed for text



(a) Before Handling Missing Values

(b) After Handling Missing Values

Figure 3: Handling Null/Missing Values:(a)Columns Before Handling Missing Values;(b) Columns after Handling Missing Values?

data and for image data we have no irrelevant or missing images to handle.

3.3.3 Handling StopWords

In text preprocessing we have performed stopwords("is", "the","and", "in") removal, tokenization and Lemmatization to normalize the data which will be efficient during implementation of ML and DL Algorithms. To handle stopwords worldcloud is used so irrelevant words are removed.NLTK (Natural Language Toolkit) is also used additionally to remove the stopwords to increase the efficiency of the data.

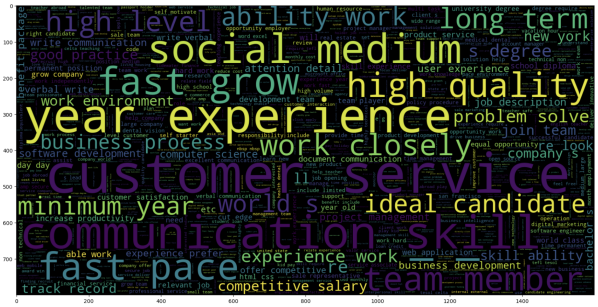


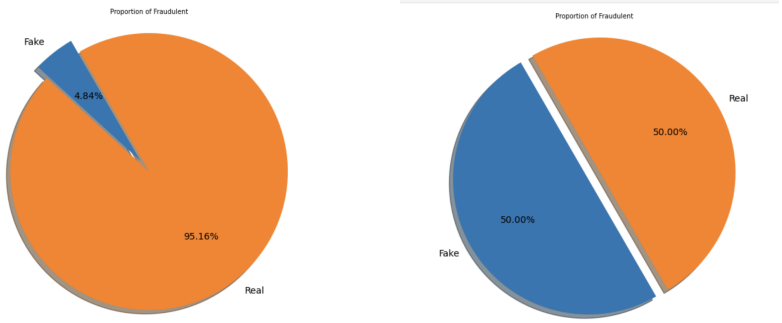
Figure 4: Stopwords

Lemmatization is also performed to improve text accuracy,Parts of Speech(POS). WordNetLemmatizer is implemented.All this combination of process make the data more efficient in case of predicting exact fake job postings.

For image Data VisualBert model is used to handle the captions in the images and it have shown the best performance.

3.3.4 Handling Class imbalance

The text data has class imbalance in target data which is very crucial for model development and prediction. Class imbalance may lead to biased results during evaluation. Here the class imbalance is handled by using RandomUnderSampler and the imbalance is handled efficiently. This removes the skewness of the data. For image data ,Data augmentation is performed to increase the efficiency of the images.



(a) Before OverSampling (b) After OverSampling

Figure 5: Handling Class imbalance:(a)Pie chart Before Handling OverSampling;(b) Piechary after Handling OverSampling ?

3.4 Model Development and Training

Model development and training is the next step implemented in the research project. When it comes to model development, selection of appropriate models is very important. In our case the aim is to do a detailed comparative analysis study on Machine Learning and Deep Learning algorithms to identify a robust model for fake news detection and this phase is called Model Selection Phase. The goal of the research question is to identify the best and most efficient deep learning and machine learning methods for detecting fake news on multimodal datasets from digital media platforms. As this is

a classification problem the classification-based algorithms have been sourced for model development and training.SVM,KNN,Decision Trees,Random Forest,Naive-Bayes,Visual-Bert,CNN,RNN,ANN,Bi-LSTM and LSTM across both are the algorithms implemented across image and Text Datasets for fake news detection. The above algorithms have been selected based on the reference of previous research works and their performance.

Figure 6 in detail shows the Model development, training and evaluation phase.

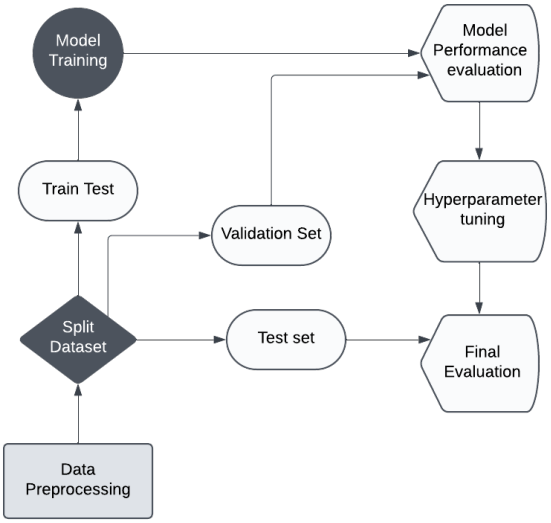


Figure 6: Model development Phase

3.5 Evaluation of Model

After model development and training it is important to measure the performance of the model which is called the evaluation phase. In this phase all the trained models will be measured using various evaluation metrics. In case of classification problem, the metrics used are Accuracy, F1-Score, Precision,Recall and Confusion matrix.Classification report can be also generated.

In the evaluation phase a detailed overview and comparison of all the metrics will be done in sequential manner.

4 Design Specification

In KDD Process Design specification is one of the initial and important steps to be considered. When it comes to design, it is all about the Machine Learning and Deep Learning algorithms that are to be implemented. Here the researcher should be clear on how each algorithm is designed and how it can be implemented more efficiently. As we have already selected the models to be implemented here, a detailed overview and description of each model and its evaluation metrics are explained. All models and metrics have been chosen based on the research question and its objective.

4.0.1 Modeling Techniques

Support Vector Machines : Support Vector Machines is a machine learning algorithm that can be used in both classification and regression implementations. It is well known

for its multi- and binary-related problems Kumar and Gopal (2010). It is more effective during the use of high-dimensional data. It is also more versatile where we can customize the usage of kernels. SVM is resistant to overfitting the data and model developed.

Decision Trees:The most popular supervised machine learning technique for problems involving classification. It creates a decision node for each characteristic, separates the dataset into smaller parts, and uses attribute selection measures to separate the data into the best attributes. All attributes have the same value, and no additional occurrences occur until one of the following conditions is satisfied.Krishna and Adimoolam (2022) This procedure is carried out repeatedly for every child: Compared to a neural network, it requires less time to train and can handle high dimension data with good accuracy.

Random Forest Classifier:To achieve more efficient and correct predictions, especially for multiclass classification problems with numerous classes or features, an ensemble machine learning algorithm known as a random forest classification is used that combines the predictions of several decision tree models. Every decision tree in a random forest is trained against multiple random parts from the subset, improving generalization and providing flexibility, feature importance, and less over fittingT S et al. (2022).

KNN Algorithm:Identifying a sample's k nearest neighbors and designating the class with the highest occurrence among them as the sample's class is the basic concept behind itSri Saranya and Juliet (2023). One advantage of KNN for multiclass classification is its ease of implementation and understanding. There are also a reasonable number of hyperparameters included, which makes tuning easy.

Naive-Bayes : This is generative based algorithm that can be used for only classification related problems. It accepts only class and category type of inputs. It is developed based on a concept called Bayes Theorem. It is also called a probabilistic classifier based on its prediction metrics. It is most used in Yes and No type of PredictionsSneha et al. (2023).

Convolutional Neural Networlk (CNN) : CNN is more widely used for image related data than text-based data. Because of their high accuracy, convolution neural networks are the preferred method for detecting objects in images. The brightness, position, color, and orientation of the object in the image have little bearing on CNN's ability to identify itGarg and Sharma (2022). Because CNN's run time is set, it is simpler to forecast when the work will be finished. Additionally, CNN's layer count has grown quickly in order to classify a vast number of objects. CNN is the greatest option for object detection in a picture because it was able to expand more quickly thanks to all the benefits.

Recurrent Neural Network(RNN) : This is a deep learning neural network technique that is used to handle sequential data formats. Compared to traditional neural networks it has more advancements in handling image and text related dataAbbas et al. (2022). Here memory state acts hidden in providing efficient results. It is an extension of LSTM Model and also has multiple RNN models incorporated. Mostly RNN is used to handle the time series data. Basic RNN performs like a feedback loop.

Artificial Neural Networks(ANN): Inspired neurons function in the human body, artificial neural networks (ANNs) are designed to solve complicated data issues such as pattern recognition, regression, and classification. ANNs employ a variety of topologies, including recurrent networks, feed-forward (shown in the figure), and back propagation. It functions as an ongoing learning process. In a feed-forward process, neurons learn the others by comprehending earlier neurons outputs. Input data is given weights for each

neuron, and non-linearity issues are resolved with the use of a transfer function. With the aid of activation functions, the hidden layer in this kind of feed-forward approach converts the input into a format that the output layer can understand.

LSTM and Bi-LSTM: Both are recurrent neural network based deep learning algorithms used in both classification and regression problemShikalgar and Arage (2023). The major difference is LSTM can only handle data in one direction whereas Bi-LSTM can handle the data in both forward and backward directions. Most of them use Bi-LSTM as it can handle both past and future data effectively. Here for text data, we have used Bi-Lstm and for Image data LSTM have used.

Visual-Bert : This model is used to handle multimodal data the data with both images and texts. BertTokenizer can be used to process the text by encoding.CNN like ResNet Jaiswal et al. (2021)along with Visual-Bert can be used to handle images.So this is known for handling the multimodal data more effectively.

4.1 Evaluation Techniques

Accuracy: The percentage of forecasts that were correct is displayed. Although the model's accuracy cannot be entirely relied upon.This can be calculated by dividing the total predictions also.

Confusion matrix: The table displays the number of predictions made for each class, allowing one to assess the model's performance for each class.

Performance Metric	Equation
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
F1-score	$\frac{2.(precision.recall)}{precision + recall}$

Figure 7: Evaluation Metrics of Classification Problem

Precision and recall: The percentage of predicted positive samples that were positive is known as precision, whereas the percentage of positive samples that were expected to be positive is known as recall. These can be used for multi-classification as well as binary classification, which is where they are commonly used.

F1 Score : The F1-Score, which is the harmonic mean of precision and recall, is frequently used as a single statistic to evaluate a classifier's performance.

5 Implementation

Implementation section includes or combines the steps that have been done during model development phase using Machine learning and deep learning algorithms on text and

image dataset. The model is checked against multiple scenarios to check its efficiency in detecting the fake news.

Developing a machine and deep learning model on multimodal data requires a number of keen procedures planned and executed steps. Carefully developing and implementing each step is essential to ensuring the model functions correctly, can be deployed, and can be used effectively in a real-world scenario.

5.1 Implementation Tools

Here the most used tool is Jupyter Notebook which comes along with Anaconda. It is a Python IDE used to run the code that has been developed. The text and Image data is stored in excel and drive for usage. The programming language used in the research is python and it has its own huge libraries. Basic to advanced Python libraries have been used for code implementation from data preprocessing, data visualisation to model development phase.

Tools Used	**Purpose**
Jupyter Notebook	It is an IDE used for running Python code
Excel Sheets, Google Drive	To store text and image data

Table 1: Tools and Their Purpose

5.2 Exploratory Data Analysis

EDA is the basic and most important step done to understand the features in dataset. In our case as we use both text and image data set EDA is designed separately text data class imbalance is identified and rectified using sapling technique during model development phase.

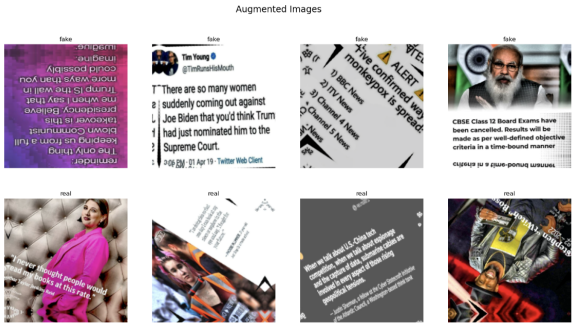


Figure 8: Augumented images

The stopwords are handled accordingly using Wordcloud to increase the accuracy of the data. Country-wise Job Posting, No. of Jobs with Experience and checking for most fake jobs based on title is done for text-based data. For image data augmentation Figure 9 is performed to increase the clarity of the images.

5.3 Data Preprocessing/cleaning

Since null values and missing values can result in an ineffective model, the first step in data cleaning is to exclude or replace them from the dataset. Dealing with duplicate values in our text data comes next. Figure 3(a)(b) shows the difference before and after handling the null values. It has no duplicates. Although eliminating outliers should be considered, the only issue we have with this classification challenge is class imbalance. Removing noise and removing special characters, which should be considered because there are no special characters in the text data.

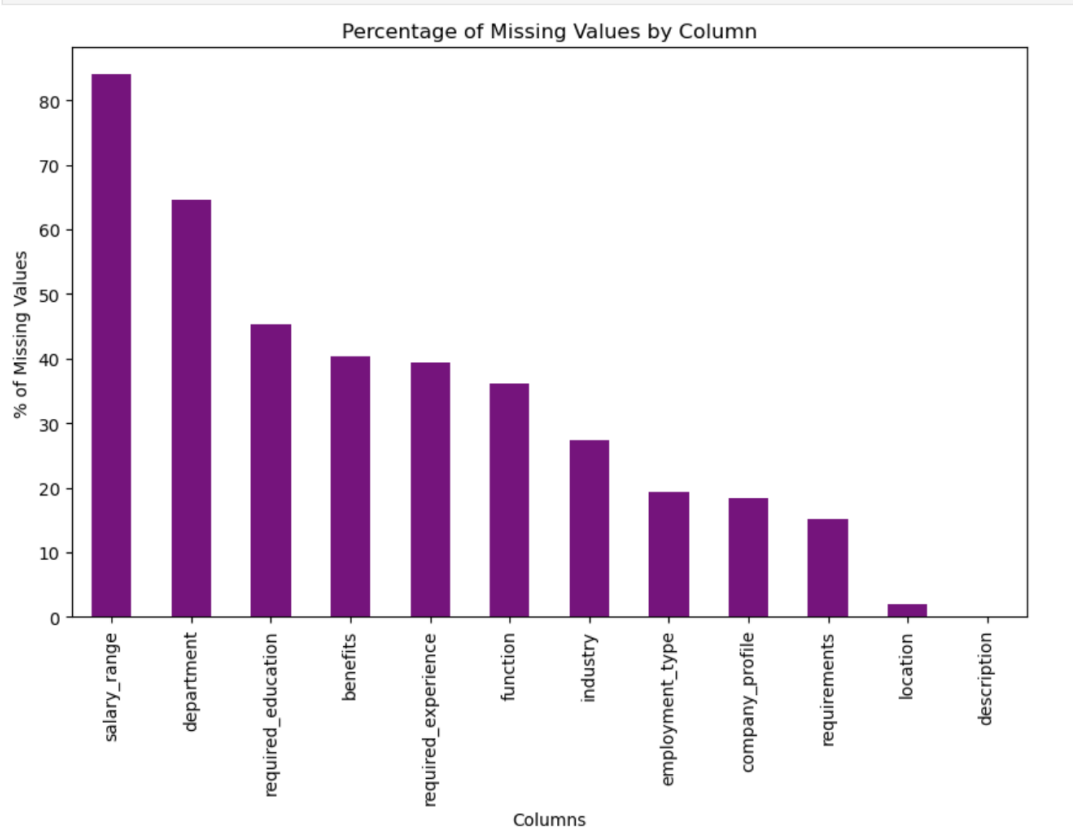


Figure 9: Percentage of missing values

Wordcloud is used to remove stopwords.Additionally, the text data undergoes tokenization and lemantization.Next, all uploaded photos are regarded as being in the.jpg format exclusively for image data. To lessen data noise, unnecessary photos are eliminated, and image resolution is appropriately maintained. The above-mentioned data cleaning procedures aid in the development.

5.4 Feature Engineering/Feature Selection

Feature engineering is done to make the raw data more effective in case of prediction for both text and image data. In our case it is handled carefully for using multimodal data. A crucial step in predictive analysis is feature selection, which lowers the number of input variables by eliminating unnecessary or redundant characteristics and then narrowing down the remaining features to just those that are necessary for the machine and deep learning algorithms. This is to see if there is a Constance feature in the data, which is

not crucial for resolving the issue statement because the variance threshold has been set to zero and every column has been examined; none was found. Here for text data, we have performed lemmatization and normalization in order to maintain the accuracy of the model. Text vectorization is performed using TF-IDF method. Word2Vec and GloVe are used to perform word embeddings. N-Gram Features is implemented to identify the patterns in fake and real news. Sentiment analysis is also considered to get the emotions from the text. As part of feature selection, the columns with less meanings were removed for model development phase. For Image data Feature engineering is done by performing Visual Feature Extraction using pretrained VGG16 model. Color and Texture Analysis is performed and adjusted accordingly. Custom Features for Domain-Specific Images have been checked, and irrelevant images are handled accordingly. Text and image pipelines are connected and implemented accordingly.

5.5 Hyper Parameter Tuning

Here SVM,KNN,Decision Trees,Random Forest,Naive-Bayes,Visual-Bert,CNN,RNN,ANN and LSTM are the deep learning and Machine Learning algorithms used to develop a robust model in detection of fake news using text and image dataset. Each algorithm has its own and effective hyper parameter tuned steps that can be implemented. A machine-learning and deep learning model's hyperparameter tuning is selecting exact and effective parameters. The model's performance and behavior are controlled by hyperparameters, which are parameters selected prior to training. In multiclass classification, hyperparameter tuning is crucial since the model's performance impacts the values of the hyper-parameters. The hyper-parameters listed below are employed in this project.

Algorithm	**Hyperparameters Tuned**
SVM	GridSearchCV, kernel, gamma, degree, class weight
KNN	GridSearchCV, n-estimators, weights, metric: euclidean, p
Decision Trees	n-estimators, max-depth, min-samples-split, bootstrap, max-features
Random Forest	n-estimators, max-depth, min-samples-split, bootstrap, max-features, min-samples-leaf
CNN/ANN/RNN and LSTM	epochs, batch-size, optimizer, learning-rate, dropout-rate, activation, units
Visual-BERT	learning-rate, batch-size, epochs, dropout, hidden-size, num-attention-heads, image-model, weight-decay, max-seq-length

Table 2: Algorithms and Their Tuned Hyperparameters

6 Evaluation

Evaluation is a critical step in the machine and deep learning pipeline that aids in determining the model's efficacy and confirming that it is functioning as intended. Carefully

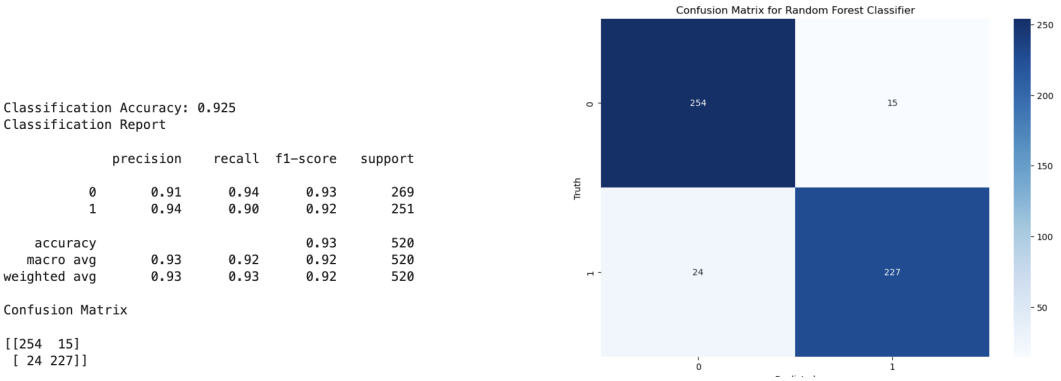
choosing the pertinent assessment metrics and assessing the model's performance are essential in order predicting the fake news. The below are mainly addressed.

Class imbalance issue: It is crucial to address the imbalanced classes while examining the category since the unequal data may significantly affect and distort the modeling results. As a result, machine learning algorithms may struggle to learn the minority class while favoring the dominant class. To deal with unequal data, this study employs a variety of strategies, including class weighting and oversampling.

Train and Test Split: As the last step before modeling, the dataset must be divided into training and test sets. The Sklearn library was used to divide the dataset into 80:20 test and train sets. For image data it has test, train and validation sets.

6.1 Case Study 1 : Fake Job Postings Prediction Using Machine Learning Algorithms

Case Study 1 is about Prediction of Fake job postings using text data taken from kaggle. KDD process is followed to obtain the results mentioned in the section. Here Machine learning algorithms like Decision tree, Random Forest, Svm,KNN, GBM.All the models have shown great performance but Random Forest, Naive-Bayes and SVM have shown high accuracy of 0.93 and 0.91 respectively and KNN algorithm has shown the least accuracy of 0.58.The model have shown better performance due to hyperparameter tuning performed by trial and error methods.By adjusting the parameters used in grid search CV tuning—which comprises building and evaluating a model for every set of hyperparameters specified in a grid—this can be prevented. The class imbalance handling also plays a great role in achieving such huge results.RandomUnderSampling and SMOTE is used to balance fake and real job class data.



Classification Accuracy: 0.925
Classification Report

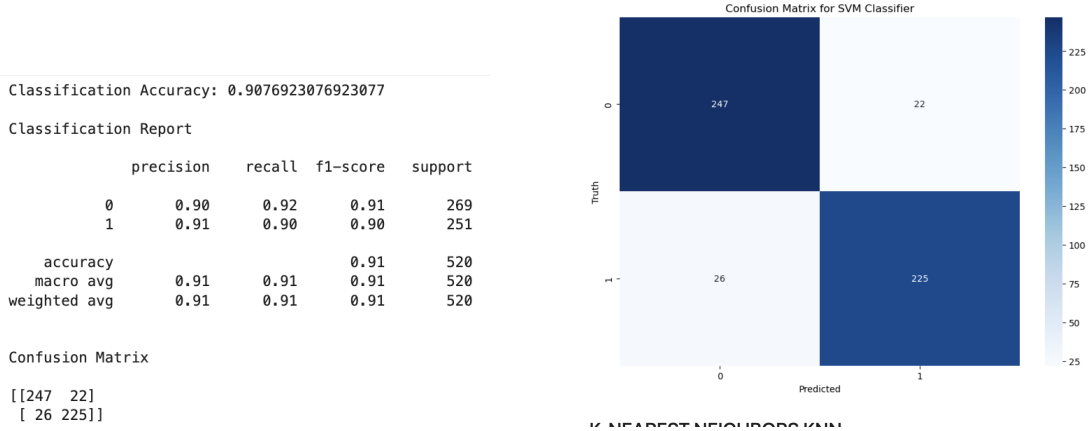
	precision	recall	f1-score	support
0	0.91	0.94	0.93	269
1	0.94	0.90	0.92	251
accuracy			0.93	520
macro avg	0.93	0.92	0.92	520
weighted avg	0.93	0.93	0.92	520

Confusion Matrix

[[254 15]
[24 227]]

(a) Classification Report (b) Confusion Matrix

Figure 10: Random Forest classifier:(a)Classification Report with performance metrics;(b)Confusion Matrix with true positives and false negatives



(a) Classification Report (b) Confusion Matrix

Figure 11: SVM classifier:(a)Classification Report with performance metrics;(b)Confusion Matrix with true positives and false negatives

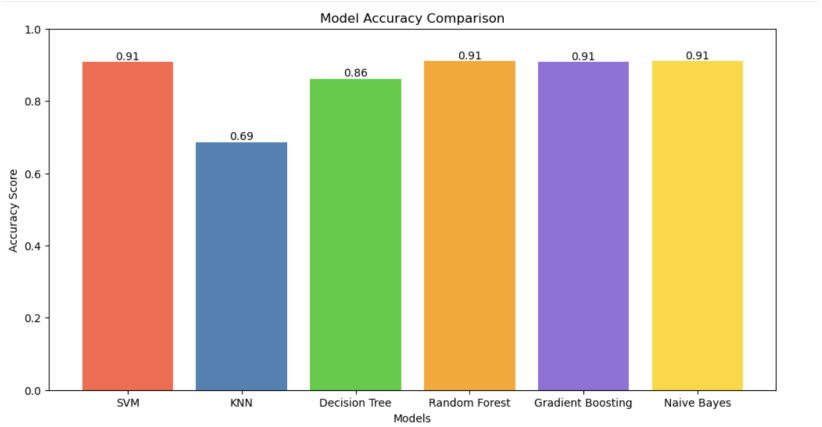


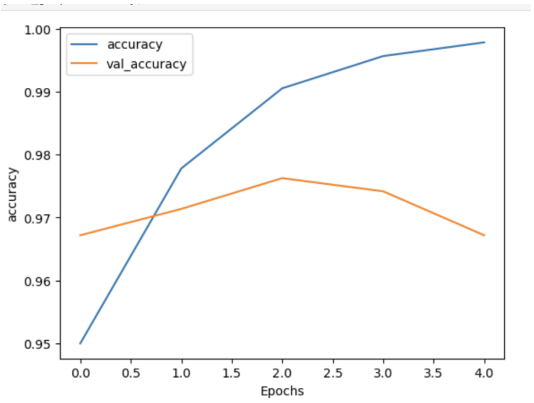
Figure 12: Comparison of Machine Learning Models

Figure.12. Depicts the overall comparison of machine learning models implemented with its accuracy and it's clear that SVM,Random Forest,Naive-Bayes and Gradient Boosting Machines Perform better and KNN has less performance comparatively that can be improved by fine tuning the model in future works. Overall, the above models help in accurate prediction of fake jobs and reduces the stress of job seekers.

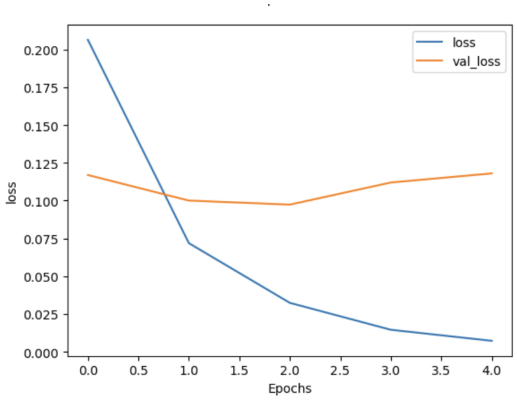
6.2 Case Study 2 : Fake Job Postings Prediction Using Deep Learning Algorithms

Case Study 2 is about Prediction of Fake job postings using text data taken from Kaggle. Knowledge Discovery Database (KDD) process is followed to obtain the results mentioned in the section. Here Deep learning algorithms like Bi-LSTM, BERT and CNN are implemented. All the models have shown great performance where CNN and BERT have achieved 0.99 percentage of accuracy in predicting the fake news and Bi-LSTM have also achieved around 0.98 percentage of predictions. Trial-and-error methods for hyperparameter adjustment have improved the model's performance. Grid search CV tuning, which includes constructing and assessing a model for each set of hyperparameters given

in a grid, can avoid this by modifying the parameters utilized. A major factor in reaching such remarkable outcomes is the management of class imbalance. Real and phony job class data are balanced using SMOTE and RandomUnderSampling.It is evident that deep learning algorithms perform better than machine learning algorithms in prediction of fake jobs from job posting website data.



(a) Accuracy Graph



(b) Loss Graph

Figure 13: Bi-Directional LSTM :(a)Train and Validation Accuracy(b)Train and Validation Loss

	precision	recall	f1-score	support
0	0.99	0.98	0.98	3414
1	0.66	0.69	0.67	162
accuracy			0.97	3576
macro avg	0.82	0.83	0.83	3576
weighted avg	0.97	0.97	0.97	3576

(a) Classification Report of Bi-LSTM

	precision	recall	f1-score	support
Not Fraudulent	0.99	1.00	0.99	3395
Fraudulent	0.89	0.75	0.82	181
accuracy			0.98	3576
macro avg	0.94	0.87	0.90	3576
weighted avg	0.98	0.98	0.98	3576

(b) Classification Report of BERT

Figure 14: Bi-LSTM and BERT:(a)Classification Report of Bi-LSTM with performance metrics;(b)Classification Report of BERT with performance metrics

Figure 13 and 14 show the performance of Bi-LSTM and BERT algorithm snippets. CNN have also shown the same performance as BERT algorithm. From Case Study 1 Fake Job Postings Prediction Using Machine Learning Algorithms and Case study 2 Fake Job Postings Prediction Using Deep Learning Algorithms, it is evident that both Machine Learning and Deep Learning algorithms perform better. But from overall Analysis Deep learning Algorithms like CNN, BERT and Bi-LSTM have shown great results in detection of fake job postings in text data.

6.3 Case Study 3 : Fake News Prediction in healthcare and social media Using Machine Learning Algorithms

Case Study 3 is about Fake News Prediction in healthcare and social media Using Machine Learning Algorithms on Image Data obtained from an open source github repository.The machine learning algorithm used for prediction are Support Vector Machines(SVM),K-Nearest neighbors,Decision Trees and Random Forest.From this five machine learning

algorithms random forest have only shown a better performance around 0.91 percentage where all other model have shown less performance in prediction of images.This shows Machine learning algorithms must be trained and pretrained more effectively to deal with large image datasets.KNN has very less performance here also due to its hypertuned parameters which be handled consistently.

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.93	0.90	0.91	41
1	0.90	0.93	0.91	40
accuracy			0.91	81
macro avg	0.91	0.91	0.91	81
weighted avg	0.91	0.91	0.91	81

(a) Classification Report of Random Forest

KNN Classification Report:

	precision	recall	f1-score	support
0	0.89	0.20	0.32	41
1	0.54	0.97	0.70	40
accuracy			0.58	81
macro avg	0.72	0.59	0.51	81
weighted avg	0.72	0.58	0.51	81

(b) Classification Report of KNN

Figure 15: Random Forest and KNN:(a)Classification Report of Random Forest with performance metrics and it has high performance;(b)Classification Report of KNN with performance metrics and it shows less performance

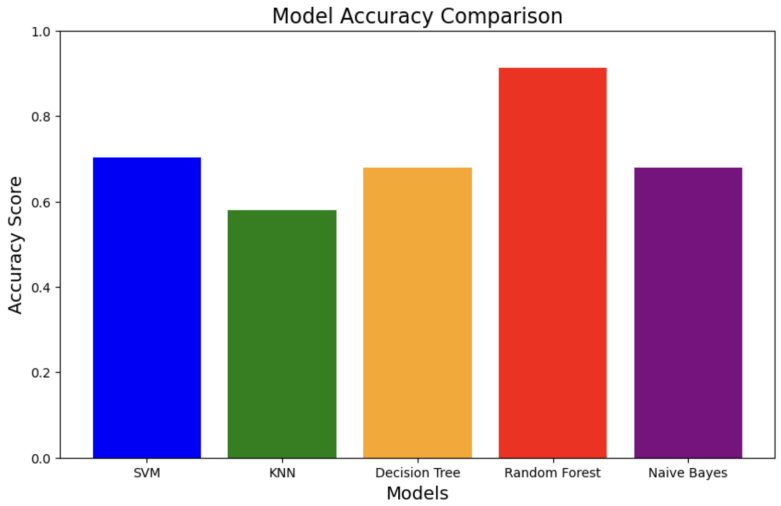


Figure 16: Comparison of Machine Learning Models on Image Data

Figure.12. Depicts the overall comparison of machine learning models implemented with its accuracy and it's clear that, Random Forest Perform better and KNN has less performance comparatively that can be improved by fine tuning the model in future works. Overall, the above models only Random Forest shows accurate prediction of fake posts where all other models Support Vector Machines (SVM), K-Nearest neighbors, Decision Trees and Naive Bayes show very less Performance.

6.4 Case Study 4 : Fake News Prediction in healthcare and social media Using Deep Learning Algorithms

Case Study 4 is about Fake News Prediction in healthcare and social media Using Deep Learning Algorithms on Image Data obtained from an open source github repository. The deep learning algorithm used for prediction Convolutional Neural Network(CNN),Recurrent Neural Network(RNN),Artificial Neural Network(ANN),Long short

term memory(LSTM) and Visual BERT. From this five deep learning algorithms Visual BERT and CNN have achieved a greater performance around 0.99 and 0.86 respectively have only shown a better performance around 0.91 percentage where all other model have shown less performance in prediction of images.RNN has shown very less performance due to its nature is only on handling sequential and time series data.This Case study shows that CNN and Visual BERT models can be used to handle the image or multimodal data more effectively.CNN is trained along with VGG16 to obtain such greater results.

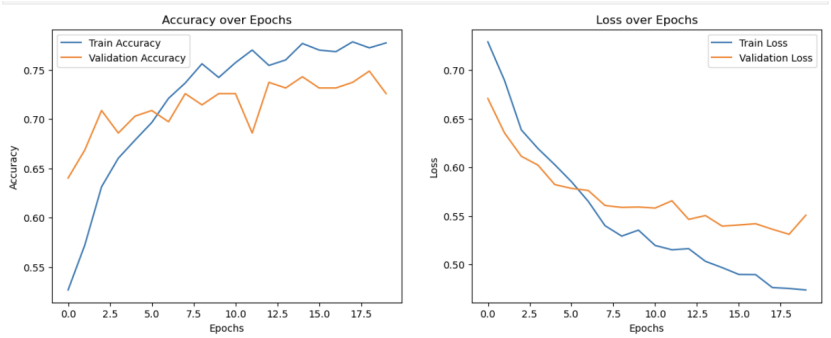


Figure 17: CNN Accuracy and Loss Over epochs

(a) Predicted Images (b) Predicted Fake images

Figure 18: Fake and Real Images Prediction:(a) Predicted Fake and Real Images;(b)Predicted Fake images using Deep Learning Techniques

6.5 Comparison Table Of Machine Learning and Deep Learning Algorithms

Comparison Table Of Machine Learning and Deep Learning Algorithms is shown in detailed in Figure 18 for all the case studies performed and this table shows that better and less performed model on both machine and deep learning model using image and text data.This models can be used by job seekers and social media users to better predict and prevent the illegal activities across Job postings,healthcare and social media websites .

FAKE NEWS PREDICTION ON TEXT DATA USING MACHINE LEARNING			
Algorithm Name	**Precision**	**Recall**	**F1-Score**
Random Forest	0.91	0.94	0.93
Decision Tree	0.87	0.88	0.87
Naive Bayes	0.92	0.90	0.91
SVM	0.90	0.92	0.91
KNN			
FAKE NEWS PREDICTION ON TEXT DATA USING DEEP LEARNING			
Bi-LSTM	0.99	0.98	0.98
BERT	0.99	1.00	0.99
CNN	0.98	1.00	0.99
FAKE NEWS PREDICTION ON IMAGE DATA USING MACHINE LEARNING			
Random Forest	0.93	0.90	0.91
SVM	0.77	0.59	0.67
Decision Trees	0.74	0.56	0.64
KNN	0.89	0.20	0.32
Naive Byes	0.76	0.54	0.63
FAKE NEWS PREDICTION ON IMAGE DATA USING DEEP LEARNING			
CNN	Accuracy - 0.86		
RNN	Accuracy - 0.58		
LSTM	Accuracy - 0.74		
ANN	Accuracy - 0.64		
VISUAL BERT	Accuracy - 1.00		

Figure 19: Comparison table for evaluation metrics

6.6 Discussion

The study uses picture and text data to predict fake news and job listings on social media, health care, and job posting websites using Machine Learning and deep learning Algorithms.In this studymore than ten algorithms have been trained and tested against text and image data and a detailed comparative case study is provided.Visual-BERT model performance is measured to be more than all other algorithm metrics.The four case study that have been discussed are Fake Job Postings Prediction Using Machine Learning Algorithms,Fake Job Postings Prediction Using Deep Learning Algorithms,Fake News Prediction in healthcare and social media Using Machine Learning Algorithms and Fake News Prediction in healthcare and Social media Using Deep Learning Algorithms In each study all the aglorithms implemented are discussed with appropriate outputs.Class imbalance,Hyperparameter Tuning and Modelling are explained with its performance.

From all the four-case study Machine Learning Algorithms Perform Better on text-based data and Deep Learning Algorithms Perform Better on Image Based Data.This

models can be used by job seekers and social media users to better predict and prevent the illegal activities across Job postings,healthcare and social media websites. All things considered, our comparative analysis study on machine learning and deep learning

techniques can be utilized to forecast false information, lessen public anxiety, and boost public trust in internet usage. The research has also addressed significant gaps in tuning the data to make the model performance more efficient.

7 Conclusion and Future Work

The research conducted to answer the research questions What are the finest and most efficient deep learning and machine learning methods for detecting fake news on multimodal datasets from various digital media platforms? and How can these methods compare different assessment metrics such as accuracy, precision, recall, and F1 score and produce the best results? have been implemented successfully and Deep learning algorithms perform better on image data and Machine Learning algorithms perform better on Text based dataset. The data collected from Kaggle and GitHub is utilized for the entire research. Traditional algorithms like KNN have shown much less performance throughout the entire research perspective where Visual-BERT,Random Forest and CNN have shown greater results throughout the comparative study. One of the main takeaways from this study is how well preprocessing techniques like feature engineering, SMOTE, and augmentation work with hyperparameter adjustment to enhance model performance.

Over all the advantages the drawback here is usage of prepresent data for model development and evaluation. In future machine learning and deep learning models should be trained and tested on real-time data from social media websites. An application where all the users can just include the news and check its truth fullness should be developed and deployed. More mechanisms and case studies will be considered to increase the trust and safety among end-users on social media, health care and job posting websites.

8 Acknowledgement

I owe a debt of gratitude to Professor Athanasios Staikopoulos for his invaluable knowledge and advice during my thesis. I also like to be grateful and thank to the National College of Ireland for giving me the chance to do this research, since his insights were very helpful in determining the path of the work. I owe my family and friends a huge debt of gratitude for their support and encouragement during my master's program. I would not have been able to complete this research assignment without their help.

References

Abbas, Q., Zeshan, M. U. and Asif, M. (2022). A cnn-rnn based fake news detection model using deep learning, *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, pp. 40–45.

Abdullah, N. A. S., Rusli, N. I. A. and Yuslee, N. S. (2024). Development of a machine learning algorithm for fake news detection, *Indonesian Journal of Electrical Engineering and Computer Science* **35**(3): 1732–1743.

Agrawal, C., Pandey, A. and Goyal, S. (2024). Multimodal fake news detection using hyperparameter-tuned bert and resnet110, *International Journal of Advanced Technology and Engineering Exploration* **11**(114): 759–772.

Al-Tarawneh, M. A. B., Al-irr, O., Al-Maaitah, K. S., Kanj, H. and Aly, W. H. F. (2024). Enhancing fake news detection with word embedding: A machine learning and deep learning approach, *Computers* **13**(9).

Babu, T., Nair, R. R., Challa, A., Srikanth, R., Aravindan, S. S. and S, S. (2023). Fake news detection using machine learning algorithms, *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, Vol. 1, pp. 1–7.

Choudhary, M., Jha, S., Prashant, Saxena, D. and Singh, A. K. (2021). A review of fake news detection methods using machine learning, *2021 2nd International Conference for Emerging Technology (INCET)*, pp. 1–5.

Dsouza, K. M. and French, A. M. (2024). Fake news detection using machine learning: an adversarial collaboration approach, *Internet Research* **34**(5): 1664–1678.

Dwivedi, S. M. and Wankhade, S. B. (2024). Deep learning based semantic model for multimodal fake news detection, *International Journal of Intelligent Engineering and Systems* **17**(1): 654 – 665.

Festus Ayetiran, E. and Özgöbek, (2024). A review of deep learning techniques for multimodal fake news and harmful languages detection, *IEEE Access* **12**: 76133–76153.

Garg, S. and Sharma, D. K. (2022). Fake news classification via cnn, *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, pp. 61–63.

Giachanou, A., Zhang, G. and Rosso, P. (2020). Multimodal multi-image fake news detection, *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 647–654.

Jaiswal, R., Singh, U. P. and Singh, K. P. (2021). Fake news detection using bert-vgg19 multimodal variational autoencoder, *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1–5.

Jouhar, J., Pratap, A., Tijo, N. and Mony, M. (2024). Fake news detection using python and machine learning, Vol. 233, p. 763 – 771.

Krishna, N. L. S. R. and Adimoolam, M. (2022). Fake news detection system using decision tree algorithm and compare textual property with support vector machine algorithm, *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, pp. 1–6.

Kumar, M. A. and Gopal, M. (2010). Binary classification using linear svm pyramidal tree, *2010 International Conference on Data Storage and Data Engineering*, pp. 54–58.

LAHLOU, Y., FKIHI, S. E. and FAIZI, R. (2024). Twitter truth: Advanced multi-model embedding for fake news detection, *International Journal of Advanced Computer Science and Applications (ijacsa)* **15**(8).

Mahara, T., Josephine, V. L. H., Srinivasan, R., Prakash, P., Algarni, A. D. and Verma, O. P. (2023). Deep vs. shallow: A comparative study of machine learning and deep learning approaches for fake health news detection, *IEEE Access* **11**: 79330–79340.

Men, X. and Mariano, V. Y. (2024). Explainable fake news detection based on bert and shap applied to covid-19, *International Journal of Modern Education and Computer Science (IJMECS)* **16**(1): 11–22.

Shikalgar, M. B. and Arage, C. S. (2023). Fake news detection using hybrid bilstm-tcn model with attention mechanism, *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 1130–1136.

Sneha, V., Sravani, V., Sathvika, V., Kumar, B. V. and Jagan, A. (2023). Fake news detection using naïve bayes and support vector machine, *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, pp. 1–6.

Sri Saranya, K. S. and Juliet, A. H. (2023). Comparison of random forest with k-nearest neighbors to detect fake news with improved accuracy, *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Vol. 6, pp. 2543–2547.

T S, S. M., Sreeja, P. S. and Ram, R. P. (2022). Fake news article classification using random forest, passive aggressive, and gradient boosting, *2022 International Conference on Connected Systems Intelligence (CSI)*, pp. 1–6.

Vu, D.-H., Nguyen, K., Tran, K. T., Vo, B. and Le, T. (2024). Improving fake job description detection using deep learning-based nlp techniques, *Journal of Information and Telecommunication* .

Wani, M. A., ELAffendi, M., Shakil, K. A., Abuhaimed, I. M., Nayyar, A., Hussain, A. and El-Latif, A. A. A. (2024). Toxic fake news detection and classification for combating covid-19 misinformation, *IEEE Transactions on Computational Social Systems* **11**(4): 5101–5118.