Fake News Detection

Rawan Sayed: **one hot encoding**

Mina Hany: **Bag of Words Features**

Mark Magdy: **Co-occurrence Matrix**

Mustafa Adel: **Term frequency-inverse document frequency**

Supervisor: Dr. Wesam Ahmed

Group Number: 3

Submission date: May 12th, 2025

Chapter 1

Introduction

In this project, we aim to classify news articles as either fake or real by building machine learning and deep learning models. To achieve accurate and reliable results, we followed a systematic approach that includes data preprocessing, feature extraction, and modeling.

1. Data Preprocessing: -

Before feeding the data into any model, we cleaned and normalized the text to improve the quality of the input.

2. Feature Extraction

After preprocessing, we transformed the text data into numerical features suitable for machine learning algorithms.

3. Modeling

We experimented with a range of models to classify the news articles:

4 Machine Learning models and 2 Deep Learning models.

Chapter 2

Code Implementation

Preprocessing:

1. All text was converted to lowercase.
2. Remove punctuation and numbers.
3. Remove stop words using NLTK library.
4. Each word was lemmatized using NLTK library (Lemmatization).
5. The text was split into individual words using NLTK library (tokens).

Visualization:

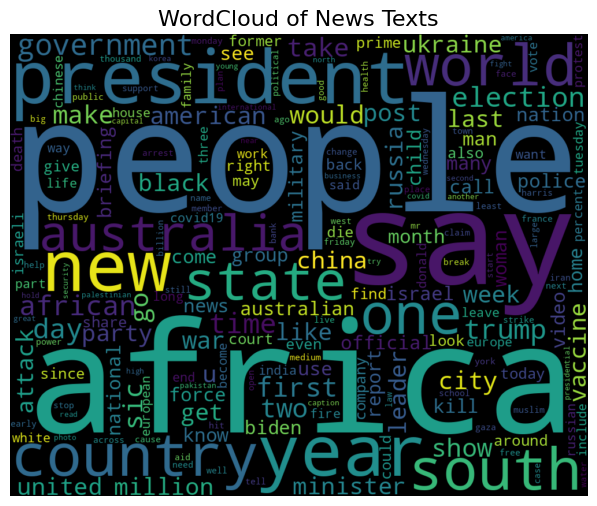


Figure 2.1: Word cloud

Text Transformation:

We used three transformation techniques: -

1. Tokenization (splitting text into tokens)

2. Converting text to lowercase

3. Lemmatization

• **One hot encoding:**

1.Logistic Regression:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 99.89% | 93.21% |  | 95.99% | 92.53% | 94.23% | 98.2% |

|  |  |
| --- | --- |
|  | A graph of a curve  AI-generated content may be incorrect. |
| Figure 1.1: Logistic Regression with one hot encode confusion matrix | Figure 1.2: Logistic Regression with one hot encode Roc curve |

2. random forest (RF) Algorithm:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| RF | 100.0% | 91.85% |  | 96.54% | 90.06% | 93.19% | 98.2% |

|  |  |
| --- | --- |
|  | A graph of a curve  AI-generated content may be incorrect. |
| Figure 1.3: random forest (RF) Algorithm with one hot encode confusion matrix | Figure 1.4: random forest (RF) Algorithm with one hot Roc curve |

3. naive Bayes algorithm:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| NB | 97.12% | 93.45% |  | 92.87% | 94.65% | 94.24% | 93.56% |

|  |  |
| --- | --- |
|  |  |
| Figure 1.5: naive Bayes with one hot encode confusion matrix | Figure 1.6: NB with BOW Roc curve |

4. SVM algorithm:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| SVM | 100% | 92.17% |  | 94.95% | 91.77% | 94.33% | 91.67% |

|  |  |
| --- | --- |
|  |  |
| Figure 1.7: SVM with one hot encode confusion matrix | Figure 1.8: SVM with BOW Roc curve |

1. Dense Neural Network:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Dense | 99.32% | 93.17% |  | 95.71% | 92.69% | 94.18% | 98.02% |

|  |  |
| --- | --- |
|  | A graph of a curve  AI-generated content may be incorrect. |
| Figure 1.9: Dense Neural Network with one hot encoding confusion matrix | Figure 1.10: Dense Neural Network Roc curve |

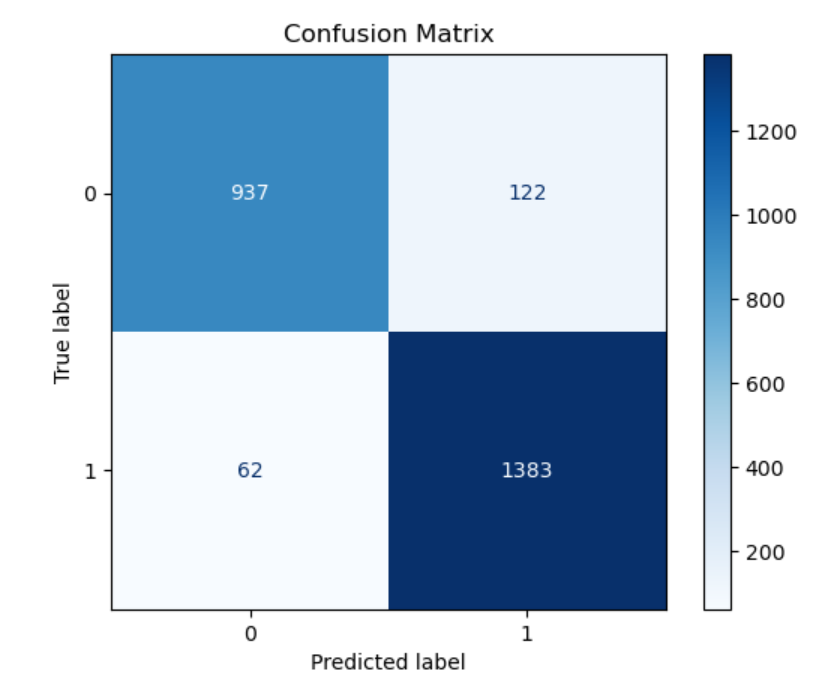
2. CNN:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| CNN | 99.43% | 93.81% |  | 94.81% | 94.48% | 94.65% | 98.58% |

|  |  |
| --- | --- |
|  | A graph of a curve  AI-generated content may be incorrect. |
| Figure 1.11: CNN with one hot encode confusion matrix | Figure 1.12: CNN Roc curve |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 99.89% | 93.21% |  | 95.99% | 92.53% | 94.23% | 98.2% |
| RF | 100.0% | 91.85% |  | 96.54% | 90.06% | 93.19% | 98.2% |
| NB | 97.12% | 93.45% |  | 92.87% | 94.65% | 94.24% | 93.56% |
| SVM | 100% | 92.17% |  | 94.95% | 91.77% | 94.33% | 91.67% |
| Dense | 99.32% | 93.17% |  | 95.71% | 92.69% | 94.18% | 98.02% |
| CNN | 99.43% | 93.81% |  | 94.81% | 94.48% | 94.65% | 98.58% |

* Bag of Words

1.Logistic Regression:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 99.89% | 92.65% |  | 95.71% | 92% | 93.76% | 98.1% |

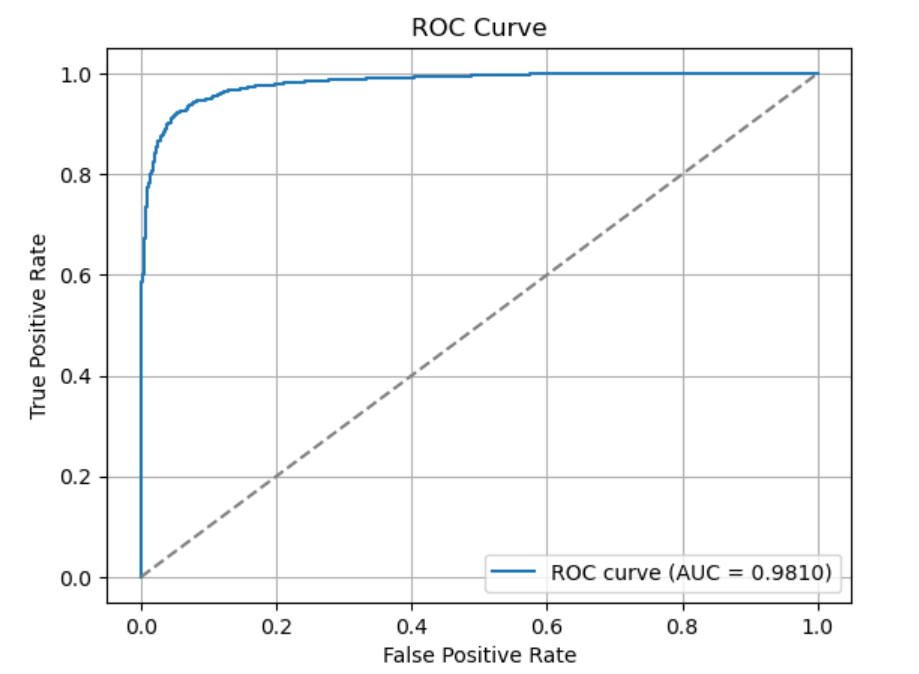


Fig 2.2 : confusion matrix for logistic regression

Fig 2.1 :ROC curve for logistic regression

2. Naive Bayes Algorithm:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Naïve Bayes | 96.71% | 92.65% |  | 91.9% | 95.00% | 93.53% | 97.68% |

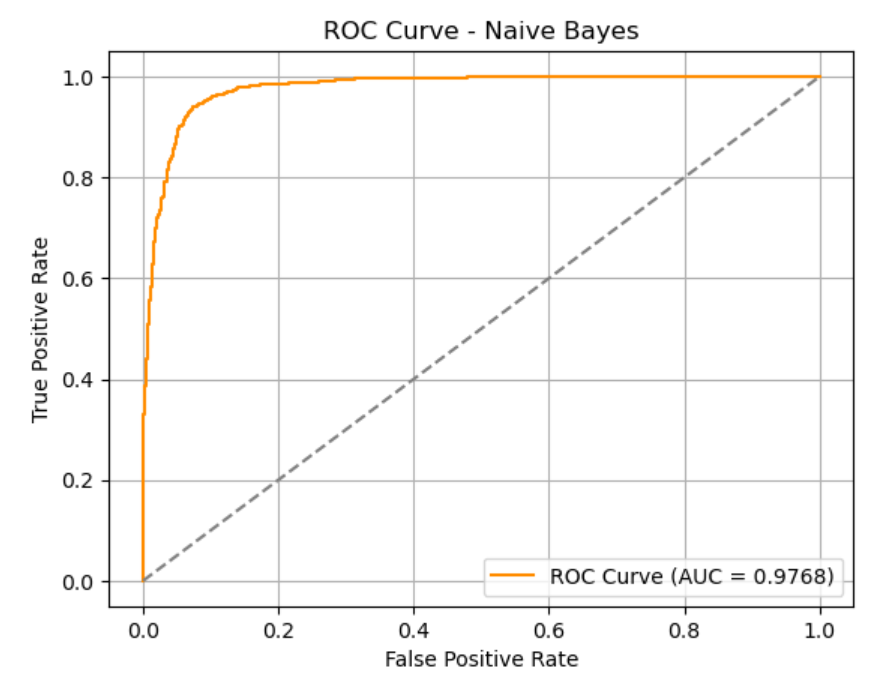
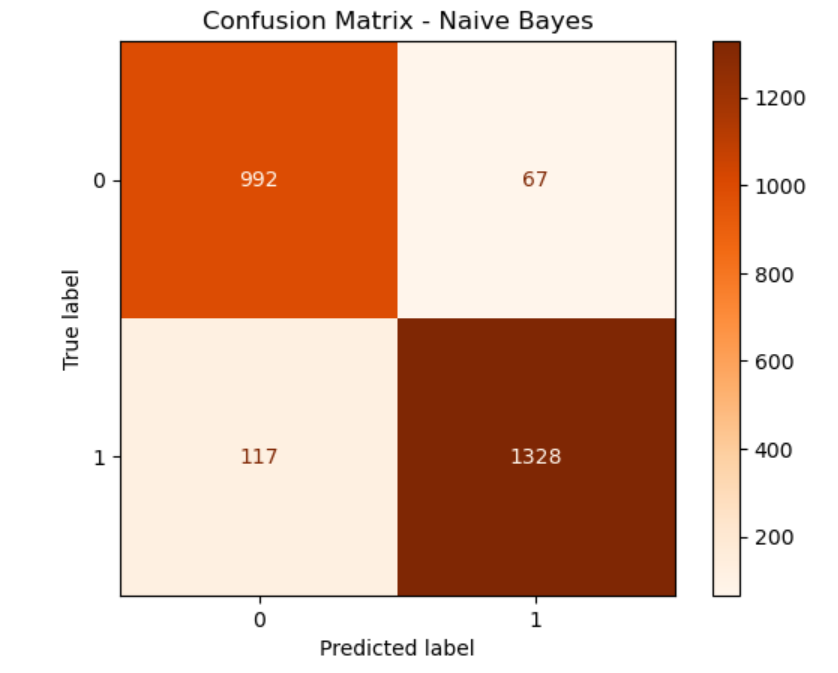
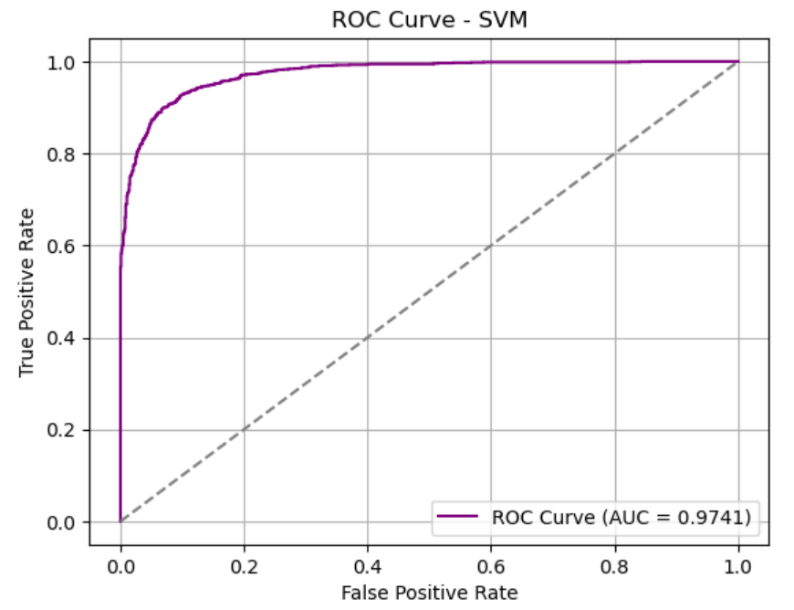


Fig 2.3 : Naïve Bayes ROC curve

Fig 2.4 : Naïve Bayes Confusion matrix

3. support vector machine:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| SVM | 99.1% | 91.61% |  | 92.6% | 93.00% | 92.72% | 97.41% |



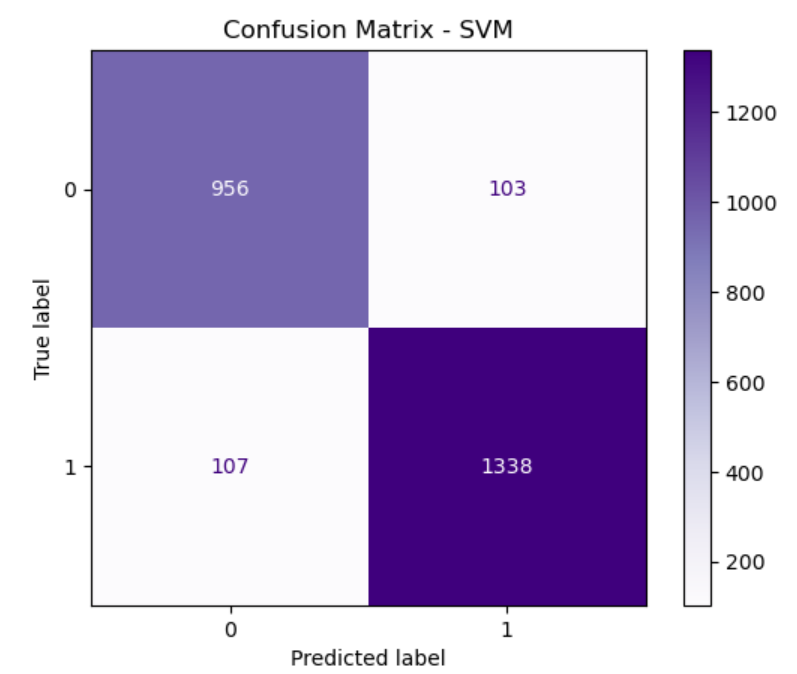


Fig 2.6 : SVM ROC curve

Fig 2.5 : SVM ROC curve

4. gradient boosting algorithim:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| SVM | 85.57% | 84.46% |  | 95.71% | 87.00% | 87.67% | 92.95% |

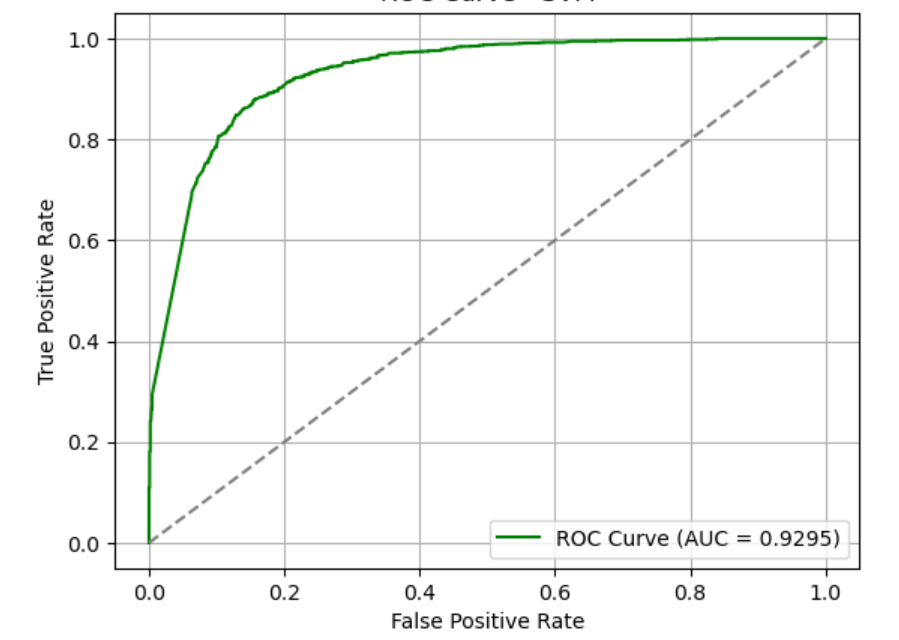
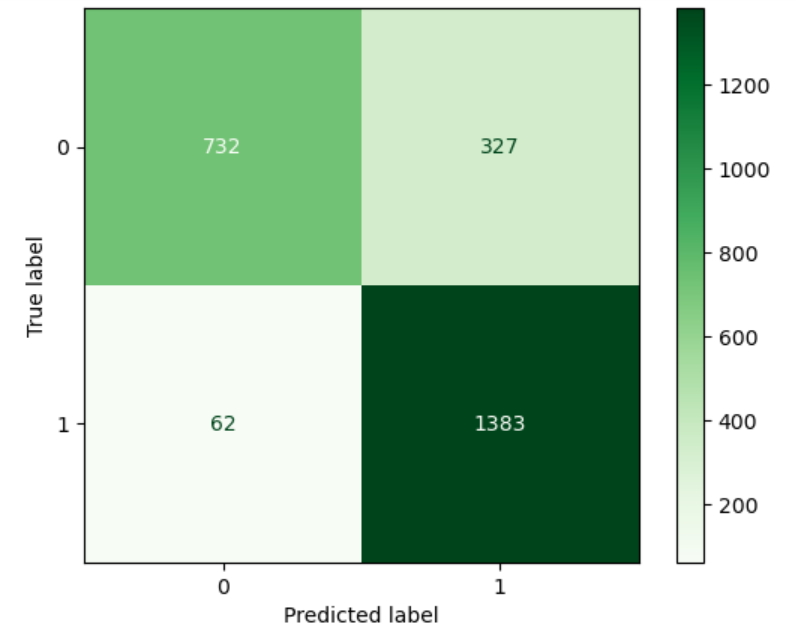


Fig 2.8 : gradient boosting confusion matrix

Fig 2.7 : Gradient Boosting ROC curve

5. feed forward neural network:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Feed forward NN | 100% | 93.73% |  | 95.85% | 94.00% | 94.64% | 97.89% |

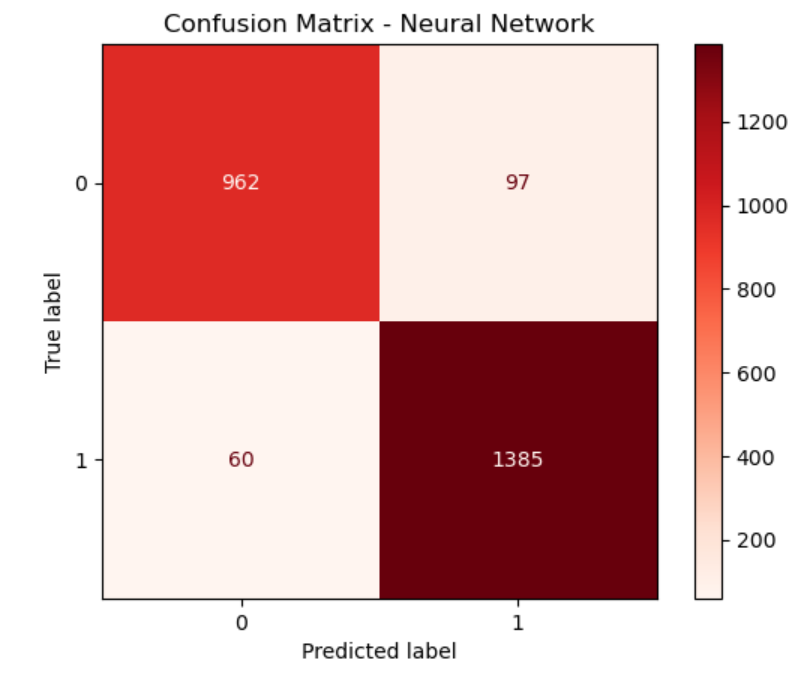
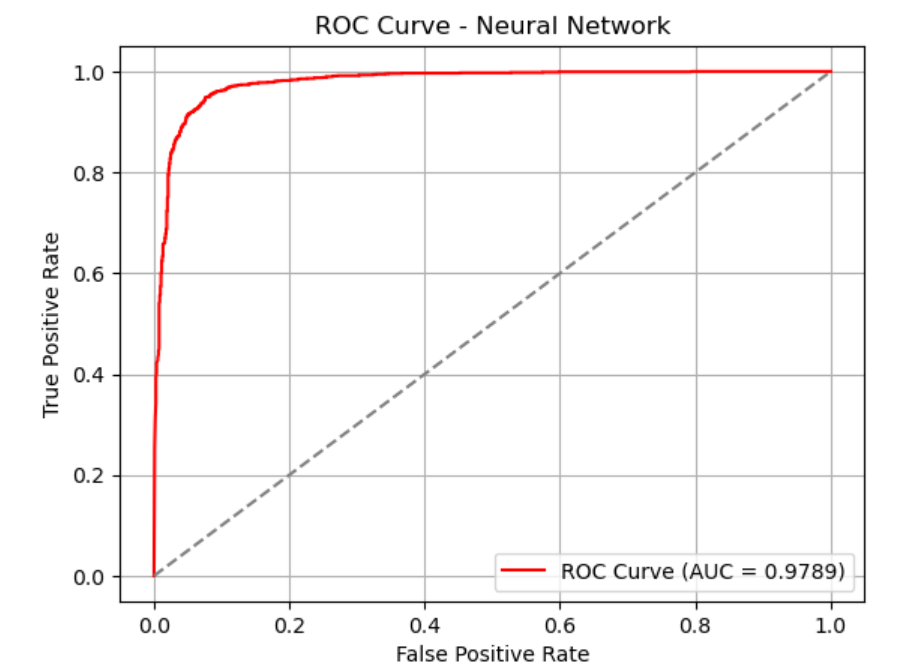


Fig 2.10 : feed forward neural network confusion matrix

Fig 2.9 : feed forward neural network ROC curve

6. CNN:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Feed forward NN | 99.20% | 92.17% |  | 92.94% | 92.00% | 93.20% | 97.68% |

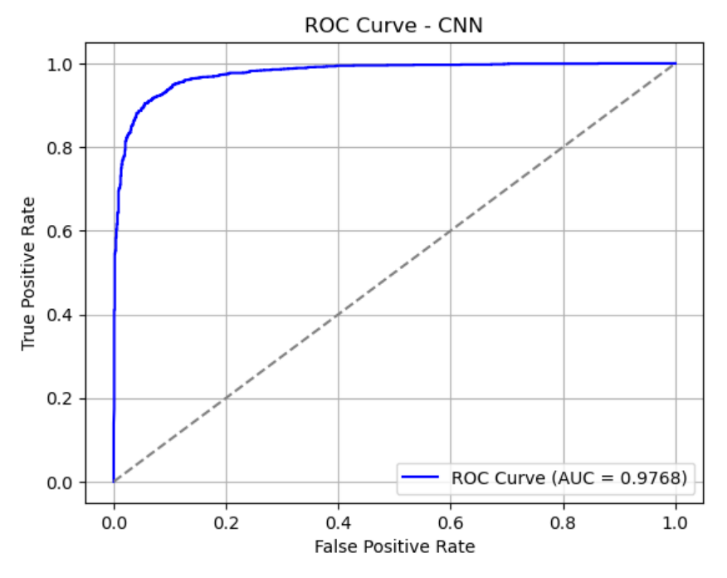
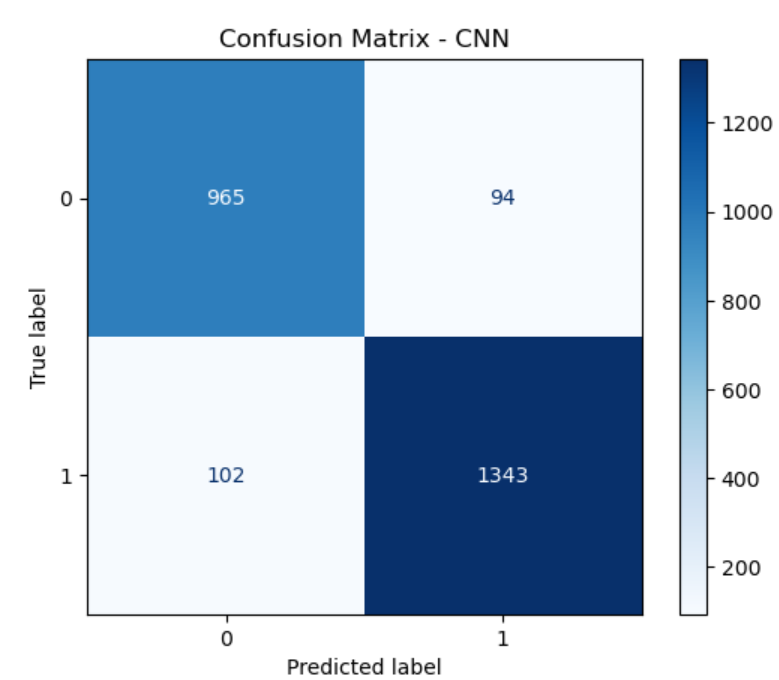


Fig 2.12: convolutional neural network ROC curve

Fig 2.11 : convolutional neural network ROC curve

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 99.89% | 92.65% |  | 95.71% | 92% | 93.76% | 98.1% |
| Naïve bayes | 96.71% | 92.65% |  | 91.9% | 95.00% | 93.53% | 97.68% |
| SVM | 99.1% | 91.61% |  | 92.6% | 93.00% | 92.72% | 97.41% |
| GB | 85.57% | 84.46% |  | 95.71% | 87.00% | 87.67% | 92.95% |
| Feed forward NN | 100% | 93.73% |  | 95.85% | 94.00% | 94.64% | 97.89% |
| CNN | 99.20% | 92.17% |  | 92.94% | 92.00% | 93.20% | 97.68% |

• **Co-occurrence Matrices:**

1.Logistic Regression:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 93.26% | 92.97% |  | 95.50% | 92.56% | 94.01% | 97.98% |

A graph of a curve

AI-generated content may be incorrect.A blue squares with white text

AI-generated content may be incorrect.

2.  Random Forest Model:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Random Forest | 93.02% | 89.02% |  | 94.81% | 87.26% | 90.88% | 95.49% |

A graph with a line and a green line

AI-generated content may be incorrect.A green squares with white numbers

AI-generated content may be incorrect.

3.  Support Vector Machine (SVM):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| SVM | 97.82% | 92.25% |  | 95.02% | 91.84% | 93.40% | 95.49% |

A graph with a line and a green line

AI-generated content may be incorrect.A green squares with white numbers

AI-generated content may be incorrect.

4.  Naive Bayes Model:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Naive Bayes | 94.00% | 91.45% |  | 91.70% | 93.38% | 92.53% | 97.18% |

A graph with a green line

AI-generated content may be incorrect.A diagram of a confusion matrix

AI-generated content may be incorrect.

1.  Dense Neural Network:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| DNN | 95.85% | 92.41% |  | 95.22% | 91.92% | 93.54% | 97.63% |

A graph with a green line

AI-generated content may be incorrect.A diagram of confusion matrix

AI-generated content may be incorrect.

2.  LSTM Model:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| LSTM | 98.45% | 92.97% |  | 95.22% | 92.78% | 93.99% | 97.85% |

A graph with a green line

AI-generated content may be incorrect.A green squares with white numbers

AI-generated content may be incorrect.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic | 93.26% | 92.97% |  | 95.50% | 92.56% | 94.01% | 97.98% |
| Random Forest | 93.02% | 89.02% |  | 94.81% | 87.26% | 90.88% | 95.49% |
| SVM | 97.82% | 92.25% |  | 95.02% | 91.84% | 93.40% | 95.49% |
| Naive Bayes | 94.00% | 91.45% |  | 91.70% | 93.38% | 92.53% | 97.18% |
| DNN | 95.85% | 92.41% |  | 95.22% | 91.92% | 93.54% | 97.63% |
| LSTM | 98.45% | 92.97% |  | 95.22% | 92.78% | 93.99% | 97.85% |

* **Term frequency-inverse document frequency**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Recall  Accuracy | Precision Accuracy | F1-measure Accuracy |
| Logistic Model | 95.89% | 92.37% | 95.94% | 90.96% | 93.38% |
| Nave Bayes Model | 94.35% | 92.49% | 93.95% | 92.76% | 93.35% |
| Support Vector Machine(SVM) |  | 93.65% | 95.94% | 92.97% | 94.43% |
| Random Forest | 100% | 90.96% | 91.6% | 91.8% | 91.7% |
| LSTM | 97.76% | 91.77% | 96.65% | 89.52% | 92.95% |
| RNN | 96.39% | 90.97% | 96.23% | 88.66% | 92.29% |

A graph with a line

AI-generated content may be incorrect.

Fig 3.1: Logistic Regression Roc curve

A graph with a line

AI-generated content may be incorrect.

Fig 3.2: Nave Bayse Roc curve

A graph with a line and a dotted line

AI-generated content may be incorrect.

Fig 3.3: SVM Roc curve

A graph with a line

AI-generated content may be incorrect.

Fig 3.4: Logistic Regression Roc curve

A graph with a line

AI-generated content may be incorrect.

Fig 3.5: LSTM Roc curve

A graph of a curve

AI-generated content may be incorrect.

Fig 3.6: RNN Roc curve