REPORT FOR INTELLIGENT DOCUMENT CLASSIFIER

1. Preprocessing Steps:

Initialization:

- The WordNetLemmatizer is initialized, which is used for reducing words to their base or root form (i.e., lemmatization).
- stopwords.words('english') is used to get a list of common English stop words (words like "the", "is", "in", etc.) that are typically removed because they don't contribute meaningful information.

Tokenization and Lowercasing:

- word_tokenize from the NLTK library is used to split the input text into individual words (tokens).
- The text is converted to lowercase to ensure uniformity (so "Dog" and "dog" are treated the same).

Stop Word Removal and Lemmatization:

- The tokens are filtered to remove stop words and to only keep alphanumeric tokens (so punctuation and non-letter characters are discarded).
- After filtering, each token is lemmatized using lemmatizer.lemmatize(token).

Rejoining Tokens:

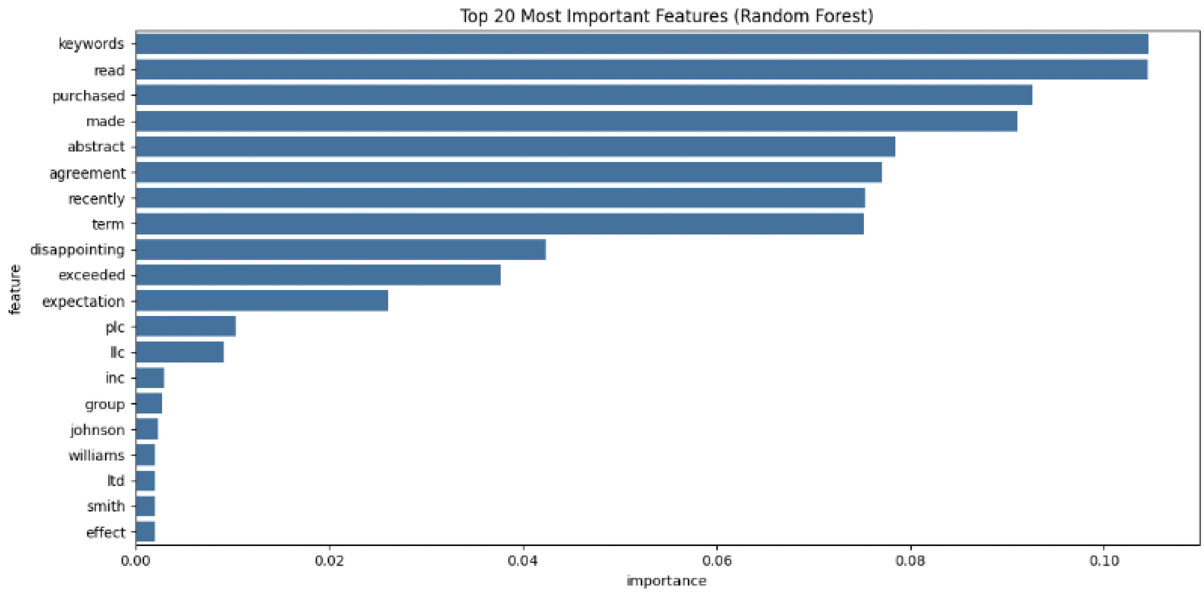
- Finally, the processed tokens are joined back into a single string, separated by spaces.

2.Architecture and Methodolgies

Summary of Architectures and Methodologies

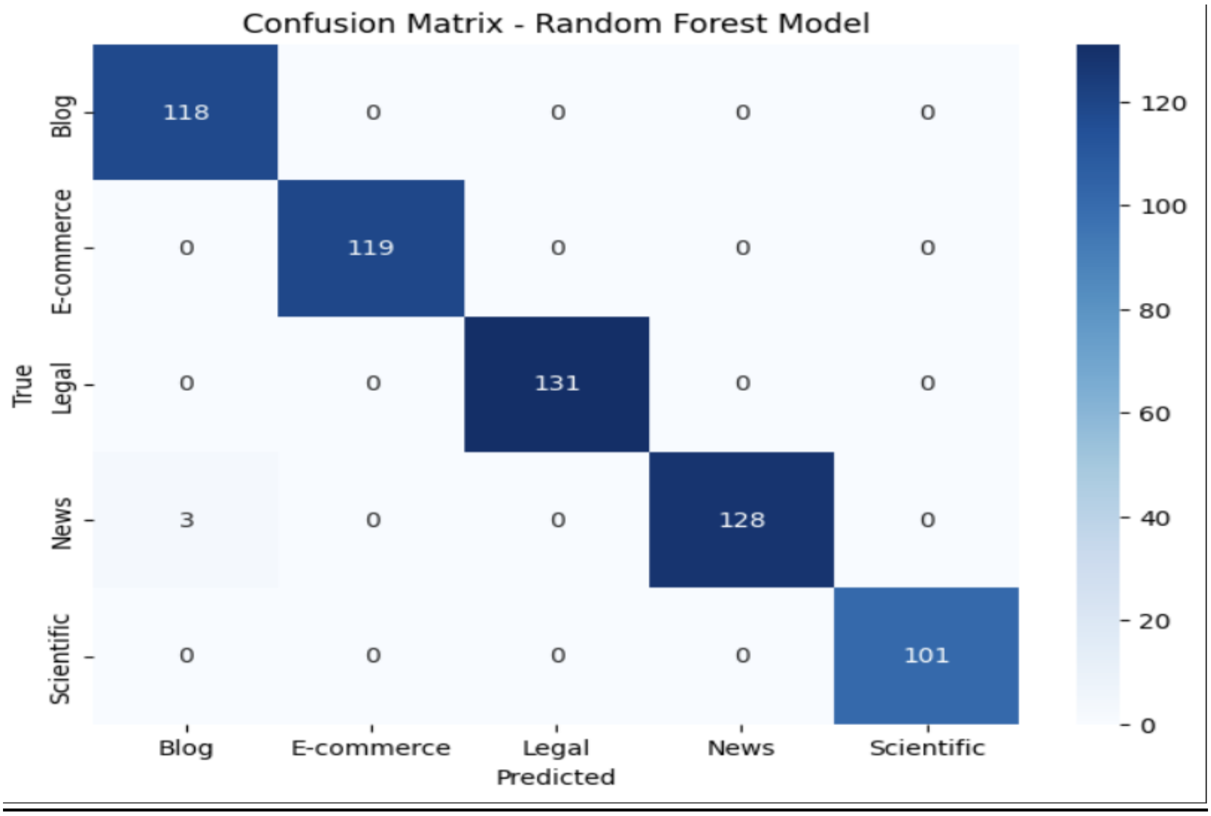
Aspect	Traditional Model (Random Forest)	Deep Learning Model (BiLSTM)
Type of Model	Ensemble Learning (Random Forest)	Recurrent Neural Network (BiLSTM)
Preprocessing	TF-IDF Vectorization, Tokenization, Lemmatization	Tokenization, Lemmatization, Padding, Embedding
Model Components	Decision Trees (Random Forest)	Embedding Layer, Bidirectional LSTM, Dense Layers
Data Representation	TF-IDF Vectors	Word Embeddings (e.g., GloVe), Sequence of Indices
Hyperparameter Tuning	RandomizedSearchCV for Random Forest	Keras Tuner for BiLSTM model parameters
Training	Fast, less computationally expensive	Computationally expensive, slow to train
Suitability	Simple, fast tasks with sparse features	Complex tasks involving sequences (e.g., text)
Evaluation	Classification report, Confusion matrix	Classification report, Confusion matrix

3.Evaluation Results for Random forest Model:

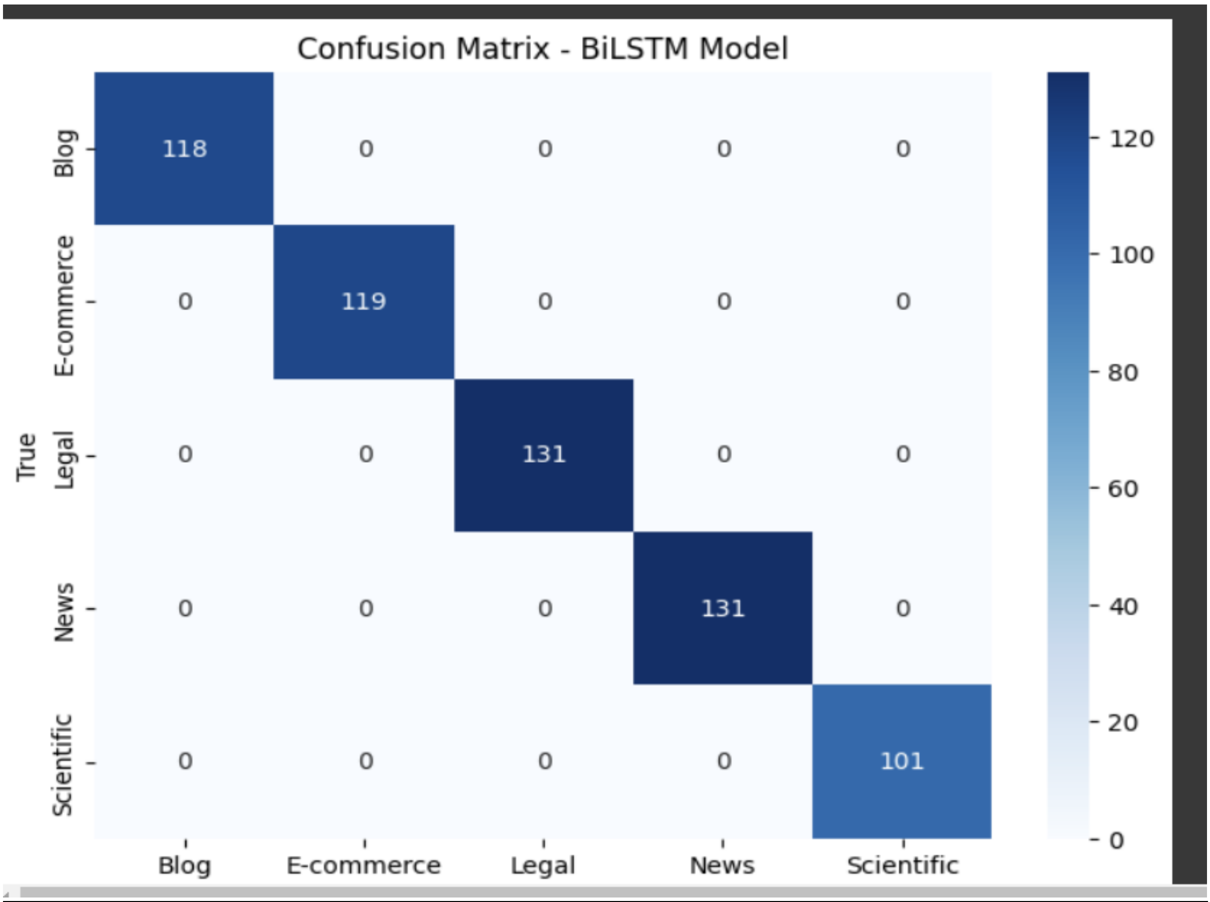


Random Forest Model Results:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	118
1	1.00	1.00	1.00	119
2	1.00	1.00	1.00	131
3	1.00	0.98	0.99	131
4	1.00	1.00	1.00	101
accuracy			0.99	600
macro avg	1.00	1.00	1.00	600
weighted avg	1.00	0.99	1.00	600


Confusion Matrix - Random Forest Model
True
Blog
E-commerce
Legal
News
Scientific
118
0
0
0
0
0
119
0
0
0
0
0
131
0
0
3
0
0
128
0
0
0
0
0
101
Blog
E-commerce
Legal
News
Scientific
Predicted
120
100
80
60
40
20
0

Evaluation results of Deep Learning Model:




```
BiLSTM Model Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       118
           1       1.00      1.00      1.00       119
           2       1.00      1.00      1.00       131
           3       1.00      1.00      1.00       131
           4       1.00      1.00      1.00       101

    accuracy                           1.00       600
   macro avg       1.00      1.00      1.00       600
weighted avg       1.00      1.00      1.00       600
```

Comparisons:

While both the models have high accuracy and precision,the traditional model showed some misclassification while the deep learning model has classified the text accurately therefore showing that the deep learning model is the prefered model but it requires time to train the model.

Challenges Faced:

1)One of the main challenges faced was incorporating and training both the models under the same class for perfect comparison and the hyperparameter tuning of both the models.At first the Randomized search in the traditional model was failing and hence required proper parameters to work without failure.

2)Classification in both the models had less accuracy and after proper tuning of both the models these errors were corrected.