Document Classification

Presented By Akshit Jain

Document Classification

Input – Two columns Text & Label

Text Column – Has all the unformatted, uncleaned text.

Label Column – 0, 1, 2, 3, 4 – these are 5 labels for 5 classes.

Preprocessing – Raw → Cleaned answer string

Output – Classifies the text into class

Total rows – 2225

Total columns - 2

Data Preprocessing

Components

Lowercase – lowered the characters and removed line breaks/tabs

Digits – Used Regex to remove digits

Special characters – Removed punctuations and special characters using Regex

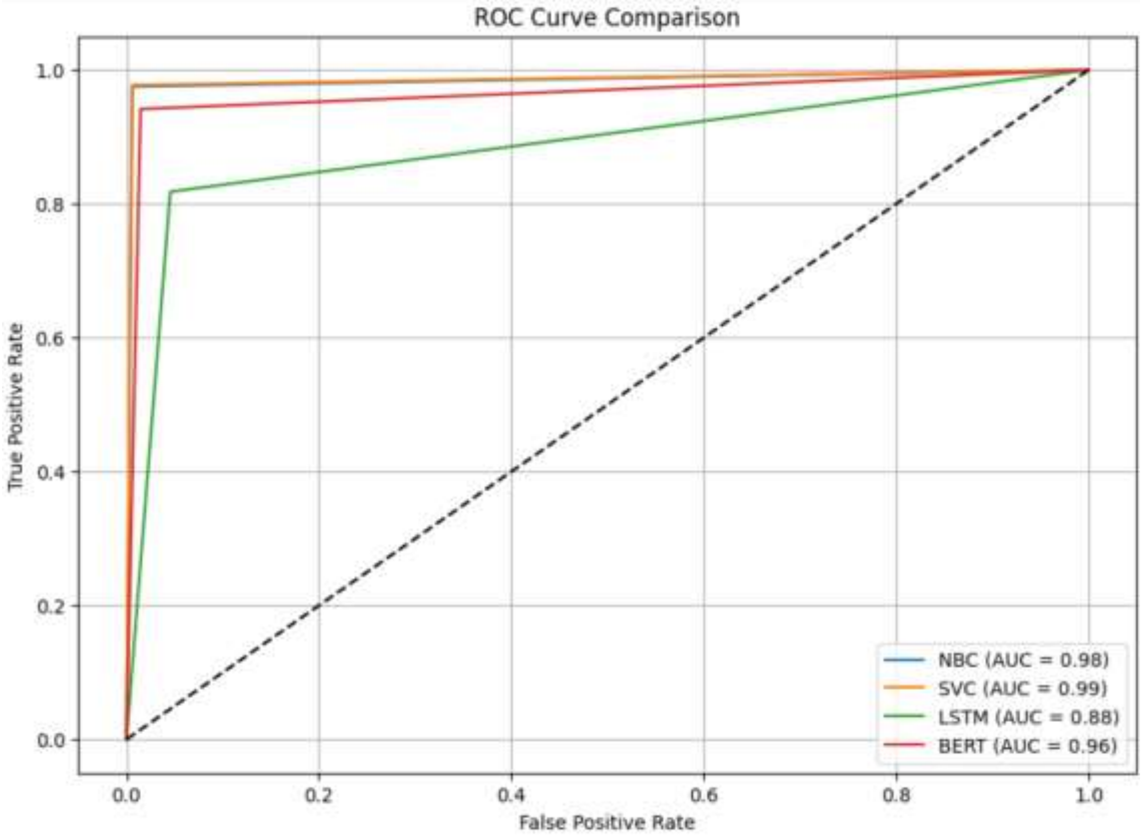
Tokenize – Tokenized words

Stop words – Removed stop words

Stemming – Applied stemming technique

Lemmatization – Applied lemmatization technique

Model Validation Accuracy



Configurations Results

Model	Accuracy	F1 Score
NBC	0.9596	0.9595
Linear SVC	0.9798	0.9797
LSTM	0.9371	0.9371
BERT	0.9415	0.9416

Evaluation

- **Training Dataset Size**: 0.8.
- **Testing Dataset Size**: 0.2.
- **Output Class**: [0,1,2,3,4]
- **Output Class Count**: 0 - 417, 1 - 511, 2 - 401, 3 - 386, 4 - 510

Conclusion

- NBC and Linear SVC performs best among all models.
- **Actionable Insights**: Useful for classifying documents.
- **Next Steps**: Fine-tuning and adding more epochs.

Thank You