Digital Forensic Analysis Using Deep Learning For Online Communication

Anirudh. H Manasa. N Preethi Thiyagarajan

Department of CSE, Sri Sivasubramaniya Nadar college of Engineering
Guide : Ms. S. Lakshmi Priya, Final Year Project, June 2022

Highlights of Proposed Model

To develop an intelligent computational model that

- Preprocesses the collected Enron dataset.
- Applies topic modelling techniques like LDA and BERT to the preprocessed dataset.
- Predicts criminal and non-criminals using LDA and Logistic Regression .
- Predicts criminal and non-criminals using *BertForSequenceClassification*.
- Compares the performance of LDA with BERT.

Challenges in Enron Dataset:

- Criminal and Non-Criminal records' irregular contribution the the dataset.
- Criminal emails not appearing to be fraud emails.
- Factors affecting the interpretability of the topics obtained from topic modelling.

Dataset Description

ATTRIBUTE	DESCRIPTION
Date	The date on which said mail was sent.
From	The sender's email address is stored here.
To	Contains the email address of the receiver.
Subject	Outlines the content of the mail body.
X-From	Contains name of the sender.
X-To	Contains name of the receiver.
Content	The body of the e-mail exchanged between the employees.
User	User of the system from which the mail originated.
Labelled	Binary value depicting criminal or non-criminal.

Table 1. Description of the dataset used.

Functional Modules and Dataset

- Data Collection
- Data Pre-processing
- Topic Modelling
  - Latent Dirichlet Allocation(LDA)
  - Bidirectional Encoder Representations from Transformers(BERT)
- Feature Selection
- Model Training

Enron Dataset:

- The corpus contains an actual 619,446 email messages that belong to 158 Enron employees and associates, including senior Enron employees and associates.
- The dates of the emails are between 1998 to 2002.

Methodology

- Email conversations are collected from the Enron dataset.
- The dataset is cleaned , preprocessed and labelled the records with 0 and 1 where 0 represents a non-criminal and 1 represents a criminal
- The preprocessed dataset is fed to the topic modelling algorithms LDA and BERT.
- The topic obtained from LDA is vectorized and fed to the Machine Learning models.
- *BertForSequenceClassification* is used to train the model and classify them as criminal and non criminal. The model is optimised using AdamW.
- The final outcome of BertForSequenceClassification is compared with LDA+Logistic Regression , CNN and CNN-LSTM.

Machine learning models used:

- Logistic Regression
- Logistic Regression SGD

Deep learning models used:

- *BertForSequenceClassification*
- CNN
- CNN-LSTM

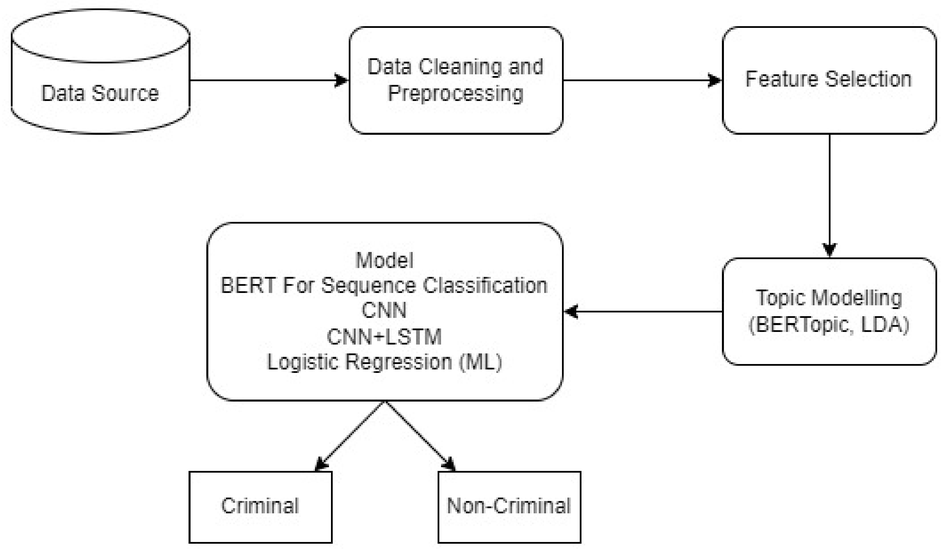

Figure 1. Architectural diagram of the proposed model

Performance Analysis

MODELS USED	ACCURACY
BertForSequenceClassification	0.97
CNN	0.925
CNN+LSTM	0.91
LDA+Logistic Regression (For comparison)	0.45
LDA+Logistic RegressionSGD (For comparison)	0.45

Table 2. The accuracy of different models

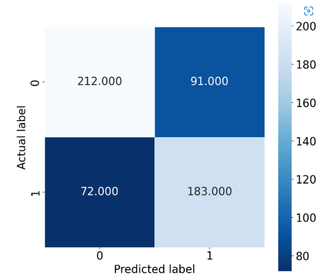

Figure 2. Heat map for actual and predicted class labels of LDA

Epoch	Training loss	Validation loss	Validation accuracy
1	0.21	0.10	0.97
2	0.11	0.08	0.97

Table 3. The decrease in loss for consecutive epochs of *BertForSequenceClassification*

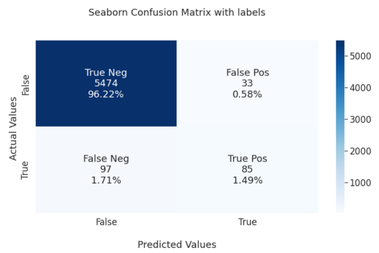

Figure 3. Heat map for *BertForSequenceClassification*

References

- Dongming, S. et al."NLP-based digital forensic investigation platform for online communication", In: Computers Security(2021),Vol 104, pp.24-30
- GitHub-Pre-processing Enron data, Website URL : https://github.com/Sun121sun/ENRON-EMAILS-ANDEMPLOYE