FinSentix: Contextual Enrichment in Financial Script through Sentiment and Term Extraction

Financial status play an important role in any country development and this status will be count via banking business or stock prices. All banking or stock related data available online in huge amount and manually analysing such huge data to extract numeric stock or banking business prices will be tedious task and to overcome from such issue we are employing automatic numeric extraction via NLP (natural language processing) and NER (name entity recognition) technique by analysing rich financial scripts or NEWS data and then extracting terms to identify sentiments. Extracting numeric values with sentiments can help business peoples to know which stock or business is giving positive or negative sentiments and base on sentiments users can invest their money and this analysis will help them in avoiding losses.

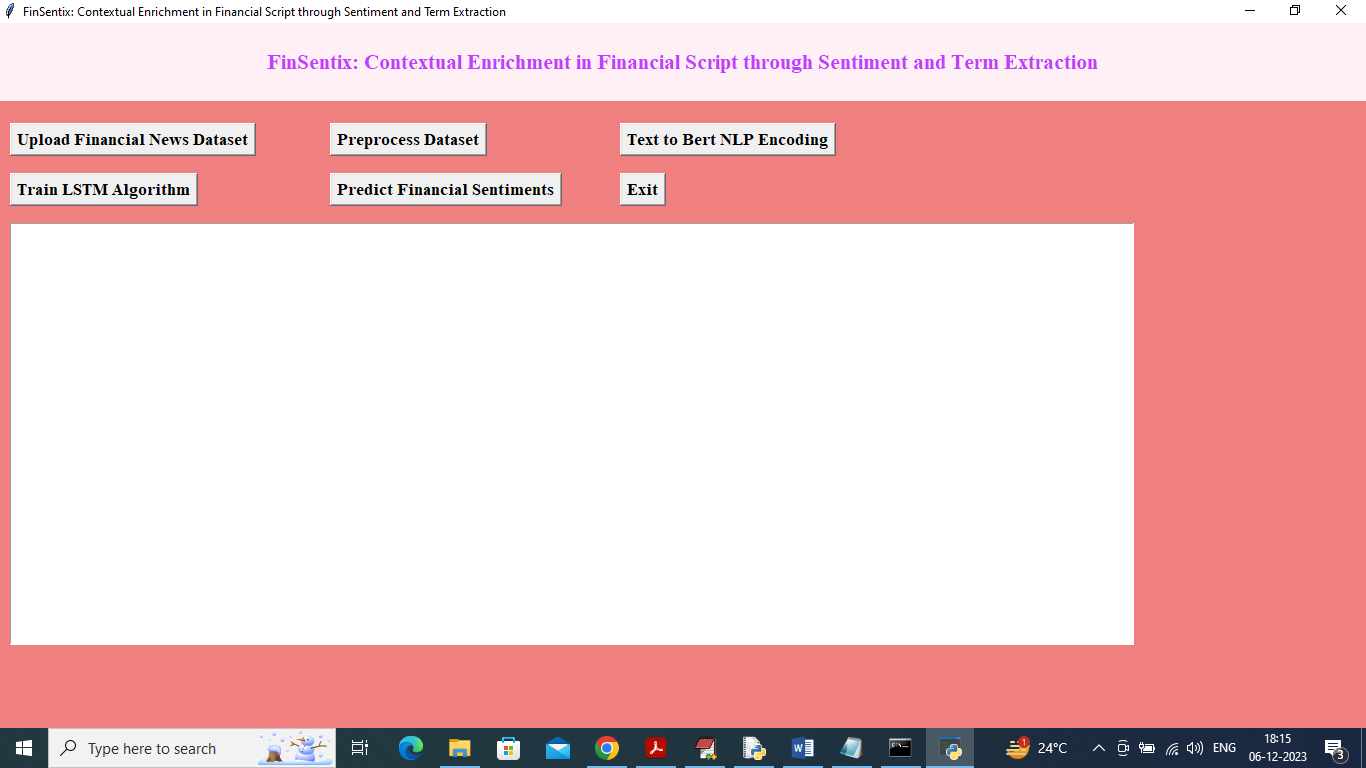
In the past many NER and financial sentiments algorithms are introduced but those algorithms are heavily dependent on manual sentiments or NER extraction which can be error prone and to avoid this problem we have automated Numeric prices or percentage extraction via NLP and NER techniques and then employ deep learning algorithm to detect sentiments

Propose algorithm consists of following modules

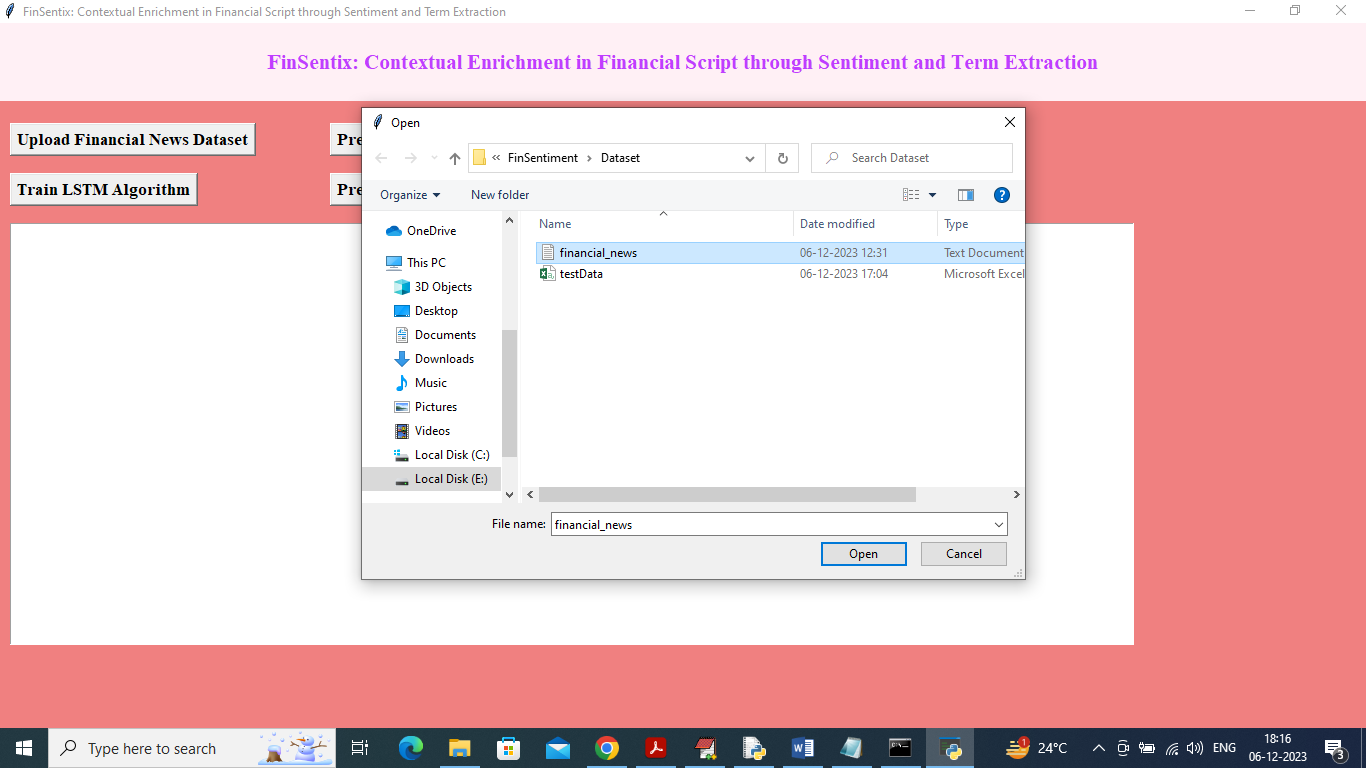
1. Data collection and processing: we have collected financial news dataset from HUGGING FACE website and then process this dataset to remove stop words, special symbols, stemming and lemmatization and then clean all business data
2. NLP Model selection: here we employ NLP regular expression algorithm to identify all numeric values from NEWS text data and then employ BERT algorithm to convert all news text data into numeric vector. BERT will replace each word with its average frequency
3. Labelling: Hugging face financial dataset already annotated with sentiments and can be download from this URL https://huggingface.co/datasets/financial\_phrasebank/blob/main/data/FinancialPhraseBank-v1.0.zip
4. Training and Validation: we have splitted BERT process data into 80 and 20 % ratio where 80% data will be used for training and 20% for testing
5. LSTM training: 80% data will be input to LSTM to train a model and this model will be applied on 20% unseen test data to calculate prediction accuracy and this model can be used to predict sentiments from NEWS financial test data
6. Real Time Prediction: we have created a model to predict prices, percentages and news sentiments from new test data and you upload this test data to predict above mention details.

SCREEN SHOTS

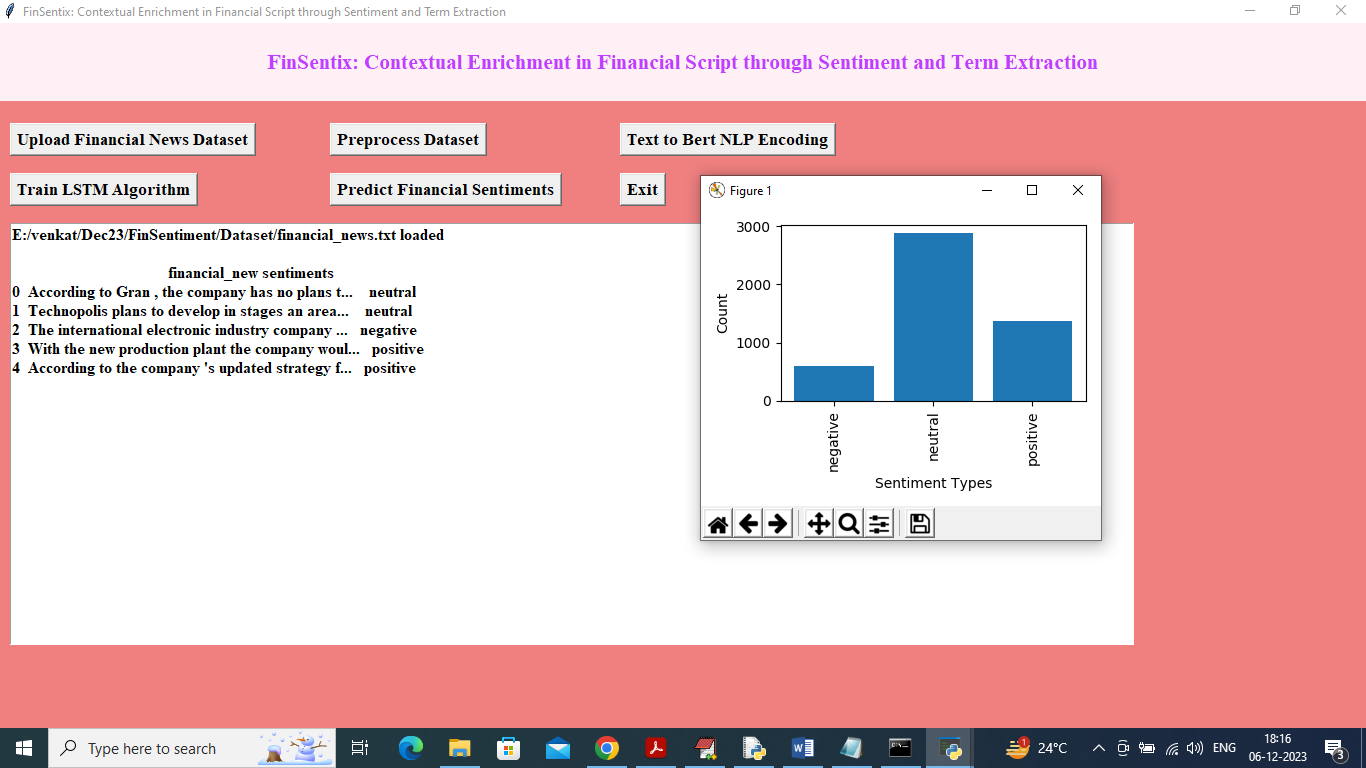
To run project double click on run.bat file to get below screen



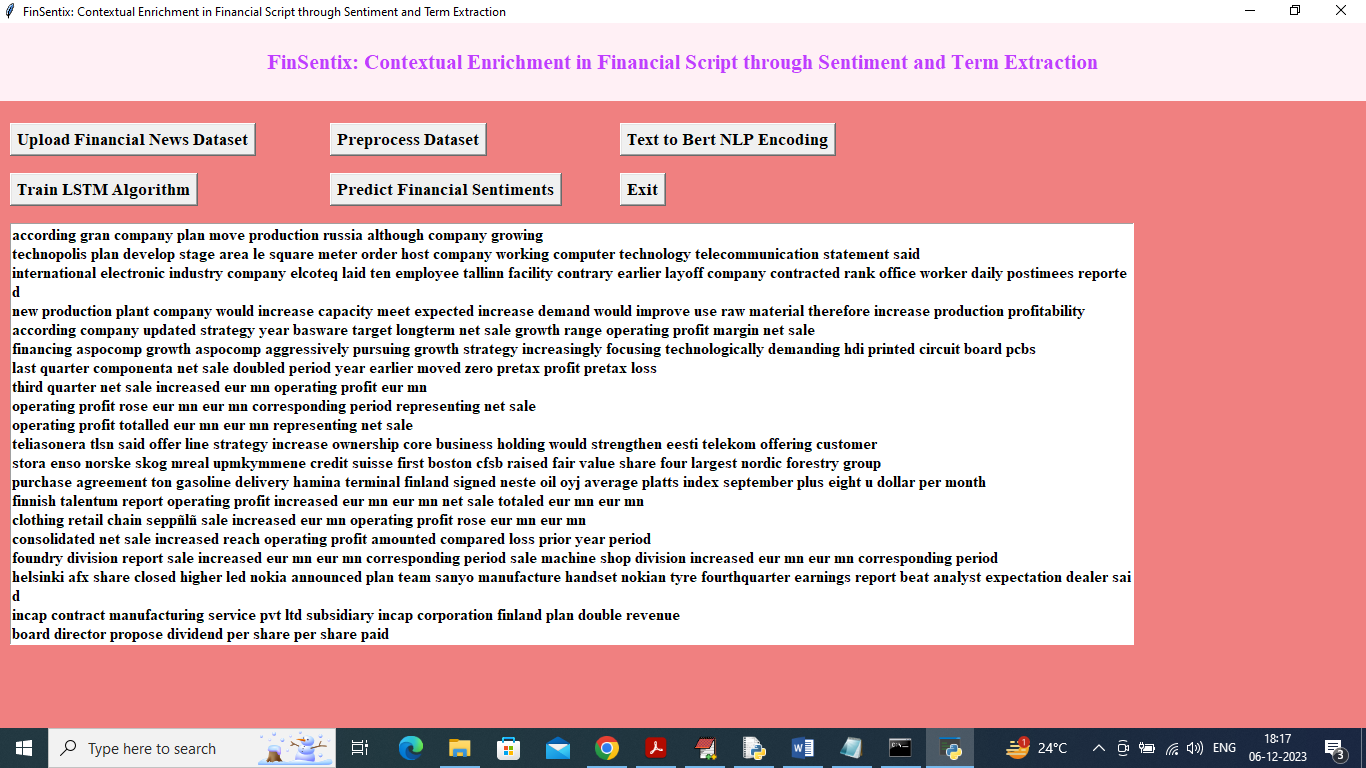
In above screen click on ‘Upload Financial News Dataset’ button to upload dataset and get below page



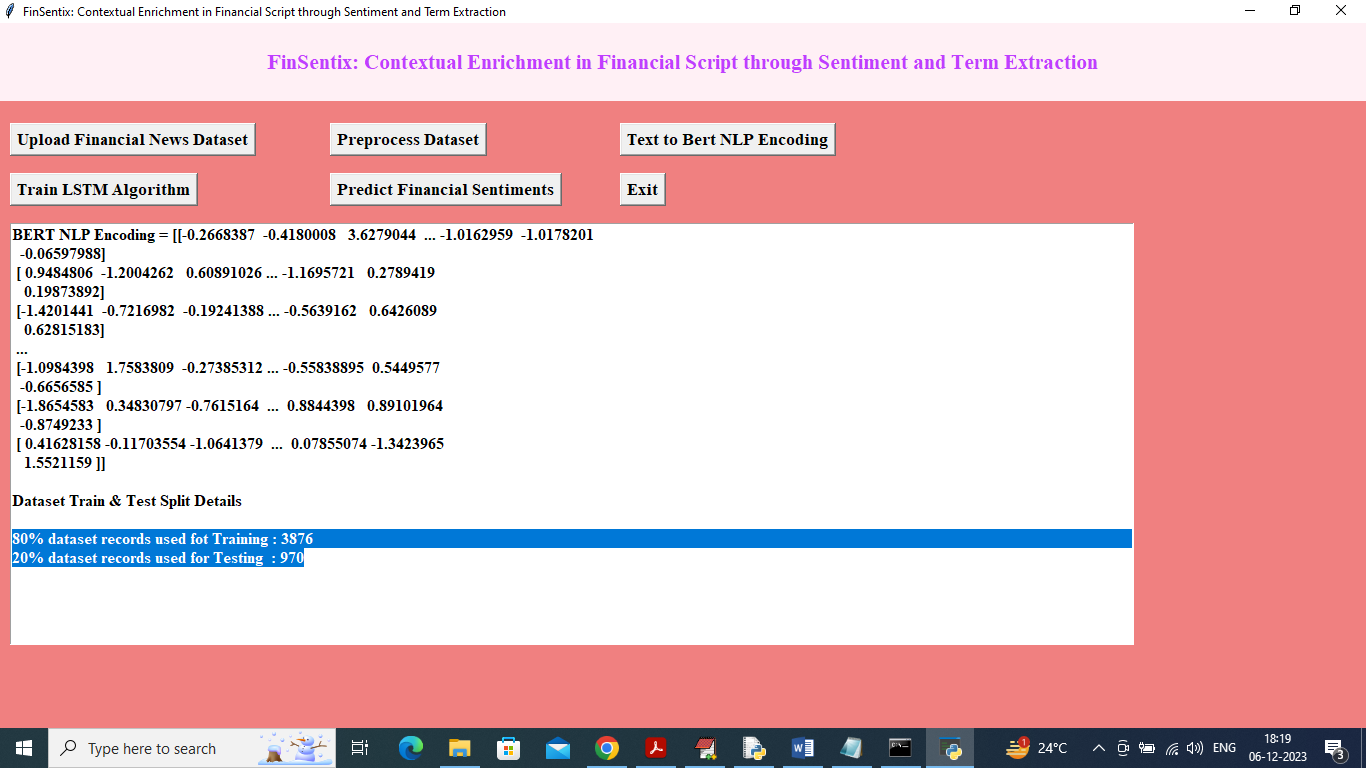
In above screen selecting and uploading financial news dataset and then click on ‘Open’ button to load dataset and get below page



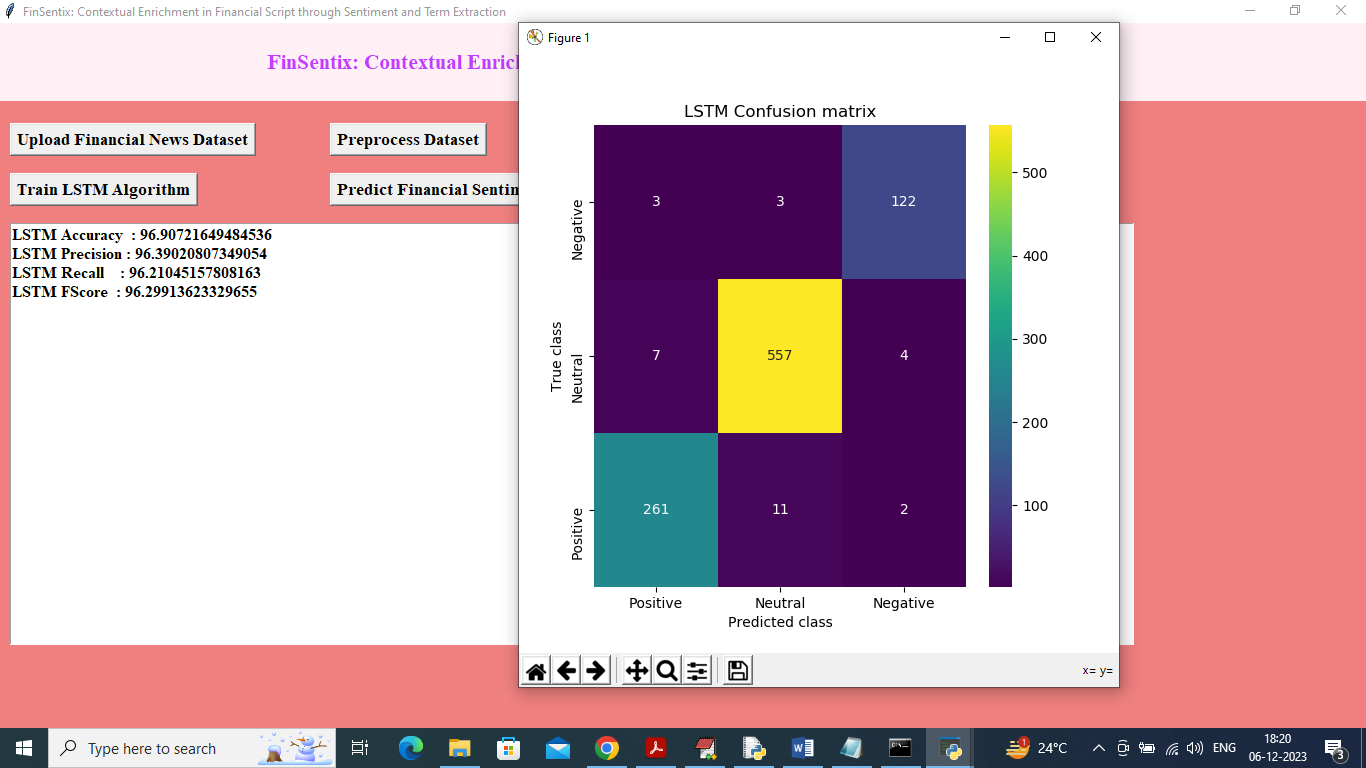
In above screen financial news data loaded and displaying few records from them and in graph x-axis represents dataset sentiments type and y-axis represents count of those records and now close above graph and then click on ‘Preprocess Dataset’ button to clean NEWS test and get below output



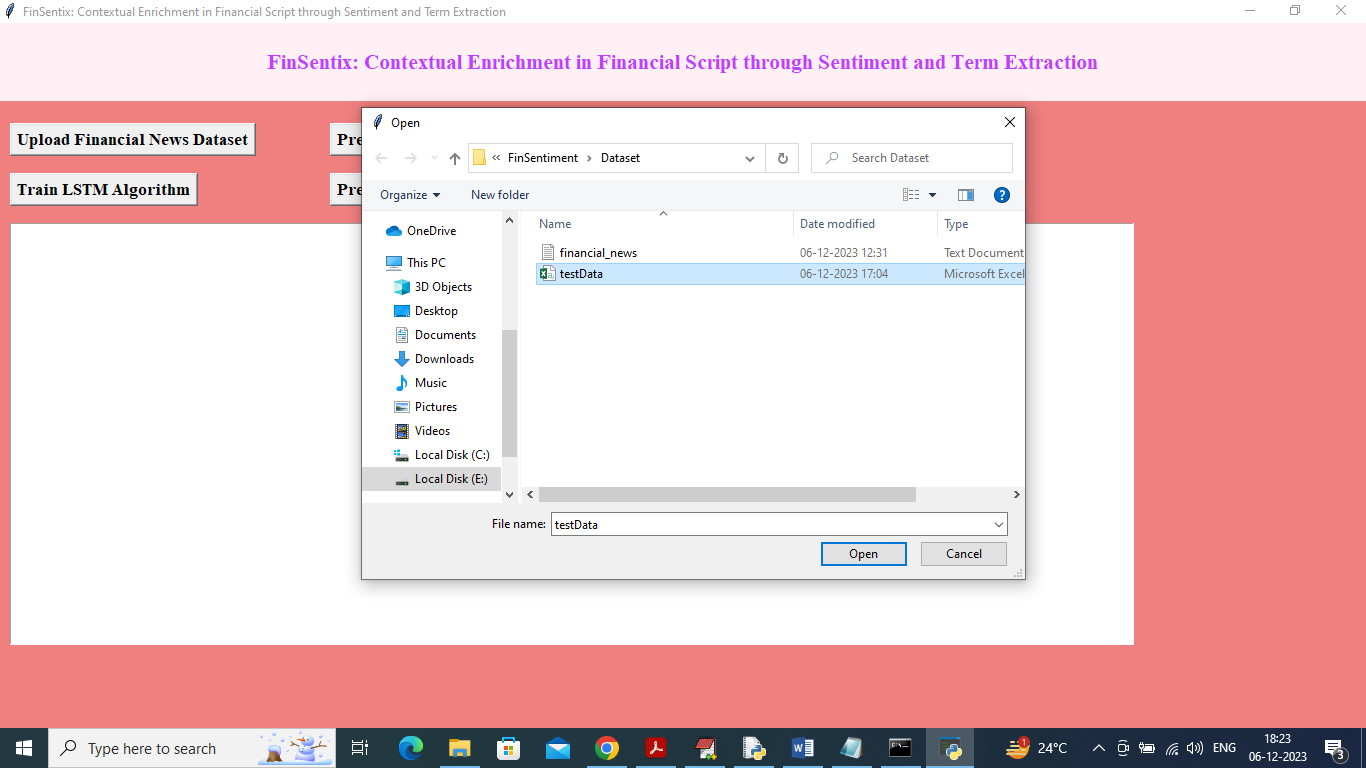
In above screen each line represents one financial news and from that line we removed special symbols, stop words. Stemming and lemmatization applied and now click on ‘Text to BERT NLP Encoding’ button to convert text news to BERT vector and get below output



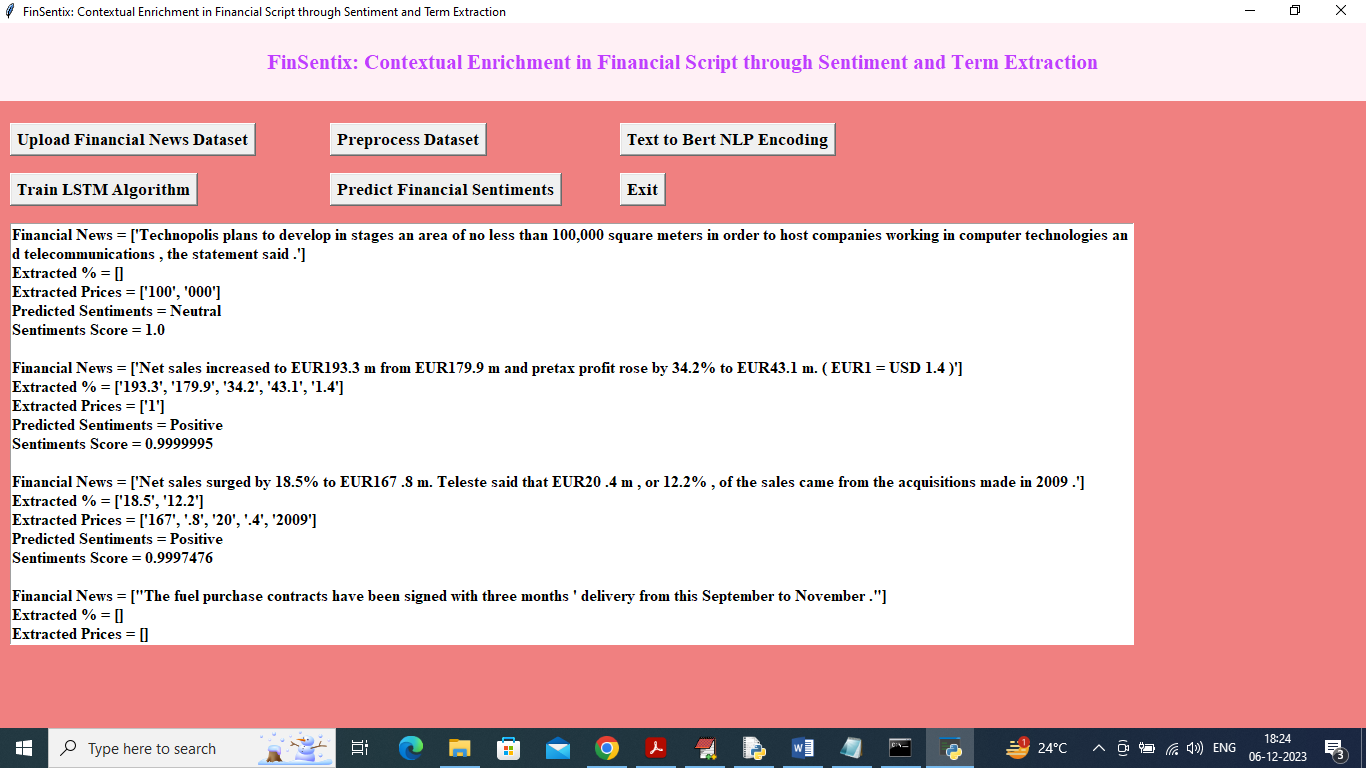
In above screen all text data converted to BERT encoding vector and then can see train and test size in blue colour and now click on ‘Train LSTM Algorithm’ button to train LSTM and get below output

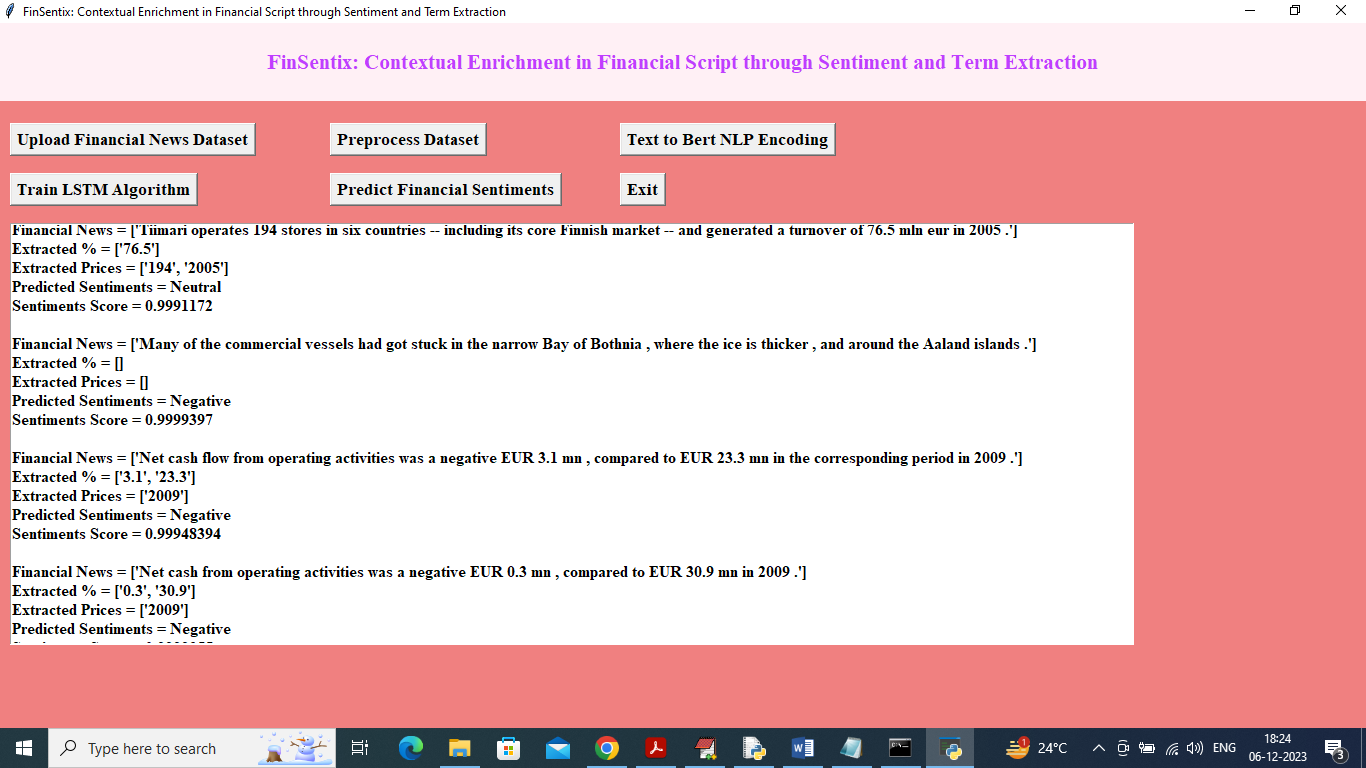


In above screen LSTM got 96% accuracy and can see other metrics like precision, recall and FSCORE and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and all different colour boxes in diagnol represents correct prediction count and remaining blue boxes represents incorrect prediction count which are very few and now close above graph and then click on ‘Predict Financial Sentiments’ button to upload test data and get below output



In above screen selecting and uploading test data and then click on ‘Open’ button to get below output





In above two screen we are displaying NEWS text from TEST data and then displaying extracted numeric %, prices, sentiments and sentiments predicted scores and this prediction you can see for each NEWS text line