Sentimental analysis in Tamil language using Deep learning

Team No- 2

20IT119 - M.VIJI

20IT048-T.MADHUMITHA

Guide Name

Dr. A. M. Abirami

Overview

- Problem Statement
- Literature Review
- Objectives
- System Design/ Diagram/ Flowchart
- Methodologies to solve the problem
- Results/ Screenshots
- Project Demo Video
- References

Problem Statement & Description

- Though Tamil language is the oldest language in the world, it lacks behind in the field of technological advancement, information sharing and understanding.
- In order to preserve endangered indigenous heritage and culture of Tamil language, we particularly provide our contribution on Sentimental Analysis in Tamil language.
- There is scarcity of large, diverse, and labeled datasets specifically annotated for sentiment analysis in Tamil. The absence of such datasets hinders the training and evaluation of sentiment analysis models for Tamil compared to most other languages.
- This project aims to address this problem to develop an effective sentiment analysis model for the unique characteristics of the Tamil language.

Literature Review

Sl. No	Authors/ Affiliations	Title	Journal Name, Year	Review Comments
1	Suba Sri Ramesh Babu	Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach	National College of Ireland, 14th August 2022	This research work aims to propose hybrid deep learning approaches such as CNN-BiLSTM, CNN-LSTM, and CNN-BiGRU. The proposed methods will be evaluated and compared on various metrics to find the best performing model among them. The result shows that CNN-BiLSTM has achieved the higher accuracy of 80.2% and highest f1-score of 0.64 when compared to other two models.

Literature Review

Sl. No	Authors/ Affiliations	Title	Journal Name, Year	Review Comments
2	Sajeetha Thavareesan Sinnathamby Mahesan	REVIEW ON SENTIMENT ANALYSIS IN TAMIL TEXTS	Eastern University, Sri Lanka. December 2018	This review paper mainly focuses on analysing the recent literature on the sentiment analysis field. On analysing various papers it is found that SVM and RNN classifiers taking TF-IDF and word2vec gives better performance than grammar rules based classification and other various classifiers
3	Vallikannu Ramanathan Meyyappan Thirunavukkarasu	Sentimental Analysis:An approach for Analysing tamil movie reviews using tamil tweets	Research gate, October 2021	In this paper three methods have been used for finding accuracy based on keywords. Term-frequency(TF) Inverse Term-frequency(IDF) Domain specific ontology(DSO) and contextual semantic sentiment analysis(CSSA). Out of this, it is found that TF along with IDF and CSSA gives better accuracy of 77.89%

Literature Review

Sl. No	Authors/ Affiliations	Title	Journal Name, Year	Review Comments
4	N.Sripriya S. Dhivya	Sentimental analysis for code-mixed tamil language	CEUR Workshop Proceedings (CEUR-WS.org), Dec 13 2021	This paper elaborates a model that generates embedding representation for the text data available. This is a multiclass classification problem that generates five different labels for the data collected from YouTube comments. It is found that Random Forest classifier builds a set of decision trees from the randomly selected subgroup of training data. This classifier is more accurate and vigorous in making decisions due to the numerous decision trees involved in the process

Literature Review

Sl. No	Authors/ Affiliations	Title	Journal Name, Year	Review Comments
5	Pavan Kumar P.H.V, Premjith B, Sanjanasri J.P and Soman K.P	Deep Learning Based Sentiment Analysis for Malayalam,Tamil and Kannada Languages	CEUR Workshop Proceedings (CEUR-WS.org), December 17-21, 2020	This paper describes the we implementation of three different Deep learning-based architectures for predicting various sentiments associated with the Dravidian CodeMix languages(Malayalam, Tamil, Kannada): 1. Deep Neural Network (DNN) 2. Bidirectional-Long Short Term Memory network (Bi-LSTM) 3. Convolution Neural network (CNN) combined with a Long Short Term Memory network (LSTM) The dataset used in this task is CodeMix text associated with the context of social media. The BiLSTM layer suits the classification of Tamil and Kannada corpus

Literature Review

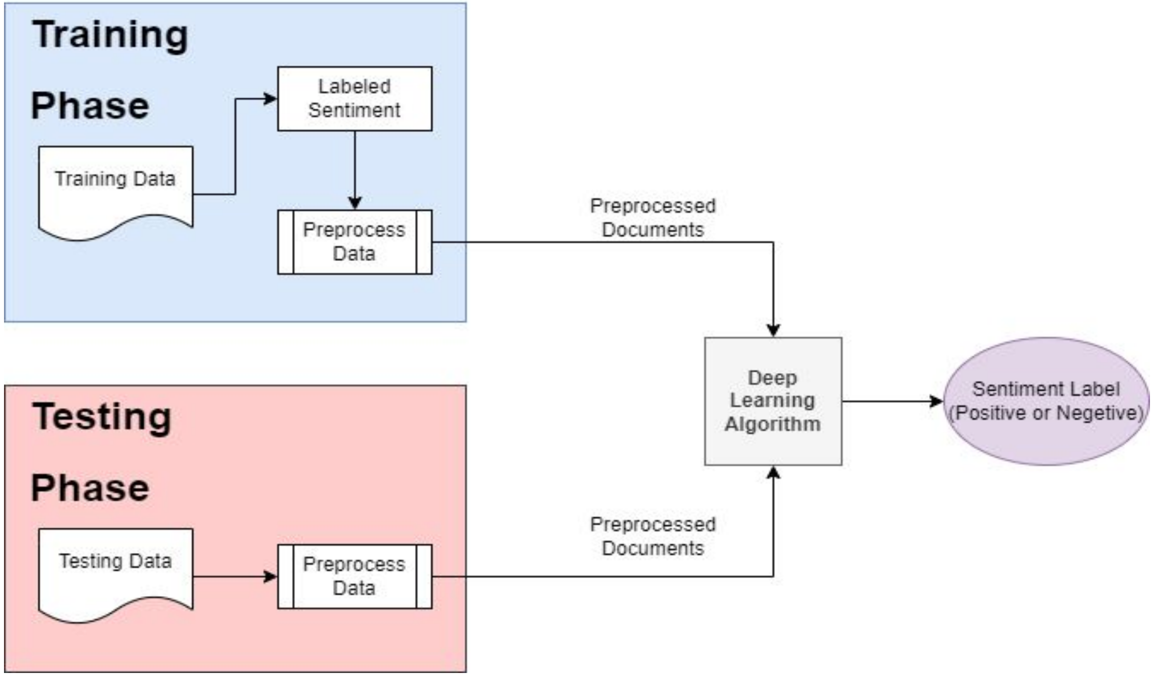
Sl. No	Authors/ Affiliations	Title	Journal Name, Year	Review Comments
6	Kaushika. N Uma.V	sentiment analysis for english and tamil tweets using path length similarity based word sense disambiguation	IOSR journal of computer engineering, june 2016	Here the data are collected from Twitter using Twitter Api.The collected tweets consist of mixed Tamil and English language which increases the ambiguity in classification.so for that Word Sense Disambiguation is employed to overcome this issue and semantic similarity between the words is predicted using Path based similarity. The shortest distance between two synset is found using Edge based similarity. The Support Vector Machine classifier is used to predict the polarity values of the tweets.

Objectives

OBJECTIVES:

- To develop web application to get input from user and analyse the sentiment with the help of back-end classifier model to return the predicted result with the polarity of the sentence.
- To perform pre-processing and feature extraction on the dataset and use them to classify the data into either positive or negative categories.
- To analyse the best performing model for sentiment classification in Tamil language by evaluating and comparing them.
- To perform sentiment analysis directly on the original Tamil text, without relying on machine translation and evaluate the developed model using appropriate metrics.

Design/ Diagram



Methodologies

Dataset collection - The dataset consists of tamil movie reviews which are collected from various sources like social media platforms,twitter,youtube etc.

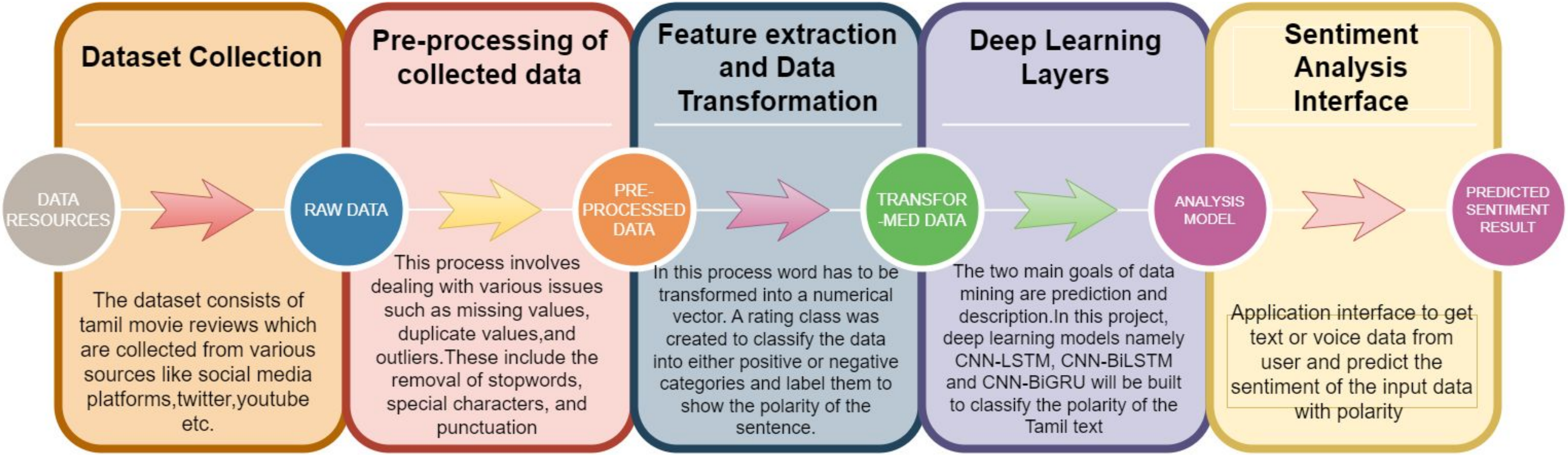
Pre-processing of collected data -This process involves dealing with various issues such as missing values, duplicate values,and outliers. These include the removal of stopwords, special characters, and punctuation using libraries such as nltk, Indic NLP which supports Tamil language.

Feature extraction and Data transformation - In this process word has to be transformed into a vector before they can start learning. A word embedding can be used to extract differences and similarities between words. A rating class was created to classify the data into either positive or negative categories and label them to show the polarity of the sentence.

Model Construction - In this project, deep learning models namely CNN-LSTM, CNN-BiLSTM and CNN-BiGRU will be built to classify the polarity of the Tamil text.

Methodologies

Model Evaluation - The performance of the model will be evaluated by taking into account the various metrics that are related to its prediction ratios.this will give the exact prediction of the data from various deep learning models



Dataset Collection

The dataset consists of tamil movie reviews which are collected from various sources like social media platforms,twitter,youtube etc.

Pre-processing of collected data

This process involves dealing with various issues such as missing values, duplicate values,and outliers.These include the removal of stopwords, special characters, and punctuation

Feature extraction and Data Transformation

In this process word has to be transformed into a numerical vector. A rating class was created to classify the data into either positive or negative categories and label them to show the polarity of the sentence.

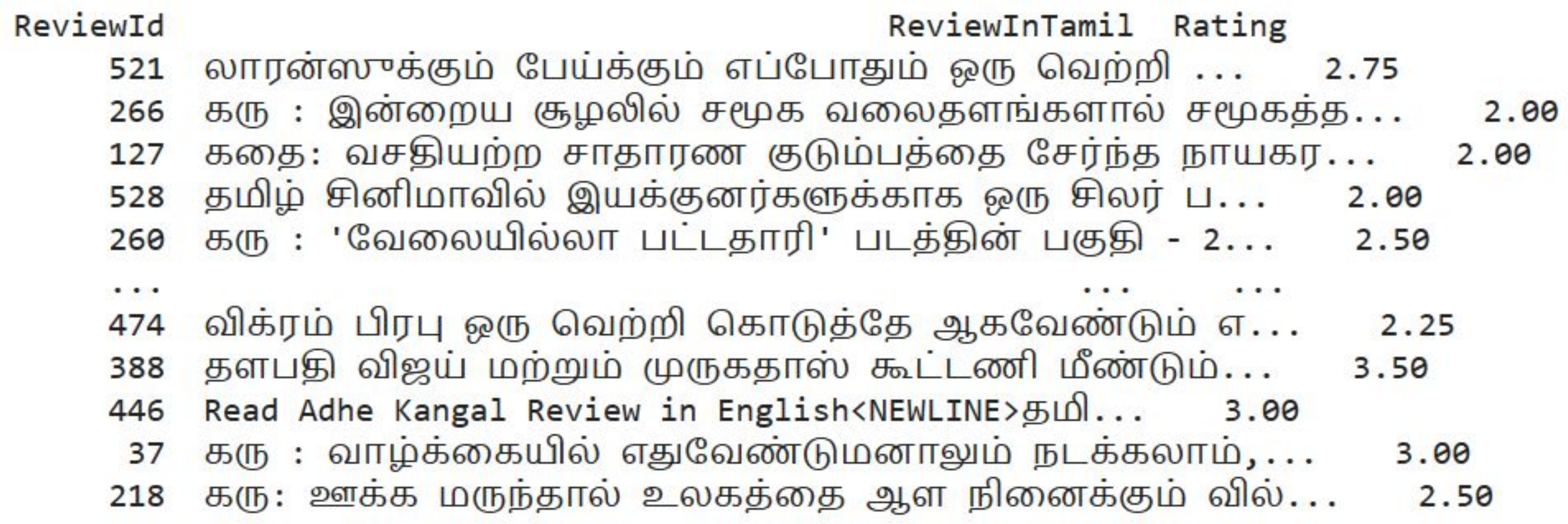
Deep Learning Layers

The two main goals of data mining are prediction and description.In this project, deep learning models namely CNN-LSTM, CNN-BiLSTM and CNN-BiGRU will be built to classify the polarity of the Tamil text

Sentiment Analysis Interface

Application interface to get text or voice data from user and predict the sentiment of the input data with polarity

Data Description



```
ReviewId                                              ReviewInTamil  Rating
     521  லாரன்ஸுக்கும் பேய்க்கும் எப்போதும் ஒரு வெற்றி ...    2.75
     266  கரு : இன்றைய சூழலில் சமூக வலைதளங்களால் சமூகத்த...    2.00
     127  கதை: வசதியற்ற சாதாரண குடும்பத்தை சேர்ந்த நாயகர...    2.00
     528  தமிழ் சினிமாவில் இயக்குனர்களுக்காக ஒரு சிலர் ப...    2.00
     260  கரு : 'வேலையில்லா பட்டதாரி' படத்தின் பகுதி - 2...    2.50
...                                                     ...     ...
     474  விக்ரம் பிரபு ஒரு வெற்றி கொடுத்தே ஆகவேண்டும் எ...    2.25
     388  தளபதி விஜய் மற்றும் முருகதாஸ் கூட்டணி மீண்டும்...    3.50
     446  Read Adhe Kangal Review in English<NEWLINE>தமி...    3.00
      37  கரு : வாழ்க்கையில் எதுவேண்டுமனாலும் நடக்கலாம்,...    3.00
     218  கரு: ஊக்க மருந்தால் உலகத்தை ஆள நினைக்கும் வில்...    2.50
```

Data Description

This dataset consists of two main columns:

ReviewId:

Each review is assigned a unique ReviewId to distinguish it from others in the dataset.

Review In Tamil:

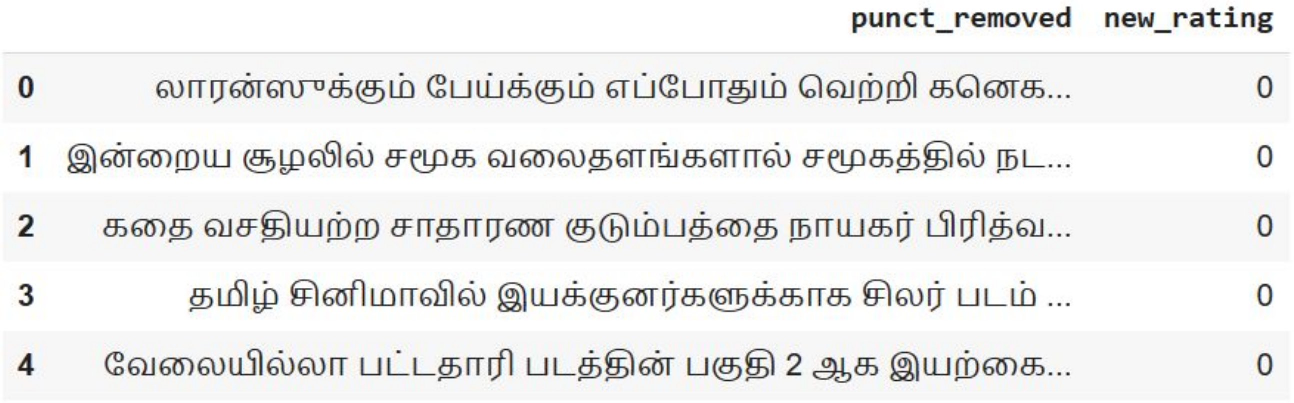
This column contains the text of the reviews written in Tamil language. The reviews may contain opinions, feedback, or descriptions of various aspects of of a particular movie.

Rating:

This column contains the ratings assigned to each review. Ratings indicate the overall satisfaction or sentiment expressed by the reviewer. (from 1 to 5)

Pre-Processed Data

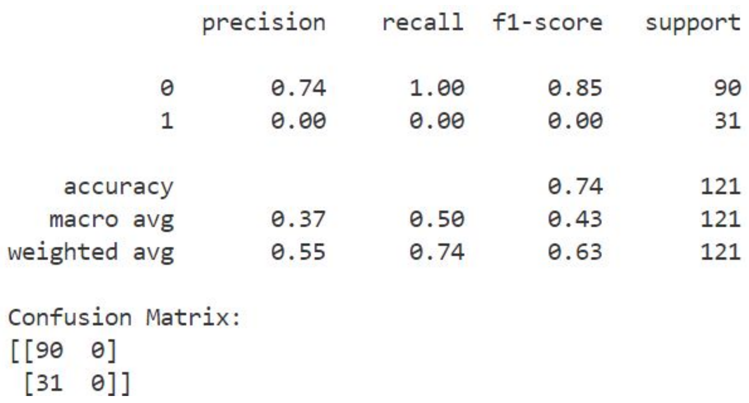
- The content of the reviews (Review In Tamil) column is preprocessed by padding, removing punctuations, newlines, tabs, and stopwords.
- Next we transformed the multi-class into binary class that is positive as 1 when the rating is above 3 otherwise, it is negative and labelled as 0 (0-Negative; 1-Positive).



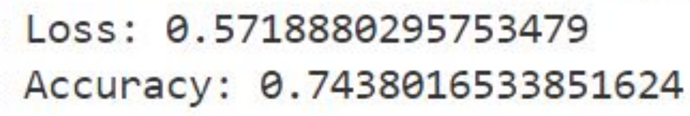
	punct_removed	new_rating
0	லாரன்ஸுக்கும் பேய்க்கும் எப்போதும் வெற்றி கனெக...	0
1	இன்றைய சூழலில் சமூக வலைதளங்களால் சமூகத்தில் நட...	0
2	கதை வசதியற்ற சாதாரண குடும்பத்தை நாயகர் பிரித்வ...	0
3	தமிழ் சினிமாவில் இயக்குனர்களுக்காக சிலர் படம் ...	0
4	வேலையில்லா பட்டதாரி படத்தின் பகுதி 2 ஆக இயற்கை...	0

Results/ Screenshots

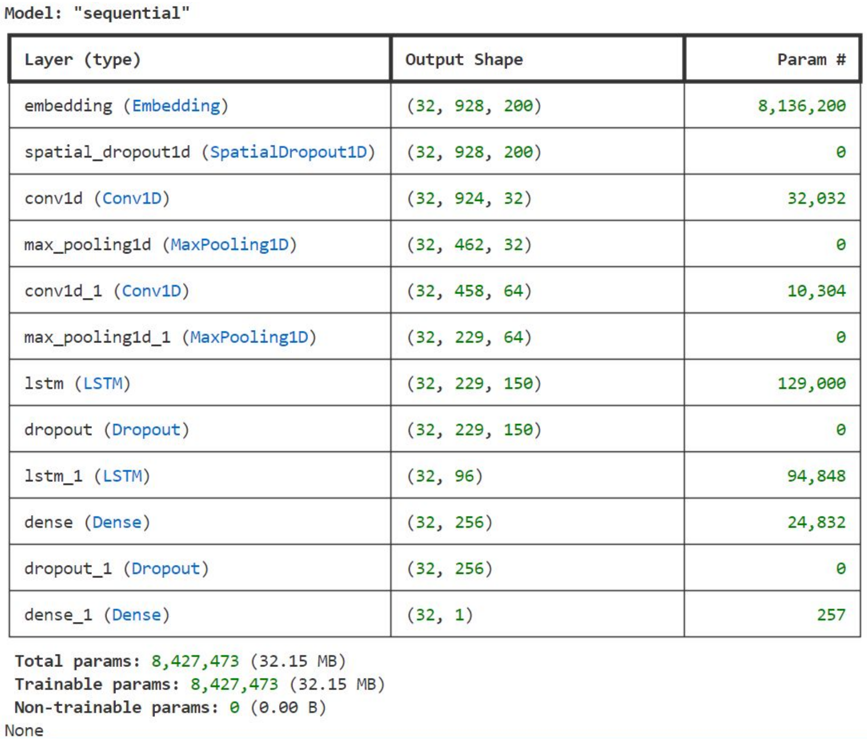
cnn_lstm



```
              precision    recall  f1-score   support

           0       0.74      1.00      0.85        90
           1       0.00      0.00      0.00        31

    accuracy                           0.74       121
   macro avg       0.37      0.50      0.43       121
weighted avg       0.55      0.74      0.63       121

Confusion Matrix:
[[90  0]
 [31  0]]
```



```
Loss: 0.5718880295753479
Accuracy: 0.7438016533851624
```

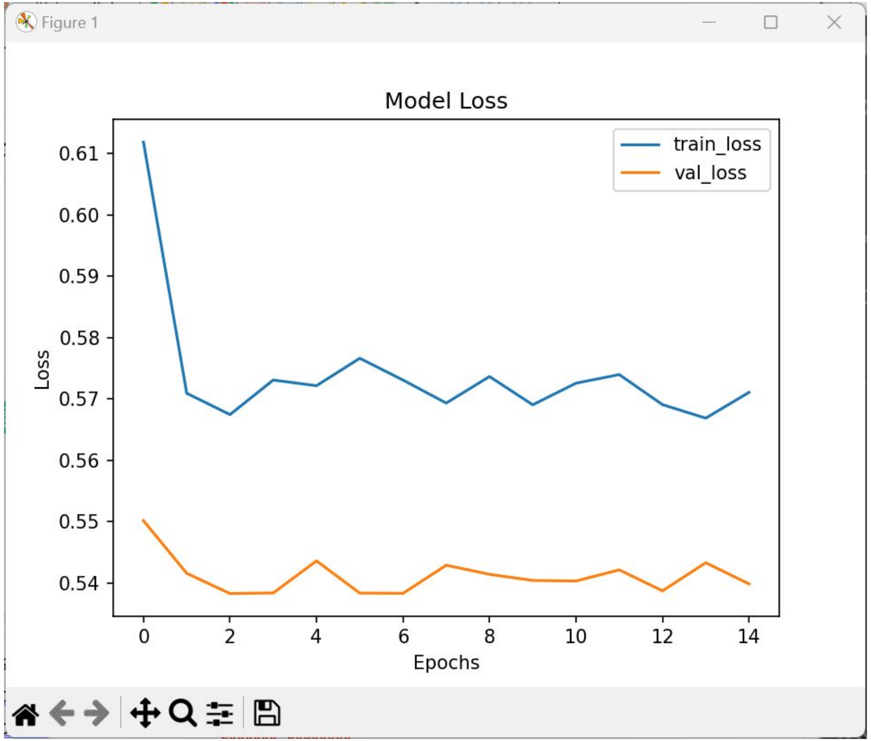


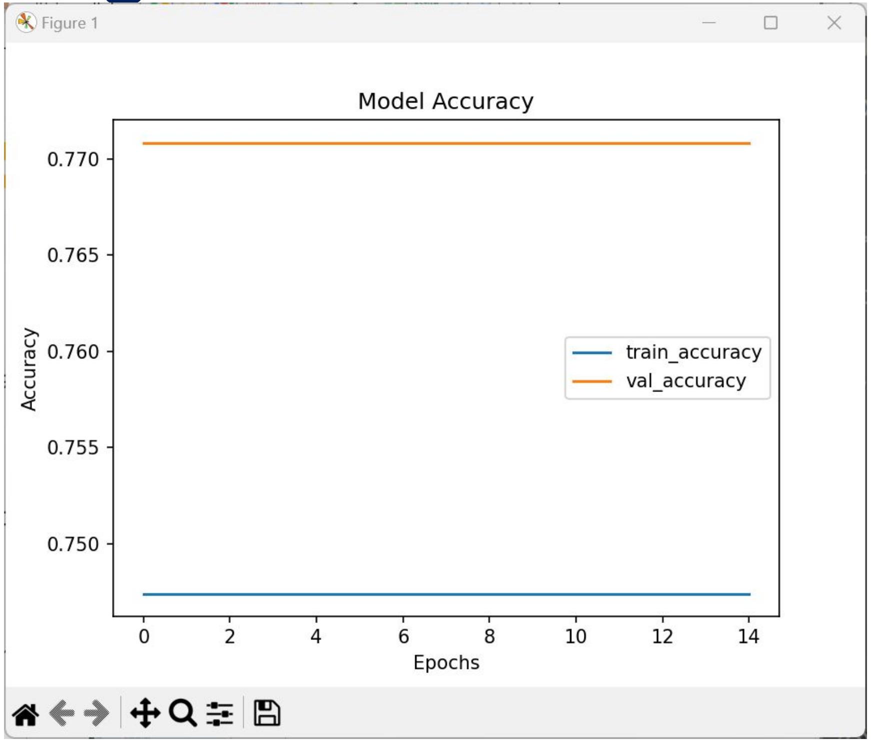
Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(32, 928, 200)	8,136,200
spatial_dropout1d (SpatialDropout1D)	(32, 928, 200)	0
conv1d (Conv1D)	(32, 924, 32)	32,032
max_pooling1d (MaxPooling1D)	(32, 462, 32)	0
conv1d_1 (Conv1D)	(32, 458, 64)	10,304
max_pooling1d_1 (MaxPooling1D)	(32, 229, 64)	0
lstm (LSTM)	(32, 229, 150)	129,000
dropout (Dropout)	(32, 229, 150)	0
lstm_1 (LSTM)	(32, 96)	94,848
dense (Dense)	(32, 256)	24,832
dropout_1 (Dropout)	(32, 256)	0
dense_1 (Dense)	(32, 1)	257

```
 Total params: 8,427,473 (32.15 MB)
 Trainable params: 8,427,473 (32.15 MB)
 Non-trainable params: 0 (0.00 B)
None
```

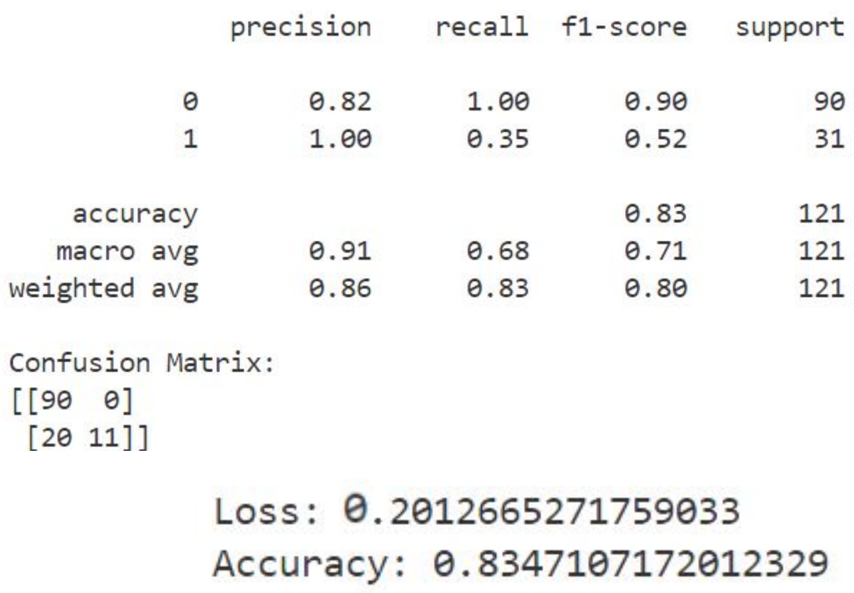
Graph of Loss and Accuracy for train and test data using CNN-LSTM





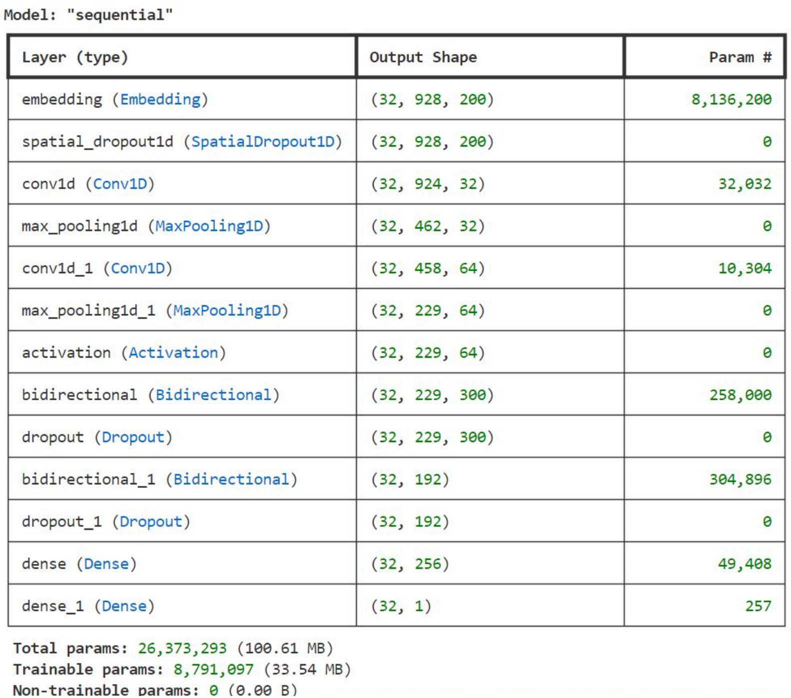
Results/ Screenshots

cnn_bilstm



```
              precision    recall  f1-score   support

           0       0.82      1.00      0.90        90
           1       1.00      0.35      0.52        31

    accuracy                           0.83       121
   macro avg       0.91      0.68      0.71       121
weighted avg       0.86      0.83      0.80       121

Confusion Matrix:
[[90  0]
 [20 11]]
```

```
Loss: 0.201265271759033
Accuracy: 0.8347107172012329
```

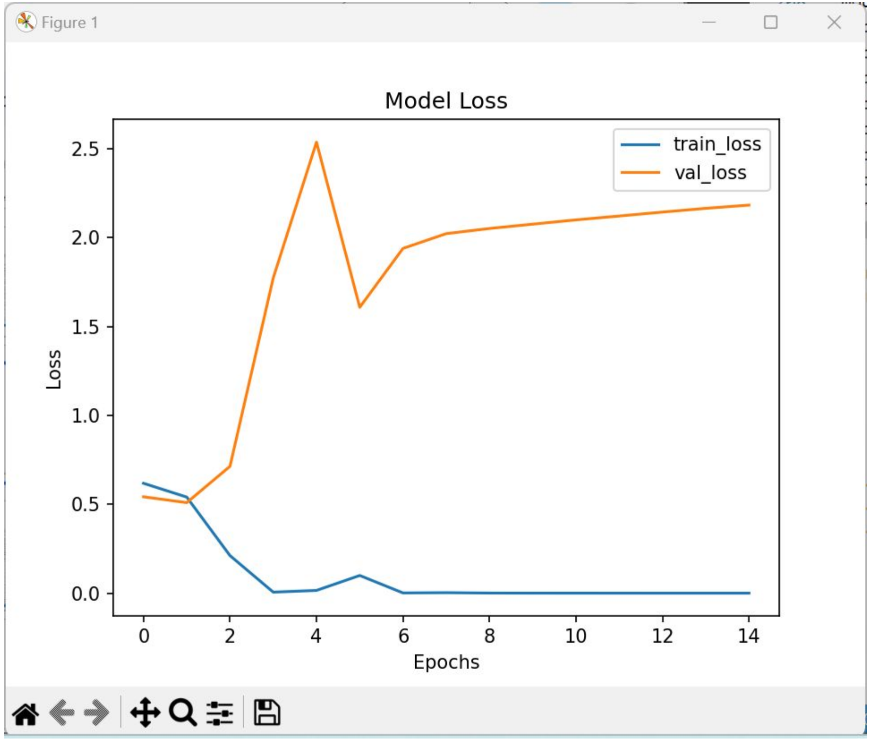


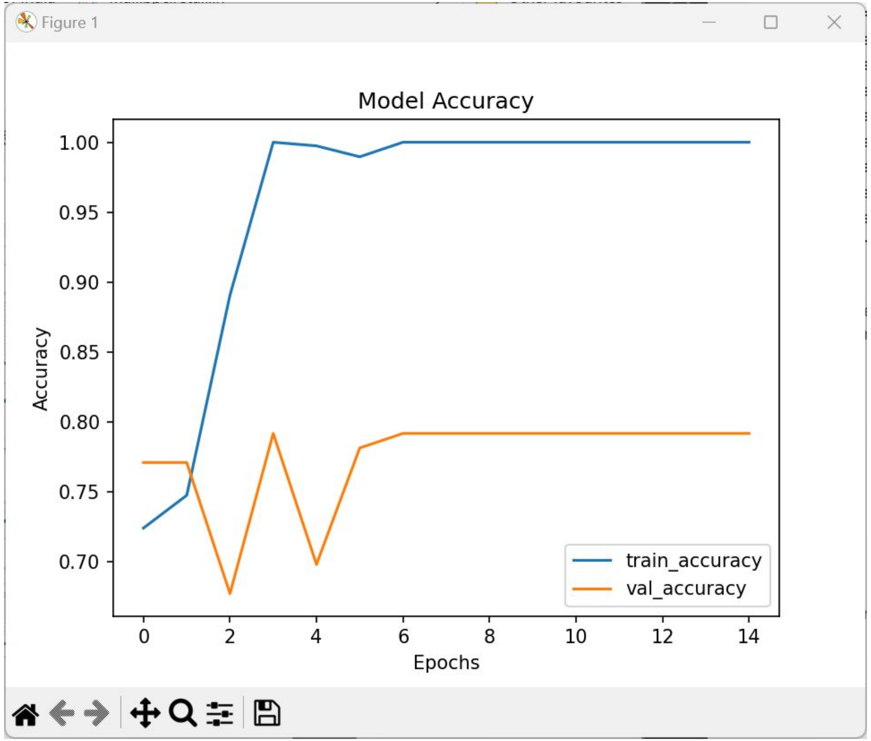
Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(32, 928, 200)	8,136,200
spatial_dropout1d (SpatialDropout1D)	(32, 928, 200)	0
conv1d (Conv1D)	(32, 924, 32)	32,032
max_pooling1d (MaxPooling1D)	(32, 462, 32)	0
conv1d_1 (Conv1D)	(32, 458, 64)	10,304
max_pooling1d_1 (MaxPooling1D)	(32, 229, 64)	0
activation (Activation)	(32, 229, 64)	0
bidirectional (Bidirectional)	(32, 229, 300)	258,000
dropout (Dropout)	(32, 229, 300)	0
bidirectional_1 (Bidirectional)	(32, 192)	304,896
dropout_1 (Dropout)	(32, 192)	0
dense (Dense)	(32, 256)	49,408
dense_1 (Dense)	(32, 1)	257

```
Total params: 26,373,293 (100.61 MB)
Trainable params: 8,791,097 (33.54 MB)
Non-trainable params: 0 (0.00 B)
```

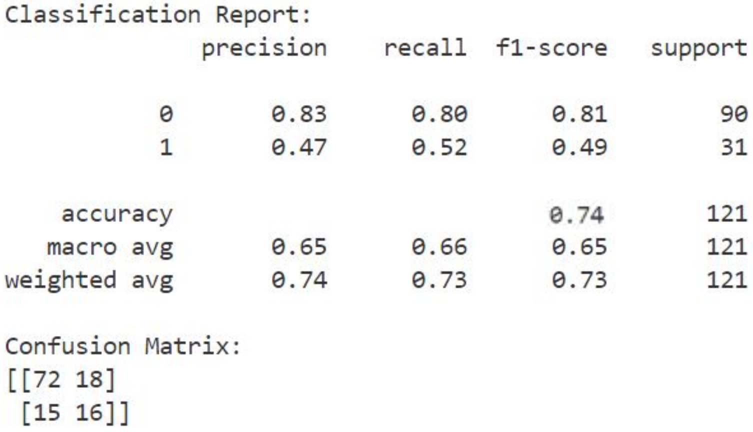
Graph of Loss and Accuracy for train and test data using CNN-BiLSTM



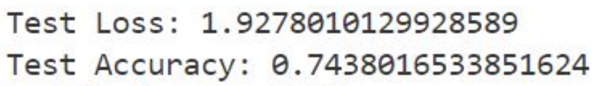


Results/ Screenshots

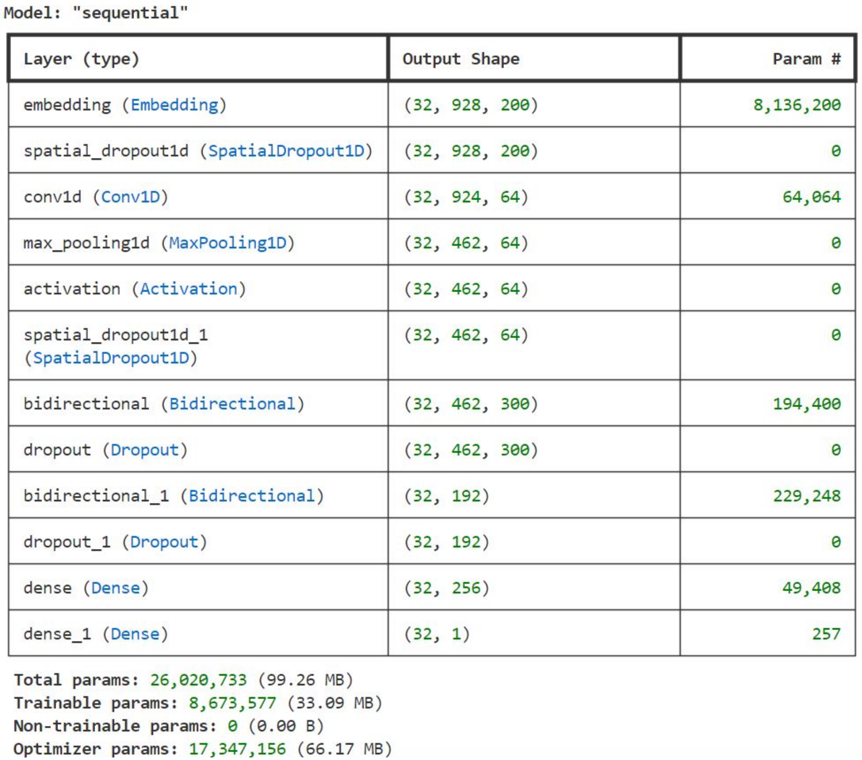
cnn_bigru



```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.81        90
           1       0.47      0.52      0.49        31

    accuracy                           0.74       121
   macro avg       0.65      0.66      0.65       121
weighted avg       0.74      0.73      0.73       121

Confusion Matrix:
[[72 18]
 [15 16]]
```



```
Test Loss: 1.9278010129928589
Test Accuracy: 0.7438016533851624
```

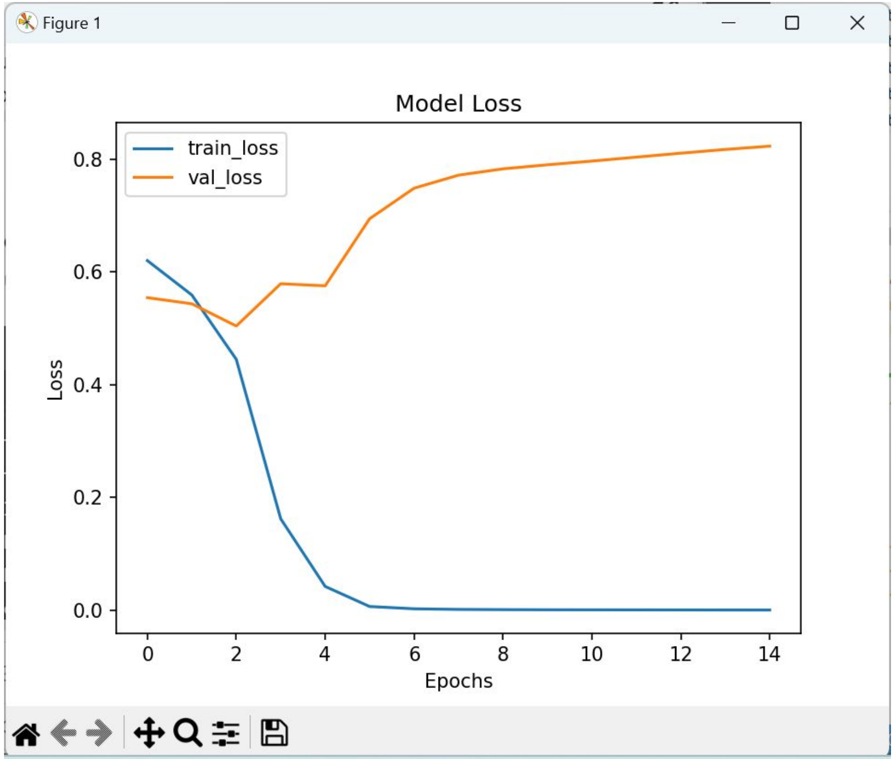


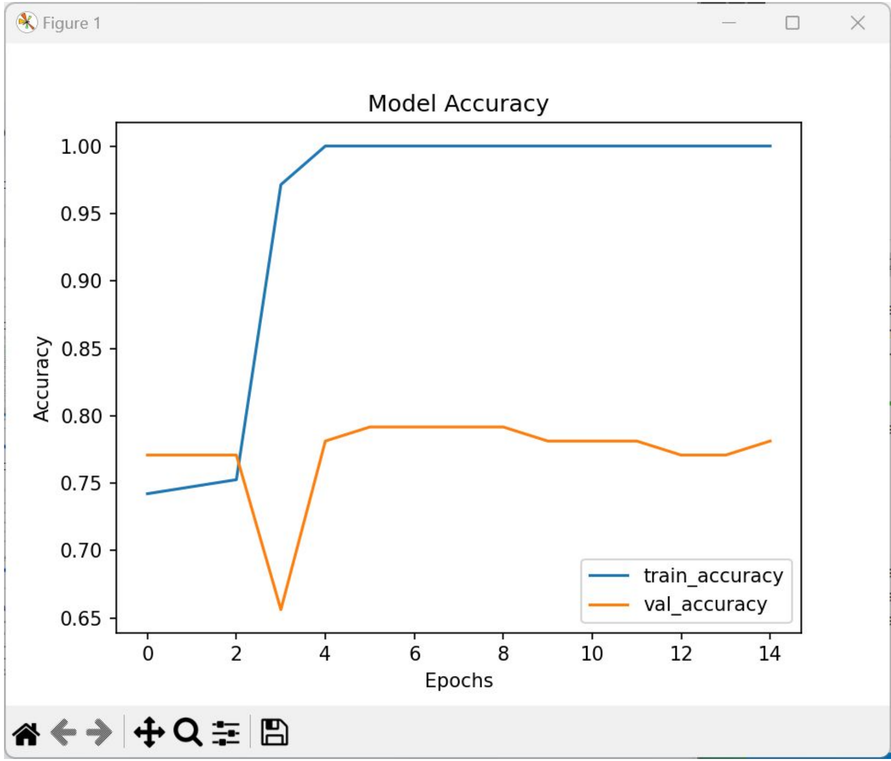
Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(32, 928, 200)	8,136,200
spatial_dropout1d (SpatialDropout1D)	(32, 928, 200)	0
conv1d (Conv1D)	(32, 924, 64)	64,064
max_pooling1d (MaxPooling1D)	(32, 462, 64)	0
activation (Activation)	(32, 462, 64)	0
spatial_dropout1d_1 (SpatialDropout1D)	(32, 462, 64)	0
bidirectional (Bidirectional)	(32, 462, 300)	194,400
dropout (Dropout)	(32, 462, 300)	0
bidirectional_1 (Bidirectional)	(32, 192)	229,248
dropout_1 (Dropout)	(32, 192)	0
dense (Dense)	(32, 256)	49,408
dense_1 (Dense)	(32, 1)	257

```
Total params: 26,020,733 (99.26 MB)
Trainable params: 8,673,577 (33.09 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 17,347,156 (66.17 MB)
```

Graph of Loss and Accuracy for train and test data using CNN-BiGRU





Model Comparison

Comparison Table:

Models	Accuracy	Loss	Precision	Recall	F1-score
CNN_LSTM	74%	0.59	0.74	0.90	0.85
CNN_BiLSTM	83%	2.20	0.82	1.0	0.90
CNN_BiGRU	74%	1.05	0.83	0.80	0.81

Model Selection:

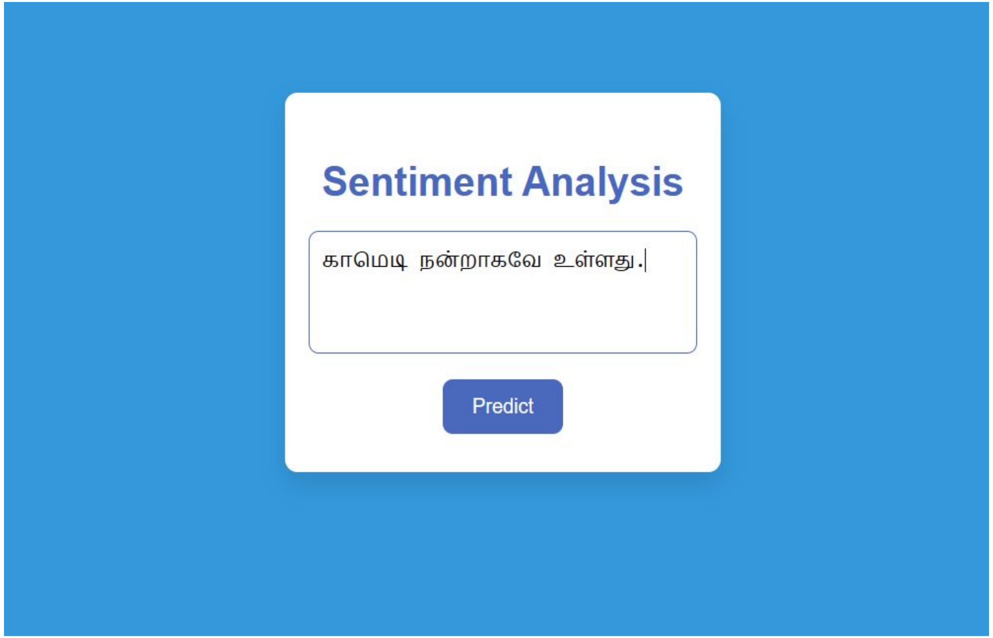
Among the three models, the CNN-BiLSTM model demonstrated the highest accuracy on the sentiment analysis task.

Therefore, the CNN-BiLSTM model was selected for deployment in the front-end web application.

Front-End Web Application

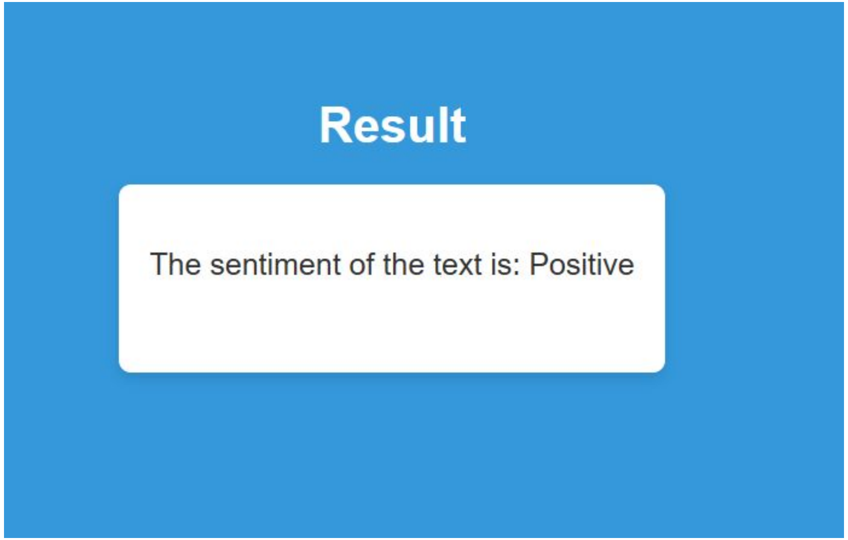
- The Flask framework in Python was used to develop the front-end web application.
- Bidirectional LSTMs allow the model to learn from both past and future contexts, enabling better understanding of the text.
- This model showed improved accuracy compared to other models, as it can capture more comprehensive contextual information.
- Users can input Tamil text into a form, and the CNN-BiLSTM model predicts the sentiment of the text.
- The application provides real-time sentiment analysis results, allowing users to quickly assess the sentiment of their text data.

Front-End Development Results



Sentiment Analysis

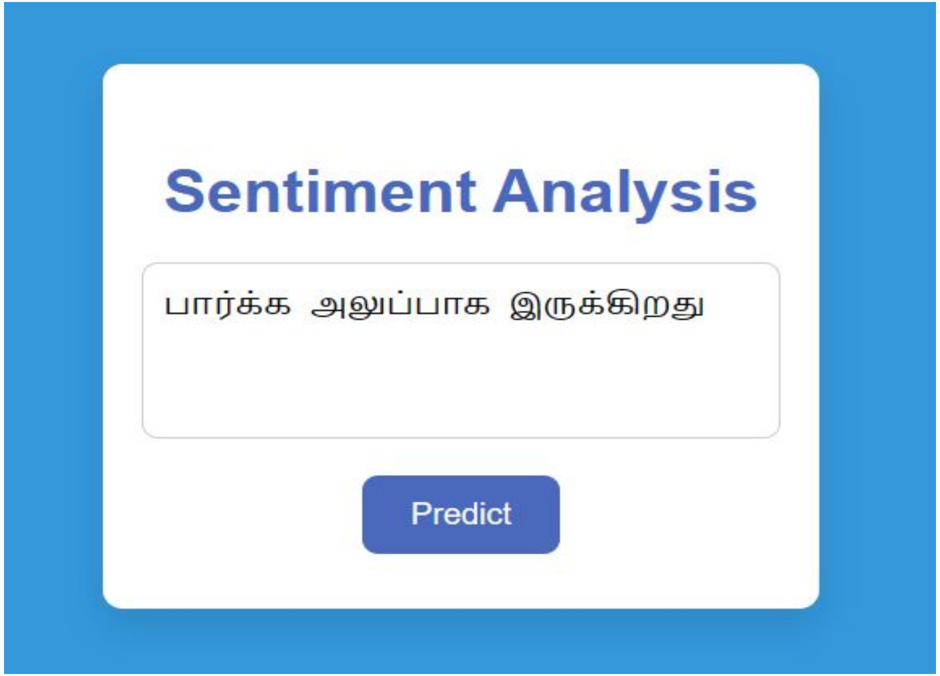
காமெடி நன்றாகவே உள்ளது.



Result

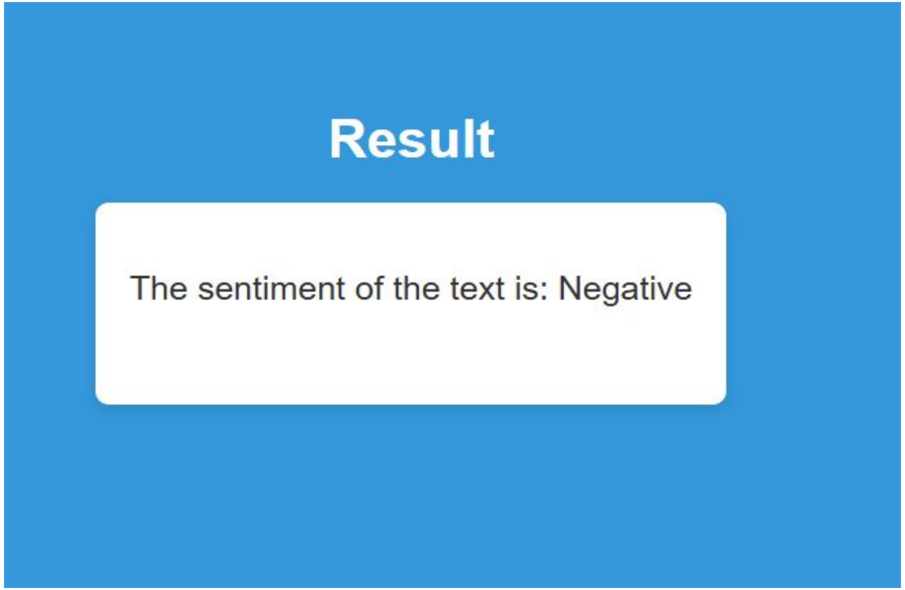
The sentiment of the text is: Positive

Front-End Development Results



Sentiment Analysis

பார்க்க அலுப்பாக இருக்கிறது



Result

The sentiment of the text is: Negative

Conclusion

- The main aim of this research work is to analyse the best performing model for sentiment classification in Tamil language by evaluating and comparing them.
- In this sentiment analysis project, we explored three deep learning models: CNN-LSTM, CNN-BiLSTM, and CNN-BiGRU, for predicting sentiment from Tamil text. These models were evaluated based on their accuracy and performance.
- The sentiment analysis project demonstrated the effectiveness of deep learning models, particularly the CNN-BiLSTM architecture, in predicting sentiment from Tamil text.
- By deploying the CNN-BiLSTM model in a user-friendly web application, individuals can easily analyze sentiment in their Tamil text data, facilitating decision-making and understanding public opinion.

References

1. Suba Sri, Ramesh Babu. Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach. National College of Ireland, 14th August 2022.
2. Sajeetha Thavareesan, Sinnathamby Mahesan. REVIEW ON SENTIMENT ANALYSIS IN TAMIL TEXTS. Eastern University, Sri Lanka. December 2018.
3. Vallikannu Ramanathan, Meyyappan Thirunavukkarasu. Sentimental Analysis:An approach for Analysing tamil movie reviews using tamil tweets. Research gate, October 2021.
4. N.Sripriya, S. Dhivya. Sentimental analysis for code-mixed tamil language. CEUR Workshop Proceedings (CEUR-WS.org), Dec 13 2021.
5. Pavan Kumar P.H.V, Premjith B, Sanjanasri J.P and Soman K.P. Deep Learning Based Sentiment Analysis for Malayalam,Tamil and Kannada Languages. CEUR Workshop Proceedings (CEUR-WS.org), December 17-21, 2020.
6. Kaushika. N, Uma.V. Sentiment analysis for english and tamil tweets using path length similarity based word sense disambiguation. IOSR journal of computer engineering, June 2016

demo link: https://drive.google.com/drive/folders/1T3QQrmRbRqxZ7zVYbciAUbgKYre02XDy