Cross-Domain Sentiment Analysis

A project report submitted in the fulfillment of the requirement for the award of the degree of Master of Computer Applications (MCA)

Submitted to:

Dr. Kumar Abhishek

Assistant Professor

National Institute of Technology, Patna

Submitted by:

Haradhan Kisku

Roll No: 2144025

MCA – CSE IV Semester

Session: 2021-2023

Master of Computer Applications

Department of Computer Science and Engineering

National Institute of Technology, Patna

Patna, Bihar – 800005

Certificate

NATIONAL INSTITUTE OF TECHNOLOGY PATNA
Department of Computer Science and Engineering

This is to certify that Mr. Haradhan Kisku, Roll No. 2144025, Enrollment No. 210474, is a registered candidate for MCA program under department of Computer Science and Engineering of National Institute of Technology Patna.

I hereby certify that he has completed all other requirements for submission of the thesis and recommend for the acceptance of a thesis entitled, “Sentiment Analysis” in the partial fulfillment of the requirements for the award of MCA degree.

_____

Supervisor

Dr. Kumar Abhishek

Assistant Professor

Department of CSE

National Institute of Technology Patna

May 2023

Declaration and Copyright Transfer

I Haradhan Kisku, Roll No. 2144025, Enrolment No. 210474, a registered candidate for MCA Programme under department of **Computer Science and Engineering** of **National Institute of Technology Patna**, declare that this is my own original work and does not contain material for which the copyright belongs to a third party and that it has not been presented and will not be present to any other University/ Institute for a similar or any other Degree award.

I further confirm that for all third-party copyright material in my project (Including any electronic attachment), I have "blanked out" third party material from the copies of the thesis/dissertation/book/articles etc; fully referenced the deleted materials and where possible, provided links (URL) to electronic source of the material.

I hereby transfer exclusive copyright for this thesis to NIT Patna. The following rights are reserved by the author: a) The right to use, free of charge, all or part of this thesis in future work of their own, such as books and lectures, giving reference to the original place of publication and copyright holding. b) The right to reproduce the thesis for their own purpose provided the copies are not offered for sale.

Signature of the candidate: ……..

Date: …….

Acknowledgement

I would like to take this opportunity to extend my heartfelt appreciation and recognize the exceptional individuals who have made this college project an extraordinary experience:

To my project supervisor, Dr. Kumar Abhishek, thank you for being an incredible mentor and guide. Your unwavering support, valuable insights, and endless patience have been pivotal in shaping this project's success. Your ability to challenge and inspire me has truly elevated my learning experience.

I am immensely grateful to the faculty members of Computer Science and Engineering at National Institute of technology, Patna for their dedication to education and their commitment to nurturing intellectual curiosity. Your passion for your respective fields and willingness to share your expertise have been a constant source of inspiration. You have ignited a thirst for knowledge that will continue to drive my academic pursuits.

I am grateful for the efforts and time contributed by mentors for helping us in successfully completing the project.

Table of Contents

1. INTRODUCTION 7

2. MOTIVATION 8

3. OBJECTIVE 9

4. RELATED WORKS 10

4.1 Transfer Learning 10
4.2 Cross-Domain Sentiment Analysis 10

5. TRANSFER LEARNING METHOD ON DEEP LEARNING 12

6. PROBLEM STATEMENT 14

7. PROPOSED METHODOLOGY 15

7.1 Dataset 15
7.2 Data Preprocessing 16
7.3 Converting words to numbers using BERT Encoder 16
7.4 Padding Sequences 16
7.5 Train Test Split 16
7.6 Creating Model 16
7.7 Fitting the model 17
7.8 Figures 17
7.8.1 Cross-Domain BERT + CNN 17
7.8.2 Bert + NN 18
7.8.3 One Hot Encoding + LSTM 18

8. RESULTS AND DISCUSSION 19

8.1 Datasets 19
8.2 Parameter Settings 19
8.2.1 Parameter setting for Cross-Domain BERT + CNN 19
8.2.2 Parameter setting for BERT + NN 20
8.2.3 Parameter setting for One Hot Encoding + LSTM 20
8.3 Experimental Results and Analysis 21
8.3.1 Cross-Domain BERT + CNN 21
8.3.2 BERT + NN 23
8.3.3 One Hot Encoding + LSTM 25

9. CONCLUSION 27

REFERENCES 28

Abstracts

The cross-domain text sentiment classification task aims to predict the sentiment of textual data across different domains. This task is particularly challenging because of variations in the vocabulary, writing style, and sentiment expressions across domains. Various techniques have been proposed to address this challenge, such as domain adaptation, transfer learning, and multi-task learning. Successfully solving this task has important applications in opinion mining, social media analysis, and customer feedback analysis. The algorithm uses BERT encoder and BERT preprocess and One Hot encoder for feature represented word vectors. Finally, by training the sentiment classifier on models like NN, CNN, LSTM, it is expected that the model can achieve better results over baseline approaches. Experiments are carried out on the Amazon product review and movie review datasets, and training and validation accuracy, training and validation loss and classification report values are used as evaluation indicators.

1. Introduction

Cross-domain sentiment analysis is a branch of natural language processing (NLP) that focuses on analyzing and understanding the sentiment or opinion expressed in text across different domains or subject areas. Sentiment analysis, also known as opinion mining, involves determining whether a piece of text conveys positive, negative, or neutral sentiment.

Traditional sentiment analysis techniques typically rely on domain-specific models trained on labeled data from a specific domain. However, these models often struggle to perform well when applied to different domains due to variations in language use, vocabulary, and sentiment expression. Cross-domain sentiment analysis aims to address this challenge by developing methods and algorithms that can effectively generalize sentiment analysis across various domains without requiring domain-specific training.

The goal of cross-domain sentiment analysis is to build models that can accurately identify sentiment polarity (positive, negative, or neutral) in text, regardless of the subject matter or domain. This is achieved by leveraging transfer learning techniques, which involve pre-training models on a large corpus of text data from diverse domains and then fine-tuning them on smaller domain-specific datasets.

In conclusion, cross-domain sentiment analysis is an important field within NLP that focuses on developing models capable of accurately analyzing sentiment across different domains. By addressing the challenges of domain adaptation and leveraging transfer learning techniques, cross-domain sentiment analysis contributes to better understanding and leveraging of sentiment expressed in text across diverse subject areas.

2. Motivation

Cross-domain NLP text classification is motivated by the need to develop models that can effectively handle text data from multiple domains or sources. In many practical applications, the text data that needs to be classified may come from different domains, such as social media, customer feedback, news articles, and scientific publications. Building models that can perform well on text data from different domains is essential to develop reliable NLP applications that can handle real-world data.

Cross-domain text classification is also motivated by the scarcity of labeled data in some domains. Labeling data can be a time-consuming and expensive process, and in some domains, there may be limited labeled data available for training a model. In such cases, it may be possible to leverage labeled data from other domains to improve the performance of a model in the target domain.

Another motivation for cross-domain text classification is transfer learning. By learning from multiple domains simultaneously, models can develop a more robust and generalizable understanding of the underlying structure of text data. This can lead to improved performance on the target domain, particularly when there is limited labeled data available.

Finally, cross-domain text classification can help to scale up the development of NLP applications. By reusing existing models and techniques across different domains, developers can save time and resources that would otherwise be spent on building and optimizing models from scratch for each domain. This can enable faster development and deployment of NLP applications for a wide range of use cases.

3. Objective

The objective of cross-domain NLP text classification is to build models that can accurately classify text data from different domains or sources, while also being able to generalize well to new and unseen data. The primary goal is to improve the performance of NLP models on text data from multiple domains, which can help to enable practical applications that require processing text data from diverse sources.

In addition to achieving high accuracy on the target domain, cross-domain text classification also aims to leverage knowledge learned from other domains to improve performance. This is achieved through transfer learning, where models are trained on multiple domains simultaneously to develop a more robust and generalizable understanding of the underlying structure of text data.

Another objective of cross-domain sentiment analysis is to reduce the amount of labeled data required to achieve high performance on a particular domain. By leveraging labeled data from other domains, models can be trained with a larger and more diverse set of data, which can help to reduce overfitting and improve generalization.

Overall, the objective of cross-domain NLP text classification is to build models that can handle text data from diverse sources and domains, while also being able to generalize well and achieve high accuracy on the target domain.

4. Related Works

4.1 Transfer Learning

In transfer learning, a pre-trained model is used as a starting point, typically trained on a large dataset from a source domain. This model has already learned general features and patterns that can be applied to a different but related target task or domain. The pre-trained model serves as a feature extractor, extracting meaningful representations from the input data.

Transfer learning has been successfully applied in various domains, including computer vision, natural language processing, and speech recognition. It has enabled breakthroughs in tasks such as image classification, object detection, sentiment analysis, machine translation, and many others.

In summary, accelerates training time, enhances generalization, and overcomes data scarcity, making it a valuable approach in machine learning and deep learning applications.

4.2 Cross-Domain Sentiment Analysis

It involves developing models and algorithms that can accurately determine the sentiment polarity (positive, negative, or neutral) of text regardless of the specific domain it belongs to.

Traditional sentiment analysis models are often trained on domain-specific datasets, which limits their applicability to new domains where labeled data may be scarce or unavailable. Cross-domain sentiment analysis addresses this challenge by developing techniques that allow sentiment analysis models to generalize across domains without the need for extensive domain-specific training.

Cross-domain text sentiment analysis has various applications. It can be used in social media monitoring to understand public sentiment towards brands, products, or events across different platforms and domains. It can assist in analyzing customer feedback from various sources, such as reviews, forums, and social media, to gain insights into customer satisfaction and sentiment trends. It is also valuable in market research, reputation management, and customer support, where understanding sentiment across diverse domains is crucial.

In conclusion, cross-domain text sentiment analysis plays a vital role in understanding sentiment across diverse domains. By employing techniques like domain adaptation, transfer learning, feature engineering, and ensemble methods, sentiment analysis models can effectively generalize sentiment analysis and provide valuable insights into sentiment trends and opinions across different subject areas.

5. Transfer Learning Method on Deep Learning

Deep learning models, such as convolutional neural networks (CNNs) for computer vision or recurrent neural networks (RNNs) for natural language processing, are typically used as the basis for transfer learning approaches. Here are some common transfer learning methods based on deep learning models:

I. **Pre-trained models**: Pre-trained models are deep learning models that have been trained on a large-scale dataset, such as BERT for word embeddings in NLP. These models have learned generic representations and capture rich features from the input data. They can be used as a starting point for a new task by either fine-tuning the entire model or using it as a fixed feature extractor.

II. **Fine-tuning**: Fine-tuning involves taking a pre-trained deep learning model and further training it on a task-specific dataset. The initial layers of the model, which capture low-level and generic features, are typically kept frozen, while the later layers are fine-tuned to adapt to the specific task. This allows the model to retain its learned representations while incorporating domain-specific knowledge.

III. **Feature extraction**: In this approach, the pre-trained deep learning model is used as a fixed feature extractor. The input data is passed through the model, and the activations of one or more intermediate layers are used as features for a new task-specific classifier. This method is especially useful when the target dataset is small and insufficient to train a deep model from scratch.

IV. **Domain adaptation**: Domain adaptation methods aim to adapt a deep learning model from a source domain to a target domain, where the two domains may have different data distributions. Techniques such as adversarial training, where the model learns to align the feature representations between the domains, can be employed to reduce the domain gap and improve performance on the target domain.

V. **Progressive neural networks**: Progressive neural networks involve training a deep learning model in a step-by-step manner, starting from a simpler task and gradually increasing the complexity. The model is trained on each task, and the knowledge learned from previous tasks is transferred and built upon for subsequent tasks. This approach allows the model to progressively learn more intricate representations.

These transfer learning methods based on deep learning models have significantly improved the efficiency and effectiveness of various machine

learning tasks, including image classification, object detection, sentiment analysis, machine translation, and many others. They enable models to leverage knowledge from large-scale datasets and overcome challenges such as limited labeled data and domain shift, leading to improved generalization and performance in different domains.

6. Problem Statement

The problem addressed by cross-domain NLP text classification is the difficulty of building models that can accurately classify text data from multiple domains or sources. This includes the challenge of achieving high performance on a target domain when there are few labeled data resources, and the problem of overfitting when models are trained on a limited data. Additionally, building models that can generalize well to new and unseen data is a significant challenge in NLP, particularly when working with text data from diverse sources. The problem statement for cross-domain NLP text classification is to develop models that can handle these challenges and achieve high accuracy on text data from multiple domains or sources.

7. Proposed Methodology

7.1 Dataset

For the experimental evaluation, two distinct datasets were chosen: the Amazon Product Review Corpus and the Movie Review Corpus. The Amazon Review Corpus encompasses a wide range of Amazon product reviews, consisting of an equal number of positive and negative texts, with 982 instances in each sentiment category. Refer to Table 1 for a detailed breakdown of the statistics related to this corpus.

Meanwhile, the Movie Review Corpus comprises reviews from various movie genres. Although specific figures for the positive and negative texts in this corpus are 982 instances in each sentiment category, it was utilized in the sentiment analysis experiment. See Table 2 for an overview of the dataset statistics.

These datasets were utilized to evaluate the effectiveness of the model in sentiment classification across different domains. The models were trained and tested on these corpora to assess their performance in accurately categorizing sentiments.

Google Drive Link for datasets:
https://drive.google.com/drive/folders/1M7RrWrl3sK66WFzdCKH7im54fcQsGGyt?usp=share_link

Table 1. Amazon Product Review

Domain	Positive Sample Number	Negative Sample Number
Amazon Review	982	982

Table 2. Movie Review

Domain	Positive Sample Number	Negative Sample Number
Movie Review	982	982

7.2 Data Preprocessing

- In this step, we preprocess the data using regex function we remove everything with other than alphabets.
- Then we convert all the words in the corpus to lowercase. So, thar same word with uppercase and lowercase can't have the different vector representation.
- Then we remove the various "STOPWORDS" found in "English" corpus such as such as "the", "a", "an", "in".

7.3 Converting words to numbers using BERT Encoder

- The words are being converted to numbers using one hot encoding.

7.4 Padding Sequences

- First, we define the sentence length
- Then we add padding to the maximum length of the sentence.

7.5 Train Test Split

- Splitting the data for training and testing with test_size = 0.33 and random state = 42.

7.6 Creating Model

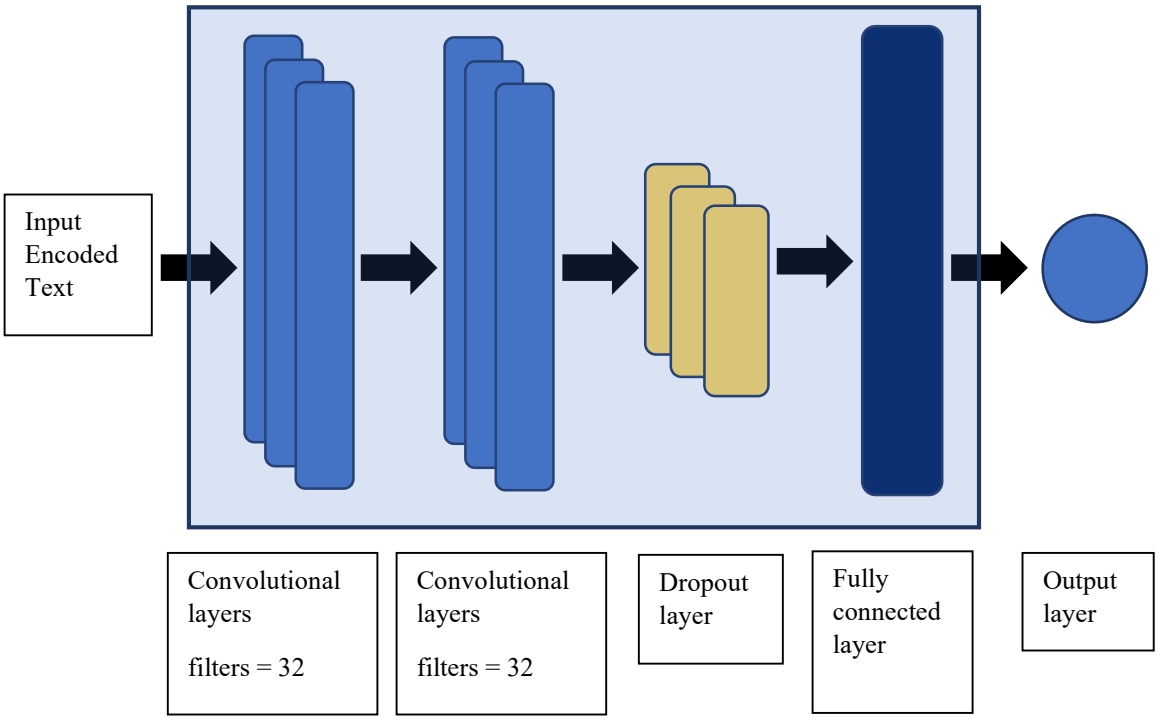
- First of all, defining model to sequential.
- Adding Embedding layer to the model.
- Adding LSTM layer.
- Adding Dense layer, activation = "sigmoid".
- Compile model.
- Print model summary.

7.7 Fitting the model

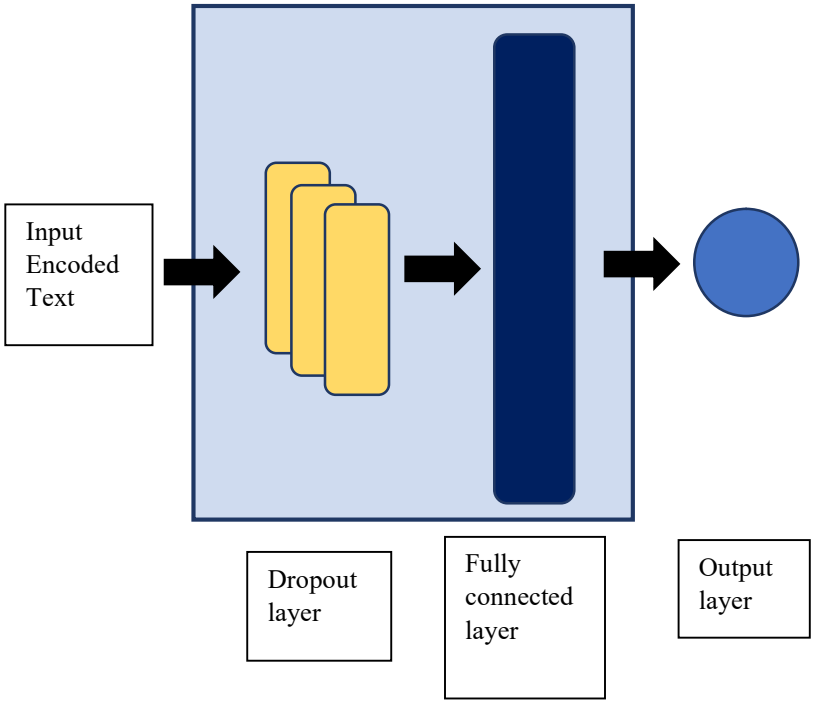
- Fitting the model with training and validation data with epochs = 20 and batch_size = 64.

7.8 Figures

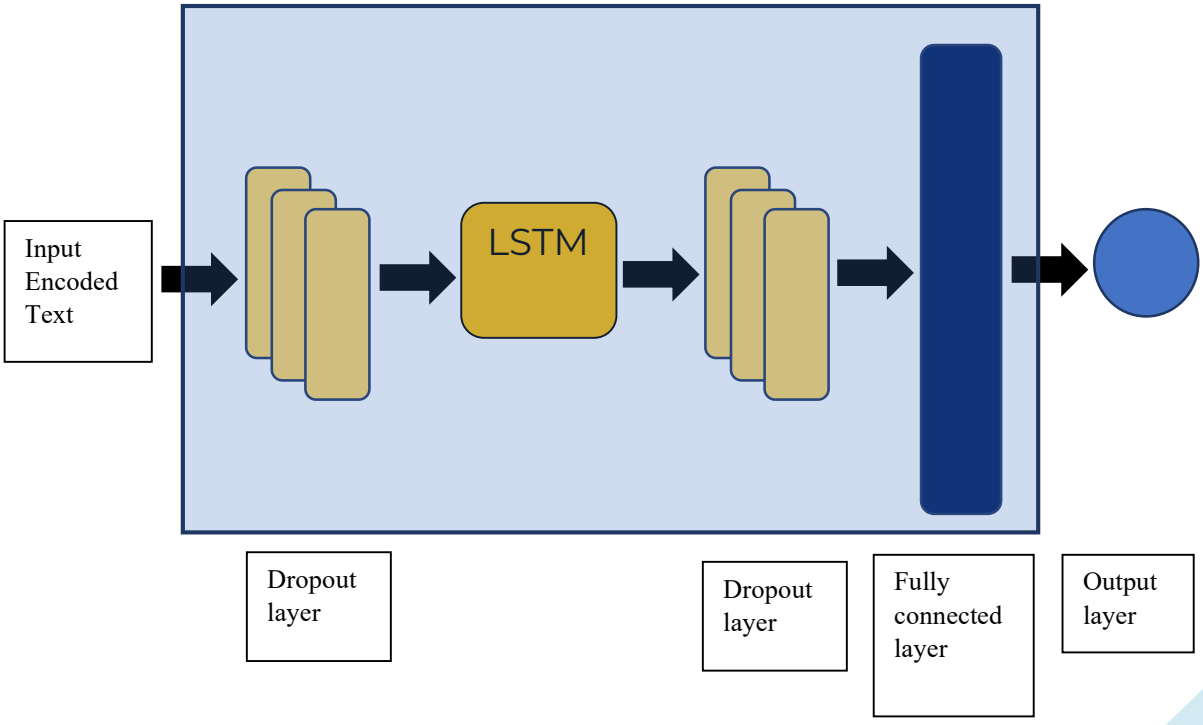
7.8.1 Cross-Domain BERT + CNN



7.8.2 Bert + NN



7.8.3 One Hot Encoding + LSTM



8. Results and discussion

8.1 Datasets

For the experimental evaluation, two distinct datasets were chosen: the Amazon Product Review Corpus and the Movie Review Corpus. The Amazon Review Corpus encompasses a wide range of Amazon product reviews, consisting of an equal number of positive and negative texts, with 982 instances in each sentiment category. Refer to Table 1 for a detailed breakdown of the statistics related to this corpus.

Meanwhile, the Movie Review Corpus comprises reviews from various movie genres. Although specific figures for the positive and negative texts in this corpus are 982 instances in each sentiment category, it was utilized in the sentiment analysis experiment. See Table 2 for an overview of the dataset statistics.

Table 1. Amazon Product Review

Domain	Positive Sample Number	Negative Sample Number
Amazon Review	982	982

Table 2. Movie Product Review

Domain	Positive Sample Number	Negative Sample Number
Movie Review	982	982

8.2 Parameter Settings

8.2.1 Parameter setting for Cross-Domain BERT + CNN

Parameter Name	Parameter Value
Embedding layer	BERT Encoder + BERT Preprocess
Convolution Layer	2

Dropout	0.1
Dense Layer	3
Activation	Sigmoid
Metrics	accuracy, precision, recall
Optimizer	Adam
Loss	Binary_crossentropy
Epochs	20

8.2.2 Parameter setting for BERT + NN

Parameter Name	Parameter Value
Embedding layer	BERT Encoder + BERT Preprocess
Dropout	0.1
Dense Layer	
Activation	Sigmoid
Metrics	accuracy, precision, recall
Optimizer	Adam
Loss	Binary_crossentropy
Epochs	20

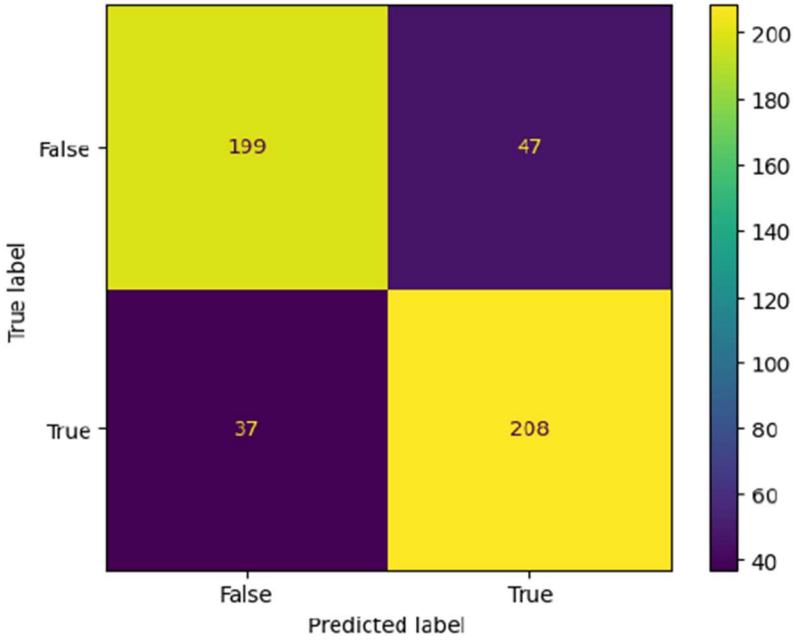
8.2.3 Parameter setting for One Hot Encoding + LSTM

Parameter Name	Parameter Value
Sentence length	20
Embedding vector features	40
Vocabulary size	5000
LSTM layers	100
Dense layer	1
Loss	Binary crossentropy
Optimizer	Adam
Metrics	Accuracy
Test_size	0.33
Random_state	42
Batch_size	64
Epochs	20

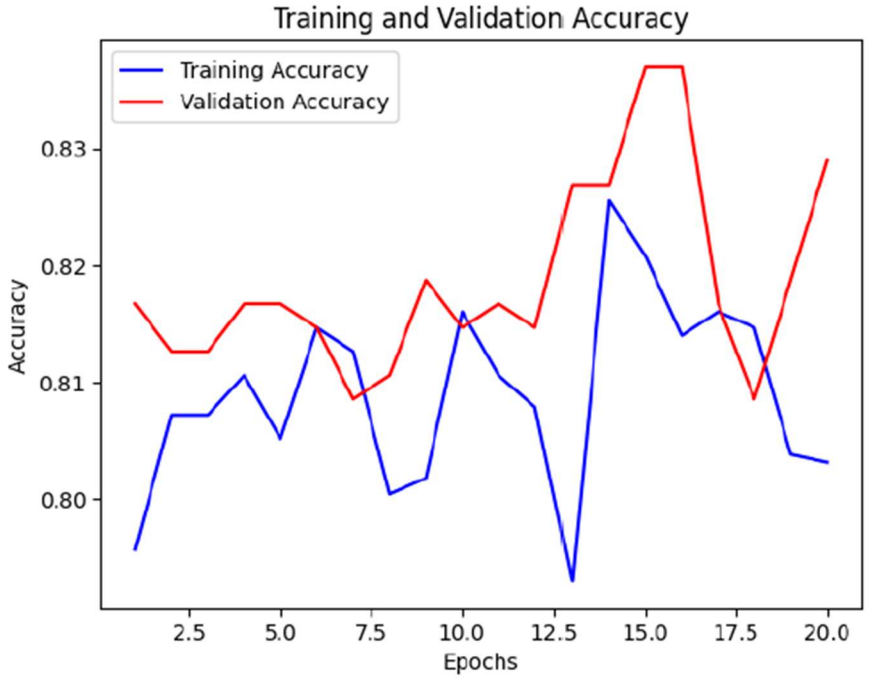
8.3 Experimental Results and Analysis

8.3.1 Cross-Domain BERT + CNN

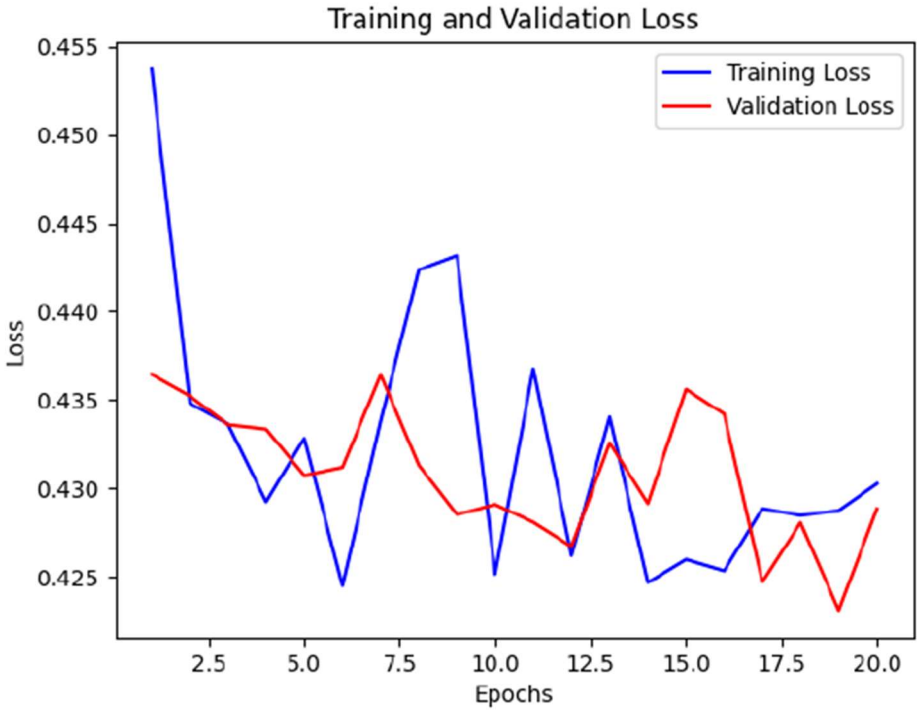
I. Confusion matrix



II. Training and Validation Accuracy



III. Training and Validation Loss

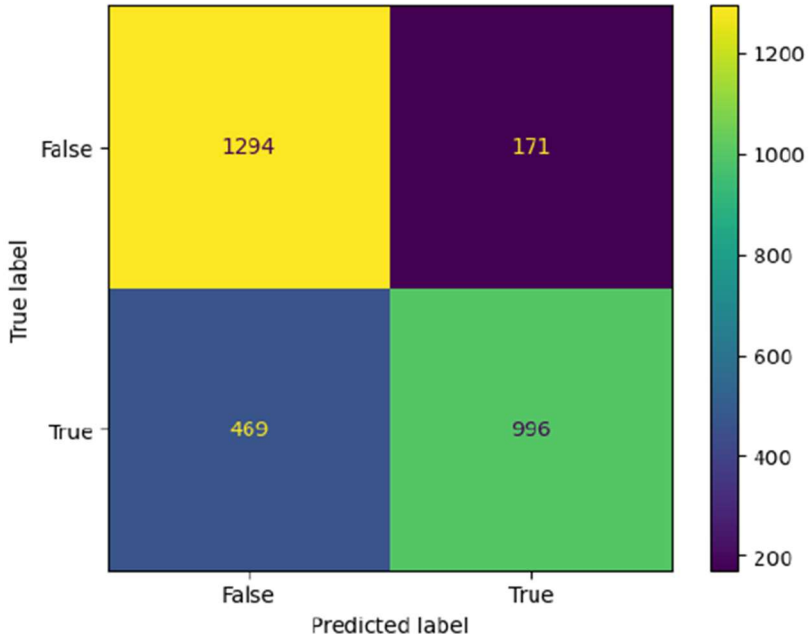


IV. Classification Report

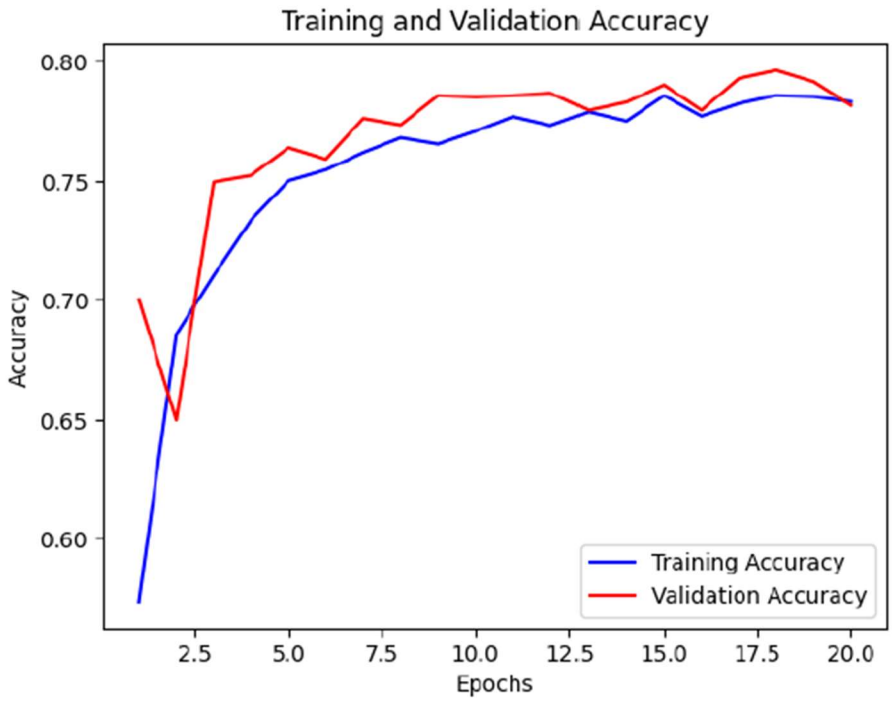
	Precision	Recall	F1-score	Support
0	0.72	0.89	0.80	246
1	0.85	0.66	0.74	245
Accuracy			0.77	491
Macro avg	0.79	0.77	0.77	491
Weighted avg	0.79	0.77	0.77	491

8.3.2 BERT + NN

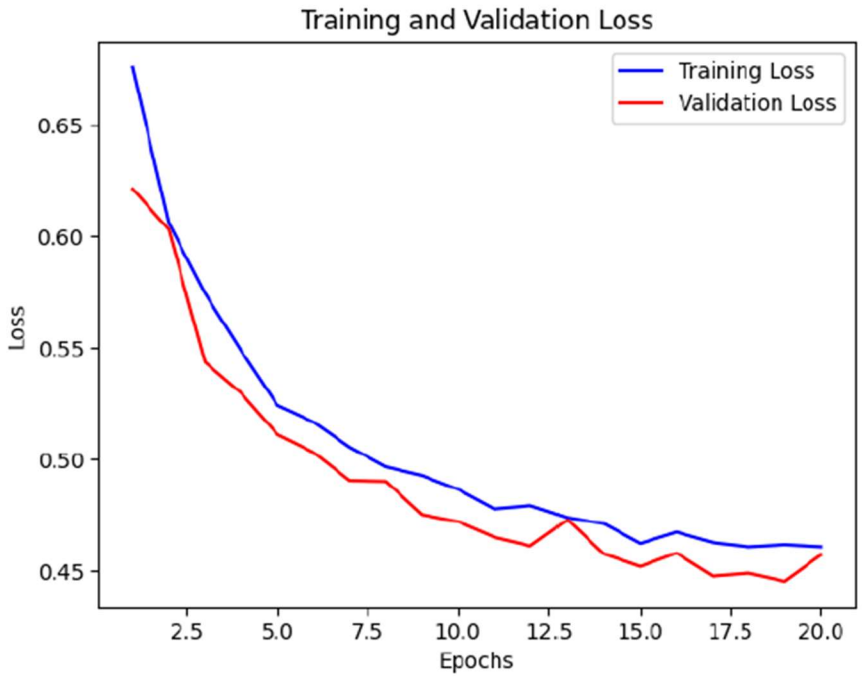
I. Confusion Matrix



II. Training and Validation Accuracy



III. Training and Validation Loss

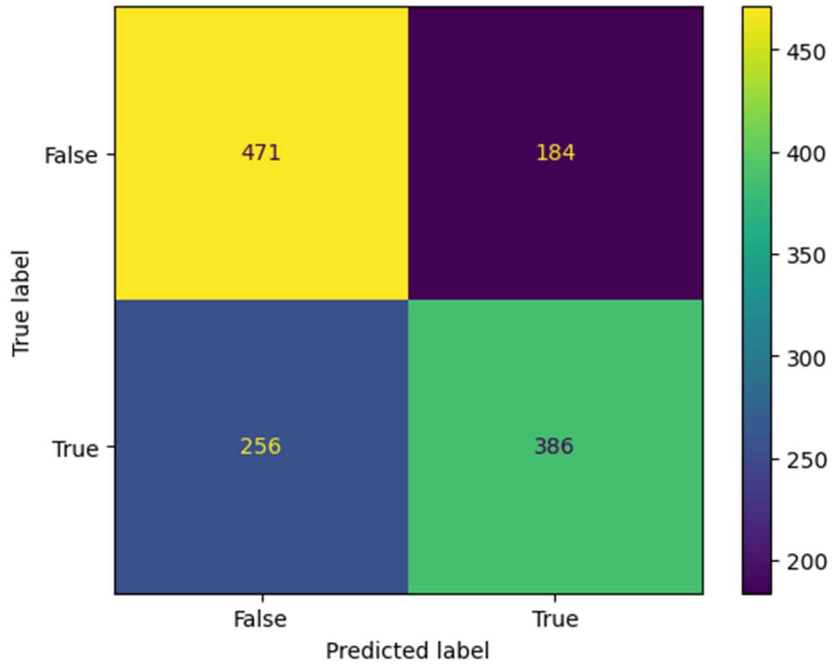


IV. Classification Report

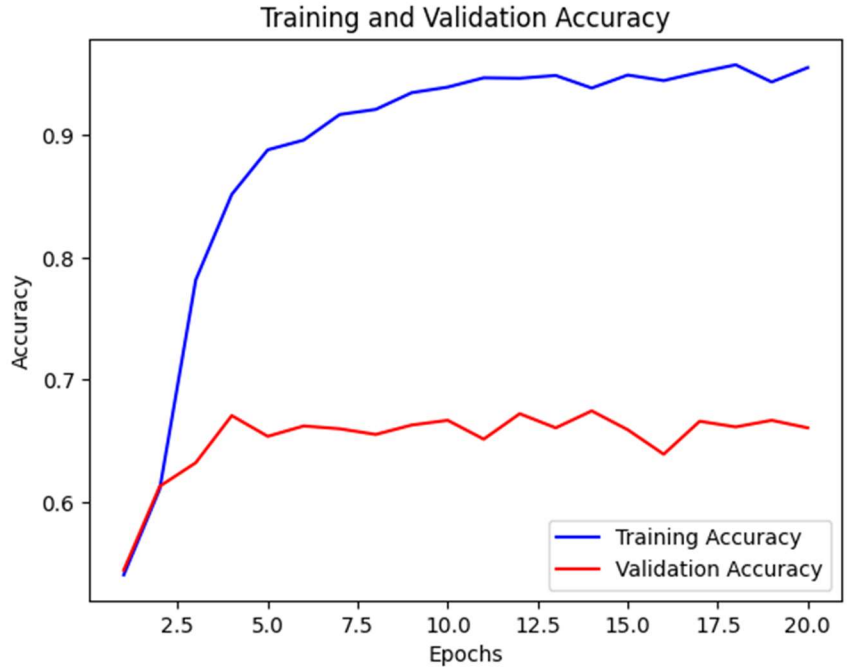
	Precision	Recall	F1-score	Support
0	0.79	0.83	0.81	1465
1	0.82	0.77	0.80	1465
Accuracy			0.80	2930
Macro avg	0.80	0.80	0.80	2930
Weighted avg	0.80	0.80	0.80	2930

8.3.3 One Hot Encoding + LSTM

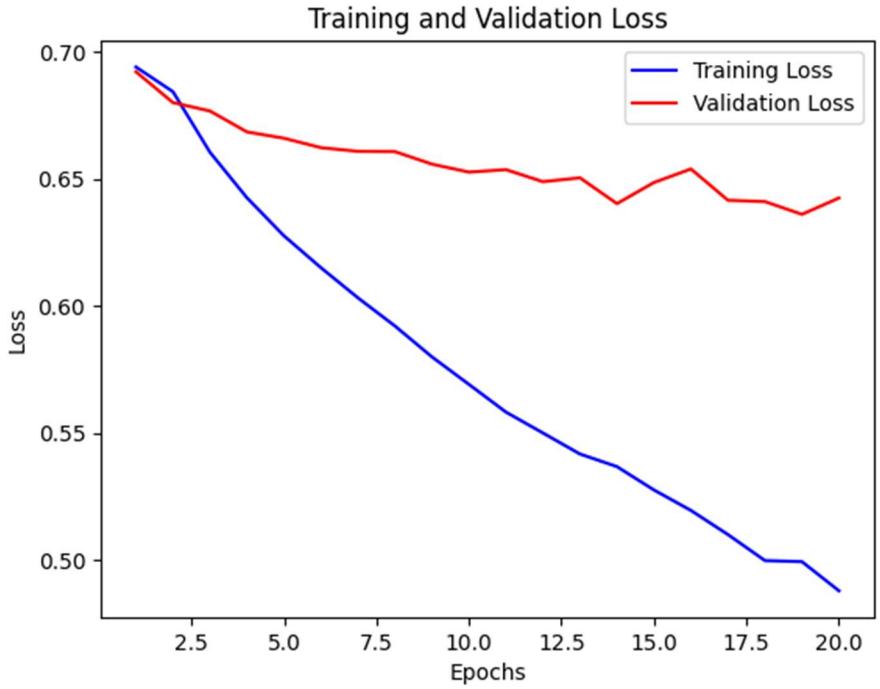
I. Confusion Matrix



II. Training and Validation Accuracy



III. Training and Validation Loss



IV. Classification Report

	Precision	Recall	F1-score	Support
0	0.69	0.60	0.64	655
1	0.64	0.73	0.68	642
Accuracy			0.66	1297
Macro avg	0.67	0.67	0.66	1297
Weighted avg	0.67	0.66	0.66	1297

9. Conclusion

In this experiment, we dive into cross-domain sentiment analysis and train model using transfer learning technique on the convolutional neural network (CNN) model. Our method aims to tackle the challenge of limited labeled data. Specifically, we train a neural network (NN) and long short-term network (LSTM) model to demonstrate its efficiency on two widely used benchmark datasets, namely Amazon review and Movie review.

Through the series of experiments, we show that how the proposed model cross-domain CNN + BERT is better over baseline approaches. Notably our approach overcomes the challenge of limited labeled data, resulting in significant improvement. By harnessing the power of transfer learning method, we tap into the knowledge gained from the source domain and apply it to enhance sentiment analysis in the target domain.

References

1. Jiana Meng, Yingchun Long, Yuhai Yu , Dandan Zhao and Shuang Liu, Cross-Domain Text Sentiment Analysis Based on CNN_FT Method, 2019

2. Nancy Kansal, Lipika Goel, Sonam Gupta, Ajay Kumar Garg Engineering College Ghaziabad, India, A Literature Review on Cross Domain Sentiment Analysis Using Machine learning, 2020

3. Ning Liu and Jianhua Zhao, A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text, 2022

Report Ver 5

ORIGINALITY REPORT

12%	10%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nitp.ac.in Internet Source	8%
2	Humayra Shoshi, Indranil SenGupta. "Hedging and machine learning driven crude oil data analysis using a refined Barndorff–Nielsen and Shephard model", International Journal of Financial Engineering, 2021 Publication	1%
3	www.ijraset.com Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	simran Garg, Devang Chaturvedi, Tanya Jain, Anju Mishra, Anjali Kapoor. "Sentiment Analysis of Twitter Data using Machine Learning: A Case Study of SVM Algorithm", Research Square Platform LLC, 2023 Publication	1%
6	nitp.irins.org Internet Source	1%

7 Jiana Meng, Yingchun Long, Yuhai Yu, Dandan Zhao, Shuang Liu. "Cross-Domain Text Sentiment Analysis Based on CNN_FT Method", Information, 2019 — 1%
Publication

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		