Cross-Domain Fake Review Detection using Deep Learning: An Adaptive Approach

Submitted in partial fulfilment of the requirements of the degree of

Bachelor of Technology (B.Tech) by

Suchith Reddy Billa (197220)

Raja Reddy Pundra (197167)

Shivaji Banothu (197275)

Supervisor:

Dr. Ramakrishnudu Tene

Associate Professor

NIT Warangal

Department of Computer Science and Engineering

NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL

2022 - 2023

APPROVAL SHEET

This Dissertation Work entitled **"Cross-Domain Fake Review Detection using Deep Learning: An Adaptive Approach"** by **Suchith Reddy Billa (197220)**, **Raja Reddy Pundra (197167)** and **Shivaji Banothu (197275)** is approved for

the degree of Bachelor of Technology (B.Tech).

Examiners

_____

_____

_____

Supervisor

_____

Dr. Ramakrishnudu Tene(Supervisor)

_____

Chairman

_____

Prof. Dr. S. Ravi Chandra(HOD)

Date : _____

Place : _____

DECLARATION

We declare that this written submission represents our ideas, and our supervisor's ideas in our own words and where other's ideas or words have been included. We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Suchith Reddy Billa - 197220

Raja Reddy Pundra - 197167

Shivaji Banothu - 197275

Date:

CERTIFICATE

This is to certify that the Dissertation work entitled **"Cross-Domain Fake Review Detection using Deep Learning: An Adaptive Approach"** is a bonafide record of work carried out by **Suchith Reddy Billa (197220)**, **Raja Reddy Pundra (197167)** and **Shivaji Banothu (197275)** submitted to the **Dr. Ramakrishnudu Tene** of "Department of Computer Science and Engineering", in partial fulfilment of the requirements for the award of the degree of B.Tech at "National Institute of Technology, Warangal" during the 2022-2023.

Prof. S.Ravi Chandra
Head of the Department
Department of CSE
NIT Warangal

Dr. Ramakrishnudu Tene
Associate Professor
Department of CSE
NIT Warangal

ACKNOWLEDGEMENT

We thank our beloved faculty supervisor **Dr. Ramakrishnudu Tene, Associate Professor, Department of Computer Science and Engineering(CSE), National Institute of Technology, Warangal**, for his persistent supervision, guidance, suggestions, and encouragement during this project. He has motivated us during the low times and given us the courage to move ahead positively.

We are grateful to **Prof. S.Ravi Chandra, Head of the Department, Computer Science and Engineering, National Institute of Technology, Warangal** for his moral support to carry out this project. We express our gratitude to the Project Evaluation Committee for their diligent efforts in assessing our project.

Suchith Reddy Billa (197220)

Raja Reddy Pundra (197167)

Shivaji Banothu (197275)

ABSTRACT

The rise in customer reviews on the Web has created a need for effective methods to retrieve valuable information hidden in these reviews. However, with the spread of opinion spam, fake review detection has become a pressing issue as they can change purchasing decisions of consumer ,From the past ten years many Machine learning and Deep learning techniques have been extensively explored. While existing literature has focused on fake review detection in one domain, this work aims to perform cross-domain fake review detection. To accomplish this, a pre-training language model BERT is used for domain adaptation. However, When transferring knowledge, it lacks domain awareness and cannot tell the differences between the traits of the source and target domains. To address this issue, we propose BERT-based domain adaptation neural network(BERT-DANN) for cross domain fake review detection.Our experiments using the Amazon real-world reviews dataset show that accuracy is significantly improved over the baseline model.

Keywords: cross-domain, BERT, Fake review, domain adaptation

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview of fake reviews:

The increasing number of customer reviews on various websites has become a valuable source of information for both businesses and potential buyers[11]. Positive reviews can bring profits, while unfavorable reviews can negatively impact sales. This has led to a trend of relying on customer feedback to reshape businesses and improve their products and services. However, the rise of social media has also led to the spread of fake reviews, where companies unfairly promote their own products or damage their competitors reputations. Fake reviews can be difficult to distinguish from genuine ones and can exert a substantial influence on customers purchasing decisions. In this context, it is crucial to develop effective methods for fake review detection and preserving the integrity of online reviews. Fake reviews, also known as spam

opinions, can be classified into three types:

- Dishonest opinions, which aim to damage or promote a product/business through negative or positive reviews, respectively. These reviews are difficult to distinguish from real reviews.
- Reviews of a brand only, which refer to comments solely on the products developed by brand.
- Non-reviews, which are irrelevant comments or advertisements that offer no genuine opinion.

1.2 Importance of cross domain fake review detection:

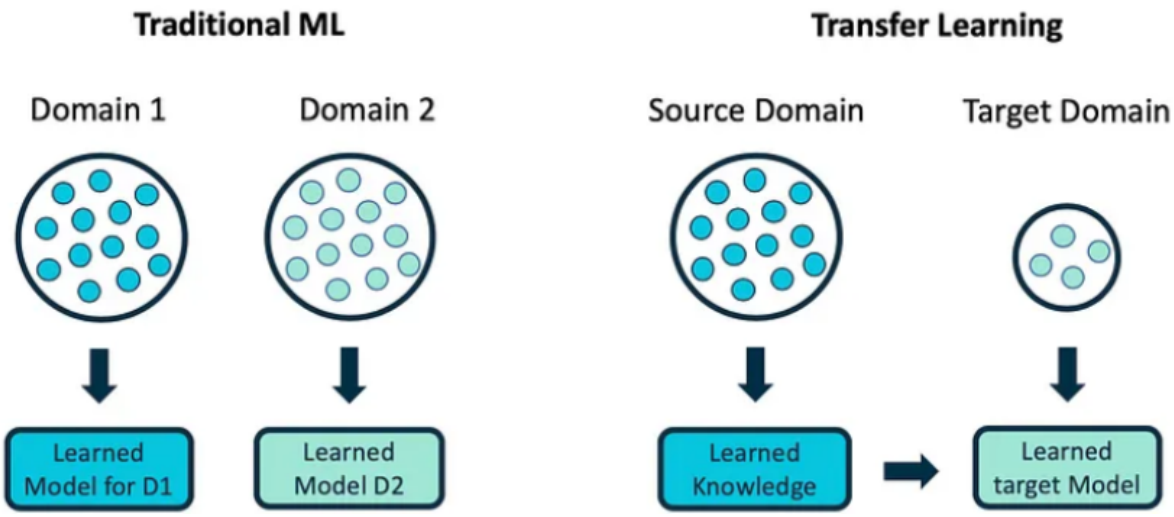
Detection of Cross-domain fake reviews is crucial for detecting and combating fake reviews across multiple domains and industries. Since fake reviews can appear in various areas, it is necessary to develop a model that can identify fake reviews across different platforms.One key advantage of cross-domain fake review detection is its scalability. By developing a model that can detect fake reviews across multiple domains, it becomes more efficient and cost-effective to combat fake reviews on a larger scale.Moreover, The effectiveness of a model trained within a specific domain may be limited when utilized in different domains, as the language and review patterns used may differ significantly. Cross-domain fake review detection addresses this issue by identifying common characteristics of fake reviews that transcend domain-specific features.

1.3 Background

1.3.1 TRANSFER LEARNING

It is used in machine learning, It uses the pre-trained model to solve a new problem. In transfer learning, a machine applies the knowledge from a previous task to generalise the new problem. For example, The Knowledge a classifier gained during predict whether an image is contained food or not is also could be used while training to predict drinks.

.



Traditionally, transfer learning issues were categorised according to how closely related the domains were to one another and whether or not labelled and unlabeled data were available.

Inductive transfer learning After learning a distinct yet interconnected concept or skill from a prior source task, the learning mechanism can enhance efficiency on the present or target work.

Transductive Transfer Learning In the situations where the domains of the target and source tasks are similar, but not identical, the transductive transfer learning approach is used.

In these situations, there is typically a lot of data that has been categorized or annotated in the source domain and less unlabeled data in the destination domain.

Unsupervised Transfer Learning It is similar to above inductive transfer learning. The distinction is that for both source and target tasks, the algorithms prioritise unsupervised jobs.We are therefore discussing the most typical scenario in which labelled data is not accessible for both the original domain and the intended domain of application..

Adaptive approach The ability of a model or algorithm which can adjust its parameters based on the arrival of new data. So that it can improve its accuracy and make more precise predictions.

Domain invariant features Features that are common across different domains.These features help to predict more accurately across different domains.

Domain classifier A model which predicts the domain of the new data or target data. It is trained in such a way that it uses domain-specific features of each domain to predict the domain of new input data.

Backpropagation A technique that calculates the gradients of the loss function in relation to the parameters and biases of the network. , propagates backwards to update the parameters in the training phase,so that the model tries to predict more accurately for test data.

1.3.2 Domain Adaptation

It is a unique instance of transfer learning. Domain adaptation is a notion that bridges the Source data distribution gap with Target data. It is the capacity to transfer a learned algorithm to a new target domain from one or more source domains. It falls under the umbrella of transfer learning. While the feature space remains consistent, the distributions of data in the source and

target domains exhibit dissimilarities.

1.3.3 General DANN architecture

A deep learning method for extracting features and a deep learning classifier for classification of labels make up the general architecture, which is a standard feedforward neural network.By incorporating a classifier for the domain, which is connected to the feature extractor through a gradient reversal layer, the gradients are multiplied by a negative scalar to enhance the training process, as illustrated in the below picture, allows for unsupervised domain adaption. The training is carried out normally in all other respects, with the loss calculated from prediction of labels(for source instances) and the classification of the domains(for all samples) being minimized. In order to produce domain-independent features,The Gradient Reversal Layer (GRL) aims to minimize the discrepancy between the feature distributions of the two domains (Indiscernible to the domain classifier).

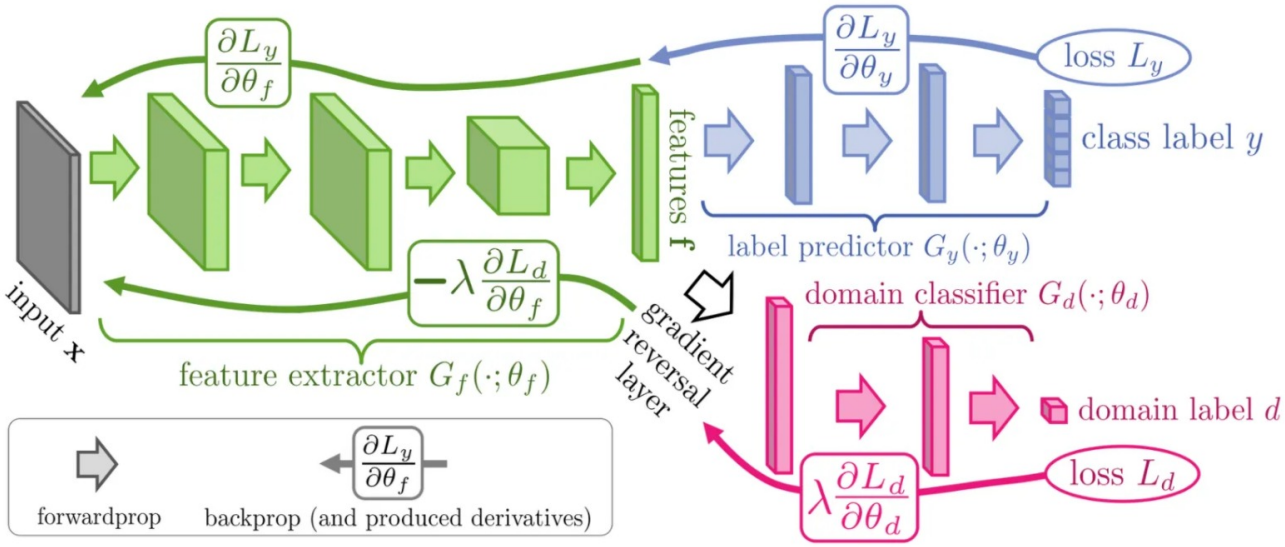


Figure 1.1: General DANN architecture [1]

let $F_f(.,\theta_f)$ be the feature extractor which is a feed forward neural network of D-dimension with parameters θ_f , $F_y(.,\theta_y)$ be the label predictor of DANN with θ_y as parameters and $F_d(.,\theta_d)$ be the domain predictor with parameters θ_d

$$Prediction\ loss\ :\ Ly^i(\theta_f,\theta_y)\ =\ Ly(F_y(F_f(xi;\theta_f);\theta_y),y_i),$$
$$Domain\ loss\ :\ Ld^i(\theta f,\theta_d)\ =\ Ld(F_d(F_f(xi;\theta_f);\theta_d),d_i)$$

Training DANN Training the DANN consists in optimizing

$$E(\theta_f,\theta_y,\theta_d) = \frac{1}{n}\sum_{i=1}^{n} Ly^i(\theta_f,\theta_y) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} Ld^i(\theta f,\theta_d) + \frac{1}{n'}\sum_{i=n+1}^{N} L_d^i(\theta_f,\theta_d)\right)$$

by finding the optimum point $\hat{\theta}_f,\hat{\theta}_y,\hat{\theta}_d$ such that

$$(\hat{\theta}f,\hat{\theta}_y)\ =\ \arg\min_{\hat{\theta}_f,\hat{\theta}_y} E(\theta f,\theta_y,\hat{\theta}_y) \tag{1.1}$$

$$\hat{\theta}d\ =\ \arg\max_{\theta d} E(\hat{\theta}_f,\hat{\theta}_y,\theta_d) \tag{1.2}$$

The saddle points in the equation (1.1 - 1.2) can be obtained by following gradient updates

$$\theta_f = \longleftarrow\ \theta_f\ - \mu\left(\frac{\partial Ly^i}{\partial\theta f} - \lambda\frac{\partial Ly^i}{\partial\theta f}\right) \tag{1.3}$$

$$\theta_y\ \longleftarrow\ \ \theta_y - \mu\frac{\partial Ly^i}{\partial\theta y} \tag{1.4}$$

$$\theta_d\ \longleftarrow\ \ \theta_d - \mu\lambda\frac{\partial Ld^i}{\partial\theta d} \tag{1.5}$$

$\mu \longrightarrow\ \ learning\ rate$

Gradient Reversal Layer

We can simplify the model by using a gradient reversal layer (GRL), which is a layer that does not have any associated parameters. At the time of forward pass, the GRL behaves like an identity function, leaving the inputs unchanged. However, during the backward pass, the GRL reverses the sign of the gradients from the subsequent layer before passing them to the

preceding layer. In other words, the gradients are multiplied by -1. This contributes to achieve the desired reduction in complexity of the model.

Mathematical pseudo function of GRL $(R(x))$

$$\mathbf{R}(\mathrm{x}) = \mathrm{x} \tag{1.6}$$

$$\frac{d\mathbf{R}}{d\mathrm{x}} = -\mathbf{I} \tag{1.7}$$

The final equation for obtaining saddle points $(\hat{\theta}f, \hat{\theta}y, \hat{\theta}d)$ by running the stochastic gradient

$$\begin{aligned} \tilde{E} = & \frac{1}{n}\sum_{i=1}^{n} Ly(F_y(F_f(xi;\theta_f);\theta_y), y_i) \\ & - \lambda \left(\frac{1}{n}\sum_{i=1}^{n} Ld(F_d(\mathbf{R}(F_f(xi;\theta_f));\theta_d), d_i) \right. \\ & \left. + \frac{1}{n'}\sum_{i=n+1}^{N} Ld(F_d(\mathbf{R}(F_f(xi;\theta_f));\theta_d), d_i). \right. \end{aligned} \tag{1.8}$$

We can implement the updates $(1.3 - 1.5)$ using stochastic gradient descent (SGD) for equation (1.8). This process results in the development of attributes that are both domain-independent and distinctive. After training, the label predictor $F_y(F_f(x;\theta_f);\theta_y)$ can be used to classification of the labels from the domain samples of target, as well as the source domain.

1.3.4 Models

Bidirectional Encoder Representations from Transformers (BERT)

BERT is a nlp model developed by Google, which has been trained to perform various language-related tasks. It employs a transformer-based architecture and can learn general

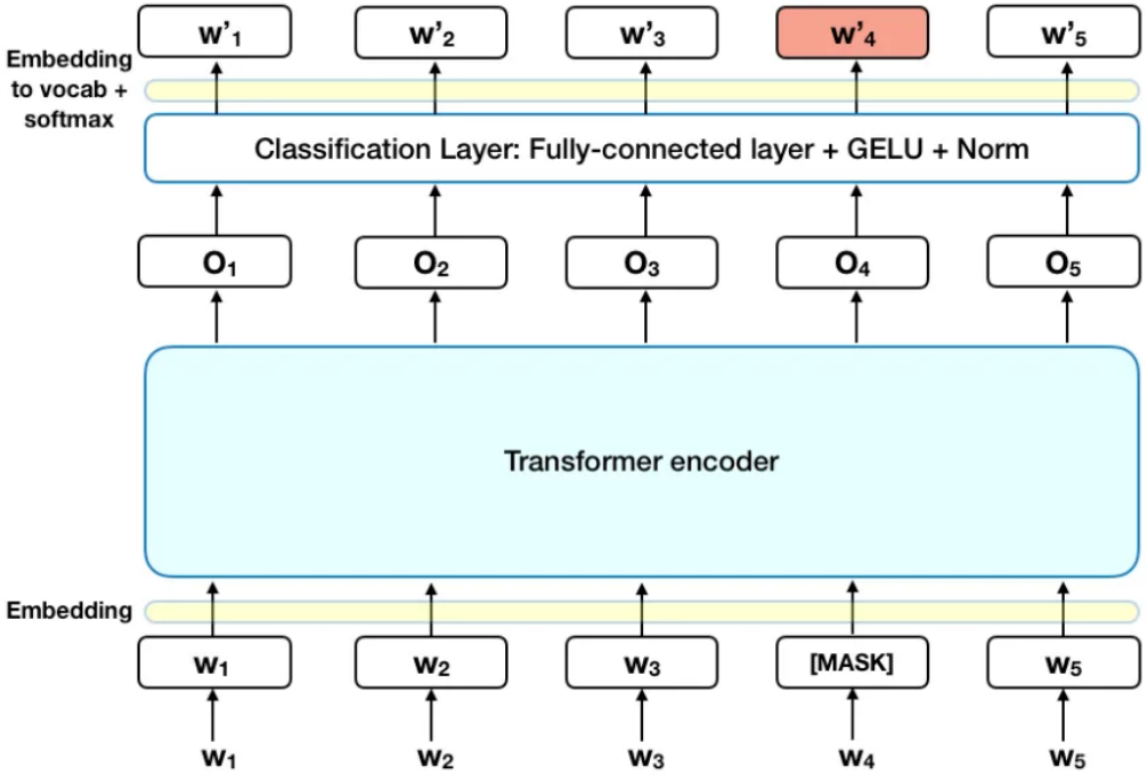


Figure 1.2: Transformer's Encoder-Decoder Model

language representations through bidirectional pre-training on large amounts of unlabeled text data. BERT's superior performance on multiple NLP tasks has made it a popular choice for fine-tuning with limited labeled data. This has led to increased research and development of transformer-based models in NLP, leveraging the successes of BERT.

Chapter 2

Review of Literature

2.1 Fake Review Detection using Machine Learning

In past few years, many machine learning and deep learning models have been explored to suppress the issue of spam detection in various domains like sms, emails, blogs and news etc. Different machine learning approaches using supervised, unsupervised and semi supervised learning techniques have been explored in recent.In [24] fake review detection is performed by feature extraction techniques like Tf-idf vectorizer, Ngram model and count vectorizer these extracted features are tested on classification techniques like Naive Bayes, Random Forest, Logistic Regression and Support Vector Machines. But due to scarcity of labelled dataset unsupervised machine learning in Fake review detection have been explored in recent years. Lau et al[25] developed an unsupervised machine learning model using semantic language

modelling. The performed experiment on amazon dataset contains 54,618 reviews using only 6% labelled information as fake. The proposed model achieved 0.9987 acc score and outperformed SVM

2.2 Fake Review Detection using Deep Learning

Neural network models on natural language preprocessing task provide great results compared to machine learning techniques. Deep learning methods capture semantic meaning using different embedding techniques(like word embedding). The features are selected using Tf-IDF, N-gram and word2vec and used on neural networks like multilayer perceptron, CNN and LSTM etc. In [18] different deeplearning classifiers are used for the detection of deceptive reviews and compared its performance are used for the detection of deceptive reviews compared its performance with machine learning modes.Li et al [26] developed a CNN based neural network model for detecting fraudulent or deceptive reviews. The developed model uses document representation techniques for classification they used word vector which is given as input at the time of training and testing.in [16] neural networks consist of a combination of CNN with BiGRU with an attention mechanism. The propsed model captures complex semantoc information and develop a document representation for it.

2.3 Cross Domain Fake Review Detection

The lock of labelled dataset is a processing issue in cross domain fake review detection many existing literature mainly focussed on fakre review detection in one domain bit cross domain problem need to be adressed effectively in [26] the model effectively addressed cross domain issue. the CNN based model using sentence weight to represent document review showed good accuracy in cross domain. Tang et al[27] used generative adversarial network model for cross

domain fake review detection. The Generative adversarial network architecture consist of 6 layer. The first 3 layers are used to get features which are accessible and normalised features. The model is tested on Yelp CHI dataset and achieved an accuracy rate of 83% , 75% in the respective hotel domain and the restaurant domain.

2.4 Adversarial Domain Adaptation

in [14] to extract to domain-independent features they have used Central Moment Discrepancy Measure for measuring the probability distribution of 2 random variables. To extract domain-independent attributes from the source data and domain-related attributes from the target domain data. CMD based discrepancy measure is used. The model is trained using labeled data from both the source domain and the target domain for sentiment classification across different domains. With labeled data from the source domain and unlabelled data from the target domain co-training is performed. the result showed a considerable improvement in accuracy. in [3] due to lack of availability of labelled data they performed cross domain crisisis data classification. The proposed Bert based Adversarial domain Adaptation model. The model extracted domain invariant information from source data using adversarial training and transferred the knowledge to target data . The domain invariant information is captured using a domain classifier which incorporated gradient reversal layer with it. The model performed the experiment using unsupervised learning

Chapter 3

Problem Statement

Online reviews play a crucial role in influencing customer decision-making processes. However, with the increased popularity of online reviews, fake reviews have also become prevalent, which can mislead customers and harm businesses. Cross-domain fake review detection is a demanding undertaking that necessitates the development of robust models that can detect fake reviews across different domains. Additionally, domain adaptation is also necessary to enhance the model's performance on new domains.

In the task of cross-domain fake review detection, let X is the input and $Y = \{0, 1\}$ are 2 possible labels. The problem at hand is involved with 2 domains D_S *source domain* and D_T *target domain*. In unsupervised domain adaptation, the learning process involves both label source samples S and unlabelled target sample T.

$$S = \{(x_i, y_i)\}_{i=1}^{N_s^l} \sim (D_s)^{N_s^l} \quad ; \quad T = \{x_i\}_{i=N_s^l+1}^{N_s^l+N_t} \sim (D_T)^{N_t} \tag{3.1}$$

N_s^l : total number of labeled source samples.

N_t : total number of unlabeled target samples.

Cross domain Fake review detection demands us to model a robustic classifier(ϕ) trained with labelled S(source) domain data to predict the reviews of T(target) domain as fake or real, with low target risk (R)

$$R_{D_T}(\phi) = Pr_{(x,y)\sim D_t}(\phi(x) \neq y) \tag{3.2}$$

R_{DT} : expected prediction error of target data

ϕ : classifier : X $\rightarrow Y$

3.1 Objective:

The main objective of this project is to develop a robust fake review detection model using Adversarial and Domain-Aware BERT for cross-domain review datasets. The specific objectives are as follows:

- To perform a comprehensive analysis of existing fake review detection methods and identify their limitations in cross-domain settings.
- To develop a BERT-based Domain Adaptation neural network model that can detect fake reviews across different domains effectively.
- To investigate the efficiency of BERT-DANN model in improving the performance on new domains.

Chapter 4

Methodology

The aim of this method is to determine whether reviews in a target dataset (T) are real or fake, by leveraging knowledge obtained from a source dataset (S) of reviews. The approach involves extracting features that are consistent across both domains using a feature extractor, while a domain classifier helps to differentiate between the two domains during training. The block diagram of the proposed method is depicted in below fig. The objective is to simultaneously optimize the label classifier by minimizing its loss and optimize the domain classifier by maximizing its loss (i.e to generate invariant features) to obtain optimized parameters for accurately classifying target data. The process involves tokenizing the reviews, obtaining BERT-based embeddings, extracting features with the feature extractor, and passing them to the classifiers. The model parameters are then optimized using stochastic gradient descent.

4.1 Proposed Architecture

In the cross-domain fake review detection model, we incorporate domain adaptation into the learning process, so that the decisions of classification results are based on attributes that are both distinctive and independent to domain changes.To achieve this, the model learns features that are both distinctive and domain-independent, using two classifiers: a label predictor that predicts class labels, and a domain classifier that distinguishes between the source and target domains at the time of training.The complete architecture for above classification is shown below

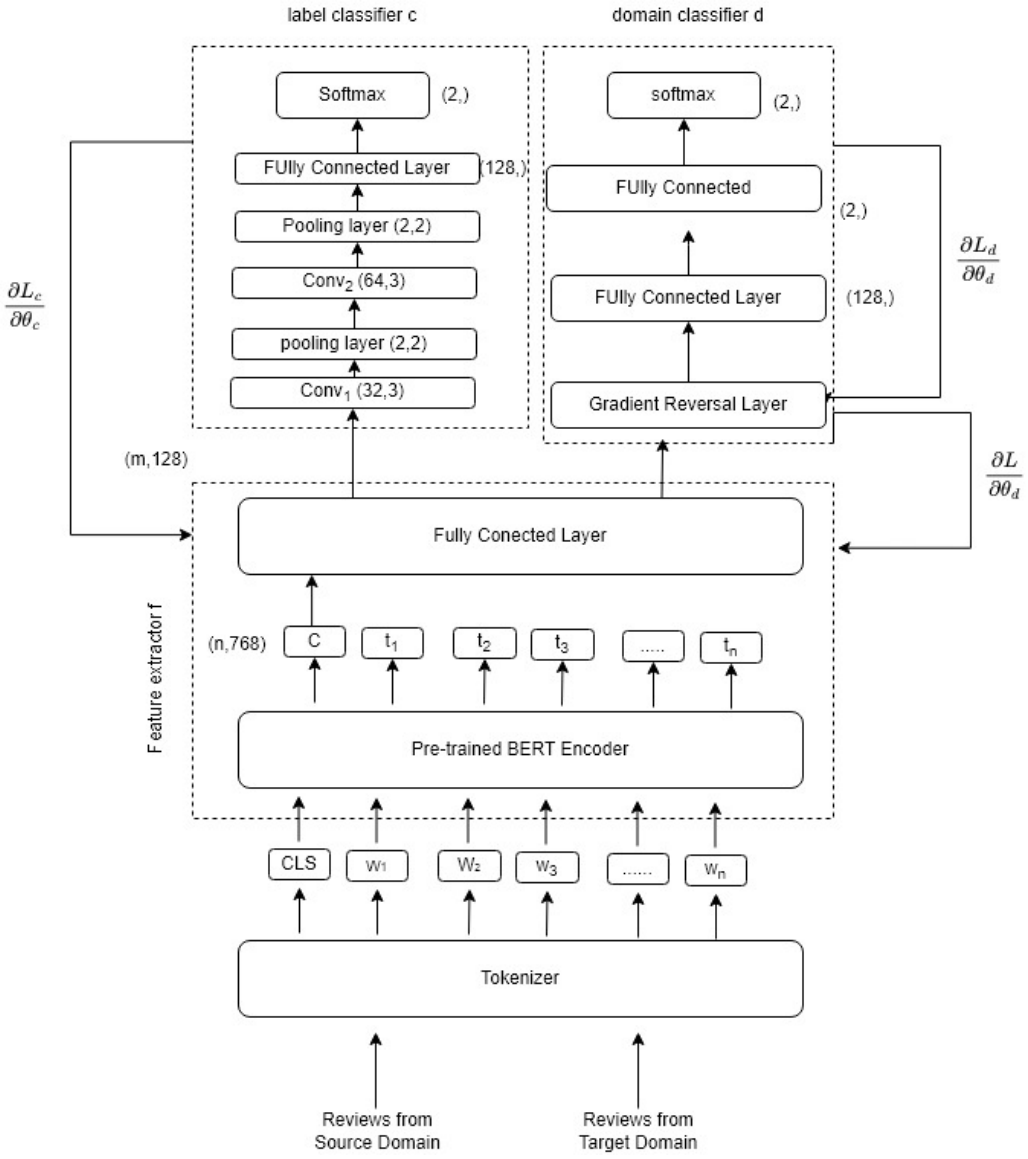


Figure 4.1: Network architecture

The proposed architecture includes a tokenizer, a feature extractor, a label classifier, and a domain classifier. The function of each component is described in detail below.

1. **Tokenizer:** Tokenization is the process of breaking into smaller components sequence into a sequence of tokens or subwords. BERT uses a WordPiece tokenizer, which breaks down words into subwords based on a pre-defined vocabulary.

   (a) **[CLS]:** In BERT, the [CLS] token is a unique token that is inserted at the beginning of all the sentences that are given input. The [CLS] token stands for "classification" and is used to represent the overall context of the sequence from input.

   (b) **[SEP]:** In BERT, the [SEP] token is a special token that is used to separate two input sequences that are concatenated together.

2. **Feature Extractor**

   The feature extractor is a critical component which consist of a pre-trained BERT encoder and fully connected neural network as shown in figure.

   Pretrained BERT Encoder: It is a neural network model that is trained on an extensive collection of corpus data. It produces contextualized representations of words which capture both meaning semantic relationship between them It uses masked language modelling and next sentence prediction during training. The encoder consist of several transformed layers, which are building block of model. Each transformer layer consist a multi-header self attention mechanism to learn contextualized representation between words. The BERT encoder produces a sequence of contextualized embeddings where each embedding represent a contextualized representation of a token in the input sequence, it produces a dimensions embeddings of each token.

   Fully connected layer : Fully connected neural network consist of 128 hidden units with relu activation function which produces 128 dimensional feature vector. The fully connected layer produces both discriminative and domain invariant features/ This is achieved by applying a series of non linear transformations to the BERT embeddings to acquire a hierarchical representation of features through the learning process that capture capture both low level and high level representation of input data. The input to the fully

connected layer 768 domensional feature vector of [cls] token and it performs non linear transformations of feature vector which produces a 128 dimensional new feature vector.

3. **Label Classifier** : label classifier is a convolutional neural network that takes input features from feature extractor which is a 128 dimensional feature vector and classifies them into real or fake reviews. The overview of CNN classifier architecture.

   Input : Input is recieved from feature extractor which is a 128 dimensional vector.

   Convolutional layer: This layer applies set of learnable features to the feature vector. The ID convolution layer produces 32 feature map of feature vector using 32 filters and the next convolution layer produces 64 feature map of feature vector using 64 filters.

   Pooling Layer : The layer reduces the size of the feature map by down sampling than the pooling operation used is max-pooling. we are down sampling because to reduce no of parameters and to prevent overfitting

   Fully Connected Layer : This layer takes output of convolution layer and flattens it into a ID vector. This layer consist of 128 hidden units with relu activation function. By incorporating non-linearity into the network, it enhances the network's capability to learn intricate features.

   Dropout : Regularization technique that is used in CNN to prevent over fitting. This technique involves dropping out a fraction of layers during training, the dropout fraction used in the architecture in 50%

   Output layer : This layer produces output of the network , which is predicted label class (fake or review) for the input feature vector. The output is normalized using a softmax activation function to produce probability distribution

4. **Domain classifier :** The domain classifier is implemented as a fully connected neural network, utilizing a ReLU activation function, is responsible for categorizing domain-invariant characteristics extracted by the feature extractor as either belonging to the source or target domain. To ensure that the feature extractor is capable of learning domain-invariant features, a Gradient Reversal Layer (GRL) is introduced between the feature extractor and domain classifier. During backpropagation, the GRL modifies the sign of the backward gradient, which encourages the feature extractor to learn domain-independent features.

   Fully connected layer 1 : This layer takes input from feature extractor. It consist of 128 hidden units with relu activation function and perform non linear transformation to feature vector and give it to next fully connected layer.

   Fully connected layer 2 : This layer has 2 hidden units with relu activation function, it takes input from fully connected layer 1 and perform non linear transformation and give to output layer

   Output layer : This layer produces output of network which is predicted label class (i.e; source domain or target domain) for the input vector. The output is normalised using softmax activation function

4.1.1 Adversarial domain adaptation

In the model we aim to learn features from the source review dataset that are independent to domain shifts and can accurately predict fake reviews in the target domain. This is achieved through an domain adaptation via unsupervisd learning task where a domain classifier is added on top of a BERT-based feature extractor, which also includes a (GRL) . During forward propagation, the GRL behaves like an identity function, and the input flows through the layer without any changes. However, during backward propagation, the

GRL takes the gradient from its succeeding layer and performs a multiplication by -1, resulting in the reversal of the gradient's sign

4.1.2 Model training

Our model's training begins with initializing the BERT component's parameters with pre-trained weights from BERT-Base, while remaining parameters are initialized as random numbers drawn from a uniform distribution. Throughout training iterations, gradients are computed , which update the feature extractor, label and domain classifier parameters. Over time, the source and target domain data distributions become more similar, enabling the feature extractor and label classifier to accurately predict target domain review labels.

Chapter 5

Experiments and Results

5.1 Dataset

The dataset used in the experiment of the proposed model is sourced from Amazon's review platform, which comprises a variety of reviews for different products. The dataset consists of both real and fake reviews, enabling the model to be tested on a diverse range of data. The reviews were collected directly from Amazon's e-commerce website, providing a realistic and extensive representation of consumer opinions. Detailed information about the dataset, including its size, and categories of products, is provided below.

category	label	
Kindle_Store_5	CG	2365
	OR	2365
Books_5	OR	2185
	CG	2185
Pet_Supplies_5	CG	2127
	OR	2127
Home_and_Kitchen_5	CG	2028
	OR	2028
Electronics_5	CG	1994
	OR	1994
Sports_and_Outdoors_5	CG	1973
	OR	1973

Figure 5.1: Dataset Details

5.2 Experimental Setup

The proposed model is developed and implemented using the Tensorflow framework in google colab platform. All the experiments are carried out on Tesla P-100 GPU with intel(R) Xeon(R) CPU @ 2.20GHz and 12GB RAM

The feature extractor takes tensors with a shape of N*D as input, where N is the batch size and D is the dimensionality of the feature vector. During training, we use a batch size of 16 and train the model over 200 batches.

5.3 Evaluation Metrics Results

The performance of the proposed model is evaluated based on the following metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.4}$$

where in

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

	Kindle_Store	Books	Pet_Supplies	Home_and_Kit	Sports_and_O	Electronics
Kindle_Store	0.6819222	0.50686496	0.701373	0.67963386	0.49656752	0.74828374
Books	0.49656752	0.46910754	0.79290617	0.58581233	0.6304348	0.694508
Pet_Supplies	0.45995423	0.7276888	0.47482836	0.4839817	0.5491991	0.7231121
Home_and_Kit	0.69908464	0.38787186	0.72196794	0.5320366	0.72997713	0.6956522

Figure 5.2: domain classification accuracy on Amazon dataset

	Kindle_Store	Books	Pet_Supplies	Home_and_Kit	Sports_and_Ou	Electronics
Kindle_Store	0.83530655	0.81327231	0.67630465	0.70857988	0.70045616	0.6607322
Books	0.84862579	0.84187643	0.7374236	0.75912229	0.74936645	0.71339017
Pet_Supplies	0.802537	0.78810069	0.80606488	0.79215976	0.78636594	0.75401204
Home_and_Kitchen	0.83530655	0.81327231	0.67630465	0.70857988	0.70045616	0.6607322

Figure 5.3: label classification accuracy using Bert model on Amazon dataset

	Kindle_Store	Books	Pet_Supplies	Home_and_Kit	Sports_and_Ou	Electronics
Kindle_Store	0.7334096	**0.8363844**	**0.70251715**	**0.8546911**	**0.81121284**	**0.8535469**
Books	0.76086956	0.82379866	**0.7402746**	**0.8020595**	**0.8455378**	**0.8501144**
Pet_Supplies	**0.84897023**	**0.8443936**	**0.84782606**	0.7723112	**0.84210527**	**0.8318078**
Home_and_Kitchen	**0.8672769**	**0.8604119**	**0.84897023**	**0.8375286**	**0.79748285**	**0.84782606**

Figure 5.4: label classification accuracy using our proposed model (BERT - DANN)

As the above 2 tables show the classification accuracy of the model tested on target domain . In each row of the above table left column value indicates the source domain and the remaining are target domain. For training the baseline model we used labeled dataset from source domain and for training BERT-DANN model we used labeled dataset from source domain and unlabelled dataset from target domain. As can be observed from the results in the above two tables, the use of BERT-DANN has led to a significant improvement in classification accuracy in most of the cases .This is due to the fact that BERT-DANN is able to capture domain invariant features, which are then used as knowledge for the target domain. By doing so, the model is able to perform better on the target domain, leading to higher classification accuracies. In comparison to the baseline model, which lacks the ability to capture these domain invariant features, the performance gains are even more pronounced.

Chapter 6

Conclusion

6.1 Conclusion

It has been observed that enhancing the BERT baseline model by capturing domain invariant features that can be leveraged as knowledge for the target domain has an improvement in the accuracy of the model. While BERT has shown good results for fake review detection, it has struggled with capturing domain invariant features in source reviews due to its task-agnostic nature. However, by integrating this Domain adaptation neural network framework with BERT, we have been able to significantly improve its accuracy. This was achieved mainly by adding a domain classifier and a gradient reversal layer to the model.

Literature Cited

[1] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[2] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online, July 2020. Association for Computational Linguistics.

[3] Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. A survey on transfer learning in natural language processing, 2020.

[4] Shubhangi Rastogi, Shabeg Singh Gill, and Divya Bansal. An adaptive approach for fake news detection in social media: Single vs cross domain. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1401–1405, 2021.

[5] Batsergelen Myagmar, Jie Li, and Shigetomo Kimura. Cross-domain sentiment classification with bidirectional contextualized transformer language models. *IEEE Access*, 7:163219–163230, 2019.

[6] V P Sumathi, S.M. Pudhiyavan, M. Saran, and V. Nandha Kumar. Fake review detection of e-commerce electronic products using machine learning techniques. In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pages 1–5, Oct 2021.

[7] Rami Mohawesh, Shuxiang Xu, Son N. Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. Fake reviews detection: A survey. *IEEE Access*, 9:65771–65802, 2021.

[8] Abrar Qadir Mir, Furqan Yaqub Khan, and Mohammad Ahsan Chishti. Online fake review detection using supervised machine learning and bert model, 2023.

[9] Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE International Conference on Data Mining*, pages 899–904, Dec 2014.

[11] Ruifeng Xu, Yunqing Xia, Kam-Fai Wong, and Wenjie Li. Opinion annotation in on-line chinese product reviews. In *LREC*, volume 8, pages 26–30, 2008.

[12] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. Binte Hassan. Spam review detection using deep learning. In *IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019*, pages 27–33, 2019.

[13] Shilpa yadav, Dr.Gulbakshee Dharmela, and Khushali Mistry. Fake review detection using machine learning techniques. In *Journal of Emerging Technologies and Innovative Research (JETIR), 2021*, volume 8, Issue 4, 2021.

[14] Sabour, S., Frosst, N. Hinton, and G. E. Dynamic routing between capsules. *Adv. Neural. Inf. Process. Syst. 30:3856–3866*, 2017.

[15] Bhagyashri Wagh, JV Shinde, and PA Kale. A twitter sentiment analysis using nltk and machine learning techniques. *International Journal of Emerging Research in Management and Technology*, 6(12):37–44, 2018.

[16] K L Santhosh Kumar, Jayanti Desai, and Jharna Majumdar. Opinion mining and sentiment analysis on online customer review. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4, 2016.

[17] Yafeng Ren and Yue Zhang. Deceptive opinion spam detection using neural network. In *International Conference on Computational Linguistics*, 2016.

[18] J. Du, Y. Dou, C. Xia, L. Cui, J. Ma, and P. S. Yu. Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[19] Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng Li. Cross-modal knowledge distillation in multi-modal fake news detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4733–4737, 2022.

[20] Kuai Xu, Feng Wang, Haiyan Wang, and Bo Yang. Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, 25(1):20–27, 2020.

[21] Shingo Kato, Linshuo Yang, and Daisuke Ikeda. Domain bias in fake news datasets consisting of fake and real news pairs. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 101–106, 2022.

[22] Shahbaz Ashraf, Faisal Rehman, Hanan Sharif, Hina Kirn, Haseeb Arshad, and Hamid Manzoor. Fake reviews classification using deep learning. In *2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT)*, volume I, pages 1–8, 2023.

[23] Bhaskar Majumdar, Md. RafiuzzamanBhuiyan, Md. Arid Hasan, Md. Sanzidul Islam, and Sheak Rashed Haider Noori. Multi class fake news detection using lstm approach. In *2021 10th International Conference on System Modeling Advancement in Research Trends (SMART)*, pages 75–79, 2021.

[24] Pratyush Goel, Samarth Singhal, Snehil Aggarwal, and Minni Jain. Multi domain fake news analysis using transfer learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1230–1237, 2021.

[25] Qi Chen, Wei Wang, Kaizhu Huang, Suparna De, and Frans Coenen. Adversarial domain adaptation for crisis data classification on social media. In *2020*

International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), pages 282–287, 2020.

[26] Raymond Y. K. Lau, S. Y. Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, and Yuefeng Li. Text mining and probabilistic language modeling for online review spam detection. *ACM Trans. Manage. Inf. Syst.*, 2(4), jan 2012.

[27] Luyang Li, Wenjing Ren, Bing Qin, and Ting Liu. Learning document representation for deceptive opinion spam detection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, Guangzhou, China, November 13-14, 2015, Proceedings 14*, pages 393–404. Springer, 2015.

[28] Xiaoya Tang, Tieyun Qian, and Zhenni You. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences*, 526:274–288, 2020.

Report

ORIGINALITY REPORT

19%	11%	12%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	docplayer.net Internet Source	2%
2	artemis.cslab.ece.ntua.gr:8080 Internet Source	1%
3	Submitted to National University of Singapore Student Paper	1%
4	Submitted to University of New South Wales Student Paper	1%
5	"ECAI 2020", IOS Press, 2020 Publication	1%
6	Submitted to University College London Student Paper	1%
7	Submitted to University of Greenwich Student Paper	1%
8	www.aiktcdspace.org:8080 Internet Source	1%
9	www.researchgate.net Internet Source	1%