SIMILARITY-BASED WEIGHTS FOR CROSS-DOMAIN SENTIMENT CLASSIFICATION OF PRODUCT REVIEWS

ADITI GUPTA

UNIVERSITI SAINS MALAYSIA

2023

SIMILARITY-BASED WEIGHTS FOR CROSS-DOMAIN SENTIMENT CLASSIFICATION OF PRODUCT REVIEWS

by

ADITI GUPTA

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

April 2023

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my main supervisor Dr. Cheah Yu-N, co-supervisor Dr. Jasy Liew Suet Yan, and for their valuable and constructive insights, patient guidance and relentless faith in me during the planning and development of this research work. Their advice and assistance helped me in keeping my progress on schedule. They are my biggest support system from my school and I could not imagine completing this dissertation without their immense knowledge and tireless support.

I am lucky to have Dr. Gan Keng Hoon, Dr. Nurul Hashimah Ahamed Hassain Malim and Dr. Noor Farizah Binti Ibrahim serve on my committee. I thank them for carefully reviewing my dissertation, providing insightful comments and challenging me with hard questions and criticism to help widen my perspective.

This research is made possible by the financial support provided by Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3and FRGS/203/PKOMP/67711796.

Last but not least, I would like to thank my family especially my husband, mother-in-law, my grandfather Mr. O.P. Gupta, parents and siblings, who stood by me and cheered me on. I also want to specially dedicate this dissertation to my kids (Akeisha and Kairav), who sometimes missed spending time with me, but provided me with the support all throughout the thesis writing process.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF TABLES ..... **v**

LIST OF FIGURES ..... **viii**

LIST OF ABBREVIATIONS ..... **ix**

ABSTRAK ..... **xi**

ABSTRACT ..... **xiii**

CHAPTER 1 INTRODUCTION **1**

1.1 Overview 1

1.2 Motivation 3

1.3 Problem Statement 5

1.4 Research Objectives 6

1.5 Research Questions 7

1.6 Research Scope 8

1.7 Research Contributions 8

1.8 Thesis Organization 9

CHAPTER 2 LITERATURE REVIEW **11**

2.1 Introduction 11

2.2 Sentiment Analysis: Overview 11

2.3 Sentiment Analysis using Deep Learning Methods 17

2.4 Cross-Domain Sentiment Classification 22

2.5 Research gap 38

CHAPTER 3 RESEARCH METHODOLOGY **41**

3.1 Introduction 41

3.2 Problem Setting 41
3.3 Dataset 42
3.4 Methodological Framework 44
3.5 Evaluation Metrics 66
3.6 Experiment Computing Resources 67
3.7 Summary 67
CHAPTER 4 DEEP LEARNING FOR CROSS-DOMAIN SENTIMENT ANALYSIS 68
4.1 Introduction 68
4.2 Deep Learning Architectures 69
4.3 Domain-related Experiments using CNN (Phase 1) 75
4.4 Comparing Different Deep Learning Models (Phase 2) 86
4.5 Summary 93
CHAPTER 5 WEIGHTED DOMAIN SIMILARITY APPROACH..... 95
5.1 Introduction 95
5.2 Comparing Similarity Scoring Measures 96
5.3 Comparing Similarity versus Relevance Source Domain Ranking 109
5.4 All Source Domains Models Weighted by Similarity Score..... 113
5.5 Selected Source Models Ranked by Similarity Scores 129
5.6 Discussion 140
CHAPTER 6 CONCLUSION AND FUTURE WORK 144
6.1 Conclusion..... 144
6.2 Research Contributions 146
6.3 Limitations and Future Work 148
REFERENCES..... 149
LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 2.1 Summary of research papers on sentiment classification 16

Table 2.2 Summary of research papers on sentiment analysis using deep learning methods. 19

Table 2.3 Summary of research papers discussed in Section 2.4.1 30

Table 2.4 Summary of research papers using the domain similarity methods ... 33

Table 2.5 Summary of research papers in-domain adaptation using deep learning methods 37

Table 3.1 List of domains used 43

Table 4.1 Hyperparameters of CNN model 72

Table 4.2 Hyperparameters of Bi-LSTM model 75

Table 4.3 Comparing baseline neural network and CNN 77

Table 4.4 Comparing results for In-Domain and Cross-Domain Sentiment Classification 79

Table 4.5 Accuracy for source domain (column) against target domain (row) 81

Table 4.6 Order of source domains according to decreasing accuracy against a target domain 82

Table 4.7 Accuracy for ablation experiment (Step 2) 85

Table 4.8 Comparing results for cross-domain sentiment classification using CNN and Bi-LSTM classifiers 87

Table 4.9 Comparing results of CNN, Bi-LSTM and Sequential CNN-Bi-LSTM models 88

Table 4.10 Comparing results of CNN, Bi-LSTM, Sequential CNN-Bi-LSTM and Sequential Bi-LSTM-CNN models 89

Table 4.11 Results for CNN, Bi-LSTM and Ensemble models 90

Table 5.1 Cosine similarity between different pairs of domains.....98

Table 5.2 Jaccard distance between different pairs of domains.....99

Table 5.3 Jensen-Shannon divergence between different pairs of domains 101

Table 5.4 Hellinger distance between different pairs of domains using LDA.102

Table 5.5 WMD scores between different domains using fastText word embeddings 105

Table 5.6 WMD scores between different domains using word2vec word embeddings 106

Table 5.7 WMD score between different domains using GloVe word embeddings 107

Table 5.8 Colour scheme for the difference in ranks between similarity and relevance order of domains 109

Table 5.9 Comparing ranking of domains obtained using similarity and relevance 112

Table 5.10 Rank-based weights for different source domains 114

Table 5.11 Results using rank-based sample weights for cross-domain sentiment classification 115

Table 5.12 Weights for score-based weighting scheme 119

Table 5.13 Results using score-based sample weights for cross-domain sentiment classification 121

Table 5.14 Average for threshold-based weighting scheme 124

Table 5.15 Assignment of weights for source domains based on the average accuracy threshold.....125

Table 5.16 Results for threshold-based weighting scheme for cross-domain sentiment classification 126

Table 5.17 Cross-Domain sentiment classification using most similar source domain.....130

Table 5.18 Cross-domain sentiment classification using K-most similar source domains.....135

Table 5.19 Results using rank-based weights for K-most similar source domains 138

Table 5.20 Results using score-based weights for K-most similar source domains 139

Table 5.21 Comparing accuracy with state-of-the-art studies..... 142

LIST OF FIGURES

Page

Figure 2.1 Sentiment Analysis Methods 12

Figure 2.2 Example of a three-layer neural network 18

Figure 3.1 Phases for the research methodology 46

Figure 3.2 Architecture of CNN model 48

Figure 3.3 Architecture of Bi-LSTM model 49

Figure 3.4 Architecture of Ensemble Model 54

Figure 3.5 Architecture of Sequential CNN-BiLSTM sentiment classifier 56

Figure 3.6 Architecture of Sequential BiLSTM-CNN sentiment classifier 58

Figure 3.7 Implementation of weights in the cross-domain sentiment classifier 63

Figure 4.1 Comparison between numbers of sentiment words 78

Figure 4.2 Comparing the accuracy of Ensemble, Sequential CNN-BiLSTM and Sequential BiLSTM-CNN models 91

Figure 5.1 Comparing different weighting schemes 129

Figure 5.2 Comparing sentiment words present in in-domain and K-most similar source domains 131

Figure 5.3 Comparing accuracy between all and K-most similar source domains 136

Figure 5.4 Training time in seconds 136

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
BOW	Bag of Words
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CRMMD	Class Refinement Maximum Mean Discrepancy
GRU	Gated Recurrent Unit
JSD	Jensen Shannon divergence
KL	Kullback-Leibler
LDA	Latent Dirichlet Model
LSTM	Long Short-Term Memory
MSDA	Multi-Source Domain Adaptation
NLP	Natural Language Processing
PMI	Pointwise Mutual Information
POS	Part of Speech
QBC	Query by Committee
RNN	Recurrent Neural Network
SCL	Structural Correspondence Learning
SCP	Significant Consistent Polarity
SDA	Stacked Denoising Autoencoders
SDAA	Selective Domain Adaption Algorithm
SFA	Spectral Feature Alignment
SO-CAL	Semantic Orientation CALculator
SVM	Support Vector Machine

TF-IDF	Term Frequency - Inverse Document Frequency
VA	Valence Arousal
WMD	Word Mover's Distance

PEMBERAT BERASASKAN KESERUPAAN UNTUK PENGELASAN SENTIMEN MERENTAS DOMAIN BAGI ULASAN PRODUK

ABSTRAK

Ketiadaan data berlabel untuk domain tertentu menimbulkan cabaran untuk melatih pengelas untuk pengesanan sentimen dalam ulasan produk. Analisis sentimen merentas domain menawarkan penyelesaian untuk melatih model menggunakan data berlabel daripada domain sumber dan menggunakannya pada domain sasaran. Walau bagaimanapun, prestasi pengelas biasanya terjejas dengan ketara apabila taburan ciri dan ekspresi sentimen domain sumber dan sasaran berbeza. Selain itu, apabila menggunakan berbilang domain sumber, setiap domain sumber tidak semestinya memberi manfaat yang sama rata kerana sesetengahnya lebih berkaitan dengan domain sasaran tertentu. Tesis ini menangani isu ini dengan membangunkan pengelas pembelajaran mendalam merentas domain dan menyiasat kesan pelbagai domain sumber terhadap latihan pengelas sentimen. Tambahan pula, kesan setiap domain sumber ke atas latihan pengelas sentimen merentas domain dan pemilihan domain sumber yang berguna dikaji untuk membangunkan kaedah baru yang memberikan pemberat kepada setiap domain sumber mengikut kepentingannya kepada domain sasaran. Metodologi tiga fasa dilaksanakan dengan Fasa 1 memberi tumpuaan pada penciptaan seni bina pembelajaran mendalam menggunakan CNN dengan hiperparameter optimum untuk tugas klasifikasi merentas domain dan diikuti dengan percubaan yang meluas untuk mencari kaitan antara pelbagai domain sumber dengan domain sasaran. Model CNN yang dibangunkan dalam fasa ini bertindak sebagai garis dasar untuk eksperimen selanjutnya. Fasa 2 bertumpu pada pembinaan model pembelajaran mendalam yang berbeza menggunakan CNN dan Bi-LSTM, dan membandingkan prestasi model tersebut untuk klasifikasi

sentimen merentas domain. Fasa 3 menggunakan ukuran pemarkahan persamaan untuk mengira dan memberikan pemberat kepada domain sumber. Model yang menunjukkan prestasi terbaik dari Fasa 2 dipilih dan dilatih menggunakan pemberat yang berbeza berdasarkan domain sumber. Keputusan menunjukkan bahawa pembelajaran mendalam boleh memanfaatkan sejumlah besar data daripada domain sumber yang berbeza dan seni bina model pembelajaran mendalam membawa kepada perbezaan, bukan sahaja kedalaman model. Ketiga-tiga skim penimbang novel tersebut menunjukkan kesan positif terhadap prestasi pengelas. Model ensembel menggunakan CNN dan Bi-LSTM menggunakan skema penimbang berasaskan ambang muncul sebagai model terbaik.

SIMILARITY-BASED WEIGHTS FOR

CROSS-DOMAIN SENTIMENT CLASSIFICATION OF PRODUCT

REVIEWS

ABSTRACT

The unavailability of labelled data for a particular domain poses a challenge for training a classifier for sentiment detection in product reviews. Cross-domain sentiment analysis offers a solution to train models using labelled data from source domains and applying it to the target domain. However, the classifier performance usually suffers significantly when the source and target domains' feature distribution and sentiment expressions differ. Also, when using multiple source domains, not all source domains are equally beneficial as some are more relevant to a particular target domain. This thesis addresses these issues by developing cross-domain deep learning classifiers and investigating the impact of multiple source domains on sentiment classifier training. Furthermore, the effect of each source domain on the training of the cross-domain sentiment classifier and selecting helpful source domains is examined. The study developed a novel method of assigning weights, to each source domain according to its importance to the target domain. A three-phase methodology is implemented, with Phase 1 focusing on creating the deep learning architecture using CNN with optimal hyperparameters for cross-domain classification tasks followed by extensive experiments to find the relevance between various source domains to the target domain. CNN model developed in this phase acted as a baseline for further experiments. Phase 2 focused on constructing different deep learning models using CNN and Bi-LSTM, and compared their performance for cross-domain sentiment classification. Phase 3 used similarity scoring measures to

calculate and assign weights to the source domains. Best performing model from Phase 2 was selected and trained using different weights for source domains and performance was compared. Results showed that deep learning can leverage a large amount of data from different source domains and the architecture of the deep learning model makes differences, not only the depth of the model. The three novel weighing schemes showed a positive impact on the performance of classifier. Ensemble model using CNN and Bi-LSTM, using threshold-based weighing scheme emerged out as the best model.

CHAPTER 1

INTRODUCTION

1.1 Overview

Easy access to the internet, the World Wide Web, and the low cost of electronic gadgets result in massive data generated every second. People post their feedback, opinions and reviews about almost everything like products, services and current issues. How people think and react to products and services is becoming essential for the product manufacturers and service providers. Before making a purchase, people want to know the feedback from the users of that product. Thus, customer feedback acts as a reliable source of information for potential customers. Analyzing customer sentiment helps service providers enhance their customer service by letting them know what makes customers happy and what not. It helps manufacturers improve products and services by fixing issues/bugs in their products and also help in letting them know how they can better improve to retain old customers while attracting new customers. Manufacturers also get to know what the customer needs and how to optimize the marketing strategies. Therefore, finding value and knowledge from the vast amount of customer review data through sentiment analysis is an important and challenging area in research.

Classification and categorization of text is a crucial part of Natural Language Processing (NLP). Various applications under its category are sentiment analysis, opinion mining, subject categorization and spam detection. Many researchers have been attracted to opinion mining and sentiment analysis in recent times(Bollegala, Weir, & Carroll, 2013; Kim, 2014; B. Liu, 2012; Poria, Cambria, & Gelbukh, 2016).Sentiment analysis can be analyzed at different levels like document-level(Rao, Huang, Feng, & Cong, 2018; Z. Yang et al., 2016), sentence-level(T.

Chen, Xu, He, & Wang, 2017; A. Khan, Baharudin, & Khan, 2011)or aspect-level(X. Chen et al., 2020; He, Lee, Ng, & Dahlmeier, 2018; Y. Wang, Huang, Zhu, & Zhao, 2016).

Different machine learning techniques like support vector machines (SVM), logistic regression, and Naive Bayes exploit shallow structured architecture for sentiment classification(Singh, Singh, & Singh, 2017; C. C. Yang, Tang, Wong, & Wei, 2010). The shallow architecture typically contains at most one or two layers of non-linear feature transformations. The NLP problems utilizing these structures are trained on sparse features with high dimensions. It also relies on manual feature engineering, which consumes significant human effort and time. Shallow architecture has proven itself in solving many well-constrained or specific problems, but their limited architecture and representation can cause difficulties when dealing with more complex real-world applications.

Recently researchers have contributed to sentiment analysis using deep learning (Bengio, Goodfellow, & Courville, 2014; Chen & Lin, 2014). Deep learning aims to extract complex features from data using minimum external interference and learn them using a deep neural network. Deep learning algorithms use large amounts of data to learn new complex features automatically. Models like Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have drawn the attention of the researchers for performing experiments on polarity detection, sentiment analysis, question answering and other NLP tasks (Kim, 2014; Rao et al., 2018).

In typical machine learning problems, most models assume that training samples are drawn from the same distribution as the unseen test samples. However, the performance of these models is highly dependent on the domain (collection of

reviews on a particular product) and need a large amount of training data. However, there are many cases in real life where the training and the test sample distributions differ. For example, there are situations when the requirement is to train a classifier for a domain (called target domain) with no labelled data or less labelled data, with one or more domains (called source domains) that have sufficient labelled data. If a classifier is trained on a random source domain, it may perform poorly as such models are known to not generalize well on the target domain.

A learning task that can handle such a situation is called domain adaptation, and the sentiment classification process using training data and test data from different domains is called cross-domain sentiment classification. Domain adaption is used for sentiment analysis while performing sentiment classification across domains (Aue & Gamon, 2005; Blitzer, Dredze, & Pereira, 2007; Pang, Lee, & Vaithyanathan, 2002; Remus, 2012).

1.2 Motivation

Sentiment analysis of product reviews is becoming essential for manufactures and potential customers. Product reviews help manufacturers make more informed decisions by providing them with insights on the aspects of a product that perform well and how to make the experience better for the customer. Product reviews also help customers decide whether to buy a product or not by letting them know about other customers' satisfaction or dissatisfaction towards the product.

A classifier is trained using labelled positive and negative user reviews in supervised binary sentiment classification. However, it is costly and infeasible to annotate reviews manually for the vast number of online products. At the same time, it is challenging to perform sentiment analysis when the training data for a particular

domain is small and insufficient. In that case, the need arises for cross-domain sentiment classification, where data from other domains is used to train the classifier to perform prediction on a target domain, particularly one with very few or no labelled examples (Blitzer et al., 2007; Bollegala et al., 2013; Hao et al., 2020; Meng, Long, Yu, Zhao, & Liu, 2019; Pan, Ni, Sun, Yang, & Chen, 2010).

One simple approach to handle cross-domain sentiment analysis is to design rules to classify the sentiment present in text. However, a challenge for using rule-based approaches in a cross-domain environment is that rule-based approaches cannot adapt automatically to domain-specific characteristics (Deshmukh & Tripathy, 2016). In addition, the rules need to be manually engineered and handcrafted. Another simple approach is to use lexicon-based methods, which would require lexical resources to be available (Barnes, Klinger, & Walde, 2018; L. Wang, Niu, Song, & Atiquzzaman, 2018). The rule-based and lexicon-based approaches do not scale well across domains as both are highly domain-dependent. Generating the rules requires deep domain knowledge, and at the same time, it is time consuming and challenging to cover all aspects of sentiment classification.

Cross-domain sentiment analysis is a challenging task as the polarity of words sometimes changes depending on the domain and context of how words are used (Y. Zhang, Hu, Li, Li, & Wu, 2015). A natural solution to this problem is to train a domain-specific classifier for each domain, but this is not practical as the labelled data in the target domain is not always sufficient for training the model. At the same time, the cost and time incurred for manual labelling of the data can be prohibitively high.

Motivated by these observations, this research developed different deep learning architectures for performing sentiment analysis on a cross-domain dataset

and learning domain-relevant features taking into consideration that not all domains are equally important for a target domain.

1.3 Problem Statement

In many situations, a domain does not have sufficient amount of labelled data for the training of the sentiment classifier. On the other hand, there are domains present with sufficient amount of training data. Therefore, there is a need for models that can leverage a large amount of labelled data from other domains (source domains) to be used for training to classify reviews for the domain with very few or insufficient labelled data (target domain). Most past studies have considered only one source domain while performing cross-domain sentiment classification (Hao et al., 2020; Peng, Zhang, Jiang, & Huang, 2018; S. Zhang, Liu, Yang, & Lin, 2015). However, combining more than one source domain can also benefit the classifier's training as the classifier would have more training data to learn from, thus improving its performance on the target domain.

Second, in past research for cross-domain sentiment classification, some labelled data from the source domain and a large amount of unlabeled data from the target domain were used to train the classifier (Blitzer et al., 2007; Bollegala, Mu, & Goulermas, 2016; Pan et al., 2010). The pivots were selected from the source and target domain by considering the feature distribution between the domains as being the same. Although the results were good, a significant drop in the classifier's performance was observed when the feature distribution and sentiment expressions in the source and target domain differed significantly. Therefore, domain adaptation focusing on the feature distribution between the source and target domains is still limited in terms of robustness.

Third, past studies have given equal importance to all the domains when using multiple source domains (Glorot, Bordes, & Bengio, 2011; Q. Liu, Zhang, & Liu, 2018). When using multiple domains for training the classifier, not every domain is equally beneficial. Some source domains are more relevant and influence the classifier's training for a particular target domain than others. Considering all available source domains to have the same impact on the training and classification of product reviews is thus not justifiable.

Over past few years, deep learning has emerged as a concept, which involves learning by examples, similarly as what humans are naturally do. It involves little human intervention by performing automatic feature engineering and self learning. Because of availability of large amount of data and computing power, many researchers used deep learning for performing sentiment analysis (Kim, 2014; Poria et al., 2016; Shrestha & Nasoz, 2019). Availability of large amount of customer reviews on products can be utilized to train a deep learning model to perform sentiment classification for a target domain which is lacking in training data.

This discussion leads us to define a research problem worth solving, which is to experiment with training data from multiple domains so that domains with insufficient training data can benefit from other domains with sufficient training data.

1.4 Research Objectives

This research aims to study the utilisation of multiple available source domains to perform sentiment classification for the target domain. The research will explore different deep learning algorithms for cross-domain sentiment classification and study methods to calculate weights for the source domains, to train cross-domain sentiment classifiers efficiently.

This thesis aims to fulfil the following objectives:

a) To design cross-domain deep learning classifiers for sentiment analysis of the reviews of a target domain with no or very little labelled data, which is insufficient for training the classifier. This can be done using multiple source domains for the classifier's training and exploiting large amounts of labelled data present in the form of reviews, to extract complex features with the help of deep learning. At the same time, the sentiment classifiers can be enhanced by considering the long-term dependencies present in the reviews and thus improve the performance.
b) To examine the effect of each source domain on the training of the cross-domain sentiment classifier for a target domain and study different similarity measures to select helpful source domains from all available source domains.
c) To calculate and assign weights to each source domain before training based on its similarity with the target domain for weighted cross-domain sentiment classification.

1.5 Research Questions

Specifically, the research study will address the following three research questions:

1. Which deep learning architecture trained with multiple source domains yields good sentiment classification performance on the target domain?
2. Which similarity measures are helpful to select the relevant source domains out of all available source domains to train a model that can generalize well on the target domain?

3. How to assign weights based on domain similarity to the selected source domains for training the model?

1.6 Research Scope

The scope of this research is focused on cross-domain sentiment classification of products reviews from Amazon.com. The source domains and target domain will be the reviews of products from Amazon.com. This research study includes 14 different domains. Other sources of reviews such as movie, hotel and restaurant reviews are not in the scope of this research. This research considers that the numbers of product reviews in both polarities are equal, and each domain has nearly the same number of labelled reviews.

1.7 Research Contributions

The main contributions of this research are summarized below:

- This research develops deep learning models, using deep learning algorithms that capture the features efficiently from a large amount of training data present in reviews and improve the accuracy of the cross-domain sentiment classification task. A novel ensemble deep learning model was developed using CNN and BiLSTM using stacking for cross-domain sentiment classification.
- It performs ablation experiments, to study the effect of each source domain in the training of multi-source cross-domain sentiment classifier. The results from ablation experiments leads to the investigation of different similarity measures for selecting the useful source domains for training the classifier for a target domain.

- It contributes three novel ways of calculating the weights based on selected similarity measures and assigning them to the source domains for training the classifier. Three methods for calculating the weights used similarity scores, threshold values and ranks of source domains. By assigning weights, improvement in performance of cross-domain sentiment classifier is observed.

1.8 Thesis Organization

The remaining chapters in this proposal are organized as follows:

a) **Chapter 2** provides an account of existing literature on sentiment classification of product reviews and user feedback with in-depth analysis and comparison. It discusses various approaches for sentiment analysis. Specifically, it reviews the existing research on cross-domain sentiment classification. It also explains the gaps between the existing literatures and provides greater context to understand the research problem being addressed in the study.

b) **Chapter 3** elaborates the research framework and methodology to solve the search questions and discusses the model performance evaluation plan. It contains the details of the three-phase methodology used to classify the product reviews according to their polarity. The deep learning model development for cross-domain sentiment classification is done in Phase 1 and Phase 2. The model is then used for further experiments using source domains selected based on similarity to the target domain for training in Phase 3.

c) **Chapter 4** addresses the first research question stated in section 1.5. It starts by optimizing the hyperparameters of deep learning models and describes the architecture for cross-domain sentiment classification. The experiment results are obtained from Phase 1 and Phase 2 of the research methodology.

d) **Chapter 5** addresses the second and third research questions stated in section 1.5. It discusses the exploration of helpful source domains to be used with the deep learning model developed in Chapter 4. The results provide an in-depth analysis of Phase 3, including experiments using similarity measures between the source and the target domains.

e) **Chapter 6** highlights the contributions of this study and presents conclusions and topics for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the background on sentiment classification in text covering the polarity detection in the reviews of the products and services. Considerable research has been conducted to study cross-domain sentiment classification over the years. However, those research studies used a single domain as the source domain, and less emphasis was given to using multiple source domains for cross-domain sentiment classification and how each source domain plays its role in the classifier's training. In line with the research objectives, the literature review is kept in the context of cross-domain sentiment classification and related approaches. This review has put forth a comparison of existing research studies. Various approaches for sentiment analysis have been identified, elaborated and analyzed. This chapter integrates the relevant pieces of past research by providing a theoretical and methodological discussion of existing work in sentiment classification of product reviews and identifies gaps in the literature. An important point to be highlighted is that many researchers used the terms “multi-domain” and “cross-domain” interchangeably to refer to using the training data (source domain) and test data (target domain) from different domains.

2.2 Sentiment Analysis: Overview

Figure 2.1 shows two different general categorical schemes in organizing sentiment analysis research. Broadly, earlier techniques for sentiment analysis can be categorized into two methods, machine learning and lexicon-based. As for the unit of

analysis, sentiment can be detected at the document, sentence or aspect levels.

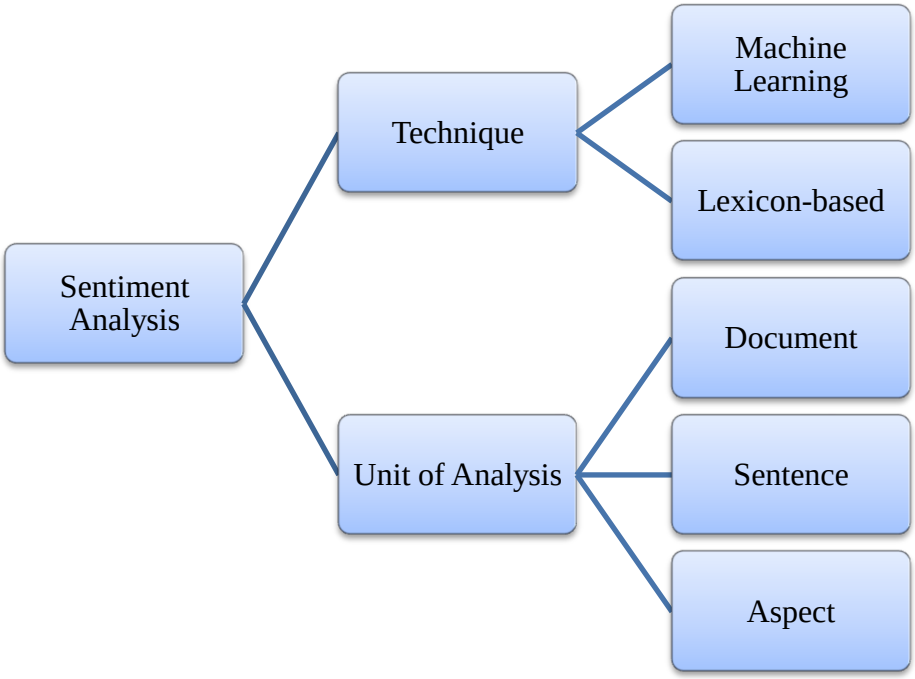

Figure 2.1: Sentiment Analysis Methods

In the machine learning category, a corpus is first annotated with sentiment before the text is transformed into a more structured representation before being fed into a machine learning algorithm for training and evaluation. Table 2.1 summarizes notable prior work on sentiment classification. In Pang, Lee, and Vaithyanathan (2002), standard bag of features were used with three machine learning methods(Naïve Bayes, maximum entropy and support vector machine) to perform sentiment classification on movie reviews. Several features like unigrams, bigrams, adjectives and position of words were used as features to test with machine learning algorithms and the best performance was achieved when unigrams were used in the SVM classifier. Jain and Dandannavar (2016) applied sentiment classification for analyzing Twitter data using Apache Spark. After pre-processing, features like n-grams, frequency counts and POS were extracted, and the model was trained using Multinomial Naïve Bayes and Decision Tree algorithms. Multinomial Naïve Bayes

did not perform as expected when supplied with a small training set, whereas Decision Tree took longer training time and less time to predict unseen data than Naïve Bayes. However, the results showed that Decision Tree performed very well with very high accuracy.

In the lexicon-based approach, dictionaries can be created manually or automatically using seed words. For example, a word-based method for extracting sentiment was proposed in Taboada et al. (2011), which used Semantic Orientation CALculator (SO-CAL) to extract sentiment from the product and movie reviews. SO-CAL used dictionaries of annotated words by polarity and strength for the sentiment classification task. It also incorporated the intensifiers and negation. The performance of this approach was consistent across domains and on new data. Rice created specialized sentiment dictionaries, one comprised of positive tokens and other negative tokens, with minimal supervision using semantic similarity, which used both the structure of the language and the text corpus. The techniques used a small set of seed words correlated with the domain and word vector representation. The polarity of the test data was calculated by weighting tokens counts by TF-IDF and multiplying weighted counts by cosine similarity, thus giving similarity-weighted positive counts and similarity-weighted negative counts for calculating the polarity. The dataset used comprised of movie reviews and US Supreme Court opinions.

Another aspect in categorizing sentiment analysis research is based on the unit of analysis, which can be divided into three levels: document-level, sentence-level and aspect-level.

In document-level sentiment analysis, a sentiment label is assigned to a document that may contain one or many sentences as illustrated in Example 2.1. The

sentiment is extracted from the review and classified based on the overall opinion as positive, negative or neutral.

Example2.1: "I bought my HP Envy laptop last week. I just loved the performance. The touch screen makes things so easy. Nice wide screen gives a superb gaming experience. Big thumbs up to configuration".

Sharma, Nigam, and Jain (2014) performed document-level sentiment analysis using WordNet. The polarity of the documents was determined based on the majority of opinion words. The polarity of the document was set as positive when there were a greater number of positive words, otherwise as negative. For an equal number of positive and negative words, the document was considered neutral. (Tripathy, Anand, & Rath, 2017) used a hybrid approach to classify the sentiment of the reviews at the document-level. SVM was used to calculate the sentiment value of each word after pre-processing of the review, and the words having sentiment value above a certain threshold were selected. The selected words were then used as input to Artificial Neural Network (ANN), and the model was assessed based on selected features for classification into positive or negative. Sentiment classification was performed on the IMDb and polarity datasets.

Sentence-level sentiment analysis segments text into sentences and further classifies them into subjective or objective sentences. The subjective sentences are then assigned a sentiment label. Some studies chose sentence-level as the unit of analysis because a document may contain more than one sentiment.

Example 2.2: "I stayed in Traders hotel last week for 3 nights. The rooms were clean and tidy. But there was a strange smell in the bathrooms. On complaint staff upgraded my room to a suite. I enjoyed my stay but wasted nearly two hours in the process".

In Example 2.2, the first sentence is an objective sentence as it just states the facts without sharing any sentiment. The second sentence expressed positive sentiment towards the cleanliness of the room while the remaining sentences contained negative sentiment about the bathroom and the unpleasant experience.

Khan et al. (2016) performed sentence-level sentiment analysis on a heterogeneous dataset comprised of movie reviews, product reviews, tweets and comments from Facebook. They used POS tagging on the pre-processed sentences to identify subjective sentences using k-nearest neighbours (kNN) model. SentiWordNet was then used to determine the strength and semantic orientation of sentiment-bearing words. Thus semantic scores obtained for sentiment words was averaged and assigned to the subjective sentence that contained those words.

In aspect-level sentiment analysis, the review is first categorized by aspects, followed by the identification of the sentiment associated to each one. In Example 2.3, ambience, food and service represent different aspects of a restaurant. For aspect-level sentiment analysis, the associated sentiment for each aspect would be positive for ambience, positive for food and negative for service.

Example 2.3: "The ambience in the restaurant was nice. Food was delicious, but the service was very slow. I had to remind the staff about my order".

Zhu et al. (2011) performed opinion polling from unlabeled textual customer reviews on restaurants. They proposed multi-aspect bootstrapping to learn terms related to each aspect. The multi-aspect segmentation model handled multi-aspect sentences, followed by an aspect-based opinion polling algorithm, which first determined the polarity of each aspect with respect to each review and then for each aspect with respect to the review set. Another solution for the multi-aspect sentiment was proposed by Sun et al. (2016). The method used the combination of semantic feature mining and lexicon-based techniques to analyze the aspect-level sentiment of online product reviews. The product aspects were extracted by the Latent Dirichlet Allocation (LDA) model, and their corresponding sentiment was calculated by the domain-lexicons developed.

Table 2.1: Summary of research papers on sentiment classification

Research paper	**Dataset**	**Approach**	**Text-features used**
Pang, Lee, and Vaithyanathan (2002)	Movie reviews	Naïve Bayes, maximum entropy, SVM	Unigram, bi-gram, adjectives, POS
Taboada et al. (2011)	Movie reviews and product reviews	Dictionary-based	Sentiment bearing words(including objectives, verbs, nouns and adverbs)
Zhu et al. (2011)	Restaurant reviews	Multi-aspect bootstrapping	Nouns, verbs, adjectives, multi-word terms
Sharma, Nigam, and Jain (2014)	Movie reviews	Dictionary-based	Opinion words
Jain and Dandannavar (2016)	Tweets	Multinomial Naïve Bayes, decision tree	n-grams, POS
Khan et al. (2016)	Movie & product reviews, Tweets and comments	kNN	POS
Sun et al. (2016)	Product reviews	LDA, Dictionary-based	Product aspects
Tripathy, Anand, & Rath (2017)	IMDb & polarity dataset	SVM, ANN	CountVectorizer, TF-IDF
Rice & Zorn (2021)	Movie reviews	Dictionary-based	Positive and negative sentiment words
Mutanov et al., (2021)	News portals	Naive Bayes, SVM, Logistic Regression, kNN, Decision Tree, Random Forest and XGBoost	TF-IDF

Mutanov, Karyukin, & Mamykova (2021) applied multi-class sentiment analysis; categorizing the data into three labels as positive, negative and neutral. They adopted one-vs.-one approach to identify as particular class, instead of one-vs.-all where multiple binary classifiers need to be trained to distinguish samples of one class from all other samples. As the training data was imbalanced in nature, resampling was performed using random undersampling, random oversampling and SMOTE. Seven machine learning algorithms were applied and performance of resampling techniques was compared.

2.3 Sentiment Analysis using Deep Learning Methods

Deep Learning is a sub-category of machine learning, which can also be used in supervised and unsupervised learning. It is based on Artificial Neural Networks (ANN), inspired by human biological neural networks. It works on the same principle on which the human brain mechanism works. It uses a cascade of multiple layers of non-linear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input. It learns multiple levels of representations that correspond to different levels of abstractions, and the levels form a hierarchy of concepts with higher-level features being derived from lower-level features.

Figure 2.2 shows a shallow neural network that uses three layers. The input layer accepts the inputs, which are word embeddings from the text. The hidden layer is commonly known as weights which are learned when the neural network is trained. The output layer gives a prediction or result of the input fed into the network.

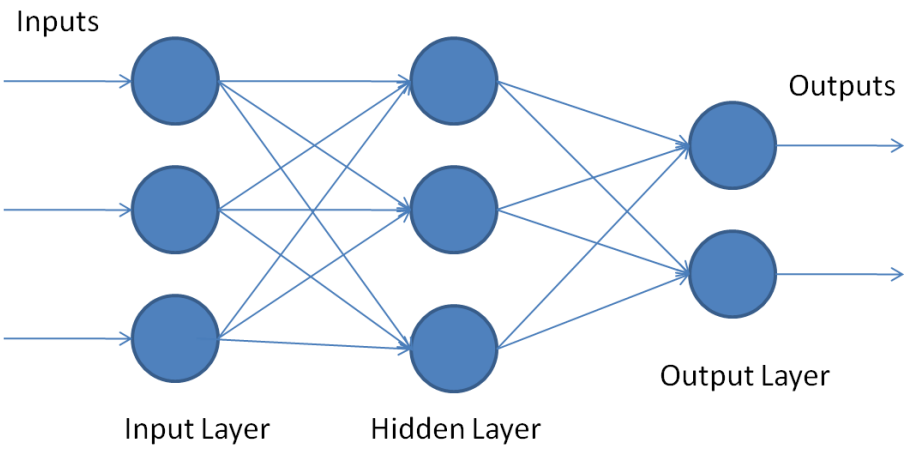

Figure 2.2: Example of a three-layer neural network

Deep learning is stacking multiple hidden layers between the input and the output layer. Recently, deep learning models have achieved remarkable results when applied to computer vision, speech recognition and natural language processing (Bengio, 2013; Chen & Lin, 2014; Severyn & Moschitti, 2015). Many researchers have experimented with convolutional neural network for classifying the reviews according to their polarity (Kim, 2014; Shi, Ushio, Endo, Yamagami, & Horii, 2017; Wei, Lin, Yu, & Yang, 2017). Table 2.2 summarizes the research on sentiment analysis using deep learning methods.

A convolutional neural network (CNN) is a common deep learning architecture in sentiment analysis. CNN can extract features automatically from data and has good classification accuracy. Kim (2014) used CNN for sentence-level sentiment classification. An improved scheme was proposed, which employed dynamically updated and static word embeddings simultaneously for sentence classification based on CNN. The model used multiple filters to obtain multiple features, which were then passed to a fully connected softmax layer. The proposed models were tested on datasets consisting of movie reviews and customer reviews.

This study which became popular is a good demonstration of the power of pre-trained word embeddings. With a relatively simple model, the authors achieved state-of-art (or comparable) results for classifying the reviews according to the polarity as positive or negative. Datasets used in the experiments comprised of movie and products reviews. The results are used in many other studies as the benchmark.

Table 2.2: Summary of research papers on sentiment analysis using deep learning methods.

Author (Year)	**Corpus**	**Representation**	**Deep Learning Model**
Kim (2014)	Movie reviews and products review	Word2Vec word embeddings	CNN
Santos and Gatti (2014)	Movie reviews and Twitter messages	Word2Vec word embeddings	CNN
Wang et al. (2016)	Stanford Sentiment Treebank and Chinese Valence Arousal Text	Word2Vec word embeddings	CNN and LSTM
Y. Wang et al. (2016)	Customers reviews from SemEval 2014 Task 4	Concatenate aspect vector into sentence hidden representations for computing attention weights	LSTM
T. Chen et al. (2017)	News resources (for training Bi-LSTM-CRF), movie reviews, Stanford Sentiment Treebank and customer reviews (for testing CNN)	Word2Vec word embeddings	BiLSTM-CRF and CNN
Hassan and Mahmood (2017)	Stanford Large IMBD Movie Review Dataset and Stanford Sentiment Treebank	Word2Vec word embeddings	CNN and LSTM
Shrestha and Nasoz (2019)	Product reviews	Paragraph vectors and product embeddings of the reviews	RNN with GRU
B. Chen, Huang, Chen, Cheng, & Chen(2019)	Micro-Blog comments	Word2Vec word embeddings	CNN and LSTM

Santos and Gatti (2014) used a deep neural network to perform sentiment analysis using character-level, word-level and sentence-level representations of the text. CNN was used to extract features from character to sentence-level. Two convolutional layers were used that allow the architecture to handle words and sentences of different sizes. The main advantage of the proposed approach was the extraction of relevant features from any part of the word without needing handcrafted inputs. Word embeddings produced using unsupervised pre-training also proved beneficial. The proposed system has tested its effectiveness on short texts from movie reviews and Twitter messages.

Wang et al. (2016) performed dimensional sentiment analysis, which provided more fine-grained sentiment analysis. The proposed model used regional CNN and Long Short Term Memory (LSTM) to predict the text's Valence Arousal (VA) ratings. The model treated individual sentences of the text as a region and extracted useful information for weighting according to its contribution to VA. Thus, the regional information was sequentially passed to LSTM for dimensional sentiment analysis and provided more intelligent and fine-grained sentiment analysis. The proposed method outperformed regression and conventional NN-based methods presented in previous studies. The datasets chosen for the experiments were Stanford Sentiment Treebank and Chinese Valence Arousal Text.

Hassan and Mahmood (2017) proposed a neural network architecture that employed CNN and LSTM on top of pre-trained word vectors. In the model, LSTM was used as a substitute for pooling layers in CNN to reduce the loss of detailed information and to capture long dependencies in sentences. Though convolutional can also capture long dependencies, it will require many layers. The Stanford Large Movie

Review Dataset IMBD and Stanford Sentiment Treebank were used for training and testing the deep learning model.

Y. Wang et al. (2016) proposed attention-based LSTMs for aspect-level sentiment classification. The hypothesis presented was that the sentiment polarity of a sentence was not only determined by the content but also highly related to the aspect concerned. Thus, the proposed model can concentrate on different parts of the sentence when different aspects are present and let aspects participate in computing attention weights for aspect-level sentiment classification. The experiment was performed on the SemEval 2014 Task 4 corpus, which consisted of customer reviews, and each review had a list of aspects and corresponding polarities.

T. Chen et al. (2017) proposed a divide-and-conquer approach that used neural networks to classify the sentences into three types according to the number of targets present in that sentence before performing the sentiment analysis. The approach used a bidirectional long short-term memory (BiLSTM) layer and a Conditional Random Field (CRF) to extract the targets from the review sentences and classify them as a non-target sentence, target sentence, and multi-target sentence. Each group of sentences is then fed into a one-dimensional convolutional neural network separately for sentiment classification. For training the BiLSTM-CRF, the dataset used comprised of news articles manually annotated with opinion target at the phrase-level. The approach is tested for sentiment classification with CNN on movie reviews, Stanford Sentiment Treebank, and customer reviews of 5 digital products containing 3771 sentences extracted from Amazon.com.

Shrestha and Nasoz (2019) used a Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU) and learned low-dimensional vector representation using paragraph vectors and product embeddings of the reviews. Fixed length feature

vectors were obtained for reviews using paragraph vectors and were grouped by products and sorted in temporal order. RNN with GRU was then trained on feature vectors. Product embeddings generated from the penultimate layer of RNN were concatenated with feature vectors and used to train an SVM for sentiment classification of product reviews. The model utilized both the semantic relationship of text and product information. Two sets of experiments were performed to compare the approach of using product embedding with paragraph vectors and using only paragraph vectors. The approach using paragraph vectors with product embeddings obtained using RNN yielded better results.

B. Chen, Huang, Chen, Cheng, & Chen (2019) used deep neural network for multi-class sentiment classification. They leveraged the benefits of both CNN and LSTM, by using two layers of CNN followed by LSTM layers. For multi-class sentiment classification, one-vs.-rest training mechanism was used. This involved the training of single classifier for each category. A drawback of this approach is that the model needs to be trained as many times as the number of categories to obtain the output.

2.4 Cross-Domain Sentiment Classification

The lack of annotated data for training the sentiment classifier motivated researchers to look for ways where labelled data from other domains can be used for training the classifier for the target domain. Manual annotation of data for every domain before training is very time consuming and labour intensive. Therefore, researchers proposed cross-domain sentiment classification, which helped classify the reviews for the domains where annotated data is insufficient or not available. Researchers explored different variations of machine learning and deep learning

approaches based on features and similarity for domain adaptation in cross-domain sentiment classification. Datasets used in various studies consisted of customers reviews about various products. Each product is considered as a domain.

2.4.1 Domain Adaptation using Feature-based Techniques

The initial approaches for domain adaptation mainly consisted of feature-based transfer techniques. Blitzer, Dredze, and Pereira (2007) proposed an algorithm that used structural correspondence learning (SCL) (Blitzer, McDonald, & Pereira, 2006), along with mutual information for classifying the reviews across domains. SCL found correspondences among different features from both source and target domain using the correlations with pivot features and non-pivot features through labelled data from the source domain and unlabeled data from both source and target domains. Pivots were the frequently occurring words in the source and the target domain. Among the shared features, the ones with the highest mutual information to the source label were selected to be used for sentiment classification. The results obtained using SCL-MI and datasets were used by many studies as state-of-the-art results for comparison.

Blitzer, Dredze, and Pereira (2007) used a general low-dimensional cross-domain representation based on co-occurrences of domain-specific and domain-independent features using mutual information. However, if the sentiment expressed in the source and target domains differed significantly, then the adaption performance might decline and lead to negative transfer (Pan & Yang, 2010).To overcome it, Pan et al. (2010) used Spectral Feature Alignment (SFA) algorithm for aligning domain-specific words from different domains into unified clusters using domain-independent words. They built a bipartite graph between domain-specific and domain-independent features and created clusters to reduce the gap between domain-specific words of the two domains, which was helpful in training the classifier for the target domain.

Another solution to the negative transfer was suggested by Li et al. (2012), which utilized an active learning approach for cross-domain sentiment classification. In the proposed approach, two individual classifiers were trained with labelled data from the source and target domains, respectively. Informative samples were selected by leveraging Query by Committee (QBC) (Freund, Seung, Shamir, & Tishby, 1997)sample selection and combination-based classifiers. This approach used the Amazon product review dataset (Blitzer et al., 2007) for experiments.

Sentiment sensitive thesaurus for cross-domain sentiment classification was proposed by Bollegala, Weir, and Carroll (2013). The sentiment sensitive thesaurus was created using labelled data from multiple source domains and unlabeled data from both the source domain and target domain. The thesaurus was used to expand feature vectors during the training and testing of a binary classifier. They used the Amazon review dataset (Blitzer et al., 2007) for four domains, i.e. Books, Electronics, DVD and Kitchen, with balanced positive and negative reviews from each domain.

Using the heterogeneous domains like Amazon reviews and TripAdvisor reviews, Bisio et al. (2013) proposed an integrated approach using sentiment-oriented metric distance to adjust the relative weights allocated to different terms, enabling for a semantic-driven Mahalanobis distance in the feature space using contextual valence shifters and WordNet-Affect, to define a feature space representing reviews. Contextual valence shifters caused shift in the original sentiment present in the lexical element. Negative shifters were used as they could flip the valence of a term of complete sentence. WordNet assigned affective labels to nouns, verbs, adjectives, and adverbs connected to sets of synonyms (synsets). The affective-labels that expressed emotional valence were used. For finding the polarity of a new review, the distance metric was used to identify in the training set the closest reviews to the new review