OPTIMIZING BERTSNN TO ENHANCE SOURCE-TARGET DOMAIN SIMILARITY SCORING FOR CROSS-DOMAIN SENTIMENT CLASSIFICATION OF PRODUCT REVIEWS

***Haitao Zhao*[1] *and Jasy Liew Suet Yan*[2*]**

[1, 2] School of Computer Sciences, Universiti Sains Malaysia,
11800 Penang, Malaysia

Emails: andrewzhao@student.usm.my[1], jasyliew@usm.my[2*]

ABSTRACT

Cross-domain sentiment analysis (CDSA) predicts sentiment polarity in a target domain using knowledge from source domains but existing CDSA methods lack effective source domain selection strategies. This study investigates BertSNN, which combines pre-trained BERT embeddings, a Siamese neural network, and various distance metrics to measure domain similarity and optimize source domain selection for CDSA. First, we experiment with document-level (DocBERT) and sentence-level (SentenceBERT) embeddings with BiLSTM and BiLSTM + CNN neural network configurations to identify the best combination for BertSNN. Second, we explore two distance metrics—Euclidean and Manhattan—alongside shifted cosine similarity to determine the most effective choice for domain similarity scoring. Using product reviews, we test on 25 target domains, examining whether using multiple top most similar source domains improves cross-domain sentiment classification compared to a single most similar source domain. Results indicate that document-level embeddings, BiLSTM and shifted cosine similarity produce the most optimal BertSNN that can select high-quality similar source domains to train a cross-domain sentiment classifier for a target domain, beating two other traditional baseline methods (i.e., bag-of-words and TF-IDF representations). Our findings also show that using top five most similar source domains (k = 5) for training generally improves cross-domain sentiment classification performance as opposed to using a single most similar source domain (k = 1). This study contributes to CDSA by advancing the understanding of embedding choices and distance metrics within a Siamese neural network for source-target domain similarity scoring and providing actionable insights on domain selection strategies to improve sentiment analysis models.

Keywords: Sentiment classification; Product review; Domain similarity; BERT embeddings; Siamese neural network; Similarity metric; Cross-domain sentiment analysis.

1.0 INTRODUCTION

Cross-domain sentiment analysis (CDSA) is essential for understanding public opinion, with applications ranging from monitoring social media trends [1] to assessing customer satisfaction through product reviews [2], [3]. The diversity and complexity of language across domains often hinder the effectiveness of sentiment analysis models, particularly in low-resource domains where labeled training data is scarce. CDSA addresses this challenge by leveraging models trained on data from related source domains (i.e., resource-rich domains) to predict sentiment in a target domain [4].

Identifying domain similarity is a critical step in CDSA, as selecting appropriate source domains can significantly improve model performance and generalizability on a target domain. Current methods for assessing source-target domain similarity often treat the task as a text similarity problem, using metrics to compute lexical, syntactic, or semantic resemblance between domain representations. However, these methods typically focus on single aspects of similarity, overlooking nuanced and domain-specific contextual language variations [5], [6]. Although recent deep learning advancements have improved text similarity modeling for short texts [7], the application of such methods to collections of long-form product reviews remains underexplored [8].

This study builds upon our prior work on BertSNN [9], a BERT-based Siamese neural network for scoring domain similarity and selecting source domains. Our primary objective is to optimize BertSNN by exploring different combinations of BERT representations, neural network architures and distance metrics. We expand the scope of cross-domain sentiment classification experiments to 25 Amazon product review target domains, significantly enhancing the evaluation framework. Additionally, we investigate whether using the top five most similar source domains improves classification performance compared to relying on a single most similar source domain.

This study contributes to CDSA by advancing domain similarity modeling with document-level embeddings, exploring diverse distance metrics for domain similarity scoring, and examining the value of multi-source domain selection strategies. Beyond its academic contributions, BertSNN also offers practical applications for businesses. By leveraging data from similar high-resource domains, BertSNN can improve sentiment analysis in low-resource domains, enabling more effective customer feedback analysis and product review aggregation. These advancements can support better decision-making and enhance customer satisfaction.

2.0 RELATED WORK

2.1 Measuring Text Similarity

Text similarity measurement is a key technique in Natural Language Processing (NLP) for assessing how closely two text units resemble each other. It plays a crucial role in various NLP applications such as search engines, text classification, and question-answering systems. Similarity can be evaluated at different levels—word, sentence structure, and overall meaning. To improve these evaluations, a range of methods have been developed, including basic techniques like character sequence comparison and word set overlap, as well as more advanced approaches involving semantic understanding. The goal of this section is to provide an overview of recent advancements in measuring text similarity, particularly in the use of BERT and Siamese Neural Network (SNN) in a more general context not limited to only CDSA.

One cutting-edge method for measuring text similarity is Semantic Textual Similarity (STS), which seeks to go beyond simple word matching to uncover the deeper meanings of text. STS leverages external knowledge sources such as dictionaries or semantic networks to refine similarity assessments. In particular, the use of neural networks, such as those developed by [10], has significantly advanced text similarity measurement through a Siamese Recurrent Neural Network (RNN) with bidirectional LSTMs and contrastive loss, thus effectively learning text similarity metrics for applications like job title normalization. This model handled semantic invariances such as typos and synonyms by projecting variable-length strings into a fixed-dimensional embedding space, outperforming traditional n-gram models.

Further advancements in deep learning have refined these approaches, with models like [11] utilizing Siamese neural networks equipped with GRU, BiLSTM, and attention mechanisms to improve the handling of smaller datasets and text variability. The GRU-based model demonstrated superior performance on the Sentences Involving Compositional Knowledge (SICK) [12] and STS2017 [13] datasets, providing a robust method for measuring STS. Similarly, [14] introduced Sentence-BERT (SBERT), a model that integrated BERT with a Siamese network and pooling strategies to generate fixed-size sentence embeddings. SBERT outperformed traditional BERT and other methods, offering faster similarity searches and higher accuracy on various STS tasks.

As the field progressed, models like the BERT-based Siamese Semantic Model (BSSM) introduced by [15] further improved text matching by using BERT for text encoding, followed by attention-based interaction layers. This approach provided higher accuracy and efficiency in semantic similarity tasks compared to earlier models such as DSSM [16] and ARC-II [17]. Additionally, [18] proposed the Siamese Multiplicative LSTM (MuLSTM), which combined mLSTM with a Siamese architecture to project sentence embeddings into a fixed-dimensional space. MuLSTM showed superior performance across multiple STS datasets, excelling in both Pearson and Spearman correlations, as well as Mean Squared Error (MSE).

Traditional approaches compress semantic information into two-dimensional vectors, leading to the loss of hierarchical details crucial for effective downstream modeling Most recently, [19] introduced a 3D Siamese network that mapped data into a higher-dimensional space, incorporating adaptive feature extraction, spatial and feature attention, and receptive field modules. This model outperformed traditional models like Sentence-BERT on several benchmarks, offering improved scalability and robustness for large-scale text similarity tasks.

Incorporating these sophisticated models into the process of measuring text similarity, especially within the framework of Semantic Textual Similarity (STS), has enhanced the ability of models to capture complex semantic relationships between texts. These advancements continue to push the boundaries of NLP, allowing for a deeper understanding of text relationships and improving the performance of applications such as search engines, recommendation systems, and machine translation, but yet to be thoroughly explored in CDSA.

2.2 Measuring Domain Similarity

The pursuit of effective cross-domain sentiment analysis (CDSA) has inspired diverse methods to measure domain similarity, facilitating knowledge transfer to enhance target domain performance. Traditional approaches often focused on lexical-level metrics, such as Jensen-Shannon divergence on unigram distributions to quantify source-target relationships [20]. Advanced metrics, such as Chameleon Words Similarity and Entropy Change, measured alignment in word polarity and entropy shifts across domains, outperforming baselines in selecting informative instances for target domains [21]. Sentiment-sensitive thesauri captured co-occurrences of lexical and sentiment elements to enhance feature vectors in CDSA [22], while UniSent applied word embeddings to detect domain drift across languages, achieving significant improvements [23]. On the other hand, [24] proposed EmbLexChange, an unsupervised method that utilized word embeddings to identify temporal lexical-semantic variations. Pivot words were selected based on frequency and contextual consistency, and the embeddings of the pivot words were compared across time frames using cosine similarity to detect semantic shifts.

Semantic methods have also advanced CDSA. For example, autoencoder representations, combined with similarity metrics, outperformed traditional distributions and embeddings in multi-tiered data selection strategies [25]. Proxy A distance iteratively selected similar source subsets, yielding superior results on datasets such as tweets and reviews. Sentence embeddings like Universal Sentence Encoder (USE) exceled in embedding text across domains, further enhancing CDSA through effective integration with similarity features [26]. Similarly, Canonical Correlation Analysis (CCA) aligned embedding spaces by maximizing correlations, achieving high performance in domain similarity tasks across datasets [27]. TCMS-Stack measured domain similarity with transfer covariance matrices, achieving state-of-the-art accuracy on Chinese emotion corpora [28].

DistanceNet [29] was introduced to incorporate distance-based metrics into an auxiliary loss function to enhance unsupervised domain adaptation. DistanceNet employed a multi-armed bandit controller for dynamic multi-source domain switching, mitigating source-target disparities and achieving superior sentiment analysis across multiple domains. This method demonstrated significant performance improvements over competing baselines, emphasizing its effectiveness in multi-domain scenarios.

Building on these foundations, [30] proposed a novel methodology for addressing the cold-start problem in sentiment analysis, wherein labeled data could be scarce in new domains. The method integrated semantic (ccLDA), syntactic (POS patterns), and lexical features into a unified cosine similarity measure to effectively quantify source-target relationships for improving training data selection. This multi-faceted strategy significantly enhanced model performance in low-resource domains, demonstrating the potential of feature integration in domain adaptation.

Efforts to incorporate deep learning into domain similarity have yielded promising results but faced challenges in explainability. For instance, Bayesian optimization has been used to learn optimal data selection measures, outperforming traditional approaches but required further validation across domains [31]. Multi-Source Domain Adaptation with Joint Learning (MDAJL) applied CNNs, BiGRUs and Class Refinement Maximum Mean Discrepancy (CRMMD) for cross-domain sentiment classification. Although effective, such black-box models obscure the degree of source-target domain similarity, thus limiting interpretability.

Despite advancements, obtaining accurate domain similarity scores often relies on simple methods that inadequately account for contextual and semantic differences. Current methods lack a comprehensive framework for domain similarity assessment in CDSA. In summary, while traditional methods like lexical metrics and embedding-based techniques have laid a strong foundation for domain similarity measurement, recent innovations in distance-based metrics, feature integration, and deep learning architectures are advancing CDSA. Approaches such as pre-trained embedding methods and 3D Siamese networks exemplify the potential of combining traditional methods with deep learning to achieve more precise, scalable, and interpretable solutions for cross-domain sentiment analysis. These advancements address critical limitations in current methods and pave the way for more effective and efficient source-target domain similarity matching techniques. In this study, we explore how to optimally construct a domain similarity measurement model using a combination of BERT and SNN trained on general text similarity data but applied specifically for domain similarity scoring.

3.0 METHODOLOGY

The methodology comprises two parts: 1) BertSNN domain similarity scoring model to identify the most similar source domain as the training set for a target domain, and 2) cross domain sentiment classification model to evaluate the performance of the source domain identified by BertSNN as training data for sentiment classification in the target domain.

3.1 Datasets

We utilize the Semantic Textual Similarity (STS) dataset [32], [33], [34], [35], [36], [13] to train our Siamese neural network model for domain similarity scoring. Sourced from the Multilingual Text Embedding Benchmark

(MTEB) [37], the STS dataset consists of sentence pairs annotated with human-judged similarity scores ranging from 0 to 5. It encompasses diverse textual sources such as image captions, news headlines, and user forums, providing a robust foundation for training and testing models in semantic textual similarity tasks.

As shown in Table 1, the dataset originally includes 2,230 sentence pairs for training from STS12 and 12,800 sentence pairs for testing across STS12 to STS17. This division, as provided by the MTEB, reflects the standard setup. However, for our experiment, we combine the original training and testing sentence pairs (15,030) into a single dataset and then split the dataset based on the train-to-test ratio of 9:1, resulting in a more balanced and comprehensive distribution across our training and validation sets. This adjustment reduces population differences from different STS sets to enable more effective model training and hyperparameter tuning. Our final BertSNN model is trained with the most optimal hyperparameter settings leveraging all 15,030 sentence pairs. BertSNN then takes long product review documents, representing a domain, as input and outputs domain similarity scores for source-target domain pairs.

Table 1: STS subsets selected from the STS dataset to train a domain similarity scoring model (n/a indicates not applicable).

Subset Name	**Train Amount**	**Test Amount**	**Similarity Score Range**
mteb/sts12-sts	2,230	3,110	0-5
mteb/sts13-sts	n/a	1,500	0-5
mteb/sts14-sts	n/a	3,750	0-5
mteb/sts15-sts	n/a	3,000	0-5
mteb/sts16-sts	n/a	1,190	0-5
mteb/sts17-sts	n/a	250	0-5
TOTAL	2,230	12,800	0-5

Table 2: Class distribution of product reviews in MDSD labeled with positive or negative sentiment from 25 domains.

Domain Name	**Positive *Amount***	**Negative *Amount***	**Total**
AP: apparel	1000	1000	2000
AU: automotive	584	152	736
BB: baby	1000	900	1900
BE: beauty	1000	493	1493
BK: books	1000	1000	2000
CP: camera & photo	1000	999	1999
CS: cell phones & service	639	384	1023
CV: computer & video games	1000	458	1458
DV: dvd	1000	1000	2000
EL: electronics	1000	1000	2000
GF: gourmet food	1000	208	1208
GR: grocery	1000	352	1352
HP: health & personal care	1000	1000	2000
JW: jewelry & watches	1000	292	1292
KH: kitchen & housewares	1000	1000	2000
MZ: magazines	1000	970	1970
MU: music	1000	1000	2000
MI: musical instruments	284	48	332
OP: office products	367	64	431
OL: outdoor living	1000	327	1327
SW: software	1000	915	1915

SO: sports & outdoors	1000	1000	2000
TH: tools & hardware	98	14	112
TG: toys & games	1000	1000	2000
VI: video	1000	1000	2000
TOTAL	**21972**	**16576**	**38548**

For cross-domain sentiment classification, we utilize all 25 distinct domains (i.e., product types from Amazon) available in the Multi-Domain Sentiment Dataset (MDSD) [2]. Each domain contains product reviews labeled as either positive or negative as detailed in Table 2. Among these domains, nearly half exhibit a balanced sentiment class distribution, while others show a skewed distribution. For instance, domains such as *apparel* [AP] and *toys & games* [TG] are evenly distributed, each containing 1,000 positive and 1,000 negative reviews. In contrast, domains like *musical instruments* [MI] and *tools & hardware* [TH] are highly imbalanced, with significantly fewer negative reviews.

Overall, the dataset comprises 21,972 positive reviews and 16,576 negative reviews, totaling up to 38,548 reviews. This broad dataset captures a wide range of sentiment patterns across diverse domains, offering a robust foundation for training and evaluating cross-domain sentiment classification models. By leveraging all 25 domains as the source domains and target domains, we aim to ensure comprehensive coverage and realistic representation of domain variability in the cross-domain sentiment classification task.

3.2 BertSNN Domain Similarity Scoring Model Architecture

BertSNN adopts a Siamese neural network architecture augmented with a pre-trained BERT model to produce high-dimensional text embeddings, as depicted in Fig. 1. Designed to extract semantic patterns from STS sentence pairs, the model is adapted for domain-level similarity scoring on the MDSD product review dataset. The pre-trained BERT model encodes input text into three-dimensional embeddings with dimensions (1, N, 768), where '1' represents the batch size, 'N' is the token length, and '768' denotes the feature dimensionality. These embeddings encapsulate both local and global contextual word relationships, enabling the network to capture intricate semantic nuances effectively.

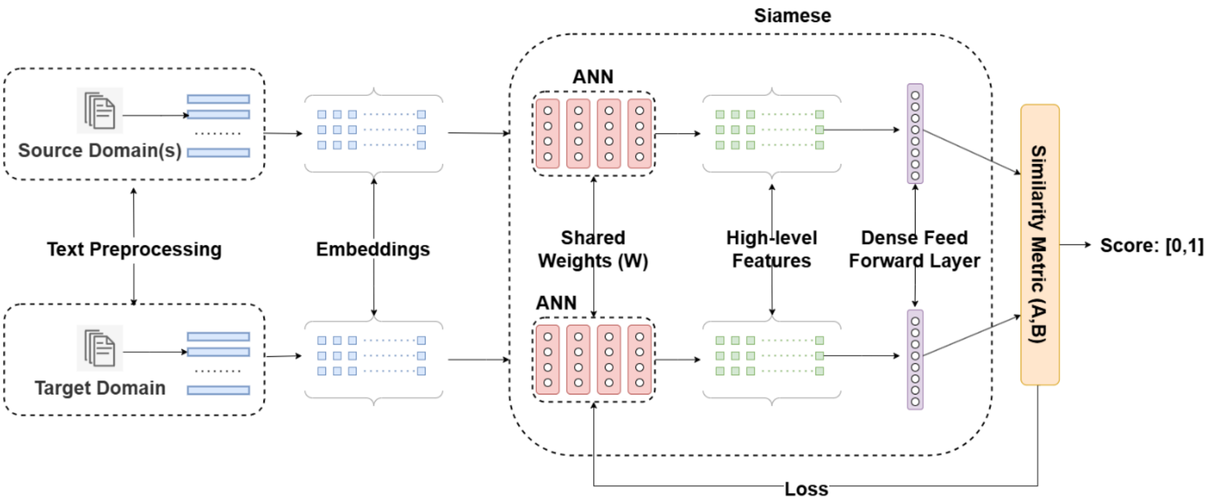

Fig. 1: BertSNN domain similarity scoring model using Siamese network architecture.

The Siamese neural network architecture consists of two parallel branches with shared weights, ensuring symmetric learning during training. This symmetry is essential for consistent feature extraction and reliable comparisons between inputs. Input embeddings are processed through a four-layer structure. The input layer captures semantic-rich representations generated by BERT, while the first hidden layer enhances these features for domain similarity assessment. The subsequent hidden layers refine the features further to prepare for the final source-target comparison. In the output layer, embeddings from the two branches are passed through a dense feed-forward layer, which computes a similarity score. Any similarity metric ranging from 0 to 1 can be used to compute the similarity score. Additionally, any distance metric, which is the inverse of a similarity metric can also be used for the similarity score computation as the model is capable of employing a distance-to-similarity transformation. For example, if Euclidean distance is selected in the similarity metric component, the network first calculates the Euclidean distance between the embeddings, which is then converted into a similarity score ranging from 0 to 1 by applying a scaling function. Lower distance values correspond to higher similarity scores, reflecting closer semantic alignment between source-target domains.

BertSNN is specifically engineered to handle three-dimensional embeddings, making it capable of processing longer texts such as MDSD reviews, which often exceed 510 tokens (excluding special tokens like [CLS] and [SEP]). This is achieved by stacking 3D tensor blocks (1,Sequence_Length,768), enabling the network to accommodate diverse sentence lengths and structures. This adaptability ensures robustness across varied textual datasets, positioning the model as a robust tool for domain similarity scoring.

The training process incorporates a bidirectional LSTM module with an input size of 768, a hidden size of 64, and four layers, augmented by a dropout rate of 0.5 to reduce overfitting. The model is optimized using the Adam optimizer with a learning rate of 0.00001 across 50 epochs, employing mean squared error as the loss function. Training is conducted in batches of 64, with early stopping applied based on validation loss to prevent overtraining. By optimizing a similarity metric, the network learns to represent and compare text embeddings accurately, capturing subtle semantic relationships across different domains. This carefully designed architecture enables the model to perform detailed and reliable domain similarity evaluations, even with complex and lengthy text inputs.

3.3 Cross-Domain Sentiment Classification Model

Our cross-domain sentiment classification model, illustrated in Fig. 2, is a neural network designed to effectively analyze sentiment across diverse domains. The architecture integrates XLNet for feature extraction, leveraging its ability to capture deep semantic representations of text. XLNet processes input text to produce contextualized embeddings, which are subsequently fed into a Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM, composed of four layers with each hidden layer comprising 128 neurons, captures sequential dependencies and sentiment nuances embedded within the order of words and phrases. This configuration enables the model to handle complex textual patterns and contextual information critical for sentiment analysis.

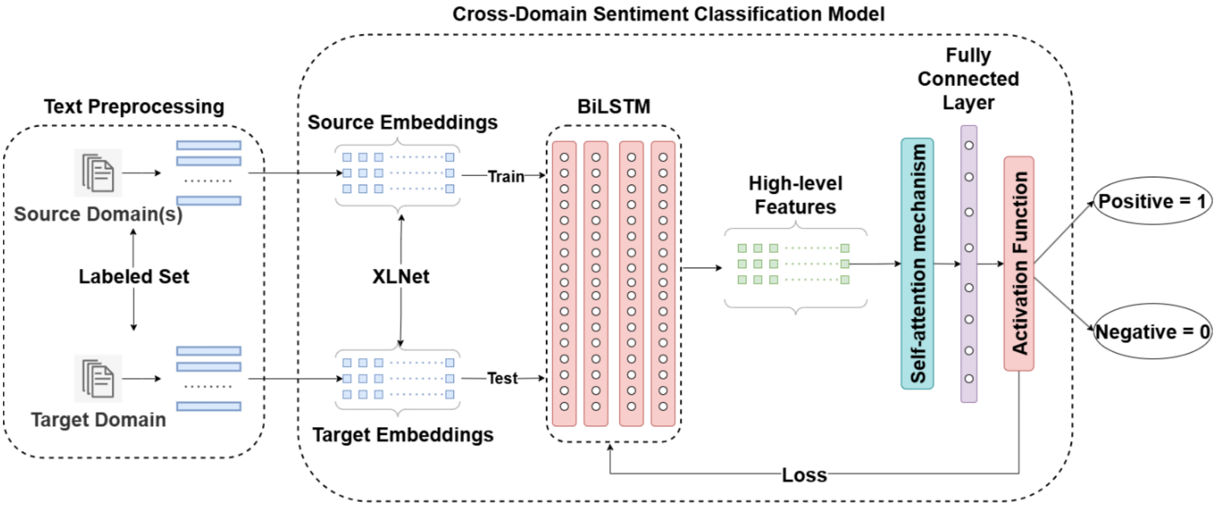

Fig. 2. Cross-domain sentiment classification model.

To further enhance feature extraction, the architecture incorporates a multi-head self-attention mechanism with four attention heads. This mechanism allows the model to prioritize relevant segments of text, focusing on key phrases and terms that signal sentiment, even within lengthy or complex reviews. Additionally, a dropout rate of 0.3 is employed throughout the model to mitigate overfitting, ensuring robust performance across diverse datasets.

The output from the BiLSTM and attention layers is passed through a fully connected layer that produces two logits, corresponding to the sentiment classes (positive and negative). These logits are then processed by a softmax activation function, converting them into probability values that indicate the likelihood of the input text belonging to each sentiment class. This probability-based approach ensures class probabilities for each model prediction sums up to 1.

The model is trained using the Adam optimizer with a notably low learning rate of 0.00001 to ensure stable convergence. Training is conducted over five epochs with a batch size of 32, striking a balance between computational efficiency and model accuracy. This carefully designed architecture combines the strengths of

XLNet, BiLSTM, and multi-head attention for cross-domain sentiment classification to maximize the classifier's capability to capture sentiment across a wide variety of domains and textual contexts.

3.4 Experiments

Our experiments are designed to evaluate and optimize the performance of BertSNN in a cross-domain similarity scoring and sentiment classification pipeline. The setup comprises two interconnected tasks: the upstream task involves the domain similarity scoring model, and the downstream task focuses on the cross-domain sentiment classification model. For each target domain, the BertSNN similarity scoring model processes reviews from a source domain through one input branch and reviews from the target domain through the other branch to compute a similarity score for each source-target domain pair. The similarity score ranges from 0 to 1, where higher values indicate greater similarity. The top k (i.e., k indicates the number of desired domains) most similar source domains to a target domain are then selected to train the cross-domain sentiment classification model and then evaluated using all samples in the target domain. The primary performance metric for evaluation is the cross-domain sentiment classification accuracy score, which serves as a proxy to assess the effectiveness of the domain similarity scoring model in identifying the most similar domain(s) to the target domain across the 25 domains in the dataset.

To investigate the impact of various architectural and methodological choices, our experiments include several key comparisons in the domain similarity scoring task.

- **Pre-trained Embeddings**: We compare the performance of two pre-trained embedding models, SentenceBERT [14] and DocBERT [38], as feature extractors within BertSNN. SentenceBERT generates sentence-level embeddings, whereas DocBERT provides document-level embeddings. This comparison helps determine the text representation best suited for capturing the semantic nuances needed for domain similarity scoring.

- **Neural Network Architectures**: We assess the effect of incorporating different neural network layers by evaluating two configurations: (a) a BiLSTM-based architecture (4 BiLSTM layers) and (b) a hybrid BiLSTM + CNN architecture (4 BiLSTM layers + 1 CNN layer). The BiLSTM layers capture sequential dependencies and long-term relationships in text, while the CNN layer in the hybrid model extracts local features and create a hierarchy of representations for the BiLSTM layers to process. This combination leverages both local and sequential information, potentially enhancing the model's ability to represent and compare text from different domains comprehensively.

- **Similarity or Distance Metrics**: Our default similarity metric in BertSNN is shifted cosine similarity. To refine the similarity scoring process and optimize model performance, we explore two other distance metrics within BertSNN with the most optimal text embedding and neural network architecture combination: 1) Euclidean distance, and 2) Manhattan distance. Each metric introduces a distinct perspective on how similarity is computed between source-target domain representations. Shifted cosine similarity measures the angular difference between vectors, emphasizing directionality, while Euclidean and Manhattan distances focus on spatial differences, emphasizing magnitude and path length, respectively. This comparison identifies the most effective metric for quantifying domain similarity.

- **Baselines**: Two baseline models based on Bag-of-Words (BOW) and TF-IDF representations are implemented for comparison with the best BertSNN model. For both baseline models, product review text from each domain undergoes preprocessing, including stop-word removal, punctuation elimination, and lemmatization using NLTK. A joint vocabulary is created for each source-target domain pair to ensure aligned vector dimensions. Each product review is then transformed into a document vector, and the aggregated vector for each domain is obtained by summing all document vectors within that domain. Finally, cosine similarity is computed between source and target domain vectors. These baseline models provide a traditional, interpretable framework for domain similarity evaluation.

We select the best BertSNN model using cross-domain sentiment classification accuracy on only the a single most similar source domain (k = 1) to a given target domain to make experiments more feasible to run. The overall evaluations are designed to comprehensively analyze the effectiveness of BertSNN across various configurations and baselines, providing critical insights that optimize its architecture and performance. By integrating both classical and neural approaches, the experiments establish BertSNN's efficacy not only as a robust framework for domain similarity scoring but also for cross-domain sentiment analysis, demonstrating its advantages over classical text similarity baseline models.

In the cross-domain sentiment classification task, we extend our experiments to compare the sentiment classifier's performance based on top k most similar source domains identified by BertSNN to a target domain.

- **Single Source Domain (SSD) versus Multiple Source Domains (MSD)**: To extend the evaluation, we compare the cross-domain sentiment classification performance using the reviews from a single most similar source domain (k = 1) against the top five most similar source domains (k = 5) for each target domain across 25 domains. This comparison highlights the impact of leveraging information from multiple source domains on downstream sentiment classification performance.

4.0 RESULTS AND DISCUSSION

For each target domain, the top k most similar source domains are identified from the remaining 24 candidate domains, as outlined in Table 2. We divide the discussion of our findings into four parts. Section 4.1 highlights the optimal BertSNN architecture by exploring combinations of pre-trained language models for text embeddings (SentenceBERT and DocBERT) and neural network configurations (BiLSTM and BiLSTM + CNN). Section 4.2 analyzes the role of distance metrics, comparing shifted cosine similarity to Euclidean distance and Manhattan distance, to refine similarity computations within BertSNN. These experiments determine the configuration best suited for BertSNN in source-target domain similarity scoring. Section 4.3 then benchmarks the best BertSNN model against two classical baselines, Bag-of-Words (BOW) and TF-IDF. Finally, Section 4.4 examines cross-domain sentiment classification performance using one single most similar source domain (SSD, k = 1) versus multiple source domains (MSD, k = 5), providing insights into the impact of leveraging additional source domains in training the sentiment classifier for each target domain.

4.1 BertSNN Architecture Configuration

Table 3 shows sentiment classification accuracy based on a single most similar source domain identified by different combinations of text embeddings and neural network configurations. Shifted cosine similarity is used as the default similarity metric in this set of experiments.

DocBERT + BiLSTM achieves the highest accuracy in *kitchen & housewares* [KH] (92.90%), *grocery* [GR] (93.12%) and *office products* [OP] (94.66%), which reflect the model's adaptability to transfer learning from source domains with both balanced (e.g., *health & personal care* as the source domain for *grocery* and *baby* as the source domain for *kitchen & housewares*, with roughly equal positive and negative reviews) and imbalanced (e.g., *jewelry & watches* as the source domain for *office products*, with more positive reviews) class distributions to the target domain. The robustness of BiLSTM combined with DocBERT's ability to capture global context, enables it to handle both balanced and skewed source class distributions effectively during training. In imbalanced source domains, it likely benefits from its capacity to emphasize the minority class through sequential dependencies. Overall, DocBERT + BiLSTM tends to avoid picking the top most similar source domain with low frequency, which makes the accuracy scores more stable across the 25 target domains, and outshines the others in 8 target domains.

DocBERT + CNN-BiLSTM outshines others only in *baby* [BB] (92.05%) and *tools & hardware* [TH] (92.86%) but obtained poor accuracy scores in 5 target domains due to its tendency of picking source domains with low frequency for model training. We observe similar findings from SentenceBERT + CNN-BiLSTM, wherein mean accuracy is drastically pulled down because of 4 target domains respectively selecting a source domain with low frequency. On the other hand, SentenceBERT + BiLSTM also outshines others only in *automotive* [AU] (91.44%) and *toys & games* [TG] (91.30%). Although its overall performance is still consistent across all the target domains, its limitations in handling long documents cause it to fall short in comparison with DocBERT + BiLSTM. SentenceBERT is optimized for sentence-level representations, so it may fail to capture the holistic context required for analyzing longer, more complex product reviews. DocBERT's ability to capture document-level contextual information allows it to generalize better across a variety of domains.

Notably, DocBERT + BiLSTM achieves the highest mean accuracy of 89.89% across all 25 target domains, outperforming all other configurations. We can conclude that DocBERT +BiLSTM produces the best combination in finding the most similar source domain to a target domain. Therefore, DocBERT + BiLSTM is chosen as the most optimal configuration for our BertSNN model.

Table 3: CDSA accuracy scores of BertSNN based on four combinations of text embeddings and neural network configurations (bracket shows the top one most similar source domain to the target domain).

Target Domain	**Accuracy (%)**			
	SentBert		**DocBert**	
	BiLSTM	**CNN-BiLSTM**	**BiLSTM**	**CNN-BiLSTM**
AP	**92.65 (SO)**	**92.65 (SO)**	92.30 (BB)	92.30 (BB)
AU	**91.44 (SW)**	88.59 (EL)	90.49 (CV)	88.45 (VI)
BB	87.58 (TG)	52.42 (MI)	91.32 (EL)	**92.05 (SO)**
BE	89.95 (CP)	**91.63 (SO)**	90.35 (DV)	86.20 (JW)
BK	87.40 (SW)	51.20 (OP)	**91.75 (DV)**	**91.75 (DV)**
CP	91.65 (EL)	91.65 (EL)	**92.15 (SO)**	88.59 (TG)
CS	91.10 (SO)	**91.69 (CP)**	81.43 (TG)	84.07 (MU)
CV	84.98 (KH)	89.64 (SW)	83.26 (BE)	68.31 (MI)
DV	86.65 (CV)	90.95 (BB)	**91.55 (BK)**	88.55 (HP)
EL	91.35 (CP)	90.00 (HP)	**91.80 (SO)**	49.95 (MI)
GF	89.81 (KH)	**92.05 (JW)**	80.79 (TH)	90.48 (VI)
GR	92.16 (BE)	90.98 (JW)	**93.12 (HP)**	92.23 (BB)
HP	**91.70 (EL)**	**91.70 (EL)**	89.55 (GR)	49.50 (TH)
JW	91.49 (EL)	89.94 (MZ)	**92.18 (BB)**	90.02 (KH)
KH	92.35 (SO)	**92.90 (BB)**	**92.90 (BB)**	86.85 (JW)
MZ	82.59 (GF)	87.01 (CS)	86.90 (GR)	50.61 (MI)
MU	**89.75 (MZ)**	63.50 (AU)	**89.75 (MZ)**	89.30 (TG)
MI	85.24 (AU)	83.73 (TG)	88.55 (HP)	93.37 (CV)
OP	90.02 (HP)	88.63 (BK)	**94.66 (JW)**	87.01 (AU)
OL	84.70 (MZ)	92.16 (SO)	**91.41 (GR)**	**91.41 (GR)**
SW	91.85 (SO)	91.59 (CV)	**91.96 (EL)**	84.60 (TG)
SO	**90.75 (AP)**	**90.75 (AP)**	90.45 (CP)	50.10 (TH)
TH	80.36 (AU)	85.71 (BB)	87.50 (GF)	**92.86 (SO)**
TG	**91.30 (BB)**	49.95 (MI)	89.60 (DV)	89.70 (MU)
VI	86.55 (CV)	86.45 (SW)	**91.50 (DV)**	82.25 (GF)
Mean	89.01	84.3	**89.89**	82.02

4.2 Assessing Distance Metrics for BertSNN

Table 4 shows accuracy scores from the cross-domain sentiment classification model trained on a single most similar source domain from BertSNN (DocBERT + BiLSTM) based on shifted cosine similarity against two distance metrics: 1) Manhattan distance, and 2) Euclidean distance. Our default BertSNN model uses shifted cosine similarity as the similarity metric but we attempt to optimize BertSNN based on other choices of distance metrics to quantify domain similarity across 25 domains.

Shifted cosine similarity, which measures angular differences between domain representations, consistently achieves the highest mean accuracy (89.89%) across all the domains. Its focus on directionality allows it to effectively capture alignment in sentiment patterns, particularly in domains such as *office products* [OP] (94.66%) and *grocery* [GR] (93.12%). Shifted cosine similarity performs well even in domains in which distance metrics are weaker in such as *jewelry & watches* [JW] (92.18%) and *music* [MU] (89.75%). Shifted cosine similarity is a clear winner in 11 target domains.

Manhattan distance, emphasizing path-length spatial differences, achieves slightly lower overall performance with a mean accuracy of 88.98%. Nevertheless, it excels in certain domains such as *health & personal care* [HP] (91.80%), *automotive* [AU] (92.35%), and *software* [SW] (92.43%), thus indicating magnitude differences in embeddings to be informative. By computing the sum of absolute differences across all dimensions, Manhattan distance emphasizes equal contributions from all features while avoiding the squaring effect of Euclidean distance. This characteristic reduces its sensitivity to outliers and makes it particularly effective for high-dimensional data such as text embeddings. Manhattan distance has demonstrated superior performance for only a handful of domains (i.e., 5 target domains).

Euclidean distance, which focuses on overall spatial magnitude differences, produces the lowest mean accuracy (87.82%) among the metrics. However, it has demonstrated strength in specific domains such as *musical instruments* [MI] (93.98%) and *cell phones & service* [CS] (91.10%). As a metric measuring the straight-line distance between data points in a multidimensional space, Euclidean distance inherently emphasizes the overall magnitude of differences. However, the Euclidean distance's sensitivity to large deviations may have limited its effectiveness in this high-dimensional space, where embeddings are less dense and potentially unevenly distributed. Its limitations become evident in domains with more complex or overlapping sentiment patterns, wherein directionality appears to be more critical than magnitude. Euclidean distance is only a clear winner for 3 target domains.

Overall, shifted cosine similarity emerges as the most effective similarity metric for BertSNN, offering superior performance across a majority of the 25 domains. Thus, we conclude that the most optimal BertSNN is composed of DocBERT + BiLSTM with shifted cosine similarity, which is henchforth used in all our subsequent experiments.

Table 4: CDSA accuracy scores of BertSNN (DocBERT + BiLSTM) based on shifted cosine similarity against two distance metrics (bracket shows the top one most similar source domain to the target domain).

Target Domain	**Accuracy (%)**		
	Cosine	**Manhattan**	**Euclidean**
AP	**92.30 (BB)**	91.56 (GF)	84.65 (SW)
AU	90.49 (CV)	**92.35 (BE)**	88.59 (EL)
BB	**91.32 (EL)**	90.76 (KH)	**91.32 (EL)**
BE	**90.35 (DV)**	89.29 (GR)	86.2 (SW)
BK	**91.75 (DV)**	89.15 (GR)	87.40 (SW)
CP	**92.15 (SO)**	87.64 (SW)	87.64 (SW)
CS	81.43 (TG)	87.92 (GR)	**91.10 (SO)**
CV	83.26 (BE)	**89.30 (EL)**	84.64 (BB)
DV	**91.55 (BK)**	91.35 (CP)	86.05 (KH)
EL	**91.80 (SO)**	86.65 (CV)	90.95 (BB)
GF	80.79 (TH)	87.74 (TG)	**90.81 (SO)**
GR	**93.12 (HP)**	85.43 (MI)	92.16 (BE)
HP	89.55 (GR)	**91.80 (BE)**	90.65 (MZ)
JW	**92.18 (BB)**	91.35 (CS)	75.23 (OP)
KH	**92.90 (BB)**	**92.90 (BB)**	89.00 (DV)
MZ	**86.90 (GR)**	**86.90 (GR)**	81.93 (OL)
MU	**89.75 (MZ)**	83.75 (CP)	81.50 (GF)
MI	88.55 (HP)	88.10 (GR)	**93.98 (SW)**
OP	**94.66 (JW)**	88.63 (BK)	**94.66 (JW)**
OL	**91.41 (GR)**	87.77 (EL)	84.70 (MZ)
SW	91.96 (EL)	**92.43 (CP)**	88.62 (MZ)
SO	**90.45 (CP)**	86.02 (JW)	84.65 (GF)
TH	**87.50 (GF)**	84.45 (CP)	**87.50 (GF)**
TG	89.60 (DV)	**91.25 (GF)**	90.00 (EL)
VI	**91.50 (DV)**	90.05 (BK)	**91.50 (DV)**
Mean	**89.89**	88.98	87.82

4.3 Comparing BertSNN Against Baselines

Table 5 summarizes the accuracy scores from the cross-domain sentiment classification model trained on a single most similar source domain from our most optimal BertSNN model (DocBERT + BiLSTM) against two other traditional baseline methods: 1) TF-IDF, and 2) Bag-of-Words (BOW).

Table 5: CDSA accuracy scores of BertSNN (DocBERT + BiLSTM) against two domain similarity scoring baselines (bracket shows the top one most similar source domain to the target domain).

Target Domain	**Accuracy (%)**		
	BertSNN	**TF-IDF**	**BOW**
AP	**92.30 (BB)**	84.45 (GF)	92.25 (HP)
AU	**90.49 (CV)**	87.77 (BE)	89.67 (HP)
BB	**91.32 (EL)**	90.79 (KH)	89.26 (MZ)
BE	90.35 (DV)	**90.76 (GR)**	86.20 (SW)
BK	**91.75 (DV)**	51.20 (OP)	**91.75 (DV)**
CP	**92.15 (SO)**	87.64 (SW)	91.65 (EL)
CS	81.43 (TG)	**86.02 (GR)**	69.31 (AU)
CV	83.26 (BE)	**89.30 (EL)**	84.64 (BB)
DV	**91.55 (BK)**	91.35 (CP)	90.65 (BK)
EL	**91.80 (SO)**	86.65 (CV)	88.55 (HP)
GF	80.79 (TH)	85.43 (TG)	**92.14 (GR)**
GR	**93.12 (HP)**	91.35 (CS)	90.53 (GF)
HP	89.55 (GR)	**91.80 (BE)**	89.55 (GR)
JW	**92.18 (BB)**	91.25 (CS)	75.23 (OP)
KH	**92.90 (BB)**	92.90 (BB)	49.75 (TH)
MZ	**86.90 (GR)**	**86.90 (GR)**	**86.90 (GR)**
MU	89.75 (MZ)	88.10 (CP)	**91.00 (BK)**
MI	**88.55 (HP)**	87.05 (GR)	80.72 (OP)
OP	**94.66 (JW)**	88.63 (BK)	83.76 (TG)
OL	91.41 (GR)	**91.56 (EL)**	88.92 (BB)
SW	91.96 (EL)	**92.43 (CP)**	83.34 (GF)
SO	**90.45 (CP)**	89.05 (CS)	88.75 (KH)
TH	87.50 (GF)	89.29 (CP)	**94.64 (SW)**
TG	**89.60 (DV)**	83.75 (GF)	50.50 (OP)

VI	**91.50** **(DV)**	49.35 (TH)	90.05 (BK)
Mean	**89.89**	85.79	84.39

BertSNN demonstrates superior performance across most domains, with its mean accuracy of 89.89% outperforming TF-IDF (85.79%) and BOW (84.39%). The result emphasizes the effectiveness of the contextualized embeddings in BertSNN, which are better equipped to capture the semantic and syntactic nuances of the text across domains. BertSNN consistently excels in a majority of the target domains (i.e., 14 target domains). The baseline models, while trailing in overall performance, offered competitive results in specific target domains. TF-IDF is a clear winner in only 6 target domains and generally outperforms BOW. BOW, on the other hand, exhibits exceptional performance only in 3 target domains. The result demonstrates that traditional representations may still be advantageous in scenarios where simple lexical patterns dominate. Therefore, we can conclude that CDSA performance increases with the use of text embeddings that are semantically-richer for domain similarity scoring, an observation that is consistent with [25].

Our results also reveal challenges in 3 target domains in which BertSNN yields comparatively the lowest performance. For instance, BertSNN selected *beauty* [BE] as the top one most similar domain to *computer & video games* [CV] and CDSA's low accuracy (83.26%) could be explained by the counterintuitive source-target domain match. The two other counterintuitive source-target domain matches are observed for *gourmet food* [GF] and *tools & hardware* [TH]. These two source-target domain matches are symmetrical, which indicate these two domains are ill-matched as being "similar" by BertSNN as evidenced by CDSA's low accuracy scores. One possible explanation could be these domains contain domain-specific words that are out-of-vocabulary (OOV) words in the pre-trained BERT models, thus causing the similarity matching between domains to be dominated by domain-independent words.

Overall, our results support the hypothesis that contextualized embeddings, as employed in BertSNN, are more effective for cross-domain sentiment classification than traditional vector-based methods (BOW and TF-IDF), an observation that is similar with [21] and [25], which examined other contextualized embeddings (e.g., ELMo and autoencoder representations). The ability of BertSNN to adapt to domain shifts and capture deeper relationships between domains underlines its potential for real-world applications. However, the performance gaps observed in certain challenging domains indicate the need for additional research into finetuning contextualized models for domain-specific adaptation.

4.4 Single Source Domain (SSD) Versus Multiple Source Domains (MSD)

Our earlier experiments utilize only a single most similar source domain (k = 1) selected by BertSNN for cross-domain sentiment classification model evaluation. We extend our experiments to test if selecting a greater number of top most similar source domains to train the cross-domain sentiment classifier for a target domain would yield better performance. From Table 6, the evaluation of cross-domain sentiment classification using a single most similar source domain (k=1) and the top five most similar source domains (k=5) reveals several important insights into the impact of leveraging multiple source domains. Across the 25 target domains, the results consistently demonstrate that incorporating information from multiple similar source domains enhances performance, as evidenced by an mean accuracy improvement from 89.89% (k=1) to 92.77% (k=5).

The most significant gains are observed in the challenging domains, whereby the model performance on these target domains using only a single most similar source domain shows poor performance because of counterintuitive source-target domain matches. Apparently, the effect of these "similarity" mismatches could be reduced by adding more similar source domains to train the CDSA classifier for a target domain. For example, in *computer & video games* [CV], accuracy increases dramatically from 83.26% (k = 1) to 90.88% (k = 5), an improvement of 7.62%. Similarly, *gourmet food* [GF] benefits from an increase in accuracy from 80.79% (k = 1) to 92.30% (k = 5), an improvement of 11.51%. These findings emphasize the benefit of leveraging multiple similar source domains to capture diverse sentiment patterns, especially in complex or less predictable domains.

Even in high-performing target domains, leveraging multiple source domains further boosts accuracy albeit having smaller increments. For instance, in *baby* [BB], accuracy improves from 91.32% (k = 1) to 94.37% (k = 5), while *grocery* [GR] increases from 93.12% (k = 1) to 94.82% (k = 5). These results highlight adding more similar source domains (i.e., 5 similar source domains) continues to add value in model training for a target domain, even when a single most similar domain already provides high performance.

Interestingly, certain domains exhibit minimal or no improvement. One such example is *outdoor living* [OL], whereby accuracy remains constant at 91.41% between k = 1 and k = 5. This suggests that additional source domains contribute little to none when a single most similar domain is already highly representative of the target domain. Unexpectedly, a slight accuracy decrease is observed in *office products* [OP], from 94.66% (k = 1) to 92.58% (k = 5). This may indicate possible redundancy or noise introduced by unrelated source domains in this specific case.

Finally, our results strongly suggest that using k = 5 is generally more effective than k = 1 for cross-domain sentiment classification. The mean accuracy improvement of 2.88% across all domains highlights the benefit of utilizing diverse similar source domains to enhance robustness and accuracy. However, careful domain selection is crucial to avoid potential issues, such as unwanted noise, which may arise as source domains become more dissimilar to a target domain. We found the most optimal k is not necessarily the same across all target domains, an interesting discovery that would require further research in the future.

Table 6: CDSA accuracy scores between SDD (k = 1) and MSD (k = 5) based on BertSNN.

Target Domain	**Accuracy (%)**	
	BertSNN (DocBert + BiLSTM + Cosine)	
	K = 1	**K = 5**
AP	92.30	**93.45**
AU	90.49	**90.90**
BB	91.32	**94.37**
BE	90.35	**93.30**
BK	91.75	**93.05**
CP	92.15	**93.85**
CS	81.43	**93.35**
CV	83.26	**90.88**
DV	91.80	**94.35**
EL	91.55	**92.90**
GF	80.79	**92.30**
GR	93.12	**94.82**
HP	89.55	**93.05**
JW	92.18	**93.50**
KH	92.90	**93.00**
MZ	86.90	**91.12**
MU	89.75	**91.70**
MI	88.55	**93.37**
OP	**94.66**	92.58
OL	**91.41**	**91.41**
SW	91.96	**93.21**
SO	90.45	**93.30**
TH	87.50	**91.07**
TG	89.60	**92.20**
VI	91.50	**92.25**
Mean	89.89	**92.77**

5.0 CONCLUSION AND FUTURE WORK

Our study highlights the efficacy of leveraging a Siamese neural network architecture for selecting the most similar source domain(s) to improve cross-domain sentiment classification performance. Our best performing BertSNN model, which integrates document-level embeddings (DocBERT) with BiLSTM and shifted cosine similarity, achieves a mean accuracy of 89.89% across 25 target domains, significantly outperforming traditional baseline methods such as Bag-of-Words (BOW) and TF-IDF, by at least 4%. This underscores the advantages of contextual embeddings combined with sequential pattern recognition in capturing complex linguistic features and semantic relationships, which traditional methods struggle to address.

Our experimental results reveal the versatility of BertSNN, which performed consistently well across a variety of domains including the challenging ones. Additionally, the model demonstrates robust adaptability when trained with source domains of varying class distributions, further affirming its suitability for diverse real-world scenarios. These findings demonstrate BertSNN's ability to significantly enhance CDSA by selecting optimal source domains for knowledge transfer to a target domain. Our study employs the performance metric from a specific classification task (i.e., cross-domain sentiment classification) as a proxy to measure BertSNN's domain similarity scoring performance. A more direct and task-independent metric to evaluate BertSNN's domain similarity scoring performance could be developed in the future.

Future work can also explore the capability of BertSNN in several possible directions. First, incorporating hybrid models that fuse BiLSTM with attention mechanisms or transformer-based architectures may enhance the model's ability to capture long-range dependencies and improve domain similarity scoring. Second, performing systematic experiments to uncover the most optimal number of similar source domains (k) to select based on a sizeable set of target domains could contribute to the development a set of heuristics or best practices for practitioners. Third, expanding the evaluation of BertSNN to a wider variety of data sources will help test the generalizability of BertSNN across a greater number of complex, real-world scenarios. For instance, future work could test BertSNN trained on product reviews to analyze sentiment on social media posts revolving around mentions of particular product domains.

ACKNOWLEDGMENT

This research is funded by the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3.

REFERENCES

[1] Q. A. Xu, V. Chang, and C. Jayne, “A Systematic Review of Social Media-based Sentiment Analysis: Emerging Trends and Challenges”, *Decision Analytics Journal*, vol. 3, p. 100073, Jun. 2022, https://doi.org/ 10.1016/j.dajour.2022.100073.

[2] J. Blitzer, M. Dredze, and F. Pereira, “Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification”, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, A. Zaenen and A. van den Bosch, Eds., Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 440–447.

[3] R. Bose, R. K. Dey, S. Roy, and D. Sarddar, “Sentiment Analysis on Online Product Reviews”, in *Information and Communication Technology for Sustainable Development*, M. Tuba, S. Akashe, and A. Joshi, Eds., in Advances in Intelligent Systems and Computing, vol. 933, Singapore: Springer Singapore, 2020, pp. 559–569. https://doi.org/ 10.1007/978-981-13-7166-0_56.

[4] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, “State of the Art: A Review of Sentiment Analysis based on Sequential Transfer Learning”, *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749–780, 2023.

[5] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, “Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review”, *IEEE Access*, vol. 5, pp. 16173–16192, 2017, https://doi.org/ 10.1109/ACCESS.2017.2690342.

[6] N. Singh and U. C. Jaiswal, “Cross Domain Sentiment Analysis Techniques and Challenges: A Survey”, Jul. 31, 2022, *Social Science Research Network, Rochester, NY*: 4292052. https://doi.org/ 10.2139/ssrn.4292052.

[7] T. Kenter and M. de Rijke, “Short Text Similarity with Word Embeddings”, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, in CIKM ’15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1411–1420. https://doi.org/ 10.1145/2806416.2806475.

[8] A. Jha, V. Rakesh, J. Chandrashekar, A. Samavedhi, and C. K. Reddy, “Supervised Contrastive Learning for Interpretable Long-Form Document Matching”, *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 2, pp. 1–17, Apr. 2023, https://doi.org/10.1145/3542822.

[9] H. Zhao and J. S. Y. Liew, "BertSNN: A Domain Similarity Scoring Siamese Neural Network for Cross-Domain Sentiment Analysis", presented at the *16th IEEE International Conference on Knowledge and Systems Engineering (KSE 2024),* Kuala Lumpur, Malaysia: IEEE, 2024.

[10] P. Neculoiu, M. Versteegh, and M. Rotaru, "Learning Text Similarity with Siamese Recurrent Networks", in *Proceedings of the 1st Workshop on Representation Learning for NLP*, P. Blunsom, K. Cho, S. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, and S. W. Yih, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 148–157. https://doi.org/10.18653/v1/W16-1617.

[11] T. Ranasinghe, C. Orasan, and R. Mitkov, "Semantic Textual Similarity with Siamese Neural Networks", in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019),* R. Mitkov and G. Angelova, Eds., Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 1004–1011. https://doi.org/10.26615/978-954-452-056-4_116.

[12] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment", in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014),* P. Nakov and T. Zesch, Eds., Dublin, Ireland: Association for Computational Linguistics, 2014, pp. 1–8. https://doi.org/10.3115/v1/S14-2001.

[13] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds., Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1–14. https://doi.org/10.18653/v1/S17-2001.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

[15] F. Xu, S. Zheng, and Y. Tian, "Bert-based Siamese Network for Semantic Similarity", *Journal of Physics: Conference Series,* vol. 1684, no. 1, p. 012074, Nov. 2020, https://doi.org/10.1088/1742-6596/1684/1/012074.

[16] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data", in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, in CIKM '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 2333–2338. https://doi.org/10.1145/2505515.2505665.

[17] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional Neural Network Architectures for Matching Natural Language Sentences", in *Advances in Neural Information Processing Systems,* Curran Associates, Inc., 2014.

[18] C. Lv, F. Wang, J. Wang, L. Yao, and X. Du, "Siamese Multiplicative LSTM for Semantic Text Similarity", in *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence,* Sanya China: ACM, Dec. 2020, pp. 1–5. https://doi.org/10.1145/3446132.3446160.

[19] J. Zang and H. Liu, "Improving Text Semantic Similarity Modeling Through a 3D Siamese Network", in *ECAI 2023*, vol. 372, in Frontiers in Artificial Intelligence and Applications, vol. 372, IOS Press, 2023, pp. 2970–2977. https://doi.org/10.3233/FAIA230612.

[20] R. Remus, "Domain Adaptation Using Domain Similarity- and Domain Complexity-Based Instance Selection for Cross-Domain Sentiment Analysis", in *2012 IEEE 12th International Conference on Data Mining Workshops,* Brussels, Belgium: IEEE, Dec. 2012, pp. 717–723. https://doi.org/ 10.1109/ICDMW.2012.46.

[21] A. Sheoran, D. Kanojia, A. Joshi, and P. Bhattacharyya, "Recommendation Chart of Domains for Cross-Domain Sentiment Analysis: Findings of A 20 Domain Study", in *Proceedings of the Twelfth Language Resources and Evaluation Conference,* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4982–4990.

[22] D. Bollegala, D. Weir, and J. Carroll, “Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1719–1731, Aug. 2013, https://doi.org/10.1109/TKDE.2012.103.

[23] E. Asgari, F. Braune, B. Roth, C. Ringlstetter, and M. Mofrad, “UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages”, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4113–4120. Accessed: Dec. 14, 2024.

[24] E. Asgari, C. Ringlstetter, and H. Schütze, “Unsupervised Embedding-based Detection of Lexical Semantic Changes”, May 16, 2020, *arXiv*: arXiv:2005.07979. [Online]. Available: http://arxiv.org/abs/2005.07979

[25] S. Ruder, P. Ghaffari, and J. G. Breslin, “Data Selection Strategies for Multi-Domain Sentiment Analysis”, Feb. 08, 2017, *arXiv*: arXiv:1702.02426. [Online]. Available: http://arxiv.org/abs/1702.02426

[26] N. Pogrebnyakov and S. Shaghaghian, “Predicting the Success of Domain Adaptation in Text Similarity”, in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, Online: Association for Computational Linguistics, 2021, pp. 206–212. https://doi.org/ 10.18653/v1/2021.repl4nlp-1.21.

[27] A. Beyer, G. Kauermann, and H. Schütze, “Embedding Space Correlation as a Measure of Domain Similarity”, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 2431–2439.

[28] P. Wei, R. Sagarna, Y. Ke, Y.-S. Ong, and C.-K. Goh, “Source-Target Similarity Modelings for Multi-Source Transfer Gaussian Process Regression”, in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Jul. 2017, pp. 3722–3731.

[29] H. Guo, R. Pasunuru, and M. Bansal, “Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits”, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, Art. no. 05, Apr. 2020, https://doi.org/ 10.1609/aaai.v34i05.6288.

[30] W.-C. Hsiao and H. C. Wang, “Cross-domain Corpus Selection for Cold-start Context”, *Journal of Information Science*, p. 01655515241263283, Jul. 2024, https://doi.org/10.1177/01655515241263283.

[31] S. Ruder and B. Plank, “Learning to Select Data for Transfer Learning with Bayesian Optimization”, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 372–382. https://doi.org/10.18653/v1/D17-1038.

[32] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, “SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity”, in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, and D. Yuret, Eds., Montréal, Canada: Association for Computational Linguistics, Jul. 2012, pp. 385–393.

[33] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, “*SEM 2013 Shared Task: Semantic Textual Similarity”, in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, M. Diab, T. Baldwin, and M. Baroni, Eds., Atlanta, Georgia, USA: Association for Computational Linguistics, Jun. 2013, pp. 32–43.

[34] E. Agirre *et al.*, “SemEval-2014 Task 10: Multilingual Semantic Textual Similarity”, in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, P. Nakov and T. Zesch, Eds., Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 81–91. https://doi.org/10.3115/v1/S14-2010.

[35] E. Agirre *et al.*, “SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability”, in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval*

2015), P. Nakov, T. Zesch, D. Cer, and D. Jurgens, Eds., Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 252–263. https://doi.org/10.18653/v1/S15-2045.

[36] [1] E. Agirre et al., “SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation”, in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, and T. Zesch, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 497–511. https://doi.org/10.18653/v1/S16-1081.

[37] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “MTEB: Massive Text Embedding Benchmark”, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148.

[38] A. Adhikari, A. Ram, R. Tang, and J. Lin, “DocBERT: BERT for Document Classification”, Aug. 22, 2019, *arXiv*: arXiv:1904.08398. [Online]. Available: http://arxiv.org/abs/1904.08398