NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Nanyang Technological University

School of Computer Science and Engineering

Cross-Domain Sentiment Classification with Domain-Adaptive Neural Networks

Neural Networks and Deep Learning

Authors:

Jose Antonio Franca Ibanez - N2304058L
Divyansh Bhutra - U2023744B
Elisia Brispalma Widawati - U2021960L

November 2023

Abstract

In the field of sentiment analysis, accurately interpreting and classifying user-generated content is invaluable for businesses and researchers alike. Sentiment analysis models are typically trained on vast datasets from specific domains. However, the challenge arises when a model, trained on one type of data such as movie reviews, is required to perform with equal accuracy on a different type of data, like restaurant reviews. Our project seeks to bridge this gap in Cross-Domain Sentiment Classification (CDSC) using domain adaptation techniques. This project introduces a novel approach, Adversarial Domain Adaptation (ADA), for enhancing sentiment analysis models' robustness across diverse domains. The proposed methodology addresses the challenge of domain shift by training a domain classifier to distinguish between source and target domains, coupled with an adversarial training strategy. The architecture comprises three key phases: Domain Classifier Training, Adversarial Training, and Sentiment Classifier Training. Experimental results demonstrate the efficacy of ADA in mitigating domain shifts, achieving improved sentiment classification accuracy across disparate datasets without the need for labelled target domain data. ADA presents a promising solution for real-world sentiment analysis applications where training and test domains differ significantly.

1 Introduction

This project centers on the development of a neural network tailored for sentiment analysis, specifically focusing on its adaptability across different domains. In light of recent advancements in deep neural networks and pre-trained language models, sentiment analysis has witnessed significant performance improvements. However, prevailing methodologies are hindered by their reliance on substantial amounts of labelled training data, leading to resource-intensive processes (Socher et al., 2013). To address this challenge, the project endeavours to harness knowledge acquired from a labelled source domain, the IMDb movie review dataset, and extend its application to a distinct target domain, YELP restaurant reviews. Our strategy for domain adaptation entails a two-step process: initially, the model undergoes primary training on a comprehensive IMDb movie review dataset to grasp sentiment expressions in film critiques. To enable domain adaptability, we integrate a balanced subset of YELP restaurant reviews during training, utilising a gradient reversal layer to foster learning domain-independent features. For this purpose, we will design a neural network framework with the following integral components:

- **Feature Extractor**: A component engineered to convert input data into a feature vector. This could be implemented via a convolutional, recurrent neural network, or a transformer-based model such as BERT. In particular, we will use BERT.
- **Sentiment Classifier**: A classifier that deduces sentiment from the feature vector produced by the extractor.
- **Domain Classifier (Discriminator)**: This classifier ascertains the domain of the input data based on the feature vector, distinguishing between source and target domains.

In the initial phase, the discriminator is conditioned to accurately identify domain-specific nuances, fostering a deeper understanding of feature representations from distinct datasets. This clarifies the purpose behind incorporating a portion of YELP data into the training set — to enhance the discriminator's ability to recognize domain-specific features and, by extension, to improve the robustness of the model's domain adaptation capabilities. The critical adversarial training phase employs a gradient reversal strategy, adjusting the feature extractor to confuse the domain classifier. This encourages the generation of domain-agnostic features crucial for consistent performance across varied domains. Simultaneously, sentiment classifier fine-tuning using labelled source domain data ensures the model maintains proficiency in sentiment analysis during domain adaptation.

The primary objective of this project is to develop a neural network capable of adapting knowledge from one domain and effectively applying it to a different domain. Subsequent sections will delve into the intricacies of the architecture, training methodologies, and experimental results, offering insights into the effectiveness of the proposed approach in navigating challenges associated with domain diversity.

2 Literature Review

Previous approaches to cross-domain sentiment classification (CDSC) can be broadly categorised into three groups: traditional baseline methods, deep neural network-based techniques, and pre-trained language models.

Traditional baseline methods such as Structural Correspondence Learning (SCL) rely on manual feature selection to learn a mapping to pivot feature space (Blitzer et al. 2006). The effectiveness of SCL heavily relies on the manual selection of common features, and poor choices in this regard can have a detrimental impact on its performance. Enhancement was subsequently made with the Structured Correspondence Learning-Mutual Information (SCL-MI) to select pivot features based on mutual information between features (unigram or bigram) and source domain labels, improving performance (Blitzer et al., 2007). Nonetheless, SCL-MI may be unsuccessful in identifying pivot features when there is limited correlation between features. Another attempt to bridge the gap between domains was proposed through the Spectral Feature Alignment (SFA) algorithm, which aligns domain-specific words of different domains aided by domain-independent words as a connecting link (Pan et al., 2010). The SFA has the drawback of strongly relying on labelled datasets. Since constructing high-quality large-scale labelled datasets proves to be challenging, deep transfer learning and its capacity for representation learning can markedly alleviate the need for labelled data in the target domain. Deep neural networks (DNNs) also eliminate the need for manual feature engineering given their ability to automatically learn relevant features and adapt to domain shifts.

DNNs possess the remarkable ability to autonomously learn vital low-dimensional features from text. Glorot et al. introduced the Stacked Denoising Autoencoder (SDA) with sparse rectifier units, which is further extended by marginalised denoising autoencoders (mSDA) to achieve better speed and scalability with high-dimensional data (Glorot et al., 2011; Chen et al., 2012). Furthermore, hierarchical attention networks (HAN), hierarchical attention transfer network (HATN), and interactive attention transfer network (IATN) were introduced to improve sentiment classification in a cross-domain context (Yang et al., 2016; Li et al., 2018; Zhang et al., 2019). These approaches excel in capturing intricate contextual relationships, enabling better sentiment classification across domains. Adversarial training methods such as domain-adversarial neural network (DANN) and adversarial memory network (AMN) were introduced to encourage learnt representation to be domain-invariant and later to select relevant features automatically (Ganin et al., 2016; Li et al., 2017). These adversarial techniques address the challenge of domain discrepancy in CDSC by extracting similar distributions of features from the source and target domains. While DNN-based methods have attained results unparalleled to the traditional methods, the transition to pre-trained language models, such as BERT, is even superior.

Pre-trained models started with the computer vision task and have now gained widespread adoption in NLP. In pre-trained language models (LM), lower layers of the network are trained with a vast amount of unlabelled text corpora, after which fine-tuning can be applied according to specific downstream tasks. Notably, Howard and Ruder introduced the Universal Language Model Fine-Tuning (ULMFiT). It employs ASGD Weight-Dropped LSTM (AWD-LSTM) in the LM general domain pre-training, followed by a multi-stage fine-tuning process: target task LM fine-tuning and target task classifier fine-tuning. Additionally, it proposed novel fine-tuning techniques such as discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing to prevent catastrophic forgetting (Howard and Ruder, 2018). Unlike DNN-based models that are trained from scratch and require a substantial amount of labelled data, pre-trained LMs, such as ULMFiT, leverage pretraining on a large and diverse text corpus that captures a wider range of linguistic features and general knowledge. ULMFiT hence established the foundational concept of transfer learning in NLP.

Transfer learning by applying pre-trained LMs to downstream tasks can be divided into two existing strategies: feature-based approach and fine-tuning approach. Feature-based approaches mainly saw pre-trained representations being used as fixed features in a downstream task-oriented architecture. Feature extractors such as Embeddings from Language Models (ELMo) proposed by Peters et al. was designed to generate deep contextualised representations that address the challenge of modelling syntax, semantics, and polysemy (Peters et al., 2018). On the other hand, the fine-tuning approach, as adopted by ULMFiT, entails fine-tuning a pre-trained model on the source domain with a small amount of target domain data. Radford et al. introduced a generative pre-trained transformer (OpenAI GPT) that is able to acquire a strong natural language understanding by generative pre-training of a language model on a diverse corpus of unlabelled text, then discriminative fine-tuning on each specific task (Radford et al., 2018). In a more recent work, Devlin et al. proposed the Bidirectional Encoder Representations from Transformers (BERT) model that augments the power of pre-trained language representations in fine-tuning approaches by pre-training deep bidirectional representations (Devlin et al., 2019). BERT alleviates the unidirectionality constraint by using a pre-training objective called masked language model (MLM) and understands sentence relationships by pre-training on a binary ”next sentence prediction” task (Devlin et al., 2019). A fine-tuned BERT model on the CDSC task was able to obtain a new state-of-the-art outcome (Myagmar et al., 2019). Overall, fine-tuning approaches offer the attractive benefit of requiring minimal learning of parameters from the ground up and minimal task-specific architecture modifications. Therefore, in our study, we select a fine-tuning-based approach of transfer learning for the CDSC task, namely domain adaptation, by fine-tuning a pre-trained BERT model. In our study, we used BERT as a feature extractor, a domain adversarial neural network to make the feature representations from different domains indistinguishable, and performed fine-tuning for sentiment classification.

3 Methodology

Preparing the Dataset:

To establish a robust foundation for our model, we will commence by rigorously preprocessing the training data, drawn from both the IMDB dataset, featuring 50,000 entries, and a carefully selected subset of 10,000 entries from the YELP dataset. It is important to note that the inclusion of YELP data at this stage is strategic, setting the stage for the adversarial training component to be introduced subsequently.

To maintain symmetry, we have ensured that the subset of YELP data comprises an equal number of positive and negative reviews, with 5,000 of each. This deliberate structuring, in conjunction with the already balanced IMDB dataset—comprising 25,000 positive and 25,000 negative reviews—allows us to assemble an equitable training dataset, totalling 60,000 reviews, evenly split between 30,000 positive and 30,000 negative sentiments.

The IMDB dataset categorizes reviews as either positive or negative, whereas the YELP dataset includes numerical star ratings ranging from 1 to 5. To reconcile these differing rating systems, we classify YELP reviews with 4 and 5 stars as positive, and those with 1 and 2 stars as negative. Reviews with a 3-star rating are excluded from this binary classification since they are deemed neutral.

For the testing phase, our approach will diverge from the training methodology; we will select a subset consisting of 20,000 samples from the YELP dataset at random, without imposing the label symmetry constraint employed during training. The rationale behind this decision is to evaluate the model's performance under more realistic, organic conditions that are representative of the actual distribution of sentiments in the dataset.

Before training and testing, we will perform thorough text preprocessing on both datasets. This critical step will include the removal of stopwords, which are frequently occurring words that offer minimal value in understanding sentiment. We will also eliminate special characters that do not contribute to sentiment analysis and apply stemming and lemmatization techniques to reduce words to their root forms, ensuring consistency and improving the model's ability to learn from the textual data effectively.

Gradient Reversal Innovation:

The gradient reversal layer represents a pivotal innovation in domain adaptation techniques. By introducing this layer between the feature extractor and the domain classifier, we actively invert the gradient during the backpropagation process. The layer applies a specific negative constant to the gradient, a process which is counterintuitive at first glance but serves a crucial purpose. This inversion is strategic; it encourages the feature extractor to develop a representation of data that is indistinguishable in terms of the originating domain.

Traditionally, neural networks are adept at learning features that are highly discriminative for the task at hand. However, when the objective is to ensure the generalizability of a model across different domains, the discriminative power of the network can become a hindrance. Features that are highly specific to the source domain may not perform well when the model is applied to the target domain. The gradient reversal layer mitigates this risk by penalizing the network for learning features that are too domain-specific. This forces the feature extractor to prioritize the learning of features that are common across both the source and target domains, effectively making the model's predictions domain-agnostic.

In essence, this innovative approach harmonizes the adversarial objectives of maximizing domain classification error while still retaining the integrity of feature extraction for sentiment analysis. This harmonization is key to our project as it allows us to leverage a model trained on abundant labelled data from one domain and apply it successfully to another domain where labelled data may be scarce or costly to obtain. By adopting this technique, we aim to enhance the versatility of our model and establish a robust framework for CDSC.

Training Methodology:

The training process is meticulously designed to ensure the model learns to generalize across domains while retaining high accuracy in sentiment classification. We will adhere to a structured approach comprising several distinct phases:

- **Domain Classifier Training**: The initial phase involves conditioning the discriminator to accurately distinguish between the feature representations of the source and target domains. The discriminator's goal is to identify whether a given feature vector originates from the IMDB dataset or the YELP dataset, thereby enhancing the model's ability to understand and differentiate domain-specific nuances.
- **Adversarial Training**: In this critical phase, we employ a gradient reversal strategy. The feature extractor, which is responsible for generating domain representations, is adjusted to confuse the discriminator. By reversing the gradient signal from the domain classifier's loss during the backpropagation, the feature extractor is encouraged to produce domain-agnostic features, which are essential for the model to perform consistently.
- **Sentiment Classifier Training**: Concurrent with adversarial training, the sentiment classifier is fine-tuned using the labelled data from the source domain. This step is crucial for the model to maintain a strong understanding of sentiment analysis while undergoing domain adaptation. It ensures that while the model becomes proficient in domain invariance, it does not lose its acumen in discerning sentiments.

4 Experiments and Results

In the presented research, the efficacy of our model is evaluated through a comprehensive set of metrics, including accuracy, precision, and F1-score. The model demonstrates notable performance on the IMDb dataset, achieving an accuracy of 67%. When applied to the YELP dataset without any modifications for domain differences, the model recorded an accuracy of 59%. This performance metric was notably enhanced to 63% following the application of our specialized domain adaptation technique.

Comparatively, the model exhibits superior performance in terms of F1-score when juxtaposed with several established models in the field. Specifically, our model surpasses the F1 scores of DANN (Ganin and Lempitsky, 2015) which scored 43.44, MCD (Saito et al., 2018) at 42.37, JUMBOT (Fatras et al., 2021) with 43.08, ALDA (Chen et al., 2020) at 39.84, and URAM (Wu and Huang, 2022) which scored 45.16. This comparison highlights the effectiveness of our model in achieving a higher level of accuracy in domain-adapted sentiment analysis tasks.

The computational framework for our study was anchored by a MacBook Pro equipped with an M2 Pro chip, serving as a singular node cluster. The initial training phase of the model on the source domain was completed in a span of around 2 hours. Subsequent to this, the domain adaptation phase extended for an additional duration of 7.5 hours.

In summary, the empirical outcomes from our research indicate that the model we developed, enhanced through our novel domain adaptation methodology, demonstrates a robust and adaptable nature. It exhibits a proficient capability in sentiment analysis, seamlessly transitioning across varied domains. Such flexibility is of paramount importance in practical applications of sentiment analysis, particularly in environments where domain-specific labeled data is scarce or unavailable.

5 Conclusion

In this project, we proposed an Adversarial Domain Adaptation (ADA) approach to cross-domain sentiment classification (CDSC). The proposed model's three-phased training process - Domain Classifier Training, Adversarial Training, and Sentiment Classifier Training - proves effective in simultaneously learning domain-invariant features and maintaining sentiment analysis proficiency. The integration of a gradient reversal strategy also ensures the generation of domain-agnostic features critical for consistent performance across different domains.

Despite our constructive findings, there are avenues for further exploration and refinement to improve the model's capabilities and applicability. First, we propose an extended fine-tuning phase with labelled data from the target domain to enhance its predictive accuracy. Using an expanded set of the YELP data, fine-tuning shall enhance the model's acumen in restaurant review contexts. We may even consider gaining additional insights by incorporating 3-star reviews from the YELP dataset, assigning them to either positive or negative sentiment, depending on our objectives. Further improvement may also entail hyperparameter calibration to optimise the trade-off parameter that controls the loss contribution from the domain and sentiment classifiers, crucial in balancing the domain adaptation and sentiment analysis objectives. Additionally, we may utilise more domain adaptation strategies, such as progressive adaptation and dynamic loss weighting. Progressive adaptation by gradually shifting the emphasis from sentiment classification to domain confusion, and dynamic weighting of the domain classifier's loss based on the validation set performance may help the model achieve heightened sensitivity to both source and target domains. Finally, we suggest a semi-supervised expansion, which implements self-training methods utilising unlabelled data from the target domain. By predicting sentiment for the unlabelled data and then incorporating these predictions as if they were true labels, the model learns from its own prediction, making it more robust and adaptable to the nuances of the target domain.

6 References

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, ”Recursive deep models for semantic compositionality over a sentiment treebank,” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631-1642.

J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with structural correspondence learning,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2006, pp. 120-128.

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, 2007, pp. 440-447.

S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 751-760.

X. Glorot, A. Borde, Y. Bengio, ”Domain adaptation for large-scale sentiment classification: a deep learning approach,” in Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 513-520.

M. Chen, Z.E. Xu, K.Q. Weinberger, F. Edu. ”Marginalized denoising autoencoders for domain adaptation,” in Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 1627-1634.

Y. Ganin, E. Ustinova, H. Ajakan, et al., ”Domain-adversarial training of neural networks,” in JMachLearnRes, vol. 17, pp. 2096-2030, 2016.

Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, ”End-to-end adversarial memory network for cross-domain sentiment classification,” in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 2237-2243.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, ”Hierarchical attention networks for document classification,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480-1489.

Z. Li, Y. Wei, Y. Zhang, and Q. Yang, ”Hierarchical attention transfer network for cross-domain sentiment classification,” in Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 5852-5859.

K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, and E. Chen, ”Interactive attention transfer network for cross-domain sentiment classification," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 5773-5780, DOI:10.1609/aaai.v33i01.33015773.

J. Howard, S. Ruder , "Universal language model fine-tuning for text classification," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 328-339, DOI:10.18653/v1/P18-1031.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, arXiv:1802.05365. [Online]. Available: https://arxiv.org/abs/1802.05365

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). Improving Language Understanding by Generative Pre-Training. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/language-unsupervised/language_understanding _paper.pdf

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805

B. Myagmar, J. Li, and S. Kimura, "Transferable high-level representations of BERT for cross-domain sentiment classification," in Proceedings of the International Conference on Artificial Intelligence, 2019, pp. 135-141.

Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proceedings of the 32nd International Conference on Machine Learning, 2015, ICML'15, vol. 37, pp. 1180-1189.

K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723-3732.

K. Fatras, T. Séjourné, N. Courty, and R. Flamary, "Unbalanced minibatch optimal transport: applications to domain adaptation," 2021, CoRR, abs/2103.03606.

M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, pp. 3521–3528.

Y. Wu and X. Huang, "Unsupervised Reinforcement Adaptation for Class-Imbalanced Text Classification," in Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, Seattle, WA, 2022, pp. 311-322.