A Semi-Supervised Approach for Multi-Domain Classification

Anirudh Garg
Voice Intelligence Team
Samsung Research Institute
Bengaluru, India
anirudh.g@samsung.com

Kartikey Singh
Voice Intelligence Team
Samsung Research Institute
Bengaluru, India
kartikey.s@samsung.com

Radhika Mundra
Voice Intelligence Team
Samsung Research Institute
Bengaluru, India
m.radhika@samsung.com

***Abstract*— This research paper presents an innovative approach for addressing the classification of data originating from multiple domains when there is a scarcity of labeled data. While current technologies have achieved impressive accuracy in single-domain classification, the challenge of multi-domain classification persists due to contextual variations. To tackle this issue, we introduce a multi-task based unsupervised data augmentation (UDA) approach that enables learning of domain-specific data contexts. UDA[1] is widely recognized as one of the most effective semi-supervised frameworks, as it requires only a small amount of labeled data for learning purposes. In our study, we leverage a BERT language model and train it using our proposed approach to acquire domain-aware embeddings for data assessment. By doing so, we enhance the ability to classify data from various domains accurately.**

Keywords — Semi-supervised Learning, Multi - domain Classification.

I. INTRODUCTION

Text classification is a fundamental task in natural language processing (NLP) and plays a vital role in numerous applications, such as sentiment analysis, document categorization, and spam filtering. Traditional supervised approaches, such as Support Vector Machines (SVM) and Naïve Bayes, have demonstrated remarkable success when abundant labelled data is available within a single domain. However, when confronted with multiple domains or limited labelled data, these approaches often suffer from poor generalization and performance degradation.

Multi-domain text classification presents a unique set of challenges due to the inherent variations and complexities associated with different domains. Each domain may have its own distinct vocabulary, grammar, and contextual nuances, making it difficult for traditional supervised approaches to generalize across domains. Same text can result in a positive sentiment for some domains, while leading to a negative sentiment for some others. Examples to the same are as shown in Table 1.

One solution to this problem is to maintain different models for different domains as shown in Fig. 1. However, as the number of domains increases, this approach becomes less feasible due to the requirement of training separate models for each domain for classification. Any change in the model architecture would have to be updated in the models corresponding to each domain thus making the process cumbersome and inefficient.

TABLE I

Text	***Common Text***	***Domain***	***User Experien-ce***
This trimmer is easy to use	Easy	Product Review	Positive
The ending of movie is easy to guess		Movie Review	Negative
The delivery package was too fast	Too Fast	Service Review	Positive
The roller coaster was too fast		Product Review	Negative

Similar texts resulting in different sentiments

To address the above challenges, we propose a multi-task based semi-supervised framework in this paper. Semi-supervised learning is a paradigm that exploits both labeled and unlabeled data to improve the performance of machine learning algorithms. In the context of text classification, it enables us to leverage the vast amount of unlabeled data that is often readily available across multiple domains. By combining unsupervised learning techniques, such as consistency training, with supervised learning models, semi-supervised approaches can effectively capture the underlying structures and patterns in the data, leading to improved classification accuracy and robustness.

Our approach learns domain-specific embeddings for each domain, which enables us to train a single model that can be used across multiple domains as shown in Fig 1. This reduces the number of models required, which can save time and resources.

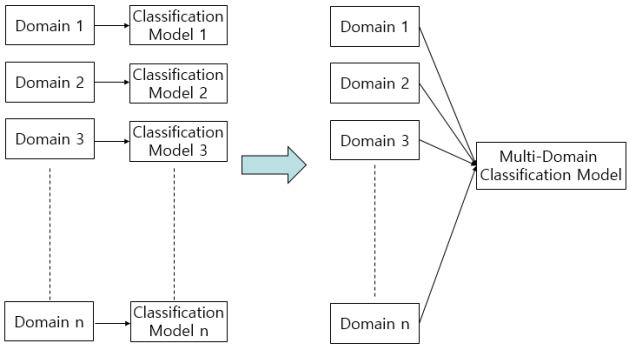

Fig. 1. Transitioning from Individual Classification Models to Multi-Domain Classification Model

Also, we extend the popular semi-supervised approach known as "Unsupervised Data Augmentation" (UDA) [1]. This allows the model to make use of a limited amount of supervised data and a large amount of unsupervised data from multiple domains to learn better representations.

In this paper, we explain our approach in detail and present the results of our experiments with 16-domain amazon product review data, which demonstrates the superiority of our method. Finally, we discuss the implications of our findings and suggest future directions for research in this area.

II. Related Work

A popular approach for multi-domain text classification is Multi-Task Learning (MTL) (Caruana, 1997) [2]. Several works employ deep learning for multi-task learning such as (Zhang et al., 2014;) [3], (Liu et al., 2016b) [4]. LSTMs, CNNs and memory networks have also been used in Chen and Cardie (2018) [5], Liu et al. (2018) [6] and Li et al. (2017) [7].

Since the advent of attention mechanisms (Vaswani et al., 2017) [8], several more works applying attention over words and sentences have tackled the problem of multi-domain classification such as Yang et al. (2016) [9], Zheng et al. (2018) [10], Yuan et al. (2018) [11].

Contextual information via generated word embeddings of language models such as ELMo (Peters et al., 2018) [12] and BERT (Devlin et al., 2018) [13] has been useful in approaching this problem too. Motivated and inspired with these works, we use the output of the BERT architecture to capture the domain and sentiment label of a text input.

III. Proposed Solution

The primary objective of our proposed MT-BERT model is to enhance its classification performance by making the model domain aware. This is particularly crucial when dealing with data that may have different labels across various domains. To achieve this, we train the model to predict both the domain and the classification labels simultaneously. By incorporating the domain prediction task, we empower the model to differentiate and treat similar data instances from different domains accordingly.

By training the MT-BERT model with parallel domain prediction, we aim to overcome the challenge of context variations across domains. This approach allows the model to capture the nuanced differences in data patterns specific to each domain, thus improving its ability to classify diverse data accurately. Ultimately, our goal is to optimize the model's classification performance across multiple domains by leveraging domain awareness as a crucial aspect of the learning process.

IV. Architecture in Detail

A. Few-Shot Learning

In multi-domain classification often, there is a paucity of labelled data (supervised data) corresponding to each domain. The supervised dataset has to be created in a long and tiresome manual process where the annotator labels each data point after careful consideration and observation. Further, the annotator has to ensure that the supervised data captures as much of the variations in the data as possible so as to make the model to be trained robust.

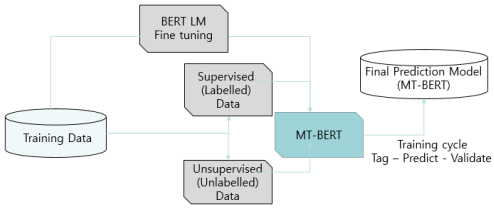

Fig. 2. MT-BERT Training Cycle

Thus, it is essential to devise a method which tackles the task of multi-domain classification requiring as little annotated data as possible. Fig. 2. depicts the ingredients required to train the MT-BERT model. To accomplish this we use the popular unsupervised data augmentation (UDA) [1] architecture. The UDA architecture significantly reduces the manual effort required in annotation by effectively utilizing the unlabelled data (unsupervised data) for consistency training. Therefore, to train the model, a much smaller amount of supervised data along with a large amount of unsupervised data is employed, thus ensuring that we utilize the maximum proportion of the training data available with least manual effort.

B. Augmentation

An important component of the UDA architecture are the augmentation techniques used to generate legible, valid and good quality augmentations. Augmentations are essential as they introduce some noise to the raw, original data so as to improve the consistency and generality of the model. Upon trying several popular augmentation techniques, a pipeline consisting of the Parrot paraphrase generator [14] and back - translation, stacked back-to-back revealed to give the best results. The back-translation of the data was done between the English and German languages.

C. Losses

The labelled data refers to a limited quantity of data that has been manually labelled and used as a basis for model predictions. These predicted labels are subsequently compared against the manually assigned labels to calculate a supervised loss. This loss is a cross-entropy loss given by the equation:-

$$L_{sup} = \sum \left[- \left(y_f^* \right) log\left(M_f(x_L) \right) \right] \qquad (1)$$

here, L_{sup} denotes supervised cross-entropy loss, y_f^* denotes true label of the input data, x_L denotes labeled input data, $M_f(x_L)$ denotes model predicted classification label.

The unlabeled data is passed through the augmentation pipeline discussed above and augmented data (data with introduced noise) is obtained. The model predicts upon both the original data and the augmented data with noise. The predictions obtained are then used to determine the unsupervised loss.

$$L_{unsup} = D_{KL} \left(M_f(x_{UL}) \| M_f(\hat{x}_{UL}) \right) \qquad (2)$$

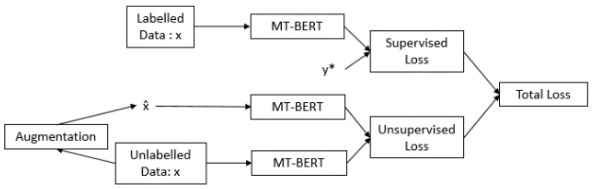

Fig. 3. Training of MT-BERT using the Unsupervised Data Augmentation Architecture

here D_{KL} denotes KL-divergence loss, x_{UL} denotes unlabeled input data, $\hat{x}_{UL}$ denotes augmentation of x_{UL} , $M_f(x_{UL})$ and $M_f(\hat{x}_{UL})$ denotes predicted classification label on unlabeled input and its augmented data respectively. Both the losses, the supervised loss and the unsupervised loss are then added to arrive at a total loss which is then back-propagated to train the model as shown Fig. 3. The supervised loss contributing to the accuracy training, while the unsupervised loss contributing to the consistency training.

Now, the model thus trained within this framework can have its own architecture to predict upon the data input to it. In our method, the model trained is the novel MT-BERT architecture where we predict domain and classification of a data instance in parallel thus enabling it to be domain-aware.

The model consists of a pre-trained BERT model fine-tuned for classification, into which an input data x is fed as shown in Fig. 4. The output of the pre-trained BERT model is then fed into a dropout layer which helps in regularization and reduction of over-fitting. Further, the output from the dropout layer is used to output the domain and label prediction of the input data x via a linear layer.

Both the output predictions are then compared with the actual labels in case of supervised data to calculate cross-entropy loss or the prediction labels from the augmentation counterpart in case of unsupervised data to calculate the KL divergent loss to result into a classification loss and a domain loss.

$$L_{sup} = \sum \left[- \left(y_f^* \right) log\left(M_f(x_L) \right) \right] + \sum \left[\left(y_d^* \right) log\left(M_d(x_L) \right) \right] \quad (3)$$

where L_{sup} denotes combined supervised cross-entropy loss, y_f^* denotes true label for classification, y_d^* denotes true label for domain, x_L denotes supervised input data, $M_f(x_L)$ denotes model predicted classification label, $M_d(x_L)$ denotes model predicted domain label.

$$L_{unsup} = D_{KL}\left(M_f(x_{UL}) \| M_f(\hat{x}_{UL}) \right) + D_{KL}\left(M_d(x_{UL}) \| M_d(\hat{x}_{UL}) \right) \quad (4)$$

where D_{KL} denotes KL-divergence loss, x_{UL} denotes unlabeled input data, $\hat{x}_{UL}$ denotes augmentation of x_{UL} , $M_f(x_{UL})$ and $M_f(\hat{x}_{UL})$ denotes predicted classification on unsupervised input and its augmented data respectively,

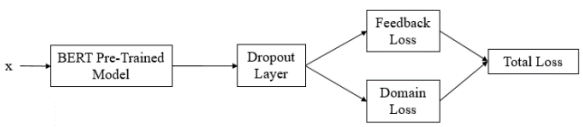

Fig 4 – MT-BERT Training Procedure

$M_d(x_{UL})$ and $M_d(\hat{x}_{UL})$ denotes predicted domain on unlabeled input and its augmented data respectively.

These two losses obtained are then further added to arrive at the total supervised loss or the total unsupervised loss as the case maybe.

$$L_{total} = L_{sup} + L_{unsup} \quad (5)$$

We have used a linear scheduler with the Adam optimizer for training our proposed model. To prevent over-fitting on the annotated text we use a method of confidence-based masking. We mask the text instances on which the model is confident above a certain pre-defined threshold to increase generality and reduce over-fitting.

V. DATASETS

We employ the dataset curated in Adversarial MT learning for text classification (Liu et al. 2017)[15] for our problem. The dataset contains reviews of products from 16 different domains. The details of the dataset are available in Table 2.We use different proportions of training data available while the validation and test sets length remains same (200 and 400 per domain respectively). In addition, we use 25000 unlabeled reviews from across all the 16 - domains to train the model for consistency loss.

TABLE II

Domain	***Train***	***Validation***	***Test***
DVD	1400	200	400
Kitchen	1400	200	400
Health	1400	200	400
Apparel	1400	200	400
Music	1400	200	400
Toys	1400	200	400
Baby Products	1300	200	400
Video	1400	200	400
Software	1315	200	400
Magazine	1370	200	400
Sports	1400	200	400
MR	1400	200	400
IMDB	1400	200	400
Camera	1390	200	400
Books	1400	200	400
Electronics	1398	200	400

Statistics of the Dataset

Fig. 5. Accuracy w.r.t % of supervised data employed

VI. Training

We utilize a vanilla BERT model [Devlin *et al.*, 2018] and fine-tune it to our requirements. The batch-size used is 8 and the ratio of supervised to unsupervised data utilized is 1:3. Adam optimizer was used with the learning rate set at 0.0002. Also, the optimizer was wrapped using a linear scheduler with no warmup steps. Confidence threshold of 0.7 revealed to give best results.

VII. Results

Along with higher accuracy, our objective was to develop a model wherein least amount of annotated data is required without compromising on the accuracy. For this, we experimented on amazon 16-domains dataset [15] with various fractions of the supervised data. Fig 5 depicts the accuracies obtained w.r.t the amount of data used. Table 3 shows the performance of our model on this dataset as compared to state-of-the-art BiLSTM model [16]. It is clear from the table that our model performs better than the BiLSTM model.

Table III

Models	***Accuracy on 16 - Domain Data***
Vanilla BERT	89.1
BiLSTM [16]	90.1
MT-BERT	91.2

Accuracy of MT-BERT model as compared to other models

VIII. Conclusion And Future Scope

Through the proposed semi-supervised architecture, we were able to build a multi-domain classification model with limited amount of annotated data. The results of our experiments proved the superiority of our proposed method over the existing state-of-the-art technologies. A promising avenue for future research using this architecture is in aspect-based classification analysis, where the objective would be to identify the classification towards specific contexts of a domain. By leveraging unsupervised data in these areas, it may be possible to improve the accuracy of classification analysis models further.

IX. Acknowledgements

This work was supported by the Voice Intelligence Team, Samsung Research Institute India, Bengaluru. We thank the anonymous reviewers for their valuable suggestions and comments. We would also like to thank our mentor Javaid Nabi for providing his valuable insights towards this endeavour.

X. References

[1] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, Quoc V. Le, “Unsupervised data augment-tation for consistency training,” in NeurIPS, 2020.

[2] Rich Caruana. Multitask learning. Machine learning, 28(1):41–75, 1997.

[3] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In European conference on computer vision, pages 94–108. Springer, 2014.

[4] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101, 2016.

[5] Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 1226–1240, 2018.

[6] Qi Liu, Yue Zhang, and Jiangming Liu. Learning domain representation for multi-domain sentiment classification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 541–550, 2018.

[7] Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2017), 2017.

[8] Ashish Vaswani, Noam Shazeer, NikiParmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all youneed. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[9] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard 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, pages 1480–1489, 2016.

[10] Renjie Zheng, Junkun Chen, and Xipeng Qiu. Same representation, different attentions: Shareable sentence representation learning from multiple tasks. arXiv preprint arXiv:1804.08139, 2018.

[11] Zhigang Yuan, SixingWu, FangzhaoWu, Junxin Liu, and Yongfeng Huang. Domain attention model for multi-domain sentiment classification. Knowledge-Based Systems, 155:1–10, 2018.

[12] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[14] Prithiviraj Damodaran. Parrot: Paraphrase generation for NLU, 2021

[15] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial MT learning for text classification. 2017

[16] Cai, Yitao, and Xiaojun Wan. "Multi-Domain Sentiment Classification Based on Domain-Aware Embedding and Attention." IJCAI. 2019.