Wireless Personal Communications (2023) 130:43–60
https://doi.org/10.1007/s11277-023-10274-x

Deep Recurrent Neural Model for Multi Domain Sentiment Analysis with Attention Mechanism

Khaled Hamed Alyoubi[1] · Akashdeep Sharma[2]

Accepted: 23 February 2023 / Published online: 22 March 2023

Abstract

The problem of multi-domain sentiment analysis is complex since meaning of words in different domains can be interpreted differently. This paper proposes a deep bi-directional Recurrent Neural Network based sentiment classification system employing attention mechanism for multi-domain classifications. The approach derives domain representation by extracting features related to description of domain from the text using bidirectional recurrent network with attention and feed it to the sentiment classifier along with the processed text using common hidden layers. We experiment with varied types of recurrent networks and propose that implementing the recurrent network with gated recurrent unit ensures that both domain-specific feature extraction and feature sharing for classification can be performed simultaneously and effectively. The evaluation of domain and sentiment modules has been conducted separately and results are encouraging. We found that using gated recurrent unit as bidirectional recurrent network in both modules gives efficient performance as it trains quickly and gives higher validation accuracy for all present domains. The proposed model also demonstrated good results for other metrics when compared with other similar state-of-the-art approaches.

Keywords Sentiment analysis · Domain attention · LSTM · GRU · RNN

1 Introduction

One of the common applications of domain dependent problems in internet research is sentimental analysis of tweets, movie reviews, news headlines or other textual data such as general web content. Sentiment analysis is widely used in varied domains for different purposes and literature mentions notable use of several techniques for this task. Traditional sentiment analysis has been accomplished-based on manual selection of features [1] and also using some

✉ Akashdeep Sharma
akashdeep@pu.ac.in

Khaled Hamed Alyoubi
kalyoubi@kau.edu.sa

[1] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[2] UIET, Panjab University Chandigarh, Chandigarh, India

evolutionary approaches like greedy heuristic [2], genetic algorithms for feature reduction [3] etc. Genetic algorithms has also been widely used in the studies of [4–7] for various other tasks related to feature reduction and optimizations. Sentiment analysis is widely considered as domain specific task and traditional sentiment analysis requires labelled data in all domains for these techniques to work efficiently that makes this problem as computationally intensive. Therefore, traditional techniques are not capable of achieving same levels of efficaciousness in the data which has more than one domain of interest. Segregation of data into different domains is itself a task of classification, resulting in higher degree of computation time and because of which researchers are trying ways for systems development that can work on multiple domains.

Multidomain sentiment analysis has been tried by many researchers recently using variety of ways like random walk-based solution by Xue et al. [8]; proposing ways for extraction of domain features automatically [9–11] and multi domain aspect-based methods [12–15]. Advances in deep learning and recurrent networks have also led to proposal of numerous studies in this field. The use of attention mechanism has also been encouraging in the field of natural language processing for extraction of the important feature for specific words, as available in recently proposed studies [16–18]. In spite of availability of these works, there has been challenges for the effective sentiment classification for multiple domains since most of these methods are based on use of machine learning tools or neural networks for feature selection which has performance issues, vanishing gradient problem and availability of labelled data. Although attention mechanism has paved the way for focussing on relevant features only and seems to be a promising option but its applicability in multiple domains is still relatively unexplored.

This paper proposes a sentiment classification model capable of working on multiple domains employing bi-directional deep recurrent neural network with attention mechanism for effective classification. Bi-directional deep recurrent network has helped to remember the contextual information and selection of independent and specific features from both directions. The attention mechanism in neural network has helped to learn important features of specific words and share information across domains. The hidden layers select domain shared and domain specific features simultaneously and next hidden layers further extracting the sentiments of the given sentence with the combination of its domain obtained from previous layers. The proposed architecture has been envisioned by experimenting various architectures and observing results by adding additional layers, changing the network architecture for domain representation and classification. The problem of vanishing gradients has been overcome by use of Gated Recurrent Networks (GRU) in network bi-directional recurrent network. More details regarding recurrent networks, LSTM and related concepts are given in "Appendix". Readers may refer to the "Appendix" for wider reference to these concepts.

The paper is organized as follows: Sect. 2 details about the related studies in the proposed field with focus on the use of recurrent networks, attention mechanism, deep neural networks, multi-stage networks etc. Section 3 describes the proposed architecture with details of sentiment and domain module described individually. Section 4 explains the experiments and required discussions while Sect. 5 concluded the experiments findings.

2 Related Works

The designing of sentiment analysis system requires extraction of useful features of interest specific to various domains. The attention mechanism is an effective approach for the detection of useful features but its usefulness for multiple domains is still unexplored.

Kim et al. [19] at Microsoft AI and Research Bloomberg LP proposed a domain attention model with an ensemble of experts. They assumed 'N' domain specific intent and slot models that are trained on the separate domains. If given a new domain, the model uses a weighted combination of feedback from the 'N' domain experts along with its own opinion to anticipate the new domain. A dynamic memory networks, proposed by Kumar et al. [20] for processing of natural language, comprises of similar procedure and uses fixed query representation as attention for selecting the relevant features respective to an intended task. However, drawback of the approach was that it was restricted to only single domain scenarios.

The use of recurrent neural networks and attention-based architectures for multi-domain data also founds mention in few works in literature. Chauhan et al. [18] proposed a technique by combining linguistic patterns with bidirectional recurrent network supported by new word embeddings. They proposed fine-tuned word embeddings helped in extraction of domain specific most relevant terms. Attention mechanism was used to capture long-term dependency between specific words. Valid aspect terms were extracted using un-supervised learning and were used to train the proposed bi-LSTM model. Lee et al. [21] proposed word attention-based mechanism to classify negative and positive sentences using CNN based weakly supervised learning algorithm. Word attention mechanism generated class activation map for impactful words and helped in producing sentence level and word level polarity scores using weak labels only.

Liu et al. [22] utilized attention mechanism over a BiLSTM based architecture in order to extract useful low-level information from hidden layers that was combined higher level phrase information extracted using a CNN layer. Fu et al. [23] made extension to an LSTM based architecture by integrating lexicon into word embeddings. Global attention mechanism was used to extract global information from the text which was combined with representational powers of words introduced by lexicon-based word embeddings. Spyridon Kardakis et al. [24] has provided a detailed analysis of various attention based studies used for sentiment analysis.

It is also prevalent from human behaviour that people frequently use some shared expressions or domain specific terms in order to express their feelings that may lead to the training of sentiment classifiers for different domains. Therefore, multi-domain sentiment analysis can be regarded as a particular application of the multi-task learning [25]. Liu et al. [17] proposed a framework for multi-task feature learning where different tasks use similar patterns of sparsity. The efficiency of these methods strongly depends on the assumptions of domain relatedness that may not be relevant in different scenarios. More important work in direction of multistage models employing recurrent networks has been done by Poria et al. [12], Rana and Cheah et al. [13], Wu et al. [14] in which they have proposed extraction of noun or noun phrases using rule-based methods in first stage. These were used in the next stage for training or feature sharing in attention based or bi-LSTM network. Do et al. [15], Yuan et al. [16], Chauhan et al. [18] proposed to use linguistic rules with attention-based approach to increase the extraction of noun phrases. Yu et al. [26] proposed a Wasserstein-based Transfer Network (WTN) to share the domain-invariant information of source and target domains and employed BERT embedding for deep level semantic information of text. Another recent work in the direction of multidomain sentiment analysis has been proposed by Yu et al. [27] by extending analysis over multimedia data and processing internal features of all modalities and correlating these modalities with each other to guage sentiment trends. The proposed recurrent architecture consists of use of BERT for extracting text features, ResNet for Image Features along with multimodal feature fusion.

After deliberating on the available studies available in literature, we infer the presence of a wider scope for technique suitable for multi-domain sentiment analysis using recurrent neural networks. Therefore, this paper proposes a sentiment classification model capable of working on multi-domain classification, employing bi-directional deep recurrent neural network with attention mechanism for effective classification. The methodology has been widely tested and experimented with usage of both Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) cells in different combinations for sentiment and domain analysis to obtain a most optimal combination.

3 Proposed Methodology

Unlike the aforementioned methods, this paper proposes a sentiment classification model for multiple domains employing attention mechanism and Bidirectional Recurrent Neural Network (BRNN) with GRU. More details about these basic building blocks of the model are available in “Appendix 1”. “Appendix 1” projects the working and basic structures of RNN, LSTM, GRU and attention mechanism.

Use of attention mechanism has helped to select independent and specific features simultaneously. The proposed architecture is shown in Fig. 1 and consists of two modules Domain module and Sentiment module both employing a BRNN.

The task of the Domain module is to predict an appropriate domain for learning representations of various domain. An attention selection is triggered by the domain representation for assemblage of most crucial features related to domain in the corresponding sentiment module. The Domain module makes use of a bidirectional GRU network for gaining the most useful representation for domain. A Bidirectional GRU runs the inputs in two modes, one from forward to backward and one from backward to forward direction, making it different from the uni-directional approach in which recurrent runs only backwards. The principle of BRNN is to split the neuron of a regular RNN into two directions, one for positive direction and one for a negative direction and by using two-time directions, input information from the past and the future can be used for the current time frame that has helped to make effective domain prediction.

We had implemented Gated Recurrent Unit(GRU) cells as our recurrent network. GRU has been preferred over LSTM network because it trains quickly and helps in retaining long term information also. We use $\pounds_{\mathrm{GRU}}(.)$ to denote the processing on embeddings.

$$
\begin{aligned}
\theta_d &= \pounds_{GRU^d}\left(X_j^i\right) \\
\theta_d &= \pounds_{GRU^d}\left(w_{x1}, w_{x2}, w_{x3} \ldots w_{xn}\right)
\end{aligned}
\tag{1}
$$

where θ_d is the output of GRU network, w_{xj} represents word representation of jth word in the text.

Since proposed architecture is a bi-directional network therefore θ_d is a combination of GRU networks, therefore θ_d will be represented as,

$$
\begin{aligned}
\theta_d &= \theta_d^{forw} \, \S. \, \theta_d^{back} \\
\theta_d &= \pounds_{GRU^d}^{forw} \left(X_j^i\right) \, \S. \, \pounds_{GRU^d}^{back} \left(X_j^i\right)
\end{aligned}
\tag{2}
$$

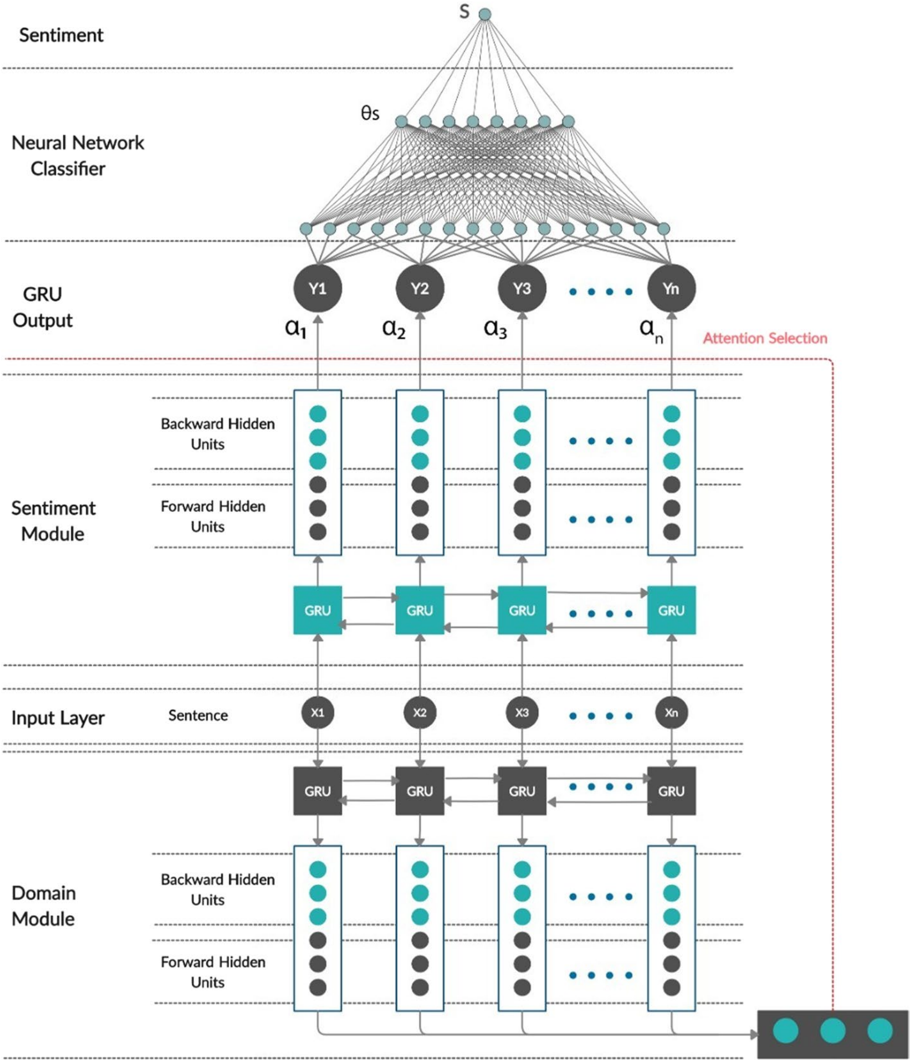


Fig. 1 Proposed architecture for multi-domain classification

where § means concatenation and θ_d is obtained by combining text processing in both forward and backward direction as in Eq. 1.

A soft-max layer at the end has been employed for domain prediction. In the experiments that follows, this label has been used to gauge the domain classification accuracy.

3.1 Sentiment Module

Another attention mechanism based bidirectional recurrent network has been used for sentiment module. The prime idea behind using attention mechanism is that the sentiment module should attend all outputs of the recurrent process, which differentiates it from the domain module. The sentiment module also implements a GRU network and θ_s denotes the outputs of sentiment module network which can be seen as the representation of the details of the text.

The representation of the domain from the domain module is also used and various attention weights are being extracted using a single feed-forward layer in sentiment module. A softmax layer converts all the attention weights to probabilistic attention weights which are then fed to sequence of two fully connected layers and a softmax layer to get final representation of sentiment for prediction of the sentiment.

The proposed network is trained on combination of cross entropy losses for both sentiment prediction and domain prediction. The sentiment module also uses Gated Recurrent Unit with attention mechanism whose output can be explained as follows:-

$$\theta_s = f_{GRU^s} = f_{GRU^s}\left(\left(w_{v1}, w_{v2}, w_{v3} \ldots \ldots .w_{vn}\right)^t\right) \tag{3}$$

where θ_s is the output of GRU network which will be a list of vectors.

Attention weights are learned using Bahdanau's model [56] with feed-forward network as follows,

$$y_i^{att} = f(w^{att}\left(\theta_d §.\ \theta_s^k\right) + b^{att} \tag{4}$$

where w^{att} and b^{att} are parameters for attention mechanism, θ_d, θ_s^k refers to domain representation(annotation) and kth sentiment vector representation.

The attention weights are calculated by normalizing output score of the feed-forward neural network as in Eq. 4

$$\alpha_i = \frac{\exp\left(y_i^{(a)}\right)}{\sum_{i=1}^{n} \exp\left(y_i^{(a)}\right)} \tag{5}$$

The final value for θ_s will be the weighted sum for all the values which will be

$$\theta_s = \sum_{k=1}^{n} \alpha_i^{\theta_s^k} \tag{6}$$

α_i represent attention weights and θ_s is the sentiment text representation.

This is further fed to dense layer and softmax layer for the final sentiment prediction.

3.2 Loss

The loss used is the sum of sentiment and domain prediction loss given by

$$L_0 = \frac{1}{N_k} \sum_{k=1}^{M} \sum_{i=1}^{N_k} L(s_i^k, p(s_i^k|x_i^k)) + \frac{\lambda}{N_k} \sum_{k=1}^{M} \sum_{i=1}^{N_k} L(d_i^k, p(d_i^k|x_i^k)) \tag{7}$$

Equation 7 is the global loss implemented for the proposed architecture, assuming the presence of M domains and N_k samples for kth domain. s_i^k, d_i^k are the true sentiment value and true domain value respectively. The probabilistic output of sentiment and domain module is given by $p(s_i^k|x_i^k)$ and $p(d_i^k|x_i^k)$.

L(.) is the cross-entropy loss function used to calculate the difference between predicted and true label with Adam as optimizer. The first L(.) term refers to the cross entropy loss observed for sentiment and the second L(.) term refers to the domain loss. L_0 is the global loss which is combination of sentiment loss and domain loss. The domain loss was weighted by λ term whose values was set to 0.04 and was extracted after experimentation with multiple values. λ term is a regularization term that control the importance of domain error.

Use of bidirectional recurrent network provides good accuracy across all the domains because bidirectional networks can understand the context better and attention mechanism in the sentiment module helps to focus on those features which differentiates one domain from other. The proposed architecture models the related domains at intrinsic feature-level rather than having separate layers for a single task. The hidden layers select domain specific and domain shared features simultaneously. The following hidden layers extract the sentiments of the given sentence with the combination of its domain obtained from previous layers.

More details are available in the next section where several models for these modules with variations in their design were built and experiments were conducted to have meaningful insights into the architecture. The explanation of the experiments helps to understand the evolvement of the proposed architecture. These experiments lay out the architectures used and the results obtained by each method in details. We conclude by proposing the architectures for multi-domain classification for which the most promising results were obtained.

4 Results and Discussions

This section details about the conducted experiments and observed results in order to derive single model that could work on different domains. All experimentation has been done on the Amazon review dataset [31] that comprises four types of reviews namely Books, Electronics, Kitchen appliances and DVDs. The dataset consists of 2000 reviews for each product with 1000 for positive and negative reviews for each of the mentioned products and a total of 8000 reviews. The reviews were suitably split into train, test and validation sets. These experiments were performed on a workstation with Processor: Intel i7-6850K CPU @ 3.60 GHz, Architecture: x86_64 and 2 Titan Xp-Pascal GPUs.

4.1 Experiment 1

We first experimented with a model inspired by the work of Yuan et al. [14]. The proposed architecture consists of a sentiment module and a domain module for sentiment and domain prediction and employs LSTM as recurrent network. We extracted large number of other parameters for the experiment during the training process, which are tabled in Table 1. These include various kinds of losses as observed on training and validation data. (For more information about parameters used in the table, readers are encouraged to refer to "Appendix 1"). Table 1 indicates significant higher values for validation metrics

Table 1 Parameters comparison for models

Parameter	Exp 1	Exp 2	Exp 3-a	Exp 3-b	Exp 4	Exp 5
Loss	0.0977	0.3634	0.2128	0.1894	0.172	0.0990
s_pred_loss	0.0835	0.3439	0.1902	0.1646	0.1421	0.127
d_pred_loss	0.3550	0.4861	0.5638	0.6212	0.7458	0.3496
s_pred_acc	0.9691	0.8771	0.9252	0.9387	0.9479	0.9708
d_pred_acc	0.8798	0.8236	0.7971	0.7482	0.6652	0.8809
val_loss	0.6086	0.4169	0.4461	0.4670	0.5402	0.4300
val_s_pred_loss	0.6868	0.4008	0.4252	0.4432	0.5112	0.4005
val_d_pred_loss	0.3731	0.4035	0.5229	0.5952	0.7264	0.3619
val_s_pred_a	0.8392	0.8323	0.8418	0.8442	0.8310	0.8397
val_d_pred_a	0.8676	0.8656	0.8185	0.7607	0.6935	0.8684

indicating a scope for further improvement. We further evaluated the model on additional number of parameters like Precision, Recall and F1-score. We further experimented and extracted domain-wise and sentiment-wise separate evaluations which were not part of the original study. These evaluations are listed in Tables 2 and 3 respectively. The obtained results of this experiment indicate that the model has shown almost same levels of accuracy for almost all domains with a marginal advantage to books, as the model was able to understand the context using a strong bidirectional network. The overall sentiment accuracy for both positive and negative sentiments depicts the same results. A receiver operating characteristic is also depicted in Fig. 2a using the test values for deriving true positives and true negatives for more insights into the model.

4.2 Experiment 2

This experiment proposes an extension to the model of previous experiment by incorporating a dense layer before the final layer with number of neurons being equivalent to number of domains. A dense layer is a fully connected layer, as in, all neurons in the previous layer are connected with all neurons in the corresponding layer. These are different from convolutional layers, since weights are reused across different sections of the vectors, whereas a dense layer has a unique weight for every neuron to neuron pair. An appropriate value of dropout is also set for regularization to prevent over-fitting. The usage of such architecture was targeted to improve the accuracy of certain domains because the dense layer might allow smaller number of reviews to get better results in their sentimental analysis. The computation is observed for the effect of incorporation of Dense Layer in the Domain Attention model and results are tabled in Tables 2 and 3. The observed values indicate that addition of dense layer resulted in an increase in values of almost all metrics with considerable decrease in validation loss. The values of different other parameters are also depicted as part of Table 1 and a considerable dip in validation loss metrics can also be observed.

4.3 Experiment 3-a

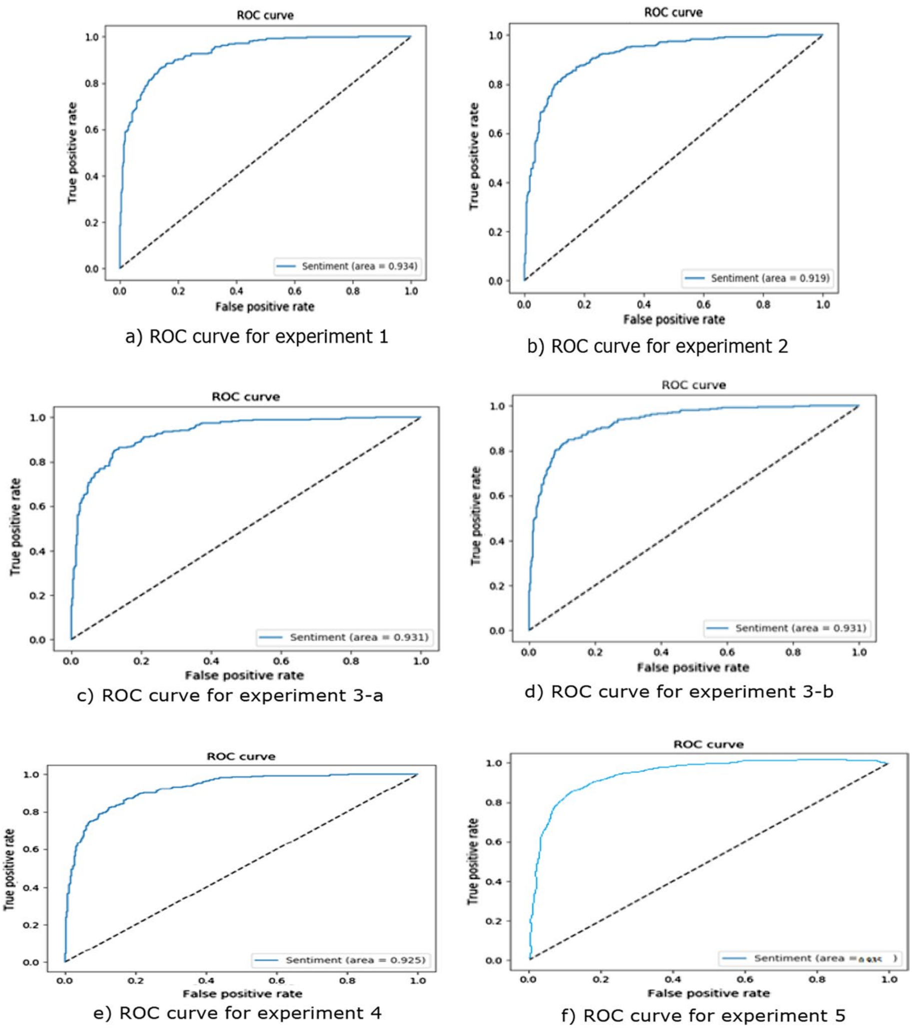
This experiment is designed to obtain a compatible structural ensemble of LSTM and GRU for hidden layers of Domain Attention model. The experiment was targeted to find a suitable recurrent network for the proposed architecture. We experimented by modifying

Table 2 Domain wise comparison of various model

Exp#	F1-Score						Precision						Recall						Accuracy					
	1	2	3-a	3-b	4	5	1	2	3-a	3-b	4	5	1	2	3-a	3-b	4	5	1	2	3-a	3-b	4	5
Books	0.906	0.901	0.864	0.881	0.819	0.916	0.883	0.841	0.862	0.866	0.798	0.912	0.931	0.970	0.866	0.896	0.842	0.941	0.886	0.911	0.881	0.886	0.871	0.886
DVD	0.889	0.854	0.846	0.834	0.772	0.916	0.949	0.987	0.865	0.862	0.792	0.942	0.837	0.752	0.827	0.807	0.752	0.891	0.842	0.837	0.812	0.817	0.812	0.872
Kit	0.853	0.873	0.797	0.695	0.636	0.851	0.906	0.864	0.797	0.702	0.620	0.855	0.807	0.881	0.797	0.688	0.653	0.847	0.842	0.861	0.847	0.842	0.847	0.885
Elec	0.875	0.878	0.820	0.720	0.622	0.853	0.810	0.851	0.805	0.703	0.642	0.851	0.950	0.906	0.837	0.738	0.604	0.876	0.866	0.861	0.866	0.861	0.856	0.8896
Avg	0.880	0.878	0.831	0.782	0.699	0.884	0.886	0.886	0.832	0.783	0.712	0.887	0.881	0.878	0.832	0.782	0.713	0.884	0.859	0.868	0.852	0.862	0.847	0.884

Table 3 Sentiment wise comparison of various models

Exp#	F1-Score						Precision						Recall					
	1	2	3-a	3-b	4	5	1	2	3-a	3-b	4	5	1	2	3-a	3-b	4	5
Positive	0.856	0.867	0.846	0.849	0.843	0.866	0.876	0.868	0.878	0.864	0.861	0.888	0.837	0.866	0.817	0.834	0.827	0.857
Negative	0.862	0.868	0.856	0.854	0.850	0.876	0.844	0.867	0.829	0.840	0.833	0.839	0.881	0.869	0.886	0.869	0.866	0.886
Avg	0.860	0.868	0.851	0.851	0.846	0.871	0.860	0.867	0.853	0.851	0.847	0.864	0.859	0.868	0.851	0.851	0.846	0.872



a) ROC curve for experiment 1

b) ROC curve for experiment 2

c) ROC curve for experiment 3-a

d) ROC curve for experiment 3-b

e) ROC curve for experiment 4

f) ROC curve for experiment 5

Fig. 2 ROC plots of various experiments

the architecture of previous experiment by using LSTM cell as bidirectional RNN in the domain module and GRU cells in the sentiment module. Domain module makes use of a LSTM network to gain the domain representation. The sentiment module is bidirectional Gated Recurrent Unit (GRU) network with attention mechanism. Unlike LSTM, GRU consists of only three gates and does not maintain an internal cell state. GRU uses reset gate and update gate which act as vectors, deciding what information is to be passed to the output to avoid vanishing gradient problem of a standard RNN. Usage of such architecture rendered notable results with a lower training period time since GRU tends to train quickly as compared to LSTM. Here as well, sentiment module attends to all outputs unlike the domain module. The extracted results and other parameters for the experiment are shown in Tables 1,2 and 3. The observed values indicate an improvement over the previous observed values.

4.4 Experiment 3-b

In this experiment, recurrent network units of both the modules are reversed i.e. GRU is used as the bidirectional RNN in the domain module to gain domain representation and LSTM in the sentiment module. The use of GRU units helps to determine the amount of information gained from previous steps that needed to be passed further using the update gates. This step makes it possible for the model to decide whether to keep all the information gained from the past steps and helps to vanquish the risk of vanishing gradients. The observed values in Tables 1,2 and 3 indicate comparable results from previous experiments in various metrics but marginal increase in accuracy values was observed across all the domains because of the fine compatibility of GRUs.

4.5 Experiment 4

We performed experiments to compute and analyze the usage of GRU cells for the bidirectional RNN in the domain module as well as in the sentiment module. Domain module, employs a GRU network to gain the domain representation and update gates to decide which information from previous steps is to be passed. The architecture's sentiment module is also GRU network with attention mechanism. This model was successful in providing good accuracy across all domains because the modules gelled together quickly. The model gave highest accuracy in most of the domains and had less training time as well. Corresponding results are available in Tables 1,2 and 3.

4.6 Experiment 5: Proposed Architecture

The last experiment was performed to improve the results further by adding an additional dense layer for higher degree of information processing. The additional dense layer helps in better retainment of specific information; prevents over-fitting and improve accuracy. The impact of addition of dense layer resulted in increase in values of each utilized parameter and these evaluations are presented in Tables 1, 2 and 3. The overall values of the proposed architecture were better than the rest of experiments due to addition of new layer to the model which enabled to learn more comparative dissimilarities amongst both domain and sentiment modules thereby leading to final results improvements. The resulted values of precision, recall and F-1 score were also better than all experiments for both domain and sentiment modules respectively. The training of the network was also quick. This was true for every model and the time taken for training of various networks didn't have much variations since pre-trained models were employed and the dataset was relatively not that big. The proposed model was also tested on low-end devices like Nvidia Jetson Nano [28] but the observant lag was also negligible.

In order to dig deeper into comparisons, we plotted the Receiver Operating Characteristics (ROC) curves for the various experiments since the classes to be tested are not imbalanced therefore ROC curve was preferred over Precision-Recall curves. Comparison of ROC curve plots of all experiments is shown in Fig. 2 which further illustrates the success of the proposed model. This is also indicative by the values of Area under Curve (AUC) for the proposed model in Fig. 2f illustrating that the final model was successfully able to make correct predictions.

Table 4 Comparison with other state of art methods

Domain	Books	DVD	Electronics	Kitchen
LS(least square) -single	77.8	77.88	81.63	84.33
SVM-single	78.56	78.66	83.03	84.74
LS-all	78.4	79.76	84.67	85.73
SVM-all	79.16	80.97	85.15	86.06
MDSC-Com [51]	79.07	80.09	83.77	85.54
RMTL [52]	81.33	82.18	85.49	87.02
MTL-Graph [15]	79.66	81.84	83.69	87.06
CMSC [8]	81.16	82.08	85.85	87.13
LSTM-single [53]	78.51	78.74	80.72	81.8
LSTM-all [53]	83.74	79.7	82.87	83.56
MTL-CNN[54]	80.2	81	83.4	83
MTL-DNN [54]	79.7	80.5	82.5	82.8
MTL-SM [55]	82.8	83	85.5	84.3
ASP-MTL [20]	84	85.5	86.8	86.2
NeuroSent [50]	79.66	80.9	86.41	86.86
Proposed	88.6	87.2	88.5	88.96

To summarize, the average parameters for all metrics tend to be highest for proposed model experiment, i.e. experiment 5, indicating a better overall choice for multiple domains. The proposed model outperformed other architectures in all domains demonstrating its effectiveness not only for a specific domain class but in each class as well. Experiment 2 achieved high F1-scores for Kitchen and Electronics domain and higher accuracy values for Books domain while experiment 1 also demonstrated comparable results but the proposed architecture performed well in terms of validation loss and validation accuracies (Table 1). For instance, the proposed model has observed lowest validation loss for domain prediction and other types of loss values are small. Overall, it can be concluded that the proposed architecture using the attention mechanism and employing GRU as recurrent network with addition of dense layer has demonstrated improvements over the existing models available in literature. This is also true as observed from the ROC plot comparisons of all experiments as presented in Fig. 2.

We further compared the results of proposed model with similar state-of-the -art studies in literature which is represented in Table 4. The results in Table 4 indicate that the proposed system is capable of outperforming all methods outlined in the table. It has convincingly outperformed methods working on individual domains like LS and SVM; methods working on multiple domains (*RMTL, MTL-Graph* and *CMSC)*; methods on combination of sentiment classifiers like MDSC also. This is possible because the proposed method is capable of extracting common information across domains and attention mechanism helps it to learn and retain the important information effectively.

5 Conclusions

The paper proposes a deep learning-based sentiment analysis system employing domain attention mechanism for successfully predicting sentiment values across multiple domains. The proposed architecture consisted of employing GRU cells for Sentiment and Domain Module which are implemented as bi-directional recurrent network with an additional dense layer to improve accuracies. A number of experiments for evaluation of the proposed architecture were conducted and it was found that using GRU layer as the bidirectional RNN gave an efficient performance over other tested models. It not only took less training and testing time but also provides higher validation accuracy in most of the domains. The predictions of the model across different domains were efficient and better than the rest of the experiments. The proposed architecture also demonstrated higher validation accuracy in the sentiment module. The evaluation of domain and sentiment modules was conducted separately using large number of parameters like Precision, Recall, F1-score etc. Overall, it can be concluded that employing of GRU as sentiment and domain module and a dense layer leads to increase in prediction accuracy and be a better classifier. The future extensions to work are possible by extending its training over multilingual dataset and building a web-utility for sentiment prediction for public use. This will also help to test the proposed model for scalability and deployment on low-end devices.

Appendices

Appendix 1: Recurrent Networks and Attention Mechanism

Recurrent Neural networks (RNN) are supervised deep learning neural networks in which the neurons are connected through time with each other and therefore allows exhibition of temporal behaviour. It is a class of network that contains hidden layer loops to hold the information at the old step to predict the value of the current time step. It forms connections between units in directed loops and remembers previous inputs through its internal state as seen in the diagram (Fig. 3).

Here, “x” is the input layer, “h” is the hidden layer, and “y” is the output layer. A, B, and C are the network parameters that are used to improve the output of the model. At any given time t, the current input is a combination of input at x(t) and x(t-1). The output of RNN in time step t-1 affects the output at time step t which allows RNN to establish a current sequence of time. RNNs consider only the short-term dependencies because of vanishing and exploding gradient problems.

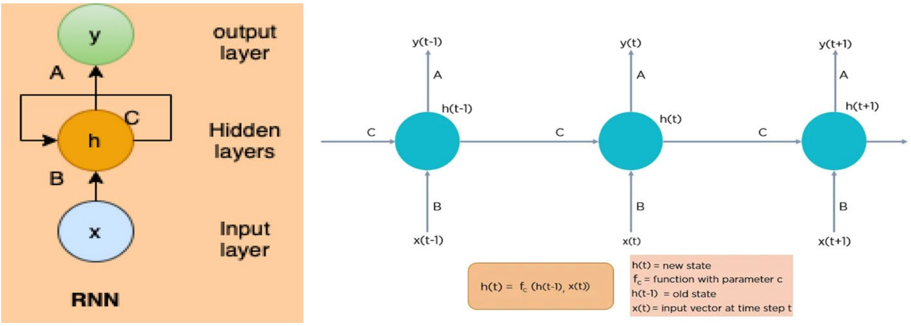


Fig. 3 RNN and its illustration

Long Short-Term Memory Networks (LSTM)

LSTM is an updated version of RNN which is specifically designed to eliminate gradient disappearance and gradient explosion issues. It has a memory cell at the top which helps to carry the information from a particular time instance to the next time instance in an efficient manner and is able to remember the information from previous states as compared to RNN.

As shown in the above Fig. 4, LSTM network is fed by input data from the instance of present time and output of hidden layer from the previous time instances. LSTMs have "cells" in the hidden layers of the neural network, which have three gates–an input gate, an output gate, and a forget gate. These gates control the flow of information which is needed to predict the output in the network.

Gated Recurrent Unit (GRU)

GRU is a well-accepted variant and used hidden states to regulate information instead of using a "cell state". GRU has two gates instead of three gatesa reset gate and an update gate. GRU combines a forgotten gate and input gate into a single update gate. The reset and update gates control how much and which information to retain which is similar to the gates within LSTMs. The final model is simpler than the standard LSTM model and given a set of observations, the learned model can provide the corresponding control output vector. It has been found that GRUs are equally efficient as LSTMs.

Attention

Attention is proposed as a solution to the limitation of the encoder-decoder model which encodes the input sequence to one fixed length vector from which output is decoded by the decoder at each time step. Attention mechanism makes sure that the model searches for relevant information in the source sentence in order to predict next word and assigns more weight to that information. This is achieved by creating a context vector for each step rather than a fixed context vector. As can be seen in the Fig. 5, there is no change in the working of the encoder but the decoder's hidden state is computed with a context vector, the previous output and the previous hidden state and also it has separate context vector for each

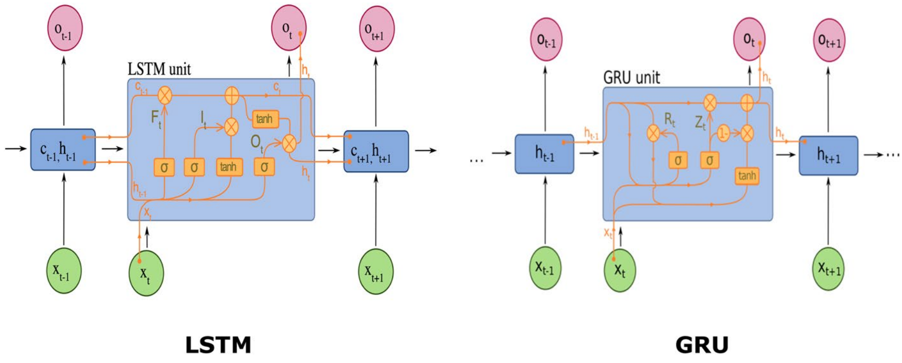


Fig. 4 LSTM and GRU architecture

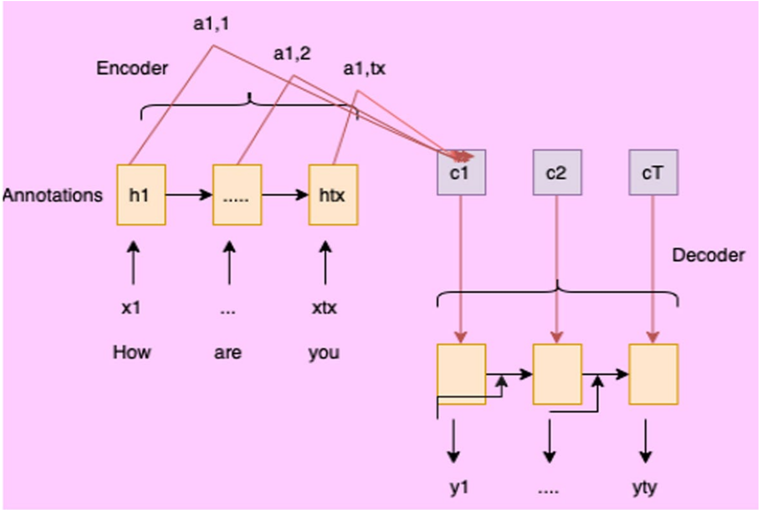


Fig. 5 Sample attention mechanism

target word. These context vectors are computed as a weighted sum of activation states in forward and backward directions and alphas and these alphas denote how much attention is given by the input for the generation of output word.

Appendix 2: Details of Used Various Parameters

d_pred_acc: Accuracy is the ratio of correctly predicted observation to total observations and here it means accuracy of the model to predict the domain.

d_pred_loss: It is the loss in the model to predict the domain and it indicates the level of inaccurate prediction by the model on a single example.

loss: It depicts how well specific algorithm models the given data and we had used cross entropy loss for training of our models.

s_pred_acc: The accuracy of the model to predict the sentiment.

s_pred_loss: The loss in the model to predict the sentiment.

val_d_pred_acc: The accuracy of the model to predict the domain during validation.

val_d_pred_loss: The loss in the model to predict the domain during validation.

val_loss: The complete loss from both the domains during validation.

val_s_pred_acc: The accuracy of the model to predict the sentiment during validation.

val_s_pred_loss: The loss in the model to predict the sentiment during validation.

Acknowledgements This research work was funded by Institutional Fund Projects under grant no. (G:1299-611-1440). Therefore, the authors gratefully acknowledge technical and financial support from Ministry of Education and Deanship of Scientific Research (DSR), King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

Author's Contribution All authors contributed towards planning, literature survey, experimentations, manuscript writing, editing and proof-reading.

Funding This study was funded by Deanship of Scientific Research (DSR), King Abdulaziz University Jeddah, under Grant No. (G:1299-611-1440).

Data Availability The data may be made available on request with permissions.

Code Availability The code may be made available on request with permissions.

Declarations

Conflict of interest The authors declare that they have no conflict of interests.

References

1. Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems, 53*(4), 704–711.
2. Gokalp, O., Tasci, E., & Ugur, A. (2020). A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. *Expert Systems with Applications, 146*, 113176.
3. Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access, 7*, 14637–14652.
4. Momani, S., Abo-Hammour, Z. S., & Alsmadi, O. M. (2016). Solution of inverse kinematics problem using genetic algorithms. *Applied Mathematics & Information Sciences, 10*(1), 225.
5. Abo-Hammour, Z., Arqub, O. A., Alsmadi, O., Momani, S., & Alsaedi, A. (2014). An optimization algorithm for solving systems of singular boundary value problems. *Applied Mathematics & Information Sciences, 8*(6), 2809.
6. Abo-Hammour, Z., Abu Arqub, O., Momani, S., & Shawagfeh, N. (2014). Optimization solution of Troesch's and Bratu's problems of ordinary type using novel continuous genetic algorithm. *Discrete Dynamics in Nature and Society*, *2014*.
7. Abu Arqub, O., Abo-Hammour, Z., Momani, S., & Shawagfeh, N. (2012). Solving singular two-point boundary value problems using continuous genetic algorithm. In *Abstract and applied analysis* (Vol. 2012). Hindawi.
8. Xue, S., Lu, J., & Zhang, G. (2019). Cross-domain network representations. *Pattern Recognition, 94*, 135–148.
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
10. Du, Y., He, M., Wang, L., & Zhang, H. (2020). Wasserstein based transfer network for cross-domain sentiment classification. *Knowledge-Based Systems, 204*, 106162.
11. Pan, S. J., Ni, X., Sun, J. T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web* (pp. 751–760).
12. Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42–49.
13. Rana, T. A., & Cheah, Y. N. (2017). A two-fold rule-based model for aspect extraction. *Expert Systems with Applications, 89*, 273–285.
14. Wu, C., Wu, F., Wu, S., Yuan, Z., & Huang, Y. (2018). A hybrid unsupervised method for aspect term and opinion target extraction. *Knowledge-Based Systems, 148*, 66–73.
15. Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications, 118*, 272–299.
16. Yuan, Z., Wu, S., Wu, F., Liu, J., & Huang, Y. (2018). Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems, 155*, 1–10.
17. Liu, P., Qiu, X., & Huang, X. (2016). Deep multi-task learning with shared memory. arXiv preprint arXiv:1609.07222.
18. Chauhan, G. S., Meena, Y. K., Gopalani, D., & Nahta, R. (2020). A two-step hybrid unsupervised model with attention mechanism for aspect extraction. *Expert Systems with Applications, 161*, 113673.
19. Kim, Y. B., Stratos, K., & Kim, D. (2017). Domain attention with an ensemble of experts. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 643–653).
20. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning* (pp. 1378–1387). PMLR.

21. Lee, G., Jeong, J., Seo, S., Kim, C., & Kang, P. (2018). Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems, 152*, 70–82.
22. Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing, 337*, 325–338.
23. Fu, X., Yang, J., Li, J., Fang, M., & Wang, H. (2018). Lexicon-enhanced LSTM with attention for general sentiment analysis. *IEEE Access, 6*, 71884–71891.
24. Kardakis, S., Perikos, I., Grivokostopoulou, F., & Hatzilygeroudis, I. (2021). Examining attention mechanisms in deep learning models for sentiment analysis. *Applied Sciences, 11*(9), 3883.
25. Riemer, M., Khabiri, E., & Goodwin, R. (2017). Representation stability as a regularizer for improved text analytics transfer learning. arXiv preprint arXiv:1704.03617.
26. Ji, Y., Wu, W., Chen, S., Chen, Q., Hu, W., & He, L. (2020). Two-stage sentiment classification based on user-product interactive information. *Knowledge-Based Systems, 203*, 106091.
27. Yu, B., Wei, J., Yu, B., Cai, X., Wang, K., Sun, H., & Chen, X. (2022). Feature-guided multimodal sentiment analysis towards Industry 4.0. *Computers and Electrical Engineering*, *100*, 107961.
28. Jetson Nano Developer Kit, Available at: https://developer.nvidia.com/embedded/jetson-nano-developer-kit. Last access 12th April, 2022.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Khaled Hamed Alyoubi is an Associate Professor of Computer Science at Faculty of Computing and Information Technology in King Abdulaziz University. He received his Ph.D. degree in Computer Science from Birkbeck University of London, UK. His research interests include Data sciences, Machine Learning and Deep Learning, IR, Data Analytics and Data mining.

Akashdeep Sharma is working as Assistant Professor in Computer Science and Engineering at University Institute of Engineering & Technology, Panjab University, Chandigarh, India. He is Ph.D. in Computer Science and Engineering from Guru Nanak Dev University Amritsar. He has received his Masters degree from Punjabi University Patiala. His areas of interest include QoS in wireless networks, Traffic Profiling, Soft computing and optimization. He is fond of designing applications using Neural Networks and Fuzzy Logic. He has more than 30 publications to his credit with reputed publishers like Elsevier, Springer, Science Direct etc. He has delivered lectures and presented several papers on Machine Learning and Deep Learning in areas of wireless networks and video analytics during workshops and international conferences.