sciendo

VOLUME 17, N° 1 2023
Journal of Automation, Mobile Robotics and Intelligent Systems

A MODEL OF CONTINUAL AND DEEP LEARNING FOR ASPECT BASED IN SENTIMENT ANALYSIS

Submitted: 10th January 2023; accepted: 18th February 2023

Dionis López, Fernando Artigas-Fuentes

DOI: 10.14313/JAMRIS/1-2023/1

Abstract:
Sentiment analysis is a useful tool in several social and business contexts. Aspect sentiment classification is a subtask in sentiment analysis that gives information about features or aspects of people, entities, products, or services present in reviews. Different deep learning models that have been proposed to solve aspect sentiment classification focus on a specific domain such as restaurant, hotel, or laptop reviews. However, there are few proposals for creating a single model with high performance in multiple domains. The continual learning approach with neural networks has been used to solve aspect classification in multiple domains. However, avoiding low, aspect classification performance in continual learning is challenging. As a consequence, potential neural network weight shifts in the learning process in different domains or datasets.

In this paper, a novel aspect sentiment classification approach is proposed. Our approach combines a transformer deep learning technique with a continual learning algorithm in different domains. The input layer used is the pretrained model Bidirectional Encoder Representations from Transformers. The experiments show the efficacy of our proposal with 78 % F1-macro. Our results improve other approaches from the state-of-the-art.

Keywords: *Continual Learning, Deep Learning, Catastrophic Forgetting, Sentiment Analysis.*

1. Introduction

Sentiment analysis is a useful tool in several social and business contexts, such as: social networks, online shops (Amazon[1], Alibaba[2]), blogs, etc. It is an important task in natural language processing (NLP) and natural language understanding (NLU) [15]. Aspect based sentiment analysis (ABSA) is a fundamental subtask in sentiment analysis where users and decision-makers can obtain more information about sentiments in reviews [8]. An aspect term is related to features about products, services, events, and people [19].

ABSA has three essential subtasks: (i) opinion target extraction (OTE), (ii) aspect category detection (ACD), and (iii) sentiment polarity (SP) or aspect sentiment classification (ASC). Thus, OTE is mainly concerned with the extraction of aspect terms (i.e., entity or attribute), ACD is related to associate entities and attributes to a global category (i.e., comfort or cleanness in hotel domain), whereas ASC is focused on the sentiment polarity classification of aspects [8].

The ASC subtask has been studied by several researchers. They have been using deep learning approaches with better results [2, 39]. The proposed ASC models have been associated usually with a single domain; however, when they have been applied to different ones, their effectiveness decreases [2]. Suitable F-measure values were obtained with the same model when it is applied to different single domains [10]: laptops (83%), hotels (89%). Nevertheless, a retrieval system can process objects or instances from more than two domains.

For that reason, approaches such as CL (CL) that is capable of learning in an incremental learning process from more than two domains have emerged [5]. It takes advantage of the local learning of several domains by identifying the main features or patterns found in the previous learning process without losing effectiveness (i.e., the price aspect is common for restaurants, hotels, and electronic devices domains) [5, 6].

The constraint in the CL setting is that a computational model would not be able to access the data from the previous tasks; it can only access a limited amount of information [5]. This learning problem is challenging. If the same model is retrained using the current available dataset D_M, it will forget how to predict for datasets $D_m; m < M$. This is known as the catastrophic forgetting problem. This occurs when networks are sequentially trained in many tasks; for instance, in task A, network weights can be modified by the learning process of a task B [22, 22]. Several proposals have tried to improve it in image classification [9]. Nevertheless, there are few proposals to solve this challenge in the ABSA subtask [2].

In this paper, we propose a hybrid model that combines the continual and deep learning approaches for ASC. First, a text preprocess module extracts the aspect word candidates (i.e., noun, adverbs) and the proposed model classifies each aspect into one of three possible classes: positive, negative, or neutral. Our model starts from a Bidirectional Encoder Representations from Transformers (BERT) [7] model and improves the CL disadvantages based on:

- Combining a CL regularization approach in NLP (i.e., ABSA) with a gradient descent modification algorithm to preserve relevant weights in a CL scenario.

- Using the output of a pretrained BERT model to improve the results and tune the BERT model on the CL process.

The rest of this paper is organized as follows. The subsection "Related Work" describes the methods based on deep and CL in ABSA. Section 2 presents the proposed model based on deep and CL. Section 3 discusses the evaluation of the model with respect to the state-of-the-art (SOTA). Toward the end, we provide concluding remarks and future research directions.

1.1. Related Work

As mention in the previous section, CL represents a long-standing challenge for machine learning and neural network systems [9]. This is due to the tendency of learning models to catastrophically forget existing knowledge when learning from novel observations [22, 25]. The most common CL strategies [2, 6, 16, 21] are described below:

- Architectural strategy: Specific architectures, layers, activation functions, and/or weight freezing strategies are used to decrease forgetting. This adds other neural network architectures for each domain, as proposed in [14, 18].
- Regularization strategy: The DL model loss function is extended with loss terms promoting selective consolidation of the weights, which are important to retain memories. This strategy includes basic regularization techniques such as weight specification, dropout, or early stopping, as described in [1].
- Rehearsal strategy: Past information is periodically replayed to the model to strengthen connections for memories it has already learned. A common approach is part of the previous training data and interleaving them with new tasks or domains for future training, as described in [20].

The architectural and rehearsal strategies propose the creation of new structures for new domains [7,16]. In the case of rehearsal, it is necessary to preserve instances of the previous domains, which is also computationally expensive. A lower cost will be achieved by regularization [16], by not adding any architecture or additional memory during a learning process. The target of this research is the CL model with a regularization strategy.

The Elastic Weight Consolidation model (EwC) is one of the more successful regularization approaches. It tries to control forgetting by selectively constraining (i.e., freezing to some extent) the model weights, which are important for the previous tasks. The EwC regularization used a Fisher Information Matrix in a stochastic gradient descent (SGD) computation. The Fisher Information Matrix in a neural network is expensive, because of the need to preserve the all neural network weight in external memory.

Other CL strategies, such as hard attention to the task (HARD) [29] and incremental moment matching (IMM) [14], are not better than in CL for sentiment analysis [35].

The synaptic intelligence (SI) model [38] is an optimization of EwC. In the SI approach, the neural network weights are calculated online when the SGD is applied. The SI reduces the EwC computational cost. To the best of our knowledge, there are no works in ABSA with SI. The architectural and regularization 1 (AR1) [16] is an SI optimization with batch instance results and last layer weight average computation. Although, AR1 was used in image classification tasks, their architecture and the model computational result was studied in this research.

In sentiment analysis task, "lifelong learning memory" (LLM), proposed in [35], is a CL regularization approach. It incorporates the mined knowledge into its learning process, where two types of knowledge are involved named aspect sentiment attention and context sentiment effect. Because only datasets of household appliances (reviews on cameras, laptops, smartphones, etc.) are applied, the CL model does not learn from diverse domains. The model performance of 82% F1-macro suggests being taken as an element of comparison in our research.

Another CL model is "knowledge accessibility network" (KAN) proposed in [12]. Its target is to classify sentiment in sentences (i.e., not ABSA substask) in a task incremental learning (TIL) scenario (e.g., in a CL process, each new task has new instances). The KAN model is closely related to the HARD model [35] (i.e., a CL architectural strategy and more expensive regularization strategy). Although, it is not similar to our research objective, KAN is of the more recent works in sentiment analysis with a CL approach.

In the ASC subtask, there are several works that combine BERT and deep learning architecture, for instance, "local context focus with BERT" (LC) [37] receives the words that correspond to the sentence where the aspect appears and a set of words in the aspect neighborhood as input as an upper layer of the BERT model. It reaches an 82% accuracy with a laptop dataset. Another approach is the model Attentional Encoder Network with BERT(AE) [31] applies a multi-head attention architecture to the BERT model output. It has two inputs: the words of the context (a sentence) and the words that make up the aspect. It reaches an 83% Accuracy in a Restaurant dataset.

The attentional encoder network with "long short-term memory networks" (LSTM) and identified with acronym AT [31] applies an attention mechanism and concatenates the aspects and their context.

The AT enables aspects to participate in computing attention weights. It uses GloVe as input and LSTM as deep learning model. The above approaches have good classification measures; however, they are evaluated in a single domain scenario and not a CL process. Their deep learning architectures are interesting as base models in a CL framework.

The sentiment analysis models improve services such as tourism or e-commerce (e.g., analyzing sold product reviews). Building datasets for each domain is expensive (time, specialists). A model (such as the one proposed in this research) can be used in various social services (tourism, e-commerce, government) and reduces model learning costs (time and memory) and economic resources.

2. Content

A new model for ASC in multidomains based on deep and CL with the regularization approach is introduced in this section.

First, a formal definition of the problem is given. Then, we introduce the different stages of the new model and its main inputs, outputs, and activities.

In the ABSA subtask, given a sentence (sequence of words) $w^c = w_1^c, w_2^c, ..., w_n^c$, an aspect is a sequence of words, defined as $w^t = w_1^t, w_2^t, ..., w_m^t$, where w^t e is a subsequence of w^c. The goal of this task is to predict the sentiment polarity "s" of aspect w^t, where $s \in positivo, negativo, neutral$.

The new model is evaluated in a particular CL setting called domain incremental learning (DIL) [34]. Each task is from a different domain or dataset; however, the classes are the same (i.e., positivo,negativo,neutral). The DIL setting is particularly suited to ASC because in testing the system, it need not know the task/domain to which the test data belongs.

A CL model with the regularization approach has three main components [6]:

- A base machine learning model (e.g., CNN, BLSTM, or a pretrained model such as Resnet).
- The CL approach to preserve knowledge (e.g., weights in a neural network, based classifier) during the learning process from one domain to another.
- The knowledge base, which is usually the common or a new part of the model (e.g., neural network weights do not vary between learning domains, new neurons inserted to the neural network) to be added according to the CL approach.

2.1. Model Description and Main Stages

The model components are presented in 1. In this, the CL model is represented at the top (i.e., labeled with "CL model"). This model preserves the knowledge learned during model training in the previous domains. The base model is represented by the square named "Bert neural network" in the middle of the figure, and the classification neural node gives the input values to the next CL model. The bottom showns how the domains were used by the CL process.

The model training process has four stages:

Stage 1. Textual representation: This stage receives the original textual opinions and returns the vector of tokens per each sentence (applying a sentence splitter) and the aspects in a sentence (nouns, adjectives and noun phrases). The aspects are selected by a part-of-speech (POS) tagger. Although in our model word tags (i.e., noun, adjective) are used to identify aspects, there are other approaches with a Deep Learning model (or other Machine Learning approach) that are used in aspect extraction process as in [17]. If a sentence has more than an aspect, then is associated a sentence vector of tokens for each aspect. The spaCy[3] NLP tool offers the implementation and documentation for developing this stage.

Stage 2. Train the CL model: This stage learns the knowledge to be included in the knowledge base, depending on BERT and the CL training process for each current domain. This stage is divided into the following steps:

- Train the BERT neural network for each domain, where the output is a classify neuron.
- Train the CL approach with the pretrained BERT model output last layer.

The stage input is the tokens vector in a sentence and the tokens vector associated with an aspect, from Stage 1.

In this stage, the BERT embedding input layer is built. To build the input vector to the BERT pre-model, a vector of weights is obtained from a vector model of words (i.e., WordPiece) and the position of each token in the sentence.

This vector is the first layer of the deep learning base model represented by BERT. Traditional Word2Vec or GloVe embedding layers provide a single context independent representation for each token. On the contrary, in BERT, the representation of each token is related to the data obtained from the sentence used as input [7]. This allows having more information about the word context when training the models. The stage's output is a neuron associated with the classification token in the BERT last layer.

Stage 3. Knowledge base upgrade: In this stage, catastrophic forgetting is avoided through the analysis of the training process results. The input is the classification neuron value (Stage 2 output) to feed a layer with three neurons (e.g., positive, negative, neutral). The obtained loss/error is used in the weight optimization process by the regularization strategy. The output is the KB vector updated with the new weights obtained from the CL approach and BERT neural network.

Stage 4. Aspect classifier creation: This stage makes available the continual deep learning model for solving the ABSA subtask in multidomains. The final configuration of the model is obtained from the parameters in the knowledge base.

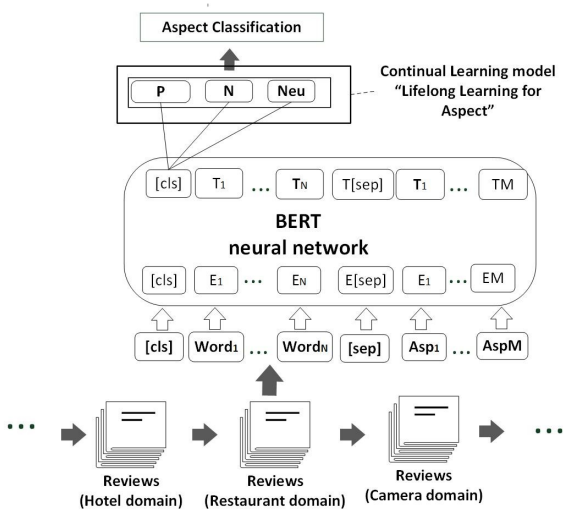

Figure 1. Input and output information representation in the CL model

A CL regularization approach has a machine learning base model for learning raw features in each task or domain (datasets). BERT Special (BSp) [31] is the base model in the new CL model proposed in this article. It is selected by the attention mechanism (i.e., the transformers) in the BERT neural network architecture. The BSp associates the words in a sentence and the aspect (represented by "BERT Neural Network" in 1).

The new CL model is general enough to replace this base model with others, as shown in Results section. The BSp method constructs the input sequence as:

$$< CLS > tokens < SEP > asp < SEP > \quad (1)$$

The first token of every input sequence is the special classification embedding ($< CLS >$), and it separates the context words associated to sentence words and aspect words (i.e., asp) with a special token ($< SEP >$). In the BERT neural network, the output of the neuron associated with the token $< CLS >$ in the last layer is the classification value in the Stage 2 learning process. The base model architecture was validated in the experiments against complex models as AE, which reaches 73% F1-macro for a single dataset about restaurant reviews [31].

2.2. The Continual Learning Model

The new proposal, named Lifelong Learning of Aspects (LLA), is inspired by AR1 [16] and the Synaptic Intelligence (SI) [38] learning process, because they achieve better classification results [24] (i.e., the LLA is represented in the 1 with " Lifelong Learning of Aspects" label).

Although both models were applied to image classification task, we adapted them for NLP and classifying aspects (e.g., words in a sentence) in three classes (positive, negative, and neutral) for different datasets during the learning process. AR1 improves SI [16] in image classification challenges; however, in our model, we combine the updated descendent gradient and weights preservation mechanism from SI in the AR1 regularization model.

The new CL model LLA combined with BERT (as base model) has the acronym (BSp_{LLA}), and it is the new computational model proposed.

In contrast to the original AR1 proposal in [21], our approach does not extend with new classes for each new domain. For each domain, the same three classes already mentioned are used. The main objective of LLA is to obtain the set of weights in the output layer $\vec{w}$ as shown in 1. One of the main settings of this algorithm is that $\vec{w}$ is initialized to 0 (as input) and does not use any random initialization as in other CL approaches, inspired in the AR1 approach. The parameters of each output layer from previous domains are stored in $\vec{w}$.

In the CL process, the base deep learning model is represented by BSp, and their output layer is the LLA CL model input (i.e., the AR1 modification for aspect classification in the NLP context). The LLA model loss function and optimization process tune the neural network weights in both models (BSp and LLA) during the learning process.

Algorithm 1 LLA

Input:
$c\vec{w} = 0$▷ The consolidated weights used for inference.
$\bar{M} = 0$ ▷ The deep learning base algorithm weights.
$M = 0$ ▷ The optimal shared weights resulting from training.
$\hat{F} = 0$ ▷ he weight importance matrix (SI algorithm).
$Text$ ▷ Domains sentences dataset.
Output: $c\vec{w}$ ▷ The trained weights used for inferencing.

1: In the $Text$ extracts, the sentence x and its candidate word aspect y are group in batches B.
2: loop ▷ For each batch in B, process all pair (x, y).
3: Train the base deep learning model with pair (x, y).
4: Learn $\bar{M}$ and $c\vec{w}$ subject to SI algorithm with $\hat{F}$ and M.
5: Save weigths in M with $M = \bar{M}$ and $c\vec{w}$.
6: Update $\hat{F}$ according to trajectories computed on the batch B_i.
7: Test the model by using $\bar{M}$ and $c\vec{w}$.
8: end loop

The BSp is trained by using each B batch of sentences in the datasets, as shown in line 3.

The CL approach is described in lines 4 and 5. Line 6 updates the parameters of the neural network using the regularization and gradient descent. The $\vec{w}$ vector is the knowledge base in the new classification domain.

The combination between $\hat{F}$ weight importance and $\vec{w}$ vector is the regularization approach to reduce catastrophic forgetting because it has the balance of the learning process in each domain and the output layer evaluation in the classification algorithm.

The BERT and LLA combination adopted the SI mechanism [38] to compute the weight importance during SGD. This mechanism is important in LLA to preserve the common neuronal weight (transfer learning) in BERT tuning and the model learning.

The loss function in LLA is defined as in [38]:

$$\tilde{L}_\mu = L_\mu + c\sum\nolimits_k \Omega_k^\mu \left(\theta_k^{'} - \theta_k\right)^2 \quad (2)$$

Where k is the neural networks weights parameters, c is a strength parameter that trades off old versus new tasks,Ω_k^μ is the parameters regularization strenght, and $\left(\theta_k^{'} - \theta_k\right)^2$ is the reference weight corresponding to the parameters at the end of the previous task and the current task.

3. Experiments and Results

Experiments were designed to compare the ASC performance of the new model LLA against the CL SOTA [6, 24]. In our experimental design, each deep learning model for ABSA was used as the base approach in each CL model (combinations between deep learning and CL models).

To verify the effectiveness of our proposal against the SOTA approaches, the following experiments were conducted:

- Compare the LLA approach with the main models of the SOTA.
- Analyze if the domain order affects the quality of the results.

3.1. Datasets and Continual Learning Scenario

To evaluate the performance of BSp_{LLA}, experiments are conducted using seven of the most used ABSA datasets, as described in Table 1. They were taken from four sources: laptops and restaurants, from SemEval-2014 Task 4 subtask 2 [26], datasets about electronic devices used in [27], and hotels reviews from TripAdvisor [32]. The considered training and testing subsets are the same as those defined by the datasets' authors.

Dataset instances are sentences and can contain more than one word tagged as aspect. During model training and testing, sentences that have more than one aspect are split into a sentence (i.e., the same sentence text) with only one aspect present.

The CL scenario is DIL, all tasks sharing the same fixed classes (i.e., positive, negative, and neutral). A factor that could influence the results of BSp_{LLA} and the SOTA is the possible semantic closeness of the domains. For instance, restaurant and hotel reviews could be semantically closed (reviews with words related to cleanliness, comfort, price).

The used datasets (domains), in the BSp_{LLA} learning process, were grouped by humans. These can be considered clusters. We estimate the domain centroid (the mean of the BERT output vectors for all sentences in each domain) and the cosine similarity of the centroids [30], with the objective to estimate semantic closeness. The results are showed in 2.

In the figure, the similarity indicates restaurant and hotel review domains are close. However, others as routers, laptops are not close to restaurants.

Other important measures to evaluate the quality of the clusters is the silhouette coefficient. It has performance measure in the interval $[-1, 1]$, and values near zero, indicate overlapping clusters [33]. The dataset using the silhouette coefficient (with cosine similarity) was -0.017. In 2, the similarity indicates restaurant and hotel review domains are close. However, others as routers, laptop are not close to restaurant.

Table 1. Labeled dataset description (Sent = Sentences, Aspect = Aspects).

Domain	Sentences	Aspects	Train	Test
Digital Cameras	597	237	477	120
Smart Phones	546	302	436	110
Routers	701	307	877	176
Speakers	687	440	549	138
Restaurants	3841	4722	3041	800
Laptops	3845	2951	3045	800
Hotels	4856	3810	3371	1485

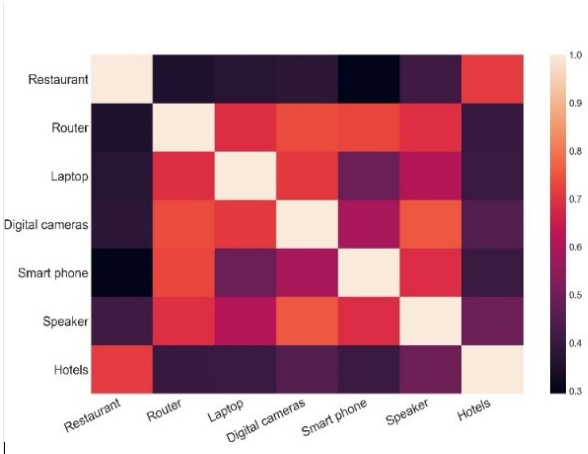

Figure 2. Cosine similarity between domain centroids

The value near zero indicates that there are some sentences close semantically among domains, and the negative value of a sample has been assigned to the wrong cluster (the datasets were created by humans and not by the same authors). This value indicates that there is common information between domains and confirms the performance of the BSp_{LLA} model. Because, it learns common aspects in ASC between different domains and reduces forgetting the past knowledge.

3.2. Compared Baselines

The new proposal BSp_{LLA} was evaluated against three SOTA strategies of CL with the regularization approach (see model descriptions in Table 3):

- "Lifelong Learning Memory" (LLM)
- "Elastic Weight Consolidation" (EwC).
- "Architectural and Regularization 1" (AR1)

The CL regularization approach employs a base model (e.g., deep learning model) for learning the raw features of datasets. During the evaluation, each of the CL models mentioned was combined with the deep learning approach used in ABSA:

- "Local Context Focus with BERT" (LC) [37].
- "Attentional Encoder Network with BERT" (AE) [31].
- "Attentional Encoder Network with LSTM" (AT) [11].

These methods were selected because they have relevant accuracy performance in ABSA, and they have as input a BERT pretrained model or GloVe input as in AT.

3.3. Hyperparameters

The pretrained uncased BERT base model was applied in the learning process.[4] The neural network architecture was 12-layer, 768-hidden, 12-heads, 110M parameters, trained on lower-cased English text. GloVe pretrained with 300 vectors was used in the Word Embedding. The weights of the LLM, EWC, LLA, and AR1 models are initialized with the Glorot initialization,[5] whereas the coeficient of L2 regularization is 10^{-5} and the dropout rate is 0.1. The BERT model was implemented by pythorch library transformers 2.1.0.[6]

Table 2. Example of imbalance between classes in used datasets.

Domain	Positive	Negative	Neutral
Restaurants	2892	1001	829
Laptops	1328	994	629
Hotels	2343	656	811

All combinations of deep learning and CL models have been trained with a batch size equal to 64 and 10 epochs, for all datasets (i.e., Each dataset (domain) trains the model for 10 epochs and uses a batch of 64 instances). The optimization function was an Adam with a 2e-5 learning rate. The training and evaluation processes were made using a 2 x Intel Xeon L5520 with 64 GB RAM on the high performance computing cluster at Central University of Las Villas, Cuba. The source code is public.[7]

3.4. Evaluation Measures

Taking into account that the selected datasets are imbalanced (see an example in Table 2), we will use F1-macro (averaged F1-score over all classes) in addition to accuracy (Accr) in the experimentation.

The performance of the proposed model was also evaluated with the Cohen-Kappa (Kappa) measure [40]. This selection is motivated because it allows considering the effectiveness of a model in imbalanced datasets [40]. Kappa is computed as shown Equation 3, where ρ_o is the model probability of the label assigned to any sample, and ρ_e is the expected label assign by annotators:

$$\mathcal{K} = (\rho_o - \rho_e)/(1 - \rho_e) \tag{3}$$

Kappa gives values in the interval $[-1, 1]$, where 0 or lower values mean not relevant model training [40].

3.5. Catastrophic Forgetting Evaluation Measure

Several authors have proposed different catastrophic forgetting measures. The definition of a standard measure is a research challenge [2, 6, 23].

The measure selected to evaluate catastrophic forgetting in this research was proposed in [12], because this is a CL work oriented to sentiment analysis in sentences, very close to the target in this research. The catastrophic forgetting measure in [12] average the result of the final classifier in the test sets of the tasks before the last one. This measure is named $OvrcForgtt$ in this research.

3.6. Evaluation Settings

Two configurations were taken into account during the experiments. Initially, a test was analyzed without adjusting the BERT architecture weights. This is because the BERT is supported by a neural network architecture, and it is a pretrained model. But this did yield low classification results (accuracy), and it was rejected. As a final configuration, the weights of BERT architecture and the deep learning model were adjusted during the backpropagation steps. Other authors as in [28] exploit this possibility of training BERT to obtain efficient classification results in ABSA for specific domains.

Table 3. Acronyms and names of the models considered in the evaluation of the new proposal.

Acronyms	Compared Baselines
LLM	Lifelong Learning Memory.
AE_{EWC}	Attentional Encoder Network with BERT and EwC.
AE_{LLA}	Attentional Encoder Network with BERT and the new model LLA.
AE_{AR1}	Attentional Encoder Network with BERT and AR1.
BSp_{EWC}	BERT Special with EwC.
BSp_{LLA}	BERT Special with the new model LLA.
AT_{EWC}	Attentional Encoder Network with LSTM and EwC.
AT_{LLA}	Attentional Encoder Network with LSTM and the new model LLA.
LC_{LLA}	Local Context Focus with BERT and the new model LLA.

Table 4. Average results using different deep learning base models and the LLA (CL approach).

Model	AE_{LLA}	AT_{LLA}	LC_{LLA}	BSp_{LLA}
Accr	0.69	0.64	0.79	**0.80**
F1	0.49	0.38	0.66	**0.73**
Kappa	0.40	0.12	0.59	**0.62**
OvrcForgtt	0.497	0.38	0.66	**0.73**

Table 5. Average results between BSp_{LLA} and other in SOTA.

Model	LLM	AE_{EWC}	AE_{AR1}	BSp_{EWC}	BSp_{LLA}
Accr	0.39	0.68	0.57	0.61	**0.80**
F1	0.23	0.50	0.33	0.62	**0.73**
Kappa	0.03	0.42	0.16	0.53	**0.62**
OvrcForgtt	0.23	0.49	0.66	0.62	**0.73**

The evaluation results are the performance measure average of all possible domain permutations in the CL process.

The BSp_{LLA} model obtains the best results against other base models with the same CL approach (i.e., LLA), as shown in Table 4. Besides, it is possible to modify or improve our base model in the future and use the LLA proposal as part of a general framework.

The results obtained by BSp_{LLA} outperform the rest of the models and demonstrate that our LLA approach can improve results in ASC during a CL of multiple domain scenario (as shown in Table 5).

In all tests (Tables 4-5), BERT-based models perform better than Word Embeddings (i.e., AT_{LLA} and AT_{EWC} have word embedding vectors as input), because BERT takes better account of the context where an aspect occurs.

In an ablation study, the BSp (Base deep learning model), without the LLA algorithm (CL approach) in the same evaluation scenario as BSp_{LLA}, as shown in Table 6. The BSp_{LLA} results were better, as an influence of the CL approach. This experimentation

Table 6. Ablation averaged experimental results between BSp_{LLA} and BSp.

Model	LLA	BSp_{LLA}
Accr	0.64	**0.80**
F1	0.52	**0.73**
Kappa	0.39	**0.62**
OvrcForgtt	0.52	**0.73**

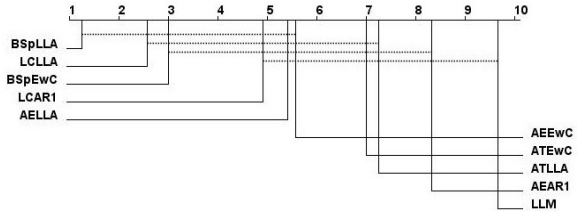

Figure 3. Ranking models with Holm's test with a significance level of 0.05 for F-measure

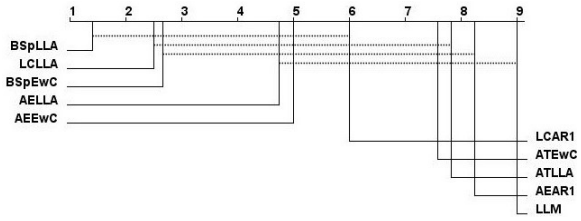

Figure 4. Ranking models with Holm's test with a significance level of 0.05 for Kappa measure

shows that the LLA algorithm cannot be eliminated without loss of effectiveness.

The Friedman test and Holm's method, for the post-hoc analysis [36], were used to verifying significant differences between the models by measuring Accuracy, Kappa, and F1 (see Figures 3 and 4). The experiments show that BSp_{LLA} has no significant differences from other SOTA approaches.

The execution time of the methods was estimated during each test. The training time of BERT-based methods (24 hours' average) was longer than Word Embedding- based methods (three hours' average). This time is associated with the complex transformer architecture and the attention mechanism learning process.

The experimentations show that a less complex architecture such as BSp_{LLA} obtains better results than others (i.e., AE_{LLA}, AE_{AR1}). The difference is that in BSp_{LLA} the input to the BERT model is the context (words in a sentence) and the aspect. The successful results are due to three main characteristics: the BERT is the base model inBSp_{LLA} and has an attention mechanism with weights obtained in huge datasets and the LLA´s regularization approach to avoid changing the values of the weights in a CL process.

3.7. The BSp_{LLA} Evaluation with a State-of-the-art Recent Proposal

The proposal presented in [13] constitutes one of the most recent state-of-the-art proposals. In these, a model is proposed that follows the Learning by Contrast strategy [4] and is named CLASSIC, modifying the BERT architecture at two points (i.e., adding two fully connected network layers) and only adjusting the weights of these new components during training.

Table 7. Results in [13] against BSp_{LLA}.

Modelo	CLASSIC [13]	BSp_{LLA}
Accr	0.90	**0.80**
F1	0.85	**0.73**

According to the authors of CLASSIC, this model performs better than LLA (see Table 7). But when the model proposal and their evaluation method was analyzed, notable differences were observed concerning those used in LLA model:

- Experimentation with 19 datasets (13 more than those used in BSp_{LLA}).
- In the CLASSIC model training, the process took five datasets randomly to estimate the experimental results. But these five datasets are not named in [13] and cannot be compared with those of LLA model training.
- In [13] the number of datasets to train the model (i.e., in a CL framework) was not declared. This value will determine if the sample size is significant.
- There are different adjustments to the CLASSIC neural network hyperparameters in the learning process, for instance, the number of 30 epochs for the electrical device datasets and 10 for the laptop and restaurant review dataset, as a consequence of instances number.
- For the all datasets, a similar hyperparameters (e.g., epoch, batch, etc.) was used by the LLA model in the training process.
- A different form of input to the BERT architecture was found by code analysis of the CLASSIC model. It is the opposite of that used in LLA input.
- To estimate catastrophic forgetting, CLASSIC was used in the measure proposed in [3], which differs from that used in LLA model experimentation.
- In CLASSIC, there is no semantic closeness analysis of the datasets. It does not determine if the learning and the final results are on close datasets or not.

The generalization (i.e., same neural network hyperparameters for all datasets) is relevant in a CL model training framework (e.g., a homogeneous training process in incremental learning). Another disadvantage of configuration changes is the need to distinguish the type of datasets to adjust the settings (e.g., epochs number) because it is necessary to use another model or external tool for this purpose (i.e., increased computational cost in training time and memory).

Based on these differences, a comparative experiment was realized for both models. The evaluated criteria were:

1) A comparison of both models (i.e., CLASSIC and BSp_{LLA}) on the same training and test datasets for Continuous Learning (The datasets used by

Table 8. The experiment results to estimate the best performance between [13] and BSp_{LLA}.

Experiment name	CLASSIC	$\boldsymbol{BSp_{LLA}}$
Same-phd	0.311	**0.316**
Same-parameters	0.182	**0.316**
Invert-input	0.311	**0.316**

the BSp_{LLA} model training process, because they have a semantic closeness study).

2) Use of the same measure to estimate catastrophic forgetting and proposed in CLASSIC.
3) Use of the same hyperparameters for all datasets as in BSp_{LLA}.
4) The model classification effectivity was estimated based on the averaged values of the F1-macro.

In the experiment with the same dataset (as shown Table 8 for "Same-PhD" results), the value of the BSp_{LLA} model did not obtain a significant difference (i.e., BSp_{LLA} has a 0.005). However, during this experiment, the CLASSIC model maintained different hyperparameter values for datasets (e.g., the dataset of the electronic device has a higher epoch and batch than the laptop or restaurant dataset).

The hyperparameter values influence the results because the model can learn better by performing a more extensive search for better solutions (depending on the type of dataset or domain). It is a disadvantage of the CLASSIC model concerning the BSp_{LLA}, as explained above.

The CLASSIC disadvantages to reducing catastrophic forgetting are shown by the result of the experiment where the same hyperparameters were kept for all datasets (See Table 8 for "Same-parameters" results). The difference between the results of these models was 0.134, and the CLASSIC model is not better than LLA because it does not generalize the hyperparameters for all sets of variables and does not selectively preserve neural network weights during CL (as BSp_{LLA}).

The experiment value "Invert-input" obtains a similar outcome to the experiment named "Same-PhD". Because in "Invert-input" the neural network hyperparameters as in CLASSIC have been maintained. The main difference in "Invert-input" is that the form of representation of the input data to the CLASSIC model was similar to the LLA training. This experiment shows that this modification did not have a results impact.

Finally, the results of the experiments (Table 8) shown that the BSpLLA model has a positive influence on avoiding catastrophic forgetting and the final results because it avoids changes in the neural network weights value during the calculation of the gradient descent. This conclusion was established by analyzing Equation 2. In this equation, terms such as Ω_k^{μ} allow to compensate or avoid weight changes on Gradient Descent.

Although omission or forgetting occurs when different instances appear in new domains, weight compensation is fundamental in a CL process. It allows previous knowledge not to be complete or partially modified.

The adjustment of the weights of a part of the BERT neural network (as was proposed by CLASSIC) is not better than a CL model based on regularization (i.e., BSpLLA), which preserves the weight values in the last layer associated with the aspect classification process.

In CLASSIC, two new components were added to the BERT network architecture, only in a specific part. Therefore, the neural network weight updating only in these components decreases computational cost in terms of execution time in training. However, this model has no compensation or regularization during neural network weight updating.

Models with BERT's output as their input or base model perform better than those that use Word Embeddings. This result is similar to those reported in SOTA [7,8] and is associated with the architecture that follows the BERT model and its learning process.

Results obtained by applying the Friedman and Holm's tests do not show significant differences from other SOTA approaches. However, the results validate the selection of the SI approach as a catastrophic forgetting reducing mechanism and it has a lower computational cost than EwC [25]. The SI weight importance update method during SGD is part of BSp_{LLA}.

Although the experimentation did not evaluate models that follow the few-shot learning approach, the BSp_{LLA} outperformed other SOTA models with high classification measure values as a result of their architecture and main components.

4. Conclusion

In state-of-art, several sentiment analysis models are trained on a single domain (i.e., restaurant or hotel reviews) or dataset. The effectiveness decreases when these models learn patterns in a new domain in a CL framework. This paper presents a novel model that combines an attentional deep learning approach with a CL model to classify aspects in the sentiment analysis context. This model allows improvement of products and services (as part of information retrieval systems) in areas such as tourism, government, and health.

The model learning process uses data from multiple domains, and retains common information patterns for a new domain with relevant results (F1-macro = 73%). The input layer was the pretrained model BERT. The CL approach was named Lifelong Learning of Aspects (LLA) with a regularization approach. The evaluation results are better than those obtained by the existing regularization approaches such as EWC, AR1, and CLASSIC.

The LLA reduces catastrophic forgetting in the multi-domain context, and it is a novel approach in the ABSA context. Although the dataset's order influence on the learning process has been evaluated, it is necessary to deepen these experiments.

There are few studies on the linguistic rule's effectiveness in classifying aspects of the Sentiment Analysis task. However, the use of the linguistic rules, in the

deep and Continual Learning models combinations, is an interesting methodology that could be evaluated given the small number of labeled datasets in multiple domains. The future work areas will be extending our approach to other language models such as Spanish and combining it with linguistic rules and few-shot learning strategies.

Notes

[1] https://amazon.com/
[2] https://alibaba.com/
[3] https://spacy.io
[4] https://huggingface.co/models
[5] Follow a uniform distribution $U(-1, 1)$ at the time of assigning the values to the initial weights of the network
[6] https://pypi.org/project/transformers/2.1.0/
[7] https://github.com/dionis/ABSA-DeepMultidomain/

AUTHORS
Dionis López* – Faculty of Engineering in Telecommunications, Informatics and Biomedical, Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba, e-mail: dionis@uo.edu.cu, www: https://www.linkedin.com/in/ ~dionis-lopez-ramos.
Fernando Artigas-Fuentes – Center for Neuroscience Studies and Image and Signal Processing, Faculty of Engineering in Telecommunications, Informatics and Biomedical, Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba, e-mail: artigas@uo.edu.cu.
*Corresponding author

ACKNOWLEDGEMENTS

This work was supported by "Center of Informatics Research," Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba.

References

[1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. "Memory aware synapses: Learning what (not) to forget", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[2] M. Biesialska, K. Biesialska, and M. R. Costa-jussà. "Continual lifelong learning in natural language processing: A survey", *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6523–6541.

[3] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr. "Riemannian walk for incremental learning: Understanding forgetting and intransigence", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–547.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations", *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[5] Z. Chen, and B. Liu. "Lifelong machine learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, 2018, pp. 1–207.

[6] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. "A continual learning survey: Defying forgetting in classification tasks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[8] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon. "Deep learning for aspect-based sentiment analysis: a comparative review", *Expert Systems with Applications*, vol. 118, 2019, pp. 272–299.

[9] R. M. French. "Catastrophic forgetting in connectionist networks", *Trends in Cognitive Sciences*, vol. 3, no. 4, 1999, pp. 128–135.

[10] M. Hoang and A. Bihorac. "Aspect-based sentiment analysis using the pre-trained language model BERT", 2019.

[11] M. Huang, Y. Wang, X. Zhu, and L. Zhao. "Attention-based LSTM for aspect-level sentiment classification", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, 2016, pp. 606–615.

[12] Z. Ke, B. Liu, H. Wang, and L. Shu. "Continual learning with knowledge transfer for sentiment classification", *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, vol. 3, 2020, pp. 683–698.

[13] Z. Ke, B. Liu, H. Xu, and L. Shu. "CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6871–6883.

[14] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang. "Overcoming catastrophic forgetting by incremental moment matching", *Advances in neural information processing systems*, vol. 30, 2017.

[15] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*, Cambridge University Press, 2020.

[16] V. Lomonaco. *Continual learning with deep architectures*. PhD thesis, Universidad de Bologna, Italia, 2019.

[17] D. López and L. Arco. "Multi-domain aspect extraction based on deep and lifelong learning", *Iberoamerican Congress on Pattern Recognition*, 2019, pp. 556–565.

[18] D. Lopez-Paz. "Gradient episodic memory for continual learning", *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[19] D. López Ramos and L. Arco García. "Aprendizaje profundo para la extracción de aspectos en opiniones textuales", *Revista Cubana de Ciencias Informáticas*, vol. 13, no. 2, 2019, pp. 105–145.

[20] A. Mallya, and S. Lazebnik. "Packnet: Adding multiple tasks to a single network by iterative pruning", *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765–7773.

[21] D. Maltoni, and V. Lomonaco. "Continuous learning in single-incremental-task scenarios", *Neural Networks*, vol. 116, 2019, pp. 56–73.

[22] M. McCloskey and N. J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem", *Psychology of learning and motivation*, vol. 24, 1989, pp. 109–165.

[23] A. Nazir, Y. Rao, L. Wu, and L. Sun. "Issues and challenges of aspect-based sentiment analysis: a comprehensive survey", *IEEE Transactions on Affective Computing*, vol. 13, no. 2, 2020.

[24] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. "Continual lifelong learning with neural networks: a review", *Neural Networks*, vol. 113, 2019, pp. 54–71.

[25] G. I. Parisi and V. Lomonaco. "Online continual learning on sequences". *Recent Trends in Learning From Data*, pp. 197–221. New York Springer, 2020.

[26] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. "Semeval-2014 task 4: aspect based sentiment analysis", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 27–35.

[27] Y. Ren, Y. Zhang, M. Zhang, and D. Ji. "Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings", *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[28] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl. "Adapt or get left behind: domain adaptation through bert language model finetuning for aspect-target sentiment classification", *arXiv preprint arXiv:1908.11860*, 2019.

[29] J. Serra, D. Suris, M. Miron, and A. Karatzoglou. "Overcoming catastrophic forgetting with hard attention to the task", *International Conference on Machine Learning*, 2018, pp. 4548–4557.

[30] R. Singh, and S. Singh. "Text similarity measures in news articles by vector space model using nlp", *The Institution of Engineers (India): Series B*, vol. 102, no. 2, 2021, pp. 329–338.

[31] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao. "Attentional encoder network for targeted sentiment classification", *arXiv preprint arXiv:1902.09314*, 2019.

[32] F. Tang, L. Fu, B. Yao, and W. Xu. "Aspect based fine-grained sentiment analysis for online reviews", *Information Sciences*, vol. 488, 2019, pp. 190–204.

[33] E. Terra, A. Mohammed, and H. Hefny. "An approach for textual based clustering using word embedding". *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, pp. 261–280. Springer, 2021.

[34] G. M. Van de Ven, and A. S. Tolias. "Three scenarios for continual learning", *NeurIPS Continual Learning Workshop*, vol. 1, no. 9, 2018.

[35] S. Wang, G. Lv, S. Mazumder, G. Fei, and B. Liu. "Lifelong learning memory networks for aspect sentiment classification", *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 861–870.

[36] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, and R. Wang, "Modality-specific and shared generative adversarial network for cross-modal retrieval", *Pattern Recognition*, vol. 104, 2020, 107335.

[37] B. Zeng, H. Yang, R. Xu, W. Zhou, and X. Han. "LCF: a local context focus mechanism for aspect-based sentiment classification", *Applied Sciences*, vol. 9, no. 16, 2019, 3389.

[38] F. Zenke, B. Poole, and S. Ganguli. "Continual learning through synaptic intelligence", *International Conference on Machine Learning*, 2017, pp. 3987–3995.

[39] J. Zhou, J. X. Huang, Q. Chen, Q. V. Hu, T. Wang, and L. He. "Deep Learning for aspect-level sentiment classification: survey, vision and challenges", *IEEE Access*, vol. 7, 2019, pp. 78454–78483.

[40] K. M. Zorn, D. H. Foil, T. R. Lane, D. P. Russo, W. Hillwalker, D. J. Feifarek, F. Jones, W. D. Klaren, A. M. Brinkman, and S. Ekins. "Machine learning models for estrogen receptor bioactivity and endocrine disruption prediction", *Environmental Science & Technology*, vol. 54, no. 19, 2020, pp. 12202–12213.