Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Gated recurrent unit with multilingual universal sentence encoder for Arabic aspect-based sentiment analysis

Mohammad AL-Smadi [a,*], Mahmoud M. Hammad [a], Sa'ad A. Al-Zboon [a], Saja AL-Tawalbeh [a], Erik Cambria [b]

[a] Jordan University of Science and Technology, Jordan
[b] Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 23 December 2020
Received in revised form 22 September 2021
Accepted 25 September 2021
Available online 13 October 2021

Keywords:
Sentiment analysis
Aspect based sentiment analysis
Deep learning
GRU
Pooled-GRU
Universal sentence encoder

ABSTRACT

The increasing interactive content in the Internet motivated researchers and data scientists to conduct Aspect-Based Sentiment Analysis (ABSA) research to understand the various sentiments and the different aspects of a product in a single user's comment. Determining the various aspects along with their polarities (positive, negative, or neutral) from a single comment is a challenging problem. To this end, we have designed and developed a deep learning model based on Gated Recurrent Units (GRU) and features extracted using the Multilingual Universal Sentence Encoder (MUSE). The proposed Pooled-GRU model trained on a Hotels' Arabic reviews to address two ABSA tasks: (1) aspect extraction, and (2) aspect polarity classification. The proposed model achieved high results with 93.0% F1 score in the former task and 90.86% F1 score in the latter task. Our experimental results show that our proposed model outperforms the baseline model and the related research methods evaluated on the same dataset. More precisely, our proposed model showed 62.1% improvement in the F1 score over the baseline model for the aspect extraction task and 15% improvement in the accuracy over the baseline model for the aspect polarity classification task.

1. Introduction

The Web has enabled a huge number of users to express their feelings, opinions, and thoughts in their daily life using different tools [1,2]. Social media websites such as Twitter, Facebook, and Blogs are considered the main source for users to express their opinions about certain services, organizations, and governments [3–5]. Recently, these platforms attracted many researchers especially the Natural Language Processing (NLP) researchers to develop models for named entity recognition (NER) [6], text generation [7], dialogue systems [8], emotion detection [9], and sentiment analysis [10].

Sentiment Analysis, also known as Opinion Mining, is the study that analyze people's written text to detect or understand their opinions, sentiments, passions, attitudes, or emotions [11]. Sentiment analysis has been utilized in various domains including healthcare, finance, product consumption, and utilized in many domains including question answering systems [12], hotels, restaurants, products, and many other services as mentioned in [13]. Sentiment analysis aims to determine the polarity (positive or negative) of a product or a service from a sentence. However, people care about various aspects of the same product and different people care about different aspects of the same product. For example, a smartphone device has many aspects such as screen resolution, battery lifetime, or the weight of the phone. However, sentiment analysis determines the sentiment polarity of a product from a text regardless to different aspects of a product. To solve this problem, ABSA has been emerged and is widely adopted in academia as well as in industry. ABSA determines the different sentiments for various aspects from the same sentence about a product. For example, a user's comment might indicate that a user likes the screen resolution of the phone but not its battery lifetime. Various ABSA approaches have been developed to determine the various aspects along with their polarities in a single review about a product [14–17].

The main difference between sentiment analysis and ABSA, sentiment analysis tends to detect the sentiment through a given text. On the other hand, ABSA is a technique that determines the various aspects in a text and the sentiment (positive, negative or neutral) of each identified aspect on that text. ABSA receives a set of texts (product reviews, comments, etc.) about a particular entity, such as Smart Phone or Laptop, then the system tries to find

* Corresponding author.
E-mail addresses: masmadi@just.edu.jo (M. AL-Smadi), m-hammad@just.edu.jo (M.M. Hammad), saalzboon16@cit.just.edu.jo (S.A. Al-Zboon), sktawalbeh16@cit.just.edu.jo (S. AL-Tawalbeh), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.knosys.2021.107540
0950-7051/© 2021 Elsevier B.V. All rights reserved.

the most frequently discussed features (screen, size, price, etc.) of that entity. The main purpose is to determine the sentiment shown in each aspect along with their summary of polarity [18]. There are four main challenges (tasks) in the ABSA research [11]:

- Aspect Term Extraction (AE): which is a process of identifying all the aspect terms in each sentence. It recognizes aspects as an information extraction task, which stated an aspect to be expressed by a noun, verb, adverb, and adjective. The aspect terms might also be expressed by multi-word entities such as “battery backup" which is much critical than single word aspects. To extract aspect terms, various features have been used like Word N-grams, Bigrams, Name List, Headword, Word cluster, Casting, POS tagging, Parse dependencies, and also the punctuation marks. For the stage of AE, many methods have been utilized such as Conditional Random Fields (CRF), Support Vector Machines (SVM), Random trees and Random Forest.
- Aspect Term Polarity or Polarity Classification (PC): this task aims to determine the polarity of each aspect term as positive, negative, neutral, or conflict. Word N-grams, Polarity of neighboring adjectives, Neighboring POS tags and parse dependencies are the most features used by researchers.
- Aspect Category Detection: a task that identifies the main categories debated in each sentence. It is based on a set of binary Maximum Entropy classifiers.
- Aspect Category Polarity: a task that depends on the information taken from the previous Aspect Category Detection task to determine the polarity (positive, negative, neutral, or conflict) of each aspect category discussed in the reviewed sentence. It is being computed by calculating the distance between the n-gram and the corresponding aspect.

In this research, an enhancement type of Recurrent Neural Networks (RNN) namely GRU is developed to tackle two tasks of the ABSA problem. The implemented GRU has been used to solve both tasks of ABSA on sentence-level: (1) AE task and (b) PC task. Sentence-level embeddings have been extracted using MUSE [19] which represents a group of sentence encoding models. Using MUSE, similar reviews have similar vector representations and hence MUSE enhances the results where obtaining opinionated aspects and their polarity relying on the reviews contextual information. Our proposed Pooled-GRU rely on the features extracted from Pooled-GRU model [20] and the embedding from MUSE. A reference of a human-annotated Arabic Hotels' reviews obtained from the SemEva2016 Task5 has been used to evaluate our model. Experimental results show that our proposed Pooled-GRU with MUSE model outperforms a baseline approach [14] with an enhancement of 62.1% in the F1 score for the AE task and an improvement of 15% in the accuracy in the PC task. Our model achieved a 93.0% F1 score in the AE task and a 90.86% F1 score in the PC task. These high accuracy corroborate the ability of our approach in determining the AE and the PC with high confidence and hence its applicability in practice.

The remainder of this paper is organized as follows. Section 2 discusses the related research efforts to our work. Section 3 describes our methodology to design and develop an accurate model to solve the AE and the PC tasks. Section 4 presents the experimental results of our implemented model for both ABSA tasks and discusses them in Section 5. Finally, the paper concludes with avenue of future work in Section 6.

2. Literature review

This section discusses the related research efforts in developing ABSA approaches targeting the English language (Section 2.1) and the Arabic language (Section 2.2).

2.1. Aspect based sentiment analysis for english and foreign languages

Various research work have been conducted to develop efficient ABSA systems. It started with earlier stages of traditional methods such as association mining as in [21,22], and Likelihood ratio as in [23] as unsupervised machine learning techniques. Moving to the supervised learning techniques with the CRF approach discussed in [24], SVM machine learning approach [25] in many studies [26], and even HMM as in [27]. Moreover, Semi-supervised learning as Double propagation–syntactic relation methods have been used in [28].

More recently deep learning approaches have been successfully applied in many NLP tasks including fine-grained tasks of the ABSA. Reference [29] designed a deep neural learning model in order to analyze the aspect-based sentiments under the SemEval'15 subtasks. They proposed a novel approach that can connect sentiments with their corresponding aspects based on the constituency parse tree. Some researchers worked on predicting sentiment from videos as in [30,31]. Our work complement these works but focuses on extracting sentiment from Arabic texts rather than videos.

A comprehensive overview of the main deep learning techniques was presented in [32]. They discussed the differences between different deep learning approaches to tackle the sentiment analysis problem at the aspect level. In their analysis, they summarized about 40 approaches and categorized them based to their major architecture and classification tasks. They found that the most used approaches include the standard and variants of Convolutional Neural Networks (CNN), GRU, and Long-Short Term Memory (LSTM). To raise the performance of models, researchers have done pre-trained and fine-tuned word embedding's. They also discussed various usage of linguistic factors such as part-of-speech and grammatical rules.

Reference [33] has been developed to expand LSTM (TD – LSTM) and target-connection LSTM (TC – LSTM) by considering the target as a feature and combined it with context features. Reference [34] leveraged weakly supervised CNN for aspect level sentiment classification. Initially, the model learns a representation of a sentence that is weakly supervised by the overall review ratings, then it applies the aspect-level labels for fine-tuning. Reference [35] suggested the use of a hierarchical and bidirectional LSTM model for aspect-level identification of sentiments that can optimize intra- and inter-sentence relationships.

Reference [36] integrated the task of defining the target into the task of classifying sentiments in order to improve the model aspect-based sentiment interaction. They demonstrated that an end-to-end machine learning architecture can solve sentiment recognition, in which a deep memory network interleaves the two subtasks.

Reference [37] used a Bi-directional LSTM RNN for aspect extraction to evaluate inputs in both directions at the same time: forward and backward. The sentence presentation was using word vectors. And for the need of preventing over fitting of the models, they used pre-trained word embeddings. For the sentiment model, they used an enhanced version of the model presented by [38]. In the output layer of the CNN model, they had a single neuron with sigmoid activation function. After that, the model was trained on a dataset of sentences along with their sentiment intensity score. And for the aspect-based sentiment model, they altered the work presented in [29] by using the dependency tree in lieu of using the constituency tree of the sentence, as they are designed to represent relationships between words in a sentence.

Reference [39] presented a neural network architecture combined with two extensions of the standard LSTM, to accomplish the task of targeted aspect-based sentiment analysis. They represented the process of inferring sentiment aspects in addition to the polarity explicitly as being a two-step attention model that can encode the target and the full sentences.

Reference [40] used two LSTM networks in order to model sentences and aspects separate manner. In addition, they used the hidden states that were generated from sentences for calculating the attentions to aspect targets through the pooling operation and conversely. For the same mission, reference [41] model utilized an Attention-over-Attention module in order to learn the main and important parts for the aspect and sentence, where that intern creates the eventual representation of the sentence. Recently, pre-trained models including BERT [42], OpenAI GPT [43], and ELMo [44] proved to be efficient in minimizing effort of feature engineering.

In order to convert ABSA from a single sentence classification task to the form of a pair classification task, reference [45] created an auxiliary sentence. They adapt the BERT system on the task of classifying sentence pairs and obtained the new state-of-the-art tests. After that, they compared the results of single sentence classification that differs from sentence pair classification using the BERT fine-tuning. Then, they evaluated the advantages of sentence pair classification to check whether their conversion methods are valid.

Reference [46] searched over the effectiveness of the BERT embedding component on the task of end-to-end aspect-based sentiment analysis (E2EABSA). In particular, they have explored pairs of the BERT embedding element with many different neural models for performing comprehensive experiments on two benchmarks datasets. The experimental results indicated the dominance of BERT-based models in catching the aspect-based sentiments and proved its robustness in over fitting.

Reference [47] reviewed ABSA research efforts in a multilingual settings that focused not only on formal but informal and scarce resource language used in social media platforms. Not like our research which focus on Arabic text, they mainly considered English-based informal languages.

Recently, some researchers leveraged an attention-based word level contextual approach for sentiment analysis as in [48,49], and [50]. On the other hand, reference [51] applied ensemble approach of symbolic (ontologies) and subsymbolic (statistical NLP) AI for polarity detection task of sentiment analysis. Others applied stacked ensemble technique to predict intensities of emotions [52].

All of the aforementioned research efforts are related to our work regarding developing ABSA techniques to solve one of the ABSA tasks. However, their models have not been trained or evaluated on an Arabic dataset. Arabic language introduces many challenges requiring models designed specifically for the Arabic language.

2.2. Aspect based sentiment analysis for Arabic language

The Arabic NLP domain has few number of resources in comparison with other languages. The majority of the research efforts discussed before were manipulated with the English language. The research on other languages on ABSA are very minimal including the Arabic language [53]. Recently, reference [54] conducted a qualitative study of the sentiment analysis for Arabic text and discussed their strengths and limitations. They suggested shifting from simplistic word-level sentiment analysis to concept-based sentiment analysis due to the complex morphology of the Arabic language. Moreover, they showed that the deep learning is not fully explored in sentiment analysis for Arabic comparing to sentiment analysis for English. Following is a discussion for further related works that conducted on the Arabic language starting from earlier traditional ABSA methods toward the most recent deep learning methods.

Reference [55] worked over the Gaza-Israel dataset, and they experimented with the POS and NER types. They found that both POS and NER are playing an important role in ABSA of the news affect evaluation. They used traditional machine learning classifiers such as CRF, decision tree (J48), Naive Bayes, and K-nearest neighbors (IBk). Also, reference [53] proposed a method that extracts n-grams features and utilizes an SVM classifier, to allow researchers to compare their system performance.

Reference [56] used several supervised machine learning approaches of a set of classifiers in addition to syntactic, morphological, and semantic features. They used the Weka tool in order to apply the SVM, IBK, and J48 classifiers. And then, they compared their results with Bayesian Networks and Naïve Bayes models.

Reference [57] studied the ABSA for Arabic Hotels reviews, and the dataset used contained 2,291 Arabic reviews. Their dataset was fitted using AraNLP [58] and MADAMIRA [59] tools, where the authors used them to extract syntactic, semantic and morphological features to improve their results. Then they utilized an RNN model using the Deeplearning4j Framework [60] to employ their solution. Their model consisted of 5 hidden layers and achieved 87% accuracy.

Moreover, reference [61] proposed a deep learning model to analyze ABSA for Arabic. Their proposed model was built based on the LSTM network. The input to the model is the text embeddings combined with the aspect embedding and the output of the LSTM layer was passed to a hidden layer in order to compute the attention weight, `vector a`, and the feature vector of the sentence and its desired aspect, `vector r`. After that, `vector a and r` were used within a softmax layer for the purpose of predicting the sentiment expressed in an aspect. The model was also tested on the Arabic Hotels reviews dataset and achieved an 82.6% accuracy.

Reference [62] used a large dataset collected from publicly available Arabic text resources including newspapers and reviews to train their word embeddings model. Then, they generated and trained word-vector embeddings from the mentioned corpus. They also trained many other binary classifiers by using their word embeddings model using vector-based features as a representation for feature detection of sentiment and subjectivity over 3 different datasets of Arabic dialect language. Their best performing model achieved 77.87% accuracy.

Reference [63] compared between different classifiers used as baseline models. For polarity reporting, they ensembled an imbalanced dataset of Arabic dialect language tweets using features learned by word embeddings instead of using hand-crafted features. They showed that applying word embedding with the ensemble combining with synthetic minority over-sampling technique (SMOTE) increases the F1-score by 15% on average. Finally, reference [64] employed word embeddings to implement ABSA on Arabic flight tweets. They utilized two pre-trained word-vector models providing vector-based features. They also showed the effectiveness of their model for the tasks of aspect detection and aspect-based sentiment detection. The vector space approach they used utilized these features without hand-crafted features to be performed in comparison with techniques that engineered features adopted manually. Their results showed that word embedding features are simple and can be used to boost the results obtained by the state-of-the-art techniques in performing ABSA over the Arabic language.

Unlike our work, all previous work are either applied traditional machine learning models as in [53,55,56] or achieved a lower accuracy than our model as in [57,61,62]. The best performing model in the mentioned work achieved an accuracy of 82% whereas our model achieved as high as 93% accuracy.

Table 1
The distribution of the train/test dataset over the various ABSA SemEval-2016-Task5 research tasks.

Task/size	Train			Test		
	Text	Sentence	Tuples	Text	Sentence	Tuples
T1: Sentence-level ABSA	1,839	10,509	8,757	452	1,227	2,604
T2: Text-level ABSA	1,839	4,802	8,757	452	1,227	2,158

Table 2
Examples of the annotated Arabic reviews of the Hotel dataset.

Sentence	Aspects	Polarity
الجودة ذاتها والخدمة في كل فنادق آيبيس The same quality and service at all Ibis hotels	الجودة ـ الخدمة ـ فنادق Hotels - Service - Quality	الجودة ـ الخدمة ـ فنادق Hotels - Service - Quality neutral - B-positive - B-A B-positive
مرافق الغرفة جيدة والخدمة ودودة ومتعاونة Room facilities are good and the service is friendly and helpful	مرافق ـ الخدمة ـ ومتعاونة Helpful - Service - Facilities	مرافق ـ الخدمة ـ ومتعاونة Helpful - Service - Facilities B-positive - B-positive - B-positive
كانت الغرفة ممتازة وكذلك الموظفون وبوفيه الإفطار The room was excellent as well as the staff and breakfast buffet	الغرفة ـ الموظفون ـ بوفيه ـ الإفطار Breakfast - Buffet - Staff - Room	الغرفة ـ الموظفون ـ بوفيه ـ الإفطار Breakfast - Buffet - Staff - Room I-positive- B-positive -B-positive - B-positive

3. Research methodology

This section discusses the research methodology we followed to develop our ABSA model. Section 3.1 describes the dataset we utilized. Next, the data pre-processing technique we leveraged is described in 3.2. Finally, the proposed Pooled-GRU model with MUSE [19] is described in Section 3.4.

3.1. Dataset

The SemEval 2016 competition Task-5 [65] has published various well-annotated datasets that contain comments and reviews from customers about six different domains (restaurants, laptops, mobile phones, digital cameras, hotels, and museums). The comments in the datasets are written in 8 different languages (English, Arabic, Chinese, Dutch, French, Russian, Spanish, and Turkish). To train and evaluate our models, we have utilized the Arabic reviews about hotels dataset [65], referred hereinafter to the dataset. According to the SemEval-ABSA16 annotation guidelines, the dataset is manually annotated by a research group of three native Arabic speakers and validated by a senior researcher.

The dataset provided with a baseline model based on a SVM model. The SVM model has been trained using the N-grams features only. The dataset contains 24,028 tuples split into two files: 19,226 instances (tuples) for training and 4,802 tuples for testing. Furthermore, the dataset has been annotated on both text-level with 2,291 reviews texts and sentence-level with 6,029 annotated sentences.

Table 1 shows the size of the dataset over the various ABSA research tasks. Each review in the sentence-level task has been annotated and represented as tuples where each tuple consists of (1) the aspect category that contains the aspect entity (E) and the aspect entity attribute (A) (E#A), (2) the opinion target expression (OTE), and (3) the aspect polarity.

The dataset contains the following entities: Hotels, Facilities, Location, Rooms, Service, Rooms Amenities, Drink, and Food. And the following attributes: General, Quality, Cleanliness, Design, Style, Comfort, Price, Options, Miscellaneous, and Features. Therefore, the E#A category could be any combination of an entity and an attribute such as `Food#Quality` or `Rooms#Cleanliness`. The considered aspect-category polarities in the dataset are `NEUTRAL`, `POSITIVE`, and `NEGATIVE`. Table 2 provides examples of annotated sentences in the ABSA dataset. This ABSA dataset has been utilized to evaluate our proposed model for the two considered research tasks namely: the aspect opinion target expression extraction and the polarity classification.

3.2. Dataset pre-processing

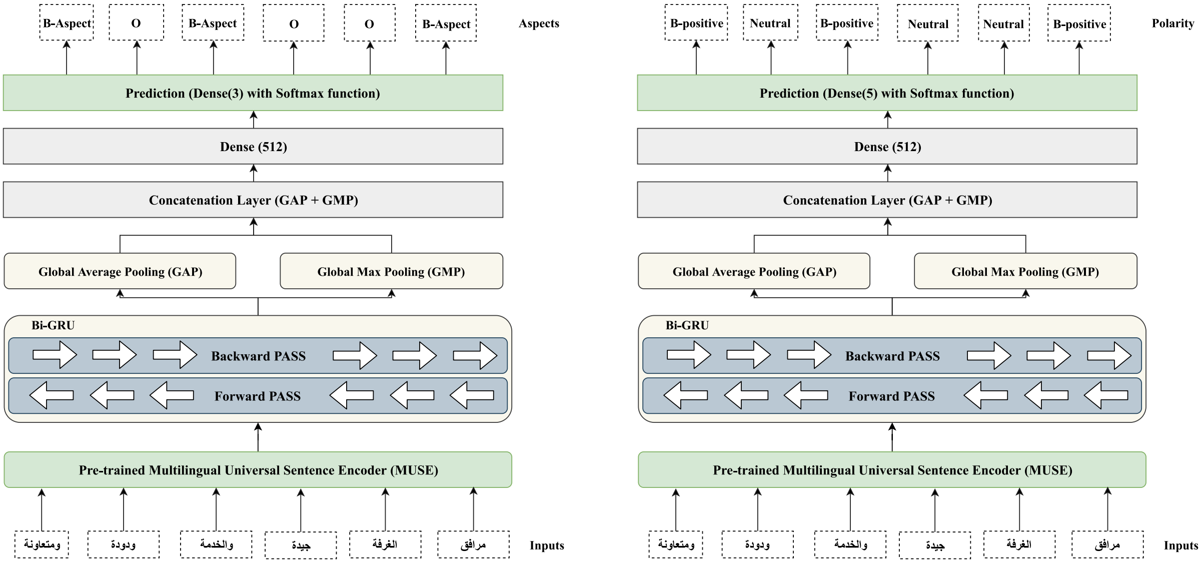
In order to analyze the potential sentences, data pre-processing is an important role to have a clean input that aims to simplify the model functionality and increases its accuracy. The cleaned version of the dataset will be used to feed the neural networks or any proposed approach. For the pre-processing step, we applied various processing techniques to reduce the unnecessary noise (i.e., punctuation and special characters (#, %, &, *, @, ؛ (,), ?, !). The dataset is available in XML format and we convert it to a CSV format using our Java code.

3.3. Multilingual universal sentence encoder (MUSE)

Word encoding or embedding is a crucial step in any NLP tasks using algorithms such as GloVe [66] or character embedding likes fastText [67]. Another technique named Universal Sentence Encoder (USE) was appeared lately related to a sentence or sequence embedding. USE is a widely used sentence encoding model released by Google in July of 2018 that provides sentence-level embedding vectors instead of word or character level embedding. USE model was implemented using two techniques: (1) a transformer [68] and (2) a deep averaging network (DAN) [69], and there are many versions of USE for English [19] and multilingual encoders (MUSE) [70]. MUSE encodes each sentence with a 512 vector representation and can handle a language-specific pretrained models since it includes several languages such as Arabic, English, Chinese, Dutch, German, French, etc.

3.4. Proposed pooled-GRU model

Vanilla Neural Networks have known limitations including the API is too constrained where networks receive a fixed-sized vector as input. Recurrent Neural Networks (RNNs) solves this problem by allowing a series of inputs while ignoring the fixed-size vector problem. RNNs tend to use previous outputs as an input state for the input layer with hidden states [71,72]. Nevertheless, RNNs suffer from gradient vanishing. In 1997 LSTMs [73] and in 2014 GRU [20] proposed as solutions to the gradient vanishing problem. Nowadays, LSTMs are widely used and adopted in deep learning approaches. In contrast, GRU started to take place in the research of NLP for several reasons: (1) the high performance of GRU with less complex structure than LSTMs, (2) high speed training where GRU consist of two gates (reset and update gate) and LSTMs consist of three gates (input, output, and forget gate). Our experimental results confirm the high performance of our GRU with MUSE approach. GRU applies update gate and reset gate where these vectors decide the selected information by previous activation s_{t-1} and candidate activation $\tilde{s}_t$ what should be


B-Aspect
O
B-Aspect
O
O
B-Aspect
Aspects
Prediction (Dense(3) with Softmax function)
Dense (512)
Concatenation Layer (GAP + GMP)
Global Average Pooling (GAP)
Global Max Pooling (GMP)
Bi-GRU
Backward PASS
Forward PASS
Pre-trained Multilingual Universal Sentence Encoder (MUSE)
ومتعاونة
ودودة
والخدمة
جيدة
الغرفة
مرافق
Inputs
B-positive
Neutral
B-positive
Neutral
Neutral
B-positive
Polarity
Prediction (Dense(5) with Softmax function)
Dense (512)
Concatenation Layer (GAP + GMP)
Global Average Pooling (GAP)
Global Max Pooling (GMP)
Bi-GRU
Backward PASS
Forward PASS
Pre-trained Multilingual Universal Sentence Encoder (MUSE)
ومتعاونة
ودودة
والخدمة
جيدة
الغرفة
مرافق
Inputs

(a) Aspects Extraction.

(b) Polarity Classification.

Fig. 1. The architecture of our implemented BiGRU model with MUSE.

passed to the output. Following are the mathematical equations that represent the GRU:

$$s_t = (1 - u_t) \odot s_{t-1} + u_t \odot \tilde{s}_t \tag{1}$$

where

$$\tilde{s}_t = tanh(W_s x_t + r_t \odot (z_s s_{t-1}) + b_s) \tag{2}$$

$$u_t = \sigma(W_u x_t + z_u s_{t-1} + b_s) \tag{3}$$

$$r_t = \sigma(W_r x_t + z_r s_{t-1} + b_s) \tag{4}$$

The update gate represents u_t Eq. (3), reset gate represent r_t Eq. (4), and $\tilde{s}_t$ Eq. (2) represents the current memory content. W_s, W_u, W_r, Z_s Z_u, Z_r, b_s these annotations represent the Weight matrices, x_t represents the vector input to the time-step t, st Eq. (1) represents the final current exposed hidden state, and $\odot$ represents the element-wise multiplication.

In this research, we proposed a Pooled-GRU with MUSE [19] to provide a new state-of-the-art results in the process of AE and PC. The BiGRU model architecture shown in Figs. 1(a) and 1(b) depicts the proposed architecture of both tasks. Each sentence encoded with a 512 vector representation using MUSE [19] as a pre-trained model. The vector of each sentence is passed to a Bidirectional with GRU layer with 250 neurons. BiGRU layers have been concatenated using Global Average Pooling (GAP) and Global Max Pooling (GMP). The intuition of using GAP and GMP is to reduces the dimensions of the previous layers while maintaining the important features. Then, The concatenated layer is passed into a Dense layer with 512 neurons. Finally, Dense layer of 3 neurons is used for the prediction layer with Softmax activation function.

Our BiGRU model has been trained using categorical cross-entropy as a loss function to reduce the amount of error that could happen. Table 3 illustrates the hyper-parameters utilized to train our proposed model for both tasks.

4. Experimentation setup and results

In this section, we present the experimentation setup and the parameters we utilized to train and evaluate our proposed pooled-gru model in 3.4. Then, we explain the evaluation measures used to evaluate the proposed models in 4.1. Finally, the results are presented in 4.2.

Table 3
Pooled-GRU Model Hyper-parameters For Both Tasks.

Model	Pooled-GRU + MUSE
Number of parameters	658,705
Number of epochs	25
Batch size	256
Optimizer	Adam
Learning rate	0.1
Embedding vector size	512

4.1. Evaluation measures

To evaluate the performance of our proposed model, we calculated the accuracy, precision, recall, and the F1-score of our model.

Following are specific definitions of the calculated measurements in the context of our ABSA problem.

- *True Positives* defined as the case in which the predicted aspect is *B-Aspect* and the actual aspect is also *B-Aspect*.
- *True Negatives* defined as the case in which the predicted aspect is *O* and the actual aspect is *O*.
- *False Positives* defined as the case in which the predicted aspect *B-Aspect* and the actual aspect is *O*.
- *False Negatives* defined as the case in which the predicted aspect *O* and the actual aspect is *B-Aspect*.

4.2. Results

Table 4 illustrates the results we obtained from the proposed pooled-gru model for both tasks of the ABSA. The results show that our pooled-gru model outperforms the baseline research model (an SVM-unigrams model), implemented in [14], and trained on the same dataset we trained our model, the SemEval2016 Task5 dataset [65]. Our model achieved 93% F1 measure comparing to the 30.9% F1 score achieved by the baseline model on the AE task, i.e., 63% improvement. Similarly, our proposed model achieved 90.86% F1 score with 91.4 accuracy where the baseline model achieved 76.4, i.e., 15% improvement, and the F1 score has not been reported in [14].

Table 4
Performance of our model (pooled-gru) comparing to the baseline model (SVM-unigram) [14] for both ABSA tasks.

Task	Approach	F1-Score (%)	Accuracy (%)	Precision (%)	Recall (%)
Aspect Extraction (AE)	Our Model Baseline	**93.00** 30.90	92.82 –	93.25 –	92.45 –
Polarity Classification (PC)	Our Model Baseline	**90.86** –	91.40 76.40	90.86 –	90.55 –

Table 5
The proposed approach results on both tasks compared to other related works.

Task	Approach	F1-Score (%)	Accuracy (%)
Aspect Extraction (AE)	Our Model Bi-LSTM \ [16]	**93.00** 69.98	92.82 –
Polarity Classification (PC)	Our Model Bi-LSTM \ [16] INSIGHT-1 \ [74] IIT-TUDA \ [75] IIT-TUDA+ Sentiment Lexicon \ [76]	90.86 – – – –	**91.40** 82.60 82.70 78.68 81.72

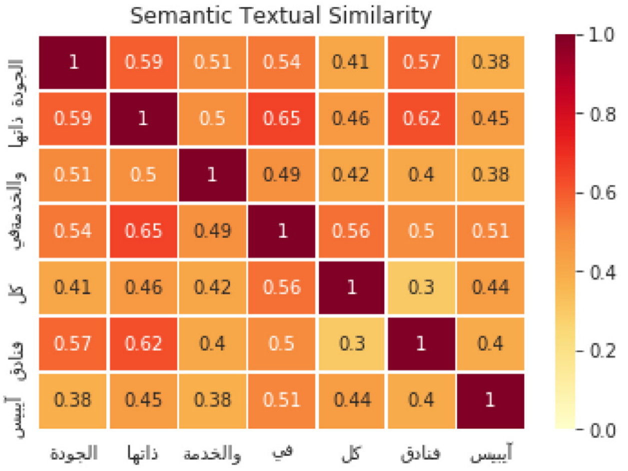


Fig. 2. Sentence similarity using MUSE with 1 aspect Similarity.

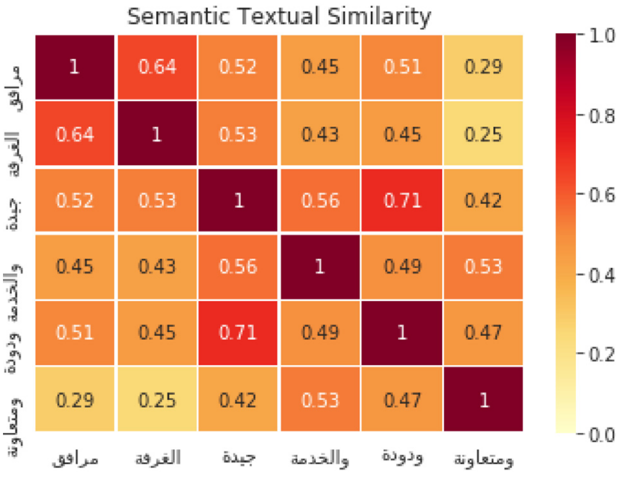


Fig. 3. Sentence similarity using MUSE with 2 aspects Similarity.

5. Discussion

Table 5 compares the results obtained from our proposed model with the results obtained from the related work that trained on the same dataset for both of the ABSA tasks. As shown in Table 5 shows that our proposed model outperforms all previous related work on both tasks On the AE task, our proposed model achieved F1 = 93.0% F1 score with an enhancement of 23.1% compared to the results obtained by [16]. Regarding to the second task, there are several submission on the ABSA SemEval2016 [65] competition. However, our proposed model achieved 91.4% accuracy with 9.68% improvement over the second best accuracy obtained by [76].

Figs. 2, 3, and 4 depict the ability of the MUSE [77] to detect semantic text similarity between two aspect words. For instance, in Fig. 3 the semantic text similarity between the two words, `Room` (second column from left) and `Facility` (first row from the top), representing the `Room Facilities` is detected with high values, i.e., 0.65, and the positive similarity represented between, `friendly` (fifth column) and `good` (third row), is also detected with high value of 0.71. This finding shows the ability of the MUSE to detect contextual-based information such as semantics relatedness including aspects and their polarities which has contributed to improve the performance of our proposed model.

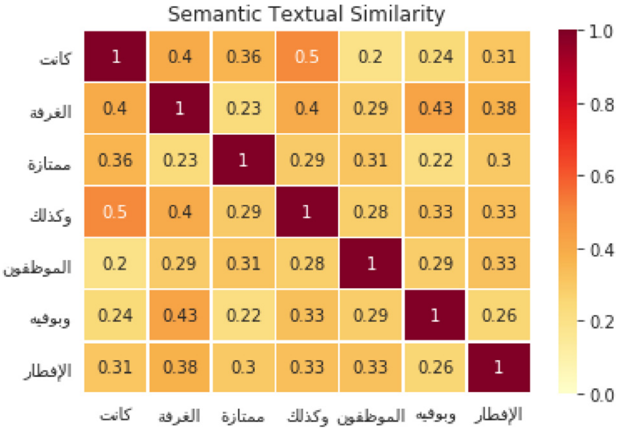


Fig. 4. Sentence similarity using USE with 3 aspects Similarity.

6. Conclusion and future work

In this paper, we have developed a deep learning model to solve two ABSA tasks: the AE task and the PC task. The proposed model is a BiGRU Recurrent Neural Networks model which provides better performance comparing to the LSTMs [20]. The Arabic Hotels' reviews [14,65] have been used to evaluate the proposed BiGRU approach. Experimental evaluation results show that our model outperformed the baseline research model as well as the models developed by the related work and evaluated on the same dataset. The implemented BiGRU depends on MUSE embedding instead of work of character embedding which shows significant improvement in the results with an enhancement of +62.1% in the F1 measure for the AE task and 15.0% in the accuracy for PC task compared to the baseline model. Our BiGRU approach achieved F1-measure of 93.0% in the AE task and 90.86% F1 score in the PC task.

For future work, we plan to use other embedding techniques that can be extracted from transformer such that, ElMo, Bert, and ULM-FiT to enhance our results.

CRediT authorship contribution statement

Mohammad AL-Smadi: Supervision, Conceptualization, Methodology, Writing – original draft. **Mahmoud M. Hammad:** Software, Validation, Writing – review & editing. **Sa'ad A. Al-Zboon:** Methodology, Software. **Saja AL-Tawalbeh:** Methodology, Software. **Erik Cambria:** Software, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is partially funded by Jordan University of Science and Technology, Research Grant Number s: 20200145 and 20190306.

References

[1] K.H. Yoo, U. Gretzel, What motivates consumers to write online travel reviews? Inf. Technol. Tourism 10 (4) (2008) 283–295.
[2] M. Grassi, E. Cambria, A. Hussain, F. Piazza, Sentic web: A new paradigm for managing social media affective information, Cogn. Comput. 3 (3) (2011) 480–489.
[3] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, J. Mar. Res. 43 (3) (2006) 345–354.
[4] X. Li, H. Xie, L. Chen, J. Wang, X. Deng, News impact on stock price return via sentiment analysis, Knowl.-Based Syst. 69 (2014) 14–23.
[5] G. Vinodhini, R. Chandrasekaran, A sampling based sentiment mining approach for e-commerce applications, Inf. Process. Manage. 53 (1) (2017) 223–236.
[6] X. Zhong, E. Cambria, A. Hussain, Extracting time expressions and named entities with constituent-based tagging schemes, Cogn. Comput. 12 (4) (2020) 844–862.
[7] Y. Li, Q. Pan, S. Wang, T. Yang, E. Cambria, A generative model for category text generation, Inform. Sci. 450 (2018) 301–315.
[8] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: AAAI, 2018, pp. 4970–4977.
[9] Y. Susanto, A. Livingstone, B.C. Ng, E. Cambria, The hourglass model revisited, IEEE Intell. Syst. 35 (5) (2020) 96–102.
[10] E. Cambria, Y. Song, H. Wang, N. Howard, Semantic multi-dimensional scaling for open-domain sentiment analysis, IEEE Intell. Syst. 29 (2) (2014) 44–51.
[11] N.K. Laskari, S.K. Sanampudi, Aspect based sentiment analysis survey, IOSR J. Comput. Eng. 18 (2) (2016) 24–28.
[12] T.H. Alwaneen, A.M. Azmi, H.A. Aboalsamh, E. Cambria, A. Hussain, Arabic question answering system: a survey, Artif. Intell. Rev. (2021) 1–47, http://dx.doi.org/10.1007/s10462-021-10031-1.
[13] T.A. Rana, Y.-N. Cheah, Aspect extraction in sentiment analysis: comparative analysis and survey, Artif. Intell. Rev. 46 (4) (2016) 459–483.
[14] M. Al Smadi, I. Obaidat, M. Al-Ayyoub, R. Mohawesh, Y. Jararweh, Using enhanced lexicon-based approaches for the determination of aspect categories and their polarities in Arabic reviews, Int. J. Inf. Technol. Web Eng. 11 (3) (2016) 15–31.
[15] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (1) (2012) 1–167.
[16] M. Al-Smadi, B. Talafha, M. Al-Ayyoub, Y. Jararweh, Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews, Int. J. Mach. Learn. Cybern. 10 (2018) 1–13.
[17] M. Al-Smadi, M. Al-Ayyoub, Y. Jararweh, O. Qawasmeh, Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features, Inf. Process. Manage. 56 (2) (2019) 308–319.
[18] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (4) (2014) 1093–1113.
[19] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R.S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, 2018, arXiv preprint arXiv:1803.11175.
[20] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[21] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 168–177.
[22] A.-M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Natural Language Processing and Text Mining, Springer, 2007, pp. 9–28.
[23] J. Yi, T. Nasukawa, R. Bunescu, W. Niblack, Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, in: Third IEEE International Conference on Data Mining, IEEE, 2003, pp. 427–434.
[24] N. Jakob, I. Gurevych, Extracting opinion targets in a single-and cross-domain setting with conditional random fields, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 1035–1045.
[25] J.S. Kessler, N. Nicolov, Targeting sentiment expressions through supervised ranking of linguistic configurations, in: Third International AAAI Conference on Weblogs and Social Media, 2009, pp. 90–97.
[26] V.J. Ashwath K, Aspect based sentiment analysis for E-commerce using classification techniques, Int. J. Intell. Sci. Eng. 1 (1) (2019) 35–42.
[27] W. Jin, H.H. Ho, R.K. Srihari, OpinionMiner: a novel machine learning system for web opinion mining and extraction, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 1195–1204.
[28] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Comput. Linguist. 37 (1) (2011) 9–27.
[29] B. Wang, M. Liu, Deep learning for aspect-based sentiment analysis, Stanford University Report, 2015.
[30] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95.
[31] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. (2021).
[32] H.H. Do, P. Prasad, A. Maag, A. Alsadoon, Deep learning for aspect-based sentiment analysis: a comparative review, Expert Syst. Appl. 118 (2019) 272–299.
[33] D. Tang, B. Qin, X. Feng, T. Liu, Effective LSTMs for target-dependent sentiment classification, 2015, arXiv preprint arXiv:1512.01100.
[34] Z. Guan, L. Chen, W. Zhao, Y. Zheng, S. Tan, D. Cai, Weakly-supervised deep learning for customer review sentiment classification., in: International Joint Conferences on Artificial Intelligence Organization, IJCAI, 2016, pp. 3719–3725.
[35] S. Ruder, P. Ghaffari, J.G. Breslin, A hierarchical model of reviews for aspect-based sentiment analysis, 2016, arXiv preprint arXiv:1609.02745.
[36] C. Li, X. Guo, Q. Mei, Deep memory networks for attitude identification, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 671–680.
[37] H. Jangid, S. Singhal, R.R. Shah, R. Zimmermann, Aspect-based financial sentiment analysis using deep learning, in: Companion Proceedings of the the Web Conference 2018, International World Wide Web Conferences Steering Committee, 2018, pp. 1961–1966.
[38] Y. Kim, Convolutional neural networks for sentence classification, 2014, arXiv preprint arXiv:1408.5882.
[39] Y. Ma, H. Peng, T. Khan, E. Cambria, A. Hussain, Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis, Cogn. Comput. 10 (4) (2018) 639–650.
[40] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, 2017, arXiv preprint arXiv:1709.00893.
[41] B. Huang, Y. Ou, K.M. Carley, Aspect level sentiment classification with attention-over-attention neural networks, in: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, 2018, pp. 197–206.

[42] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[43] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018, URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf.

[44] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018, arXiv preprint arXiv:1802.05365.

[45] C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, 2019, arXiv preprint arXiv:1903.09588.

[46] X. Li, L. Bing, W. Zhang, W. Lam, Exploiting BERT for end-to-end aspect-based sentiment analysis, 2019, arXiv preprint arXiv:1910.00883.

[47] S.L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artif. Intell. Rev. 48 (4) (2017) 499–527.

[48] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification, Int. J. Intell. Eng. Inf. 8 (1) (2020) 1–18.

[49] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1023–1032.

[50] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: AAAI, 2018, pp. 5876–5883.

[51] E. Cambria, Y. Li, F. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: CIKM, 2020, pp. 105–114.

[52] M.S. Akhtar, A. Ekbal, E. Cambria, How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes], IEEE Comput. Intell. Mag. 15 (1) (2020) 64–75.

[53] M. Al-Ayyoub, A. Gigieh, A. Al-Qwaqenah, M.N. Al-Kabi, B. Talafhah, I. Alsmadi, Aspect-based sentiment analysis of Arabic laptop reviews, in: The International Arab Conference on Information Technology (ACIT'2017), 2017.

[54] O. Oueslati, E. Cambria, M.B. HajHmida, H. Ounelli, A review of sentiment analysis research in Arabic language, Future Gener. Comput. Syst. 112 (2020) 408–430.

[55] A.-S. Mohammad, M. Al-Ayyoub, H.N. Al-Sarhan, Y. Jararweh, An aspect-based sentiment analysis approach to evaluating Arabic news affect on readers, J. UCS 22 (5) (2016) 630–649.

[56] M. Al-Smadi, M. Al-Ayyoub, Y. Jararweh, O. Qawasmeh, Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features, Inf. Process. Manage. 56 (2) (2019) 308–319.

[57] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, B. Gupta, Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews, J. Comput. Sci. 27 (2018) 386–393.

[58] M. Althobaiti, U. Kruschwitz, M. Poesio, AraNLP: a java-based library for the processing of Arabic text, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4134–4138.

[59] A. Pasha, M. Al-Badrashiny, M.T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, R. Roth, Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, in: LREC, 14, 2014, pp. 1094–1101.

[60] D. Deeplearning4j, Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0, 2016.

[61] M. Al-Smadi, B. Talafha, M. Al-Ayyoub, Y. Jararweh, Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews, Int. J. Mach. Learn. Cybern. 10 (8) (2019) 2163–2175.

[62] A.A. Altowayan, L. Tao, Word embeddings for Arabic sentiment analysis, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 3820–3825.

[63] S. Al-Azani, E.-S.M. El-Alfy, Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text, Procedia Comput. Sci. 109 (2017) 359–366.

[64] M.M. Ashi, M.A. Siddiqui, F. Nadeem, Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets, in: International Conference on Advanced Intelligent Systems and Informatics, Springer, 2018, pp. 241–251.

[65] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, G. Eryiğit, SemEval-2016 task 5: Aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEVal-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 19–30, http://dx.doi.org/10.18653/v1/S16-1002, URL https://www.aclweb.org/anthology/S16-1002.

[66] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[67] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, 2016, arXiv preprint arXiv:1612.03651.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[69] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, 2015, pp. 1681–1691.

[70] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G.H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al., Multilingual universal sentence encoder for semantic retrieval, 2019, arXiv preprint arXiv:1907.04307.

[71] A. Karpathy, The unreasonable effectiveness of recurrent neural networks, Andrej Karpathy Blog 21 (2015).

[72] C. Olah, Understanding lstm networks, 2015, URL https://research.google/pubs/pub45500/ (Last accessed Sep. 2021).

[73] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[74] S. Ruder, P. Ghaffari, J.G. Breslin, Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis, 2016, arXiv preprint arXiv:1609.02748.

[75] A. Kumar, S. Kohail, A. Kumar, A. Ekbal, C. Biemann, Iit-tuda at semeval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1129–1135.

[76] M. Salameh, S. Mohammad, S. Kiritchenko, Sentiment after translation: A case-study on Arabic social media posts, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 767–777.

[77] J. Ramaprabha, S. Das, P. Mukerjee, Survey on sentence similarity evaluation using deep learning, in: Journal of Physics: Conference Series, 1000, (1) IOP Publishing, 2018, 012070.