Comparing word embedding models for Arabic aspect category detection using a deep learning-based approach

Rajae BENSOLTANE^{1*} and Taher Zaki¹

¹RF-SIC Laboratory, Faculty of Science, Ibn Zohr University, FP-Agadir, Morocco

Abstract. Aspect category detection (ACD) is a task of aspect-based sentiment analysis (ABSA) that aims to identify the discussed category in a given review or sentence from a predefined list of categories. ABSA tasks were widely studied in English; however, studies in other low-resource languages such as Arabic are still limited. Moreover, most of the existing Arabic ABSA work is based on rule-based or feature-based machine learning models, which require a tedious task of feature-engineering and the use of external resources like lexicons. Therefore, the aim of this paper is to overcome these shortcomings by handling the ACD task using a deep learning method based on a bidirectional gated recurrent unit model. Additionally, we examine the impact of using different vector representation models on the performance of the proposed model. The experimental results show that our model outperforms the baseline and related work models significantly by achieving an enhanced F1-score of more than 7%.

Keywords: Aspect-based sentiment analysis, Aspect category detection, Word embeddings, Deep learning, BiGRU, Arabic.

1 Introduction

Sentiment analysis (SA) is one of the main tasks of natural language processing (NLP) that aims to extract opinions, thoughts, or attitudes toward a specific entity or subject. With the internet revolution, online users are creating large amounts of data every day. Analyzing these unstructured datasets is valuable for many entities (e.g., governments, companies, and customers) and applications (recommender systems). Therefore, SA has become one of the very active fields for researchers in both industry and academia.

SA can be handled on three levels, namely document level, sentence level, and aspect level. The first two levels identify the sentiment polarity of the whole document or sentence, which is not always beneficial. Users can express different opinions about different aspects or entities in the same text or review, making the aspect level, also known as aspect-based sentiment analysis (ABSA), more suitable and practical for real-life scenarios.

In the SemEval-16 task 5 [1], ABSA was divided into three main tasks: aspect category detection (ACD), opinion-target expression (OTE), and sentiment polarity identification. The objective of the first task is to identify aspect categories (E#A pairs) in a given review or

* Corresponding author: r.bensoltane@uiz.ac.ma

sentence from predefined inventories of entities and attributes. For the second task, the aim is to extract the opinion-target expression used to refer to each E#A pair. The last task aims to identify the sentiment polarity of each aspect. Fig. 1 shows an annotated review from the SemEval 2016 task-5 dataset. Aspect-based was widely studied in the English language, while research in other languages such as Arabic is still limited and requires further effort to enhance this task [2].

id = 1606:0			
text: من افضل فنادق المدينه المنوره انصح وبشده من يريد ان يشعر انه في بيته الثاني ان يذهب اليه اكل نظيف تعامل راقى			
target: اكل	category: FOOD_DRINKS#QUALITY	polarity: positive	
target: تعامل	category: SERVICE#GENERAL	polarity: positive	

Fig. 1. An annotated review from SemEval-2016 task 5 Arabic Hotel dataset.

The Arabic language is one of the four most used languages on the internet. It is the official language of 22 countries with more than 400 million people in the world [3]. Unlike English, Arabic has a rich and complex morphology and many varieties, making the tasks of SA and ABSA more difficult in this language. Most of the proposed methods in Arabic SA adopted rule-based or traditional machine-learning approaches. These methods require a laborious task of feature-engineering and the use of external resources (e.g., lexicons and NLP tools) that are unavailable or unreliable, especially in dialectical Arabic [4]. Recently, deep learning (DL) methods have demonstrated their effectiveness in handling many NLP tasks in English, including SA, without the need for manually engineered features or linguistic resources [5]. However, these techniques are under-discovered in Arabic, and a limited number of DL methods were published, especially in Arabic ABSA (AABSA). Therefore, the aim of this paper is to investigate the use of a bidirectional gated recurrent unit (BiGRU) model in handling ACD, which is, to our knowledge, the first time to apply this model to tackle this task in Arabic. Additionally, the impact of using different word embeddings models is also examined.

The key contributions of this paper are:

- Applying a BiGRU model to enhance the Arabic ACD task.
- Train a word2vec model for the Arabic Hotel domain and examine its effect on the performance of the proposed model. As far as we know, this is the first time to apply a domain-specific embedding model in Arabic ABSA.
- Comparing traditional word embeddings with contextual embeddings in Arabic ACD task.
- Outperforming the baseline and related work models evaluated on the same reference dataset.

The remainder of this paper is organized as follows: related work to the Arabic ACD task is discussed in section 2. Section 3 provides the research methodology. The dataset and baselines are presented in Section 4. Section 5 discusses the results. Conclusion and Future work directions are provided in Section 6.

2 Literature review

AABSA is receiving less attention from the research community compared to the English language. Meanwhile, the ACD task is less investigated than other ABSA tasks, namely

aspect term extraction and aspect sentiment polarity classification. For example, we couldn't find any Arabic ACD work published in 2019 or 2020. The first work in AABSA was that of Al-Smadi, Qawasmeh [6]. They provided the first annotated AABSA corpus following the annotation guidelines of SemEval 2014 task 4 [7]. They also reported a set of baseline results for each ABSA task, including ACD, which achieved an F1-score of 15.2%.

Obaidat, Mohawesh [8] tried to enhance the previous results using a lexicon-based approach. Fourteen lexicons were built for each category using seed words and Pointwise Mutual Information (PMI). The proposed method looked up for the words of a given review in each lexicon to sum up their weights and then assigned the review to the aspect category with the highest sum value. However, the improvement over the baseline's result was not impressive (23.4% vs. 15.18%).

In SemEval 2016 task 5 workshop, a new ABSA corpus was provided for the Arabic language [1, 9]. The dataset was selected from a Hotel reviews corpus [10] and was annotated on sentence-level (SB1) and text-level (SB2) for three different slots, ACD (slot 1), opinion-target expression (slot 2), and sentiment polarity (slot 3). The baseline models for ACD achieved an F1-score of 40.3% in SB1, and an F1-score of 42.7%, in SB2.

In Ruder, Ghaffari [11], the authors introduced their submitted models to the previous competition in multiple languages, including Arabic. For slot 1, they implemented a multi-label model based on a convolutional neural network (CNN) that outputted probabilities over category labels using a threshold. They initialized word embeddings using 300-dimensional GloVe vectors for English, whereas the vector representations were randomly initialized for the other languages. The proposed model obtained an F-1 score of 52.1% in Arabic.

Tamchyna and Veselovská [12] presented their model that was also submitted to the same competition in slot 1 in many languages, including Arabic. They implemented a binary classifier for each category (E#A pair) using a Long short-term memory (LSTM) as an encoded layer to capture long-term dependencies in the data and reduce the need for feature engineering and linguistic tools, and a logistic regression classifier to output label probabilities. The model achieved an F1-score score of 47.3% for the ACD task.

The authors in Al-Smadi, Qawasmeh [13] tried to enhance the previous results by implementing two different models, one based on support vector machine (SVM) and the second one based on recurrent neural network (RNN). Both models were trained using lexical, morphological, syntactic, and semantic features. RNN achieved slightly compared results to the best model in SemEval 2016 by achieving an F-measure of 48%. Meanwhile, SVM obtained the best results and outperformed RNN by more than 45.4% in F-measure. However, the execution time of RNN was faster than that of SVM.

Another attempt to enhance the previous results was proposed in Al-Dabet, Tedmori [14]. They handled the ACD task as a multi-label classification problem using a binary relevance (BR) classification approach. The proposed model combined CNN with IndyLSTM [15] and was decomposed into a set of independent binary classifiers to train every single aspect category independently. The experimental results achieved an F-1 score of 58.1%.

Unlike the previous methods, our model is implemented based on a BiGRU model. Additionally, different vector representations are investigated, including word-level, character-level, domain-specific, and contextualized word embeddings models.

3 Methodology

3.1 Background

This sub-section presents the technical aspects of the main components used to build our model.

3.1.1 Word embeddings

Word embedding models represent a text as a numerical vector representation by mapping each word in the corpus vocabulary to a set of real-valued vectors in a vector space of dimension N . Semantically associated words tend to appear near to each other and have similar vectors representations. One of the earlier word-embedding models was proposed by [16]. They implemented two different models, namely CBOW and skip-gram. The CBOW model predicts the current target word based on its surrounding words within a fixed window, whereas the skip-gram model predicts context words within the range of a fixed window based on a given current target word. However, one of the main drawbacks of these models is that they are unable to resolve polysemy disambiguation.

Recently, contextualized word embeddings or language models have overcome this limitation by providing different representations for the same word depending on the context it occurs. One of the first contextualized models was ELMO [17], which assigns an embedding to each word using a bidirectional LSTM (BiLSTM) model to look at the entire sentence. Another famous model is BERT [18], which is a deeply bidirectional and unsupervised language representation model implemented based on the transformer architectures. It learns information from both the left and the right sides of a token's context during the training phase. The BERT model has achieved state-of-the-art results in many NLP tasks, including SA, question answering, and named entity recognition.

3.1.2 GRU network

Gated recurrent unit (GRU) is a form of RNN, an Artificial Neural Networks (ANN) that was implemented to recognize the sequential information in data and predict the next likely scenarios using patterns. The main limitation of RNN is the problem of gradient vanishing. Therefore, LSTM and GRU were introduced to overcome this shortcoming. Compared to LSTM, GRU has a less complex structure combining the input and forget gate of LSTM into a single update gate and combining the hidden and cell states into a single hidden state. Fig. 2 illustrates the architecture of GRU.

Fig. 2. The architecture of GRU.

GRU networks can only manage sequences from front to back, which leads to information loss. Therefore, bidirectional GRU was widely used in many NLP tasks, including SA, text classification, named entity recognition, and question answering, e.g., [19-22] to process data in both directions and provide complete context information.

3.2 Model overview

The dataset we used in our experiments assigned more than one category to the same review. Thus, we considered the ACD task as a multi-label classification problem. The proposed model is composed of multiple components: an embedding layer to generate the vector representations of each word, a BiGRU layer to remember long term dependencies and extract text information from both the left and right side at the same time, followed by Average and Max Pooling to reduce dimensionality, a dropout layer to reduce overfitting and prevent co-adaptation of hidden units, and a dense layer with a sigmoid activation function to output probabilities over each category. Fig. 3 illustrates the used model.

Fig. 3. The architecture of the proposed model.

The following word embeddings models are also examined:

- Word-level representations: The model we used to generate these representations is AraVec [23], an Arabic version of word2vec. The AraVec models were trained using three different dataset resources, namely, Twitter, Wikipedia, and Common Crawl webpages crawl data for both skip-gram and CBOW models.
- Character-level representations: we used FastText [24] to generate the character-level embeddings, an extension of word2vec. FastText computes embeddings for character n-grams. Words are then represented as the sum of their character n-gram embeddings. FastText has released versions for 104 languages, including Arabic. The Arabic version was trained on the Wikipedia dataset with a dimension size of 300.
- Domain-specific word embeddings: pre-trained word embeddings are generally trained on general domain corpora, which cannot capture specific local meanings within the domain. Many researchers tried to overcome this problem by training domain-specific embeddings, which has shown promising results in many NLP tasks, including SA, e.g., [25-27]. Therefore, we attempted to evaluate the impact of using domain-specific word embeddings on the whole performance of the proposed model.
- Contextualized word embeddings: these vector representations were generated using language models (LM). For the Arabic language, there exist two types of LM, namely monolingual and multilingual models. The first type is trained only on Arabic datasets such as Hulmona [28], AraBERT [29], AraGPT2 [30], whereas the second type is trained on a multilingual corpus such as mBERT, which provides sentence representations for 104 languages, including Arabic. The AraBERT model was adopted in this research to provide

the contextualized embeddings by concatenating the representations from the last four hidden layers.

4 Experiments

4.1 Dataset

The dataset used in our experiments was released by the SemEval task 5 workshop in 2016. It was selected from the Arabic Hotel reviews provided by ElSahar and El-Beltagy [10]. The annotation was performed on both text-level (2291 reviews) and sentence-level (6029 sentences). This research performed the ACD task only on the sentence level. The total number of predefined categories is 35, combining two parts: entity (E) and attribute (A), and each sentence can be assigned to more than one category. The distribution of the dataset over aspect categories is illustrated in Table 1 [9]. We notice that the dataset was split by the ABSA SemEval-16 [1] into train (4802 sentences) and test (1227 sentences) sets. Additionally, we used 10% of the training data for validation.

Table 1. Distribution of data over aspect categories [9].

Entities #Attributes	GENERAL	PRICES	DESIGN&FEATURES	CLEANLINESS	COMFORT	QUALITY	STYLE&OPTIONS	MISCELLANEOUS	Total
HOTEL	1821	434	158	326	265	320	X	179	3503
ROOMS	396	60	419	436	156	174	X	42	1683
ROOM-AMENITIES	316	15	63	143	40	213	X	75	865
FACILITIES	974	102	150	141	58	172	X	97	1694
SERVICE	2781	X	X	X	X	X	X	X	2781
LOCATION	1390	X	X	X	X	X	X	X	1390
FOOD&DRINKS	X	53	X	X	X	865	219	60	1197
Total	7678	664	790	1046	519	1744	219	453	13113

4.2 Baselines

The proposed model is compared against the following baseline models:

- SVM-based: is the baseline model released with the SemEval Hotel dataset based on an SVM classifier along with n-gram features.
- Recurrent-based: we compared BiGRU against other recurrent neural networks by replacing the BiGRU layer in the proposed model by BiLSTM, LSTM, and GRU, respectively, and evaluating them using all the examined word embeddings models.

Besides these baselines, results of the related work models already discussed in section 2 are also reported, including INSIGHT [11], UFAL [12], and C-IndyLSTM [14].

4.3 Experimental environment and parameters

In this research, the word-level embeddings were generated using an AraVec model trained on a Twitter dataset using a skip-gram model with 300 dimensions. For domain-specific embeddings, we trained a word2vec skip-gram 300 dimensions model using a dataset selected from the corpus provided by Elnagar, Khalifa [31] that contains $\approx 100k$ Arabic Hotel reviews. Additionally, The AraBERTv01 model [29] was used to generate contextualized embeddings. This model was implemented following the BERT base version with 12 transformer blocks, 768 hidden dimensions, 12 attention heads, a maximum sequence length of 512 tokens, and a total of $\sim 110M$ parameters. It was trained on two corpora [32, 33] and scrapped articles from Arabic news websites, and it has a vocabulary size of 64k tokens. Table 2 illustrates the models' parameters employed in our experiments. Additionally, the experimental environment is presented in Table 3.

Table 2. Experimental parameters.

Model	Parameter	Value
Models using BERT embeddings	Learning rate	5e-5
	Dimension dim	768
	Batch size	16
	Dropout	0.1
	Hidden GRU states	128
	Optimizer	Adam
	Loss	Binary crossentropy
Models using other pre-trained word embedding models	Learning rate	1e-3
	Dimension dim	300
	Batch size	32
	Dropout	0.1
	Hidden GRU states	128
	Optimizer	Adam
	Loss	Binary crossentropy

Table 3. Experimental environment.

Operating system	Parameters
GPU	NVIDIA Tesla K80
RAM	12 GB
Python	3.7
TensorFlow	2.2.0
Keras	2.4.3

4.4 Preprocessing

The preprocessing phase consists of removing elongation, punctuation, diacritics, numbers, and non-Arabic words.

For models with BERT word embeddings, data transformation into a compatible format to BERT was applied as follows:

- The data was tokenized to break down sentences into tokens. We notice that no segmentation was required because we used the AraBERT v01 in our experiments.
- Special characters ([CLS] and [SEP]) were added at the beginning and the end of each sentence.

We notice that other steps, including stemming, stop words removal, and normalization, were ignored because they have decreased the performance of the models, especially those using BERT word embeddings. Thus, this shows that applying these operations has led to information loss and has affected the contextual information learned by BERT.

4.5 Evaluation metric

The standard metric for evaluating the ACD task is F1-score, which is computed using the following formula:

$$F1 = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

5. Results and discussion

The experimental results are presented in Table 4. They show that the proposed model outperforms the baseline by more than 25% in F-1 score and achieves better results than related work models evaluated on the same dataset. The BiGRU-based model outperforms more complicated models (e.g., C-IndyLSTM) by more than 1% in F-1 score using the same AraVec embeddings and more than 7% using contextualized embeddings. Moreover, GRU achieves better results than LSTM and BiLSTM. This can be justified by the fact that GRU has fewer parameters than LSTM, making the learning process easier. Furthermore, GRU performs better than LSTM under small-size datasets [34], which is the case for most Arabic corpora, especially in ABSA.

Meanwhile, the BiGRU model outperforms GRU and BiLSTM outperforms LSTM, which shows that the Bidirectional models are more effective in handling this task. For the embedding models, AraVec embeddings obtain better results than FastText embeddings with all recurrent neural models. Moreover, domain-specific embeddings prove to be better than generic embeddings, even that they were trained on a small-size corpus compared to the corpora used to train the standard word embedding models. Hence, this shows the importance of integrating domain-related knowledge in learning word embeddings.

Furthermore, contextualized embeddings achieve the best results in this task (F1-score: 65.5% using BERT+BiGRU). This can be due to the ability of the BERT model to resolve polysemy and Out of Vocabulary (OOV) problems compared to free-context word embeddings. Therefore, this proves that pre-trained language models are better than context-independent word embeddings in handling the morphological complexity and the ambiguity of the Arabic language in this task.

Table 4. Results of the aspect category detection task.

	Model	F1-score
Baseline and related work models	SVM-based [1]	40.3%
	INSIGHT [11]	52.1%
	UFAL [12]	47.3%

	C-IndyLSTM [14]	58.1%
AraVec embeddings	LSTM	54.6%
	BiLSTM	57.6%
	GRU	57.7%
	BiGRU	59.7%
FastText embeddings	LSTM	52.9%
	BiLSTM	54.4%
	GRU	56.0%
	BiGRU	57.6%
Domain-specific embeddings	LSTM	58.7%
	BiLSTM	59.8%
	GRU	60.2%
	BiGRU	61.1%
Bert embeddings	LSTM	61.1%
	BiLSTM	62.3%
	GRU	63.6%
	BiGRU	65.5%

6 Conclusion and directions for future work

This work proposed a BiGRU-based model to perform the ACD task on an Arabic reference dataset. Moreover, we examined the impact of using different word embeddings, including word-level, character-level, domain-specific, and contextualized word embeddings. Experimental results show that BERT-BiGRU achieves the best F1-score (65.5%) over the evaluated models and significantly outperforms the baseline and related work on the same dataset. Moreover, domain-specific embeddings have shown promising results, even that they were trained on a limited size dataset. Therefore, future work includes investigating the effect of training word embeddings using a larger domain-specific corpus on the performance of the examined models. Moreover, we plan to further pre-train BERT on in-domain datasets to improve the achieved results. Additionally, we intend to investigate several data augmentation methods to handle the imbalanced labels issue in the used dataset and examine their effect on the proposed model's performance.

References

1. M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, and O. De Clercq. *Semeval-2016 task 5: Aspect based sentiment analysis*. in *International workshop on semantic evaluation*. 2016.
2. S. Al-Dabet, S. Tedmori, and M. Al-Smadi, *Extracting Opinion Targets Using Attention-Based Neural Model*. SN Computer Science, **1**: p. 10 (2020).
3. I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, *Arabic natural language processing: An overview*. Journal of King Saud University - Computer and Information Sciences, (2019).
4. A. Alawami, *Aspect Terms Extraction of Arabic Dialects for Opinion Mining Using Conditional Random Fields*, in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Editor. 2018, Springer International Publishing: Cham. p. 211-220.

5. D. Tang, B. Qin, and T. Liu, *Deep learning for sentiment analysis: successful approaches and future challenges*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **5**(6): p. 292-303 (2015).
6. M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider. *Human annotated arabic dataset of book reviews for aspect based sentiment analysis*. in *2015 3rd International Conference on Future Internet of Things and Cloud*. 2015. IEEE
7. M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. *SemEval-2014 Task 4: Aspect Based Sentiment Analysis*. in *COLING 2014*. 2014.
8. I. Obaidat, R. Mohawesh, M. Al-Ayyoub, A.-S. Mohammad, and Y. Jararweh. *Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches*. 2015. IEEE
9. M. Al-Smadi, O. Qawasmeh, B. Talafha, M. Al-Ayyoub, Y. Jararweh, and E. Benkhelifa, *An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study*. 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST): p. 98-103 (2016).
10. H. ElSahar and S. R. El-Beltagy. *Building large arabic multi-domain resources for sentiment analysis*. 2015. Springer
11. S. Ruder, P. Ghaffari, and J. G. Breslin, *Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis*. arXiv preprint arXiv:1609.02748, (2016).
12. A. Tamchyna and K. Veselovská. *UFAL at SemEval-2016 Task 5: Recurrent Neural Networks for Sentence Classification*. in *SemEval 2016*. 2016. Association for Computational Linguistics
13. M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, *Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews*. Journal of Computational Science, **27**: p. 386-393 (2018).
14. S. Al-Dabet, S. Tedmori, and M. Al-Smadi, *Enhancing Arabic aspect-based sentiment analysis using deep learning models*. Computer Speech & Language, **69**: p. 101224 (2021).
15. P. Gonnet and T. Deselaers. *Indylstms: Independently Recurrent LSTMS*. in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.
16. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*. Advances in neural information processing systems, **26**: p. 3111-3119 (2013).
17. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, *Deep contextualized word representations*. arXiv preprint arXiv:1802.05365, (2018).
18. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. in *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019. Association for Computational Linguistics
19. J. Zhang, F. a. Liu, W. Xu, and H. Yu, *Feature Fusion Text Classification Model Combining CNN and BiGRU with Multi-Attention Mechanism*. Future Internet, **11**(11): p. 237 (2019).
20. L. Zhou and X. Bian. *Improved text sentiment classification method based on BiGRU-Attention*. in *Journal of Physics: Conference Series*. 2019. IOP Publishing
21. J. Liu, Y. Yang, S. Lv, J. Wang, and H. Chen, *Attention-based BiGRU-CNN for Chinese question classification*. Journal of Ambient Intelligence and Humanized Computing: p. 1-12 (2019).

22. M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghaly, *Arabic aspect based sentiment analysis using bidirectional GRU based models*. arXiv preprint arXiv:2101.10539, (2021).
23. A. B. Soliman, K. Eissa, and S. R. El-Beltagy, *AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP*. Procedia Computer Science, **117**: p. 256-265 (2017).
24. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, *Bag of tricks for efficient text classification*. arXiv preprint arXiv:1607.01759, (2016).
25. P. K. Sarma, Y. Liang, and W. A. Sethares, *Domain adapted word embeddings for improved sentiment classification*. arXiv preprint arXiv:1805.04576, (2018).
26. P. K. Sarma, Y. Liang, and W. A. Sethares, *Shallow domain adaptive embeddings for sentiment analysis*. arXiv preprint arXiv:1908.06082, (2019).
27. H. Xu, B. Liu, L. Shu, and P. S. Yu, *Double embeddings and cnn-based sequence labeling for aspect extraction*. arXiv preprint arXiv:1805.04601, (2018).
28. O. ElJundi, W. Antoun, N. El Droubi, H. Hajj, W. El-Hajj, and K. Shaban. *hulmona: The universal language model in arabic*. in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. 2019.
29. F. Baly and H. Hajj. *AraBERT: Transformer-based model for Arabic language understanding*. in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. 2020.
30. W. Antoun, F. Baly, and H. Hajj, *AraGPT2: Pre-Trained Transformer for Arabic Language Generation*. arXiv preprint arXiv:2012.15520, (2020).
31. A. Elnagar, Y. S. Khalifa, and A. Einea, *Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications*, in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A.E. Hassanien, and F. Tolba, Editors. 2018, Springer International Publishing: Cham. p. 35-52.
32. I. A. El-Khair, *1.5 billion words arabic corpus*. arXiv preprint arXiv:1611.04033, (2016).
33. I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja. *OSIAN: Open source international Arabic news corpus-preparation and integration into the CLARIN-infrastructure*. in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 175-182. 2019.
34. Y. Wang, G. Huang, J. Li, H. Li, Y. Zhou, and H. Jiang, *Refined Global Word Embeddings Based on Sentiment Concept for Sentiment Analysis*. IEEE Access, **9**: p. 37075-37085 (2021).