IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 11, 2020, accepted December 27, 2020, date of publication January 6, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049626

A Novel Deep Learning-Based Multilevel Parallel Attention Neural (MPAN) Model for Multidomain Arabic Sentiment Analysis

MOHAMMED A. EL-AFFENDI[1], KHAWLA ALRAJHI[2], AND AMIR HUSSAIN[3], (Senior Member, IEEE)

[1]Department of Computer Science, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]EIAS Data Science and Blockchain Laboratory, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[3]School of Computing, Edinburgh Napier University, Edinburgh EH11 4DY, U.K.

Corresponding author: Mohammed A. El-Affendi (affendi@psu.edu.sa)

This work was supported by the EIAS Data Science and Blockchain Laboratory, Prince Sultan University. The work of Amir Hussain was supported in part by the Scottish Government Chief Scientist Office under its COVID-19 Priority Research Program under Grant COV/NAP/20/07, and in part by the U.K. Government's Engineering and Physical Sciences Research Council under Grant EP/T021063/1 and Grant EP/T024917/1.

ABSTRACT Over the past few years, much work has been done to develop machine learning models that perform Arabic sentiment analysis (ASA) tasks at various levels and in different domains. However, most of this work has been based on shallow machine learning, with little attention given to deep learning approaches. Furthermore, the deep learning models used for ASA have been based on noncontextualized embedding schemes that negatively impact model performances. This article proposes a novel deep learning-based multilevel parallel attention neural (MPAN) model that uses a simple positioning binary embedding scheme (PBES) to simultaneously compute contextualized embeddings at the character, word, and sentence levels. The MPAN model then computes multilevel attention vectors and concatenates them at the output level to produce competitive accuracies. Specifically, the MPAN model produces state-of-the-art results that outperform all established ASA baselines using 34 publicly available ASA datasets. The proposed model is further shown to produce new state-of-the-art accuracies for two multidomain collections: 95.61% for a binary classification collection and 94.25% for a tertiary classification collection. Finally, the performance of the MPAN model is further validated using the public IMDB movie review dataset, on which it produces an accuracy of 96.13%, placing it in second position on the global IMDB leaderboard.

INDEX TERMS Arabic sentiment analysis, deep learning, multilevel parallel attention, natural language processing, positioning binary embedding, power-of-two, polynomial space positioning attention.

I. INTRODUCTION

Sentiment analysis is an important subarea of natural language processing (NLP) where advanced statistical and machine learning models are used to measure sentiments, emotions and opinions in many domains, including customer satisfaction, product acceptance, market directions, and public approval of political decisions and events. Sentiment analysis applications have recently attracted more attention with the rise of social media and Web 2.0 technologies. Over the past few years, considerable work has been done in Arabic sentiment analysis (ASA). Many algorithms and machine learning approaches have been used to build sentiment analysis systems in various domains. As reported in [1], [2], the number of publications in ASA has been growing exponentially over the past few years. Traditionally, shallow machine learning is still the dominant approach in the publications surveyed by [1], [2], with SVM and naïve Bayes approaches representing more than 70% of all reported works. According to the same survey, only 2% of the reviewed publications applied deep learning in ASA. This shows that sentiment research in Arabic is still lagging other languages, where deep learning is fast replacing shallow machine learning.

Furthermore, most of the reported deep learning-based ASA studies employed convolutional neural networks (CNNs), long short-term memory (LSTM) networks or

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

combinations of both. These sequential models lack the flexibility needed to build arbitrary networks. Furthermore, most previous models were based on the word2vec embedding approach [3], [4] or variations such as fastText [5]. Such approaches are not fully ''contextualized'' and ignore the position and order of words, hence negatively impacting the resulting accuracy.

To address these problems, some researchers have designed attention-based models that search for and amplify relevant context areas in input vectors [6]–[9]. Most of these models have been derived in the context of sequence-to-sequence encoder-decoder models.

In an attempt to address the limitations of existing approaches, this article introduces a novel deep learning-based multilevel parallel attention neural (MPAN) model comprising multiple parallel channels, where each channel represents a different neural model that works in parallel with other models to perform training or testing. Outputs from the parallel channels are combined (concatenated) to produce the final predictions.

The key contributions of this article include the following:

(i) A new polynomial space positioning attention (PSPA) model that computes neural attention based on a nonlinear power-of-two polynomial representation of sentences and words is presented as a parallel space positioning alternative to the classic time sequencing encoder-decoder attention model. Despite its simplicity, PSPA has outperformed the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) [9] on public datasets.

(ii) A novel deep learning-based MPAN model is proposed to overcome the limitations of conventional noncontextualized embedding schemes. The MPAN model is strongly based on PSPA.

(iii) Experimental results show that the proposed model is highly accurate in performing benchmark ASA tasks for a collection of 34 datasets, including 21 binary classification sets and 13 tertiary classification sets. In benchmark experiments, the MPAN model is shown to achieve an accuracy of 95.61% for binary classification collection and 94.25% for tertiary classification collection, all of which are well above the current baselines reported in the literature.

(iv) Finally, the performance of the MPAN model is further validated using the public IMDB movie review dataset, producing an accuracy of 96.13%, placing it in second position on the global IMDB leaderboard above BERT [9].

The rest of this article is organized as follows: Section II provides a literature review and a survey of related models and datasets. Section III describes the novel MPAN deep learning model, including the background, proposed model parameters, configurations and embedding scheme. The experimental procedures, results, and analyses are presented in Section IV. Conclusions are offered in Section V along with future work directions.

II. LITERATURE REVIEW AND RELATED WORK: MODELS AND DATASETS

Many efforts have been made to apply machine learning in ASA using different models and approaches. The models used include classic lexicon approaches, shallow machine learning and, more recently, deep learning models. Since the focus of this article is on deep learning and how it can be used to produce state-of-the-art results in ASA, a comprehensive hierarchical literature review has been conducted to discern and review previous contributions.

The survey starts with extracting relevant articles and information from recent comprehensive ASA review articles [1], [2], [10], followed by a focused review of the collected articles to extract models, features and datasets.

According to [1], [2], [10], most of the surveyed ASA publications have adopted shallow machine learning models that have existed for many years. Support vector machines (SVMs) and naïve Bayes classifiers are the most dominant approaches, accounting for more than 70% of the reported models. Deep learning efforts represent a small fraction (approximately 2% at best) of the surveyed articles.

In [1], [11], the authors conducted two detailed literature reviews: one on the applications of deep learning in Arabic NLP at large and the other on the ASA subarea. The authors surveyed many articles in both cases and noted that deep learning approaches represent a small fraction of the reported ASA research contributions. Their studies also concluded that ASA has recently received more attention, and the number of published articles in this area is exponentially growing, but shallow machine learning is the dominant approach (within the reviewed articles, only 18 were on deep learning). These include [11]–[25].

In most of these applications, the authors employed some sort of word embeddings and distributed semantic approaches to represent words and sentences.

For example, in [17], the authors used the word2vec approach [3], [4], whereas in [18], the authors used sentiment-specific embeddings based on Facebook's fastText approach [5], [26]. In [21], the authors used doc2vec, which is based on the word2vec approach [27]. In [20], the authors created word embeddings based on 10 billion Arabic words using word2vec. The pretrained word embeddings were passed to a convolutional neural network (CNN). They demonstrated the effectiveness of their approach on different datasets, including the datasets provided in [28]–[31].

In [32], the authors provided an extended version of the annotated Arabic book review dataset (BRAD 2.0), where they added more than 200k records to the original BRAD 1.0 dataset, raising the size to 692586 annotated reviews. To verify and validate the proposed dataset, the authors applied several state-of-the-art supervised and unsupervised classifiers to categorize book reviews. The highest accuracies attained ranged between 90% and 91%.

In a related effort, the authors of [33] compiled and annotated large datasets for Arabic text classification. Again, the

TABLE 1. Summary of deep learning models used in cited articles.

Reference	System Type	Dataset	Sentiment Classes #
[36]	CNN+LSTM	SemEval 2017 ASTD, ARsas	3
[37]	Hybrid CNN LSTM	Main-AHS, Sub-AHS, Ar-Twitter, ASTD, OCLAR.	2
[16]	CNN+LSTM	Main-AHS, Sub-AHS, Ar-Twitter, ASTD.	2
[35]	CNN+LSTM	Merged datasets with more than 90,000 items	2
[38]	CNN, LSTM, and CNN+ LSTM	ASTD	3
[20]	CNN	ASTD, LABR,LARGE, Ar-Twitter	2
[39]	Character CNN	AraSenTi-Tweet	3
[40]	SVM, L2RN, RNN	ASTD, MASTD, ArSAS, GS, and the Syrian Corpus	2 & 3

TABLE 2. Summary of the datasets used in cited articles.

No	Dataset Description	Reference
1	Twitter Dataset (Ar-Twitter): The Ar-Twitter dataset is collected from Twitter's Arabic tweets and labeled manually with two sentiment classes: positive and negative. The dataset is multidomain with 2000 tweets, balanced with 1000 for each category. However, *approximately 14 records were deleted from the negative records because the documents were either empty or damaged.*	[28]
2	Arabic Sentiment Tweets Datasets (ASTD): ASTD consists of 10,006 Arabic tweets classified as objective, positive, negative, and mixed. In one experiment, we used only two classes with 1684 negative tweets and 799 positive tweets. In addition, in another experiment, we used three classes labeled negative, positive, and neutral with a count of 1684, 795, and 2483 tweets for each.	[30]
3	SEMEVAL2017-task 4-A: This dataset covers 9655 Arabic tweets: 6100 for testing, and 3555 for training. Sentiment classes used for labeling tweets are positive, negative and neutral.	[41]
4	Social Media Posts (SMP): A part of the [42] project is the SMP dataset, which was a sentiment tagging of Arabic social media with four classes: positive, both, neutral, and negative. SMP includes two subdatasets: a BNN dataset, which includes 1200 Levantine sentences, and a SYR dataset, which consists of 2000 Syrian tweets.	[42]
5	Opinion Corpus for Arabic (OCA): OCA includes 500 Arabic movie reviews. Approximately 250 reviews are identified as positive and 250 as negative.	[43]
6	LABR: LABR includes 63,000 book reviews; every review is rated on a scale of 1 to 5 stars. In this research, LABR is used for binary classification. Therefore, rates 1 and 2 are considered negative, while rates 4 and 5 are considered positive.	[44]
7	LARGE: An Arabic sentiment analysis dataset built from movies, hotels, restaurants, and product reviews. Reviews are identified by three labels: positive, negative, and neutral.	[29]
8	Arabic Health Services Dataset (AHS): This corpus is collected from health services reviews that have been gathered from Twitter. AHS consist of 2026 records, where 1398 are negative reviews and 628 are positive reviews. We used full and subset versions of AHS to compare with [37] and [16] experiments. The full dataset is called AHS-Main, and the subset is called AHS-sub.	[45]
9	Omara *et al.* [35] Dataset (OD): [35] merged known ASA datasets to train a CNN model for sentiment analysis. OD has been recalled to analyze results precisely with their results. They collected entries labeled positive or negative only. OD consists of 8 datasets: Ar- twitter, ASTD, SMP, OCA, LABAR, LARGE, AHS-Main and part of SemEval.	[35]
10	ArSAS: This dataset includes over 21,000 Arabic tweets. Entries are tagged with four classes: positive, negative, neutral, and mixed. 1302 Entries tagged as mixed were ignored.	[46]
11	OCLAR: OCLAR consists of Arabic customer reviews from multidomains, including restaurants, hotels, hospitals, local shops, etc. The corpus includes 3916 reviews with a rating of scale of 1 to 5. Following the [37] approach in distributing this dataset, the positive class reviews from values of 3, 4, and 5, while the negative class reviews from values of 1 and 2.	[47]
12	AraSenTi-Tweet: The AraSenTi dataset includes Suadi dialect tweets. This dataset identified with four classes (positive, negative, neutral, mixed). In this research, mixed tweets are ignored.	[48]

two datasets have been validated using several deep learning models.

In [2], the authors conducted a detailed systematic literature review of ASA techniques, technologies and domains of application. Their study confirmed that most of the reported efforts were based on shallow machine learning and linguistic lexicon approaches. Only six of the 108 articles surveyed were in the area of deep learning.

In a third parallel review [34], the authors conducted a detailed review using a set of 51 articles. They concluded that only 18 of these studies applied machine learning in general to ASA, of which only two were in the area of deep learning.

A good survey of ASA deep learning models was also provided in [35] in the context of their work on applying a deep learning CNN model to a constructed dataset comprising a merged collection of eight public datasets (a total of 92492 items). Their survey covered fourteen publications that applied deep learning to ASA, mostly as CNNs and, in a few cases, combined with LSTM. All models used some sort of embedding to capture the context at the word or character levels. The best accuracy achieved on the constructed dataset was 94.33%.

In this article, we extend the survey provided in [35] to cover other efforts and collect the updated versions of the reported datasets from their original sources. In the following paragraphs, we review the relevant ASA deep learning studies made while using Tables 1 and 2 to link the models and datasets to the corresponding references and provide short summaries.

In [36], the authors applied a CNN combined with long short-term memory (LSTM) to analyze sentiments in three public datasets: SemEval 2017, ASTD, and ARSAS. They used the word2vec model to generate the required word embeddings. Their classification experiments were based on three sentiment classes. Their model achieved state-of-the-art results on the Arabic datasets SemEval 2017 and ASTD and produced 92% accuracy on the ARSAS dataset.

In [37], the authors employed a hybrid CNN–LSTM model with word2vec embeddings for the binary classification of Arabic sentiments using five public datasets. The datasets used were Main-AHS, Sub-AHS, Ar-Twitter, ASTD, and OCLAR. The proposed model outperformed previous results in three out of the five datasets.

In [16], the authors applied a combined CNN-LSTM model for binary sentiment classification using four datasets: Main-AHS, Sub-AHS, Ar-Twitter, and ASTD. They conducted sentiment analysis at different levels: character level, character N-gram level, and word level. The word-level analysis produced good results on some datasets.

In [35], the researchers applied character-level deep CNNs for Arabic sentiment analysis. They evaluated their model on a merged Arabic sentiment dataset comprising eight datasets with more than 90,000 items. The best accuracy achieved was 94.33%. Their model outperformed traditional machine learning classifiers by nearly 7%.

In [38], the authors used an ensemble approach to combine the performance of a CNN with an LSTM model using the ASTD dataset with three sentiment classes: positive, negative and neutral. The accuracy produced by the CNN model was 64.3%, the accuracy produced by LSTM was 64.75%, and the combined ensemble accuracy is 65.05%, which was much lower than the results produced by other articles using the same data (for example, the accuracies produced in [36]). The authors used pretrained word embeddings to train the models.

In [20], the authors applied a binary sentiment analysis model on nine datasets, including LABR and ASTD. The datasets represented two domains: reviews and tweets. They used the CBOW and Skip-Gram variations of the word2vec model to compute intermediate word representations based on a large corpus they compiled.

They also applied a CNN on balanced and imbalanced datasets. The accuracy they reported on LABR was 86.7% for imbalanced data and 89.2% for a balanced dataset, which compared favorably with other approaches.

In [39], a recurrent neural network (RNN) model was employed for Arabic sentiment analysis with different word embeddings. They used the AraSenTi-Tweet dataset provided in [48]. They examined three available public Arabic word embeddings. Their highest accuracy was 93.5% in identifying 3 sentiment classes.

In [40], the authors applied a hybrid method based on lexicons and machine learning techniques for Arabic sentiment analysis. In [40], three main tasks were introduced: building a lexicon, classifying entries, and extracting new words and their polarities. Note that they applied classification by using shallow machine learning models, such as support vector machines (SVMs), L2–logistic regression (L2RN), and deep learning models, such as recurrent neural networks (RNNs). The researchers extended their classification experiments to tertiary datasets. They evaluated their models on five datasets (ASTD, MASTD, ArSAS, GS, and the Syrian Corpus). Although annotating and learning words is time-consuming, the highly accurate results obtained on many Arabic datasets justify the efforts and provide valuable baselines for subsequent works in Arabic sentiment analysis.

Table 1 summarizes the models and experiments performed in the cited articles. Table 2 summarizes the datasets used in the cited articles. The updated versions of these datasets have been retrieved from their original sources to compare our model with the respective works. More detailed information about the retrieved updated versions is provided in Table 3.

III. PROPOSED MODEL

A. BACKGROUND

It is clear from the literature review in Section II and Table 1 that most of the reported deep learning-based ASA studies

employed sequential-layer CNNs, LSTM or combinations of both. In sequential-layer models, information flows sequentially from one layer to the next based on their order in the sequence. They lack the flexibility needed to build arbitrary networks.

Another relevant issue is that most of the reported models are based on the word2vec embedding approach [3], [4] or variations such as fastText [5]. These approaches compute only one embedding vector per word over the whole corpus, irrespective of the number of occurrences and the different contexts in which the word occurs. Accordingly, the resulting word embedding vector is not ''contextualized'' since it is computed over many ''contexts''. Another problem with most embedding systems is that they ignore the position and order of words, thus negatively impacting the resulting accuracies.

To solve these problems, some researchers have designed attention-based models that search for and amplify relevant context areas in input vectors [6]–[9]. Most of these models have been derived in the context of sequence-to-sequence encoder-decoder models, and accordingly, attention is defined [6] as mechanism to compute a weighted context vector based on the hidden states of the encoder component of the model. The subsection below provides a brief account of sequence-to-sequence attention models

B. SEQUENCE-TO-SEQUENCE AND TRANSFORMER ATTENTION MODELS

Most current attention mechanisms are based on sequential encoder-decoder models [6], [8], where a decoder generates (predicts) a full target sequence based on the input from an encoder. The input from the encoder comes in the shape of a weighted context vector that summarizes the states of a given

TABLE 3. The MPAN model specification.

Channel	Type of Model	Embedding	Input	Loss/Acc Metrics Functions
1	Bidirectional LSTM	PBES	Context (Sentence) Embedding	Cosine loss/cosine metric, MSE loss/cosine metric
2	Convolutional	PBES	Word Embedding	
3	Convolutional	PBES	Character embedding	

input sequence. A good example of applying encoder-decoder attention models is neural machine translation (NMT), where the encoder summarizes an input sentence (in one input language) while the decoder generates the translation in another language.

Practitioners have noticed that ''context vectors'' coming from encoders may not be able to provide all the information needed to generate decoder output sequences when the sequences are relatively long. Based on this observation, some researchers have proposed new models that replace the context vectors by new nonsequential attention models that provide more detailed information about input sequences without relying on the states of sequential encoders. A good example in this direction is the ''transformer attention'' model, which has been popularized by the success of the BERT model [9].

A good example of strict sequential attention models was provided by Bahdanau *et al.* [6], which is computed in the context of the alignment process with decoder states using the encoder hidden states as follows:

1. Pass the input sequence to the encoder to compute the set of hidden states.
2. Use an MLP model to compute the alignment scores with the decoder state from the preceding step $\mathbf{s}_{t-1}$. For the case $t = 1$, the last encoder state h_t is the initial hidden state.
3. Compute the weights using SoftMax.
4. Use the weights to compute the weighted context vector.
5. Compute the decoder hidden state using the concatenated vectors s $\mathbf{s}_{(t-1)}$ and $\mathbf{c}_t$ and the previous decoder output y_{t-1}.
6. Compute the final output y_t.

Transformer attention [9] is computed using a similar process, with the following major differences:

1- The transformer model does not require the encoder hidden states and relies directly on the original input sequence x.
2- Transformer attention is based on self-attention, which is computed as a relative inter-word similarity vector between the elements of the input sequence x.
3- The transformer uses a parallel memory-based approach to compute the self-attention vectors y. Sequential encoders are not a condition in the transformer model.
4- The attention vector **y** is the only information passed to the decoder and upper layers (hence, the term ''attention is all you need'' in the title of [9]).

C. POLYNOMIAL SPACE POSITIONING ATTENTION (PSPA) MODEL

One of the distinguishing features of the proposed MPAN model is the use of a novel polynomial space positioning attention (PSPA) scheme at the lower branch levels of MPAN. This is a new alternative to the common encoder-decoder sequential attention schemes [6], [7]. While sequential attention schemes assume some sort of time dependency between the components of a sentence (w_1, w_2, w_3, ...), PSPA assumes that the components of a sentence are just neighbors in a space where the relations between words are specified by a nonlinear polynomial:

$$S = w_0 + w_1 * x + w_2 * x^2 + w_3 * x^3 + \ldots w_m * x^m \quad (1)$$

where x is a ''power-of-two'' (x is 2^n), and w_i is a binary representation of the i_{th} word. The use of a power-of-two radix ensures that the binary representation of the i_{th} word occupies exactly n binary slots at the i_{th} position in the sentence, where n is the desired length of the binary representation of the respective word.

Similarly, the relations between characters in a word W, are specified by another nonlinear polynomial at the character level:

$$W = c_0 + c_1 * y + c_2 * y^2 + c_3 * y^3 + \ldots + c_m * y^m \quad (2)$$

where y is a power-of-two radix (y is 2^n) and c_i is a binary representation of the respective character.

PSPA has been implemented using a simple Galois power-of-two positional binary embedding scheme (PBES) [50], [51], where words are transformed into base-64 binary integers (radix = 64=2^6), and sentences are transformed into base-2^{60} binary integers.

The main advantages of the GPOW2 transformation include the following:

- The ability to capture character, word, and sentence embeddings simultaneously in the same pass.
- GPOW2 preserves the positions of characters and words in the resulting embeddings (space positioning).
- GPOW2 is an invertible transformation where a transformed string can easily be recovered from its binary representation. Figs. 1 and 2 show pseudocodes of the forward and backward GPOW2 transformations.

The main significance of the built-in space positioning feature is making it possible to compute attention vectors in parallel without being restricted by sequential encoder-decoder structures and without the need to use a separate positioning algorithm such as the one used by BERT and transformer models [9].

As shown in validation Table 4, PSPA produced state-of-the-art results that beat other attention models, including BERT [9].

Convert word to base-64 representation for a word:
The simple algorithm below can be used to convert a given word w into a binary base-64 representation wbin.
The word w is first transformed into a decimal base-64 representation n64:

```
n64=0
for char in w
    k=int(char) -1569 --Unicode position of Arabic chars
    n64=n64 + 64*k
wbin =bin(n64)
```

where bin() is a method to convert decimal numbers into binary

FIGURE 1. Forward transformation GPOW2 pseudocode.

D. MULTILEVEL PARALLEL ATTENTION (MPA)

In MPA, attention is computed at multiple levels: input character sequence, input word sequence and at the document level. At each level, the attention is computed using a process similar to the approach used in the Bahdanau *et al.* [6] attention model, with the following differences:

1- MPA is computed in the context of an alignment process with decoder states such as the one used in the Bahdanau approach.
2- Unlike the Bahdanau approach, and similar to the transformer approach, MPAN attention relies directly on the input sequence and does not require the encoder hidden states – sequential encoders are just an option.
3- In the case of MPA, positioning is built in. Words and sentences are transformed into binary power-of-two integers that strictly preserve the order of characters in a word and the order of words in a sentence, irrespective of the neural model used.
4- Accordingly, no positioning step, such as the one used in BERT [9], is required.

TABLE 4. Accuracy of MPAN for IMDB movie review dataset.

Loss/Metric	Test Accuracy (Sentence Level Attention)	Test Accuracy (Word Level Attention) (%)	Test Accuracy (Char Level Attention) (%)	Overall Merged Test Accuracy (%)
Cosine Loss/ Cosine Acc	89.04	94.28	95.12	95.50
MSE Loss/ Cosine Acc	88.85	95.23	95.50	96.13
Contrastive_Loss/ Contrastive Acc	86.65	93.66	93.81	94.38

The accuracies have been obtained for max sentence length of 700 using the configurations of Table III above

5- Unlike the Bahdanau and transformer approaches, MPA uses a direct cosine similarity metric in the alignment process instead of using the indirect "dot product" between vectors.

E. PROPOSED MULTILEVEL PARALLEL ATTENTION NEURAL (MPAN) MODEL

Based on the above background information and discussion, a novel multilevel parallel attention neural (MPAN) model has been derived and tested. The MPAN model comprises independent channels that are combined at the output layer to produce predictions. Each channel represents a different neural model with a different input (see Fig. 3). Table 3 summarizes the main components and hyperparameters of the model. As seen in the table, three combinations of loss/metrics functions have been used, including contrastive loss functions commonly used in image deep learning models [52]. This is mainly due to the binary nature of the PBES embeddings.

F. ARABIC SENTIMENT ANALYSIS DATASETS

Table 5 describes the datasets used in our experiment (based on the survey in Section II above). The table provides the main attributes for each set, including the corresponding references, category, size, number of positive items, number of negative items, and number of neutral

Recovering a word w from its decimal base-64 representation, n64:

```
w=""      -- initialize w to the null string
N = logbase64(n64)
while N>=0
    code = n64//(64**N)        -- integer division to get the current char code
    w = w + char(code)         -- character based on code
    n64=n=n64%(64**N)          -- use the modulus operation to shift left
    N=N-1
```

At the end of the algorithm w is the required word

Where logbase64 is defined as:

```
Logbase64(n) = round(log2(n)/log2(64))
```

FIGURE 2. Backward transformation GPOW2 pseudocode.

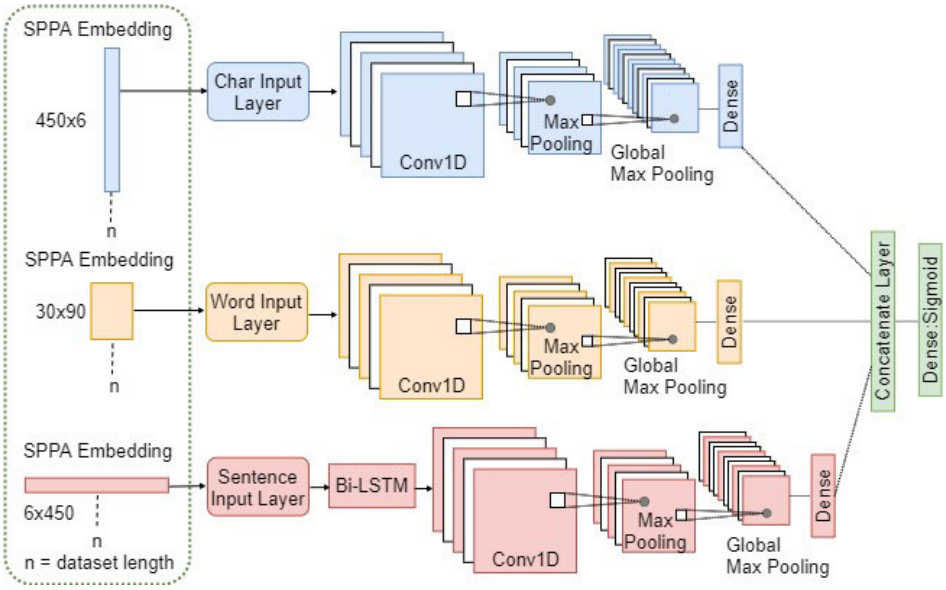

FIGURE 3. The proposed MPAN architecture has three inputs: char, word, and sentence embeddings.

TABLE 5. Arabic sentiment datasets.

Name	Domain/class	Total	Positive	Negative	Neutral
AHS (Main) [45]	Health/bin	2,026	628	1,398	-
AHS (sub) [45]	Health/bin	1,733	502	1,231	-
Ar- twitter [28]	Tweets/bin	2,000	1,000	986	-
AraSenti [48]	Tweets/ter	17,573	4,957	6,155	4,639
ArSAS [46]	Tweets/ter	21,065	4,643	7,840	7,279
ASTD [30]	Tweets/ter	10,006	799	1,684	832
LABR [44]	Book reviews	63,257	42,832	8,224	12,201
LARGE (Main) [29]	Multireviews/ter	34,492	24,948	6,650	2,894
LARGE Attraction Reviews [29]	Multireviews/bin	2,154	2,073	81	-
LARGE Hotel Reviews [29]	Multireviews/ter	15,572	10,775	2,647	2,150
LARGE Movie Reviews [29]	Multireviews/ter	1,524	969	384	171
LARGE Products Reviews [29]	Multireviews/ter	4,272	3,101	863	308
LARGE Restaurant Reviews [29]	Multireviews/ter	10,970	8,030	2,675	265
OCA [43]	Movie reviews/bin	500	250	250	-
OCLAR [47]	Multireviews/bin	3,916	3,465	451	-
OD [35]	Mixed/bin	92,492	71,398	21,081	-
SemEval Test & Train [41]	Tweets/ter	9,455	2,257	3,364	3,834
SemEval Development [41]	Tweets/ter	671	222	128	321
SMP (BBN)+(SYR) [42]	Multi/ter	3,200	719	1,761	697
SMP(BBN) [42]	BBN posts/ter	1,200	434	438	305
SMP(SYR) [42]	Tweets/ter	2,000	285	1,323	392
Full data for 3 classes	Multidomain	149,880	81,377	35,806	32,697
Full dataset for 2 classes	Multidomain	125,611	86,720	38,891	-

TABLE 6. Merged collections of the tertiary and binary arabic sentiments datasets.

Name	Type	Total Size	Positive	Negative	Neutral
Merged tertiary sets	Multidomain	149,880	81,377	35,806	32,697
Merged binary sets	Multidomain	125,611	86,720	38,891	-

items. If the dataset includes both binary and tertiary classification components, the size of each component is also provided.

Table 6 provides the merged collections of Table 5 binary and tertiary classification datasets.

G. DATA PREPROCESSING

The following preprocessing steps were used to clean the data and render them suitable for the intended tasks:

- Isolating binary from tertiary classification samples to process with different configurations.

TABLE 7. Comparing MPAN accuracies with state-of-the-art results for famous public datasets.

	Benchmark Datasets				
	IMDB	SNLI	QQP	SciTail	MultiNLI
MPAN (our model)	96.13	95.38	94.16	94.83	95.58
CA-MTL [54]		92.1			
BERT [9]	95.8				
MT-DNN [55]				94.1	
XLNet [56]			92.3		
T5-11B [57]					92.0

IMDB = Movie Reviews Dataset; SNLI = Stanford Natural Language Inference Dataset; QQP = Quora Question Pairs; SciTail = Science Entailment; MultiNLI = Multiple Natural Language Inference

TABLE 8. MPAN test accuracies for each of the binary classification datasets + the merged set.

No	Dataset	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	AHS (Main) [45]	95.99	95.99	95.98	95.21	91.58	93.36
2	AHS (sub) [45]	96.63	96.53	96.72	97.49	91.80	94.56
3	Ar- twitter [28]	94.31	94.14	94.47	90.63	91.80	91.21
4	AraSenti [48]	96.99	96.91	97.07	95.95	94.55	95.24
5	ArSAS [46]	95.93	95.84	96.01	95.40	91.59	93.46
6	ASTD [30]	93.84	93.65	94.02	91.86	87.08	89.41
7	LABR [44]	96.38	96.33	96.43	93.70	98.79	96.17
8	LARGE (Main) [29]	96.71	96.65	96.76	94.89	97.80	96.32
9	LARGE Attraction Reviews [29]	98.65	98.65	98.65	97.80	99.92	98.85
10	LARGE Hotel Reviews [29]	97.77	97.70	97.83	96.81	98.39	97.59
11	LARGE Movie Reviews [29]	92.85	92.60	93.08	87.93	96.23	91.89
12	LARGE Products Reviews [29]	96.12	96.09	96.13	93.29	98.20	95.68
13	LARGE Restaurant Reviews [29]	95.99	95.90	96.07	93.09	97.78	95.38
14	OCA [43]	90.11	89.73	90.42	79.41	90.36	84.53
15	OCLAR [47]	95.96	95.97	95.97	93.36	99.95	96.54
16	OD [35]	95.72	95.68	95.75	92.45	97.64	94.97
17	SemEval (Main) [41]	94.70	94.55	94.83	93.88	89.00	91.37
18	SemEval Development [41]	91.96	91.71	92.11	84.51	96.97	90.31
19	SMP (BBN)+ (SYR) [42]	94.13	94.02	94.23	92.57	87.85	90.15
20	SMP(BBN) [42]	91.08	90.77	91.31	84.77	84.19	84.48
21	SMP(SYR) [42]	95.95	95.89	96.00	98.01	87.09	92.23
22	Merged Set, all 21 datasets have been merged into one set	95.57	95.52	95.61	92.34	96.48	94.36

TABLE 9. MPAN test accuracies for the tertiary classification dataset + merged tertiary datasets.

No	Dataset	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	AraSenti [48]	96.43	96.38	96.48	95.25	95.03	95.14
2	ArSAS [46]	94.93	94.82	95.02	93.51	91.96	92.73
3	ASTD [30]	91.89	91.80	91.98	91.86	80.65	85.89
4	LABR [44]	94.55	94.52	94.58	94.25	90.38	92.27
5	LARGE (Main) [29]	96.18	96.15	96.20	95.29	94.33	94.81
6	LARGE Hotel Reviews [29]	96.19	96.18	96.20	95.86	93.44	94.63
7	LARGE Movie Reviews [29]	93.60	93.44	93.73	90.90	92.01	91.45
8	LARGE Products Reviews [29]	95.90	95.86	95.93	94.93	94.82	94.88
9	LARGE Restaurant Reviews [29]	96.73	96.73	96.72	94.85	96.80	95.82
10	SemEVal [41]	92.00	92.15	91.86	90.36	83.92	87.02
11	SMP (Main) (BBN)+ (SYR) [42]	92.74	92.94	92.53	90.20	85.42	87.74
12	SMP(BBN) [42]	90.96	90.84	91.04	85.88	86.42	86.15
13	SMP(SYR) [42]	93.89	93.72	94.04	94.43	84.92	89.42
14	Merged sets, all 13 datasets merged into one set	94.23	94.21	94.25	93.14	90.17	91.63

- Removing HTML tags and URLs.
- Removing non-Arabic characters.
- Removing emojis.
- Letters that were repeated more than two times were removed.
- Tokenization.
- Setting the maximum size of statements to a max value, padding statements that contain less than max words, and truncating statements that contain more than max words.

TABLE 10. Comparing MPAN results with the binary classification baselines using test accuracies.

Dataset/Study	**MPAN (%)**	[20] (%)	[15] (%)	[35] (%)	[37] (%)	[16] (%)	[40] (%)
AHS (Main) [45]	**95.98**				88.10	94.24	
AHS (sub) [45]	**96.72**				96.68	95.10	
Ar- twitter [28]	**94.47**	85.01			84.20	88.10	
ArSAS [46]	**96.01**						81.52
ASTD [30]	**94.02**	79.07			79.18	76.41	74.98
LABR [44]	**96.43**	89.60					
LARGE Attraction Reviews [29]	**98.65**	96.20					
LARGE Hotel Reviews [29]	**97.83**	91.70					
LARGE Movie Reviews [29]	**93.08**	80.70					
LARGE Products Reviews [29]	**96.13**	87.30					
LARGE Restaurant Reviews [29]	**96.07**	78.50					
OCLAR [47]	**95.97**				90.30		
OD [35]	**95.75**			94.33			
SMP(SYR) [42]	**96.00**		85.28				81.28

TABLE 11. Comparing MPN results with tertiary classification baselines.

Dataset/Study	**MPAN (%)**	[38] (%)	[36] (%)	SVM [40] (%)	RNN [40] (%)	L2R2 [40] (%)	[39] (%)
AraSenti [48]	**96.43**						93.50
ArSAS [46]	**94.93**		92	66.39	65.67	66.7	
ASTD [30]	**91.89**	65.05	66	73.11	72.3	62.94	
SemEVal [41]	**92.00**		62				
SMP(SYR) [42]	**93.89**			73.01	73.67	72.01	

IV. EXPERIMENTAL PROCEDURE AND RESULTS

Before discussing the state-of-the-art Arabic Sentiment Analysis datasets, we show how the MPAN model performs in comparison to state-of-the-art models on public datasets. In subsection A below we compare the performance of MPAN with five well-known models on five public datasets. Then, in subsection B, we present the results for ASA datasets.

A. VALIDATING THE PERFORMANCE OF MPAN USING THE IMDB AND NAUTRAL LANGUAGE INFERENCE DATASETS

To evaluate the potential of the MPAN model, several experiments are performed to compare the proposed approach with other approaches using multiple public datasets. The datasets used include: IMDB movie review dataset; SNLI, Stanford Natural Language Inference Dataset; QQP, Quora Question Pairs dataset;, SciTail, Science Entailment dataset; MultiNLI, Multiple Natural Language Inference dataset. Table 4 shows the details of the exercise for the IMDB dataset, where three combinations of loss functions and metrics have been used. The experiments are performed using the configurations given in Table 3 and Fig. 3. Table 7 shows that MPAN is beating all top ranking models across all datasets. The listed models in column 1 are the current top ranking attention models in published leaderboards. For example, the IMDB leaderboard table [53], (as of Dec 10, 2020), shows that the MPAN model is above BERT-large and L-Mixed, with a test accuracy of 96.13% (see table 4) obtained using the MSE loss and cosine accuracy metric. Given the simplicity of the MPAN model and the relative complexity of BERT, this is very encouraging.

B. PERFORMANCE COMPARISONS AND ANALYSIS FOR ASA DATASETS

The MPAN model is first applied separately to each dataset in the binary and tertiary classification collections. The motivation is to compare our results with the state-of-the-art accuracies obtained for each baseline, where available. Since all baseline models are just reporting the test accuracies for the datasets, comparisons with these modes are based on MPAN test accuracies. However, because many of these datasets are imbalanced, we computed the precision, recall and F1 metrics to confirm the reliability of the results obtained. The MPAN model accuracies for the binary classification datasets are listed in Table 8, while those for the tertiary classification datasets are listed in Table 9. The MPAN model is then applied to the merged datasets for each collection. The results for the merged datasets are given at the bottom of Tables 8 and 9.

Tables 10 and 11 show how the MPAN model accuracies compare with previous baselines provided for both binary and tertiary classification datasets. Note that, for some datasets, no reliable results are available, and accordingly, these are not included in Tables 10 and 11. It is apparent that the MPAN model beats all reported baselines with accuracies between 90.42% and 98.65% for the binary classification case and between 91.04% and 96.72% for the tertiary classification case. The results of the MPAN model are shown in bold in the first column. This performance is expected, since the MPAN model produces state-of-the-art results on international public datasets such as those given in table 7.

V. CONCLUSION

A novel deep learning-based multilevel attention neural (MPAN) model has been applied to perform Arabic sentiment analysis tasks using two collections of labeled datasets: a tertiary classification collection comprising 13 datasets (with 149,880 items) and a binary classification collection comprising 21 datasets (with 125,261 items). The accuracies achieved are 94.25% for the tertiary classification case and 95.61% for the binary classification case. It has been shown that MPAN outperforms all reported baselines for all individual datasets. The MPAN model is very promising and has great potential. At the universal level, MPAN model outperformed all known attention models, achieving top positions on leaderboards of IMDB, SNLI, QQP, SciTail and MultiNLI datasets..More work is needed to develop larger Arabic sentiment analysis datasets and further develop and optimize the MPAN model as a novel benchmark approach for both Arabic and multilingual sentiment analysis [58]. Alternatively, semi supervised MPAN models will also be explored, e.g., building on [59].

REFERENCES

[1] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 320–342, Mar. 2019, doi: 10.1016/j.ipm.2018.07.006.

[2] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Appl. Comput. Intell. Soft Comput.*, vol. 2020, Jan. 2020, Art. no. 7403128, doi: 10.1155/2020/7403128.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2013, pp. 3111–3119.

[5] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[7] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: http://arxiv.org/abs/1508.04025

[8] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] M. E. M. Abo, R. G. Raj, and A. Qazi, "A review on arabic sentiment analysis: State-of-the-art, taxonomy and open research challenges," *IEEE Access*, vol. 7, pp. 162008–162024, 2019, doi: 10.1109/access.2019.2951530.

[11] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, "Deep learning for arabic NLP: A survey," *J. Comput. Sci.*, vol. 26, pp. 522–531, May 2018, doi: 10.1016/j.jocs.2017.11.011.

[12] N. Abdelhade, T. H. A. Soliman, and H. M. Ibrahim, "Detecting Twitter users' opinions of Arabic comments during various time episodes via deep neural network," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, Cham, Switzerland, Sep. 2018, pp. 232–246.

[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," *J. Comput. Sci.*, vol. 27, pp. 386–393, Jul. 2018, doi: 10.1016/j.jocs.2017.11.006.

[14] M. Al-Smadi, B. Talafha, M. Al-Ayyoub, and Y. Jararweh, "Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2163–2175, Mar. 2018, doi: 10.1007/s13042-018-0799-4.

[15] S. Al-Azani and E. S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text," in *Proc. 7th Int. Conf. Sustain. Energy Inf. Technol. (SEIT)*, Madeira, Portugal, Jan. 2017, pp. 359–366.

[16] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A combined CNN and LSTM model for Arabic sentiment analysis," in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl, Eds. Cham, Switzerland: Springer, 2018, pp. 179–191.

[17] A. A. Altowayan and L. Tao, "Word embeddings for arabic sentiment analysis," in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, Dec. 2016, pp. 3820–3825.

[18] A. A. Altowayan and A. Elnagar, "Improving arabic sentiment analysis with sentiment-specific embeddings," in *Proc. IEEE Int. Conf. Big Data*, Boston, MA, USA, Dec. 2017, pp. 4314–4320.

[19] A. Barhoumi, Y. Estève, C. Aloulou, and L. Belguith, "Document embeddings for Arabic sentiment analysis," in *Proc. Conf. Lang. Process. Knowl. Manage. (LPKM)*, Sfax, Tunisia, Sep. 2017, Paper hal-02042060, pp. 1–9.

[20] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for Arabic sentiment classification," in *Proc. 26th Int. Conf. Comput. Linguistics Tech. Papers*, Osaka, Japan, Dec. 2016, pp. 2418–2427.

[21] A. M. El-Halees, "Arabic opinion mining using distributed representations of documents," in *Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT)*, Gaza City, Palestinian, May 2017, pp. 28–33.

[22] M. Gridach, H. Haddad, and H. Mulki, "Empirical evaluation of word representations on Arabic sentiment analysis," in *Arabic Language Processing: From Theory to Practice*, A. Lachkar, K. Bouzoubaa, A. Mazroui, A. Hamdani, A. Lekhouaja, Eds. Cham, Switzerland: Springer, 2018, pp. 147–158.

[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 Comput. Sci.*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.

[24] A. El-Kilany, A. Azzam, and S. R. El-Beltagy, "Using deep neural networks for extracting sentiment targets in Arabic tweets," in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham, Switzerland: Springer, 2018, pp. 3–15.

[25] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/mci.2018.2840738.

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[27] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, vol. 32, Jan. 2014, pp. 1188–1196.

[28] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Amman, Jordan, Dec. 2013, pp. 1–6.

[29] H. ElSahar and S. R. El-Beltagy, "Building large Arabic multi-domain resources for sentiment analysis," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Cham, Switzerland: Springer, 2015, pp. 23–34.

[30] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic sentiment tweets dataset," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 2515–2519.

[31] E. Refaee and V. Rieser, "An Arabic Twitter corpus for subjectivity and sentiment analysis," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, 2014, pp. 2268–2273.

[32] A. Elnagar, L. Lulu, and O. Einea, "An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis," *Procedia Comput. Sci.*, vol. 142, pp. 182–189, 2018, doi: 10.1016/j.procs.2018.10.474.

[33] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102121, doi: 10.1016/j.ipm.2019.102121.

[34] M. Elhag M. Abo, R. Gopal Raj, A. Qazi, and A. Zakari, "Sentiment analysis for arabic in social media network: A systematic mapping study," 2019, *arXiv:1911.05483*. [Online]. Available: http://arxiv.org/abs/1911.05483
[35] E. Omara, M. Mosa, and N. Ismail, "Deep convolutional network for arabic sentiment analysis," in *Proc. Int. Japan-Africa Conf. Electron., Commun. Computations (JAC-ECC)*, Alexandria, Egypt, Dec. 2018, pp. 155–159.
[36] I. Abu Farha and W. Magdy, "Mazajak: An online arabic sentiment analyser," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, Florence, Italy, 2019, pp. 192–198.
[37] M. Al Omari, M. Al-Hajj, A. Sabra, and N. Hammami, "Hybrid CNNs-LSTM deep analyzer for arabic opinion mining," in *Proc. 6th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Granada, Spain, Oct. 2019, pp. 364–368.
[38] M. Heikal, M. Torki, and N. El-Makky, "Sentiment analysis of arabic tweets using deep learning," *Procedia Comput. Sci.*, vol. 142, pp. 114–122, Dec. 2018, doi: 10.1016/j.procs.2018.10.466.
[39] A. Alwehaibi and K. Roy, "Comparison of pre-trained word vectors for arabic text classification using deep learning approach," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Orlando, FL, USA, Dec. 2018, pp. 1471–1474.
[40] K. Elshakankery and M. F. Ahmed, "HILATSA: A hybrid incremental learning approach for arabic tweets sentiment analysis," *Egyptian Informat. J.*, vol. 20, no. 3, pp. 163–171, Nov. 2019, doi: 10.1016/j.eij.2019.03.002.
[41] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop Semantic Eval.*, Vancouver, BC, Canada, 2017, pp. 502–518.
[42] M. Salameh, S. Mohammad, and S. Kiritchenko, "Sentiment after translation: A case-study on arabic social media posts," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Denver, CO, USA, 2015, pp. 767–777.
[43] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for arabic," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 10, pp. 2045–2054, Oct. 2011, doi: 10.1002/asi.21598.
[44] M. Nabil, M. Aly, and A. Atiya, "LABR: A large scale arabic sentiment analysis benchmark," 2014, *arXiv:1411.6718*. [Online]. Available: http://arxiv.org/abs/1411.6718
[45] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Nancy, France, Apr. 2017, pp. 114–118.
[46] A. Elmadany, H. Mubarak, and W. Magdy, "ARSAS: An Arabic speech-act and sentiment corpus of tweets," in *Proc. 3rd Workshop Open-Source Arabic Corpora Process. Tools*, Miyazaki, Japan, May 2018, p. 20.
[47] M. Al Omari, M. Al-Hajj, N. Hammami, and A. Sabra, "Sentiment classifier: Logistic regression for arabic Services' reviews in Lebanon," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Sakaka, Saudi Arabia, Apr. 2019, pp. 1–5.
[48] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-tweet: A corpus for arabic sentiment analysis of saudi tweets," *Procedia Comput. Sci.*, vol. 117, pp. 63–72, 2017, doi: 10.1016/j.procs.2017.10.094.
[49] T. Khalil and S. R. El-Beltagy, "NileTMRG at SemEval-2016 task 5: Deep convolutional neural networks for aspect category and sentiment extraction," in *Proc. 10th Int. Workshop Semantic Eval.*, San Diego, CA, USA, 2016, pp. 271–276.
[50] M. A. El Affendi and K. H. S. Al Rajhi, "Text encoding for deep learning neural networks: A reversible base 64 (Tetrasexagesimal) integer transformation (RIT64) alternative to one hot encoding with applications to arabic morphology," in *Proc. 6th Int. Conf. Digit. Inf., Netw., Wireless Commun. (DINWC)*, Beirut, Lebanon, Apr. 2018, pp. 70–74.
[51] M. A. Elaffendi, I. Abuhaimed, and K. AlRajhi, "A simple galois power-of-two real time embedding scheme for performing Arabic morphology deep learning tasks," *Egyptian Inform. J.*, Apr. 2020, doi: 10.1016/j.eij.2020.03.002.
[52] J. Loy, *Neural Network Projects with Python: The Ultimate Guide to Using Python to Explore the True Power of Neural Networks Through Six Projects*. Birmingham, U.K.: Packt Publishing, 2019.
[53] *Sentiment Analysis on IMDb*. Accessed: Dec. 10, 2020. [Online]. Available: https://www.paperswithcode.com/sota/sentiment-analysis-on-imdb
[54] J. Pilault, A. E. hattami, and C. Pal, "Conditionally adaptive multi-task learning Improving transfer learning in NLP using fewer parameters & less data," 2019, *arXiv:2009.09139*. [Online]. Available: https://arxiv.org/abs/2009.09139
[55] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," 2019, *arXiv:1901.11504*. [Online]. Available: http://arxiv.org/abs/1901.11504
[56] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*. [Online]. Available: http://arxiv.org/abs/1906.08237
[57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified Text-to-Text transformer," 2019, *arXiv:1910.10683*. [Online]. Available: http://arxiv.org/abs/1910.10683
[58] E. Cambria, S. Poria, A. Hussain, and B. Liu, "Computational intelligence for affective computing and sentiment analysis [guest editorial]," *IEEE Comput. Intell. Mag.*, vol. 14, no. 2, pp. 16–17, May 2019, doi: 10.1109/mci.2019.2901082.
[59] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018, doi: 10.1016/j.neucom.2017.10.010.

MOHAMMED A. EL-AFFENDI is currently a Professor of computer science with the Department of Computer Science, Prince Sultan University, a Former Dean of CCIS, AIDE, the Rector, a Founder, and the Director of Data Science Laboratory (EIAS), a Founder and the Director of The Center of Excellence in CyberSecurity. His current research interests include data science, intelligent and cognitive systems, machine learning, and natural language processing.

KHAWLA ALRAJHI received the B.Sc. degree in information technology from King Saud University, Riyadh, Saudi Arabia, in 2013, and the M.Sc. degree in software engineering from Prince Sultan University, Riyadh, in 2018. She is currently a Lecturer and a Research Assistant with EIAS, College of Computer and Information Sciences, Prince Sultan University, Riyadh. Her research interests include natural language process, machine learning, and artificial intelligence.

AMIR HUSSAIN (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. He held postdoctoral and academic positions at the West of Scotland from 1996 to 1998, the University of Dundee from 1998 to 2000, and the University of Stirling from 2000 to 2018, respectively. He is currently a Professor and the founding Head of the Cognitive Big Data and Cybersecurity Research Laboratory, Edinburgh Napier University, U.K. His research interests include cognitive computation, machine learning, and computer vision.