Egyptian Informatics Journal 29 (2025) 100601

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full Length Article

Benchmark Arabic news posts and analyzes Arabic sentiment through RMuBERT and SSL with AMCFFL technique

Mustafa Mhamed [a,c], Richard Sutcliffe [b,c,**], Jun Feng [c,*]

[a] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[b] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
[c] School of Information Science and Technology, Northwest University, Xi'an 710127, China

ARTICLE INFO

Keywords:
Natural language processing
Arabic sentiment analysis
SSL
RMuBERT
ANP5
ANPS2

ABSTRACT

Sentiment analysis aims to extract emotions from textual data; sentiment analysis and text recognition are two of the most common tasks associated with natural language processing. Emergent technologies have been developed and employed in various fields, including marketing, health care, and policy making. However, with the growth of social media platforms and the flow of data, especially in the Arabic language, substantial difficulties have emerged that call for the creation of new frameworks to address problems, such as the lack of datasets related to news platforms, the complicated formation of the Arabic language, and complications with classifying, and system challenges, whether in machine learning, deep learning, or online analysis tools. This paper provides a new framework that helps address ASA challenges and work on various tasks based on the state-of-the-art ASA. First, it presents a new collection named (ANP5) from Arabic news posts from several Arabic platforms, then uses SSL with AMCFFL technique to analyze the Arabic sentiment and generate a second dataset (ANPS2). Next, applied ML classifiers, RF and SVM, do the best among the other classifiers, with an accuracy of 82.00%; however, the measurement distributions for each class are different (Experiment 1). Following that, DL models, BIGRU, CNN-LSTM, LSTM, and CNN, had accuracies of 88.10%, 89.30%, 89.85%, and 90.10% (Experiment 2). Experiments 1 and 2 represent the initial benchmark classification as the first baseline. Afterward, a new RMuBERT Model was developed and compared with four transformers on the two datasets: ANPS2 accuracy (90.87%) and ANP5 (90.33%). RMuBERT performed better than the baselines (Experiment 3). Further testing of RMuBERT on various Arabic corpora with different classes, lengths, and sizes: ArSarcasm (3C), STD (2C), AJGT (2C), and AAQ (2C), revealed accuracies of 77.76%, 91.79%, 94.07%, and 93.48%, respectively. Still, RMuBERT performed better than the baselines (Experiment 4). Finally, on the largest Arabic sentiment corpora with six million Arabic tweets, the performance is up to (91.12%); RMuBERT works efficiently with less training time (Experiment 5).

1. Introduction

Digital networks offer an exceptional framework for big data analytics in various real-world applications [1]. When individuals submit their views or ideas while conversing on social media sites such as MySpace, Facebook, and Twitter, a tremendous quantity of data is created regularly [2]. Social data is one of the three significant data characteristics: velocity, heterogeneity, and high volume [3]. It is created via numerous social channels [4]. Aside from these features, it has a semantic feature, which refers to the fact that it is humanly made and includes symbolic information with intrinsic subjective value [5,6]. Because of this unique feature of big social data, emotion recognition faces several obstacles and possibilities, such as application [7], machine learning [8], semi-automatic learning [9], and identifying and categorizing the individual's personality [10].

Sentiment classification, opinion mining, sentiment mining, and review mining are some terms used in sentiment analysis [11]. It's a method for forecasting how people react to various items or social organizations based on their feelings [12]. Advanced techniques were used for emotion recognition [13] and detecting harassment discourse [14].

[*] Corresponding author.

[**] Corresponding author at: School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.
E-mail addresses: mustafamhamed2099@gmail.com (M. Mhamed), rsutcl@essex.ac.uk (R. Sutcliffe), fengjun@nwu.edu.cn (J. Feng).

https://doi.org/10.1016/j.eij.2024.100601
Received 6 October 2024; Received in revised form 1 December 2024; Accepted 25 December 2024
Available online 15 January 2025

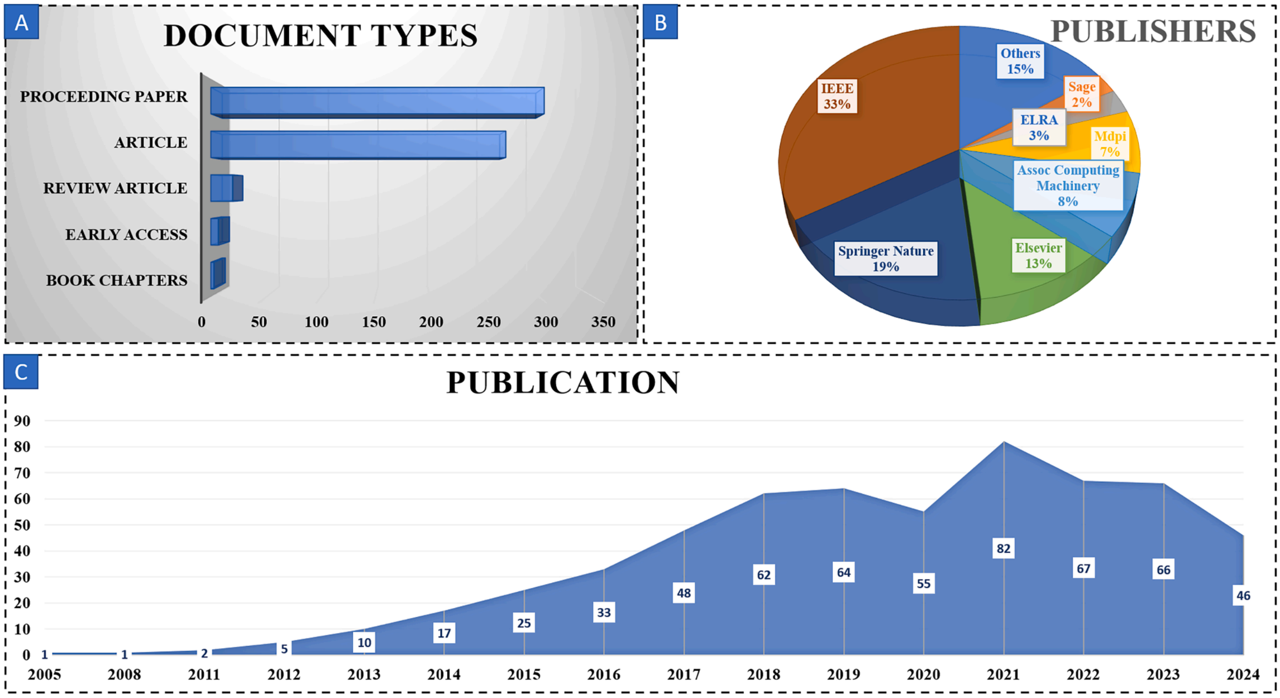


Fig. 1. Represents the global ASA growth and distribution via time. (A) document kinds that have been published and their distribution, (B) the allocation of resources according to the publishers, and (C) articles analysis with timelines.

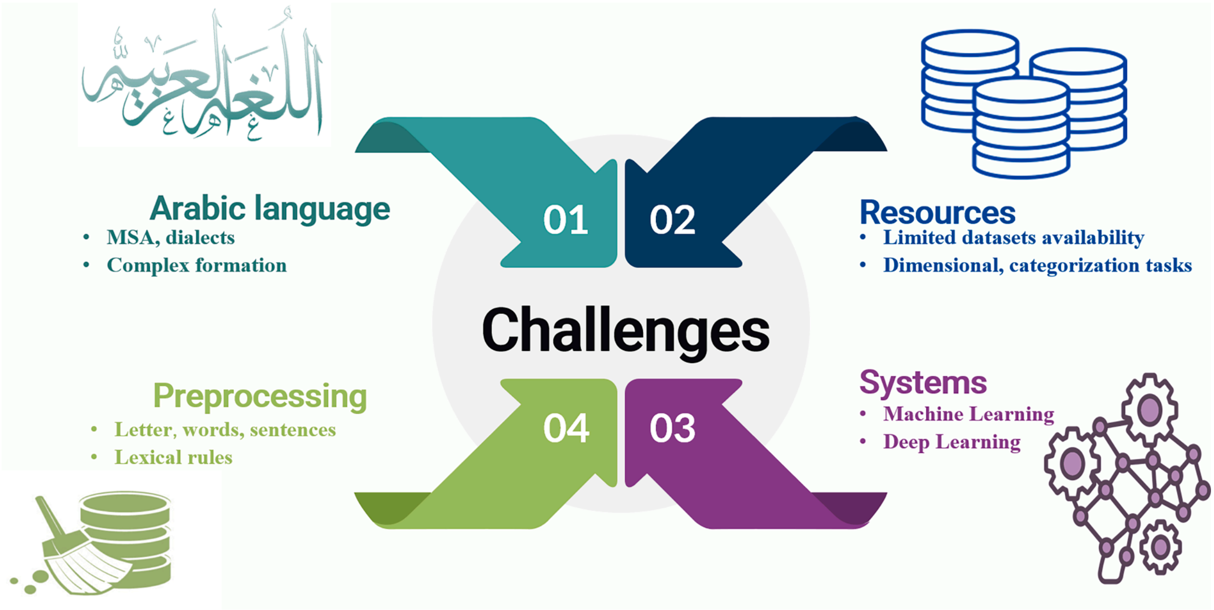


Fig. 2. Describes the current research challenges.

Textual and visual representations are often used as sources for sentiment analysis, aspect-based sentiment analysis, and fuzzy decision [15]. The emotions expressed on social media cover over 23 distinct languages, and more than 11 platforms are valuable for developing business strategies to help the company reach its objectives [16]. It's often used to manage a product's or brand's internet reputation and knowledge discovery via the public [17].

Further, owing to the abundance of channels and sources for Arab news, it is now challenging to identify and validate sources due to the enormous amount and growth of data. Analysis of feelings for comments and tweets also presents challenges [18]. Fig. 1, a graph and analysis, shows the worldwide ASA growth and distribution throughout time. Most of the information selected for this investigation was derived from the Web of Science database. The following keywords are primarily employed to locate relevant scientific publications: Arabic sentiment analysis and Arabic sentiment resources.

Arabic NLP tools, such as SA, have gained popularity in recent years due to the rise of Arabic online content, notably on social media, and the region's seismic political, climatic, and economic changes [19]. Recent supervised learning (DL) developments have significantly progressed in various NLP tasks, such as synthetic text generation impact [20], recognizing the best features [21], comprehensive text processing [22],

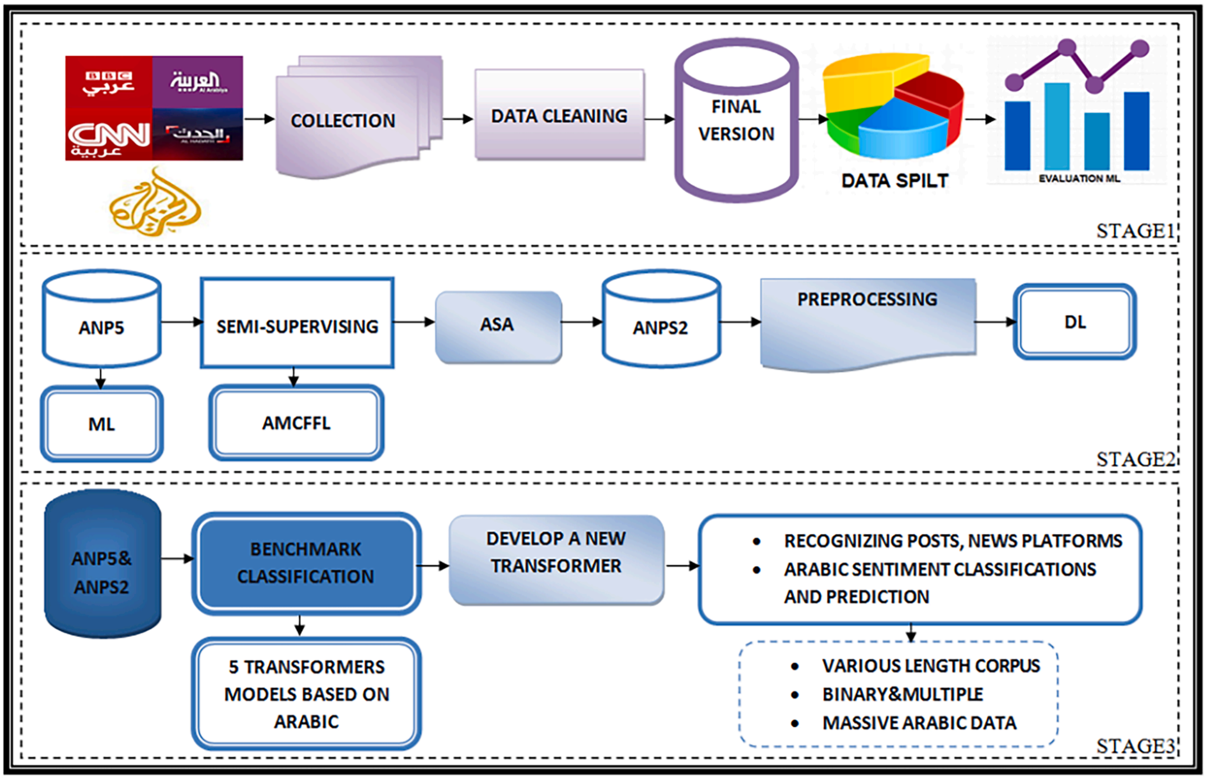


Fig. 3. The general description of the work.

and the effective categorization of relational ontological frameworks for user feedback [23].

Despite the rapid development of artificial intelligence technologies applied in Arabic sentiment analysis, the research community faces challenges (see Fig. 2). These include a lack of readily available dataset resources, the complicated formation of the Arabic language, the difficulties associated with classifying Arabic emotions, and system challenges, whether they pertain to machine learning, deep learning, or online analysis tools.

In this paper, we introduce a new collection of Arabic News Posts (ANP5) and present a new framework (see Fig. 3) with various techniques that help solve issues in Arabic in the three main tasks: Platforms Posts Recognition (PPR) or (Recognize Platforms of the Posts or news), Sentiment classifying (SC), and Arabic Sentiment Prediction (ARP).

We design a semi-supervised learning approach for categorization that can benefit from unlabeled data and enhance classification performance using transformer models. The key contributions of this paper are listed below:

- Following state-of-the-art ASA, this study develops a framework that helps solve the ASA issues and works on numerous tasks such as SSL, PPR, SC, and ARP.
- Gather data from Twitter-based Arabic news sources to generate a unique 5-class dataset of Arabic news postings (ANP5).
- After cleaning and preprocessing the collection, ML classifiers were applied to the corpus as initial benchmarking.
- Classifying and examining the sentiment from data using semi-supervising and the newly proposed technique (AMCAL).
- Employing DL models for initial evaluation of the second newly generated corpus (ANPS2).
- Transformers were applied using state-of-the-art techniques on two datasets, outperforming baselines in performance.
- Developed a new RMuBERT Model based on MARBERT and applied it to two datasets; RMuBERT performs best.
- Further, RMuBERT utilizes various Arabian datasets with different classes, lengths, and sizes; second, it uses massive data. RMuBERT was more efficient and provided the optimum results in the shortest training time.
- Finally, the datasets and the investigations are provided free to the public at.[1]

Table 1
Recent works in sentiment analysis on Arabic and non-Arabic.

Paper	Dataset	Language	Model	Result
[34]	Amazon (2C)	English	BERT	88.48 %
[35]	IMDB (2C)	English	BERT	94.00 %
[36]	Wikipedia (2C)	Chinese	BERT + FC	92.65 %
[37]	Chin_Corpus (3C)	Chinese	BERT	86.4 %
[38]	SemEval (4C)	Arabic	ArabicBERT	72.00 %
[39]	Italian XXL (2C)	Italy	BERT	93.32 %
[40]	TripAdvisor (5C)	Spanish	BERT	56.72 %
[41]	Fre_Corpus (2C)	French	BERT	82.3 %
[42]	ArSarcasm-v2 (2C)	Arabic	MARBERT	86.00 %
[43]	Sarcasm (3C)	Arabic	AraBERT	70.00 %

The rest of this paper is organized as follows: Section 2 reviews previous work on sentiment analysis with the latest deep-learning technique. Section 3 describes the data collection, statistics, and methodology; Section 4 contains experiments and findings; Section 5 offers the discussions; Section 6 limitations and future research; and Section 7 provides the conclusion.

2. Previous work

The tremendous growth in the utilization and popularity of social media platforms over the last several decades has resulted in a massive

[1] https://github.com/mustafa20999/-ANP5-ANPS2-datasets.

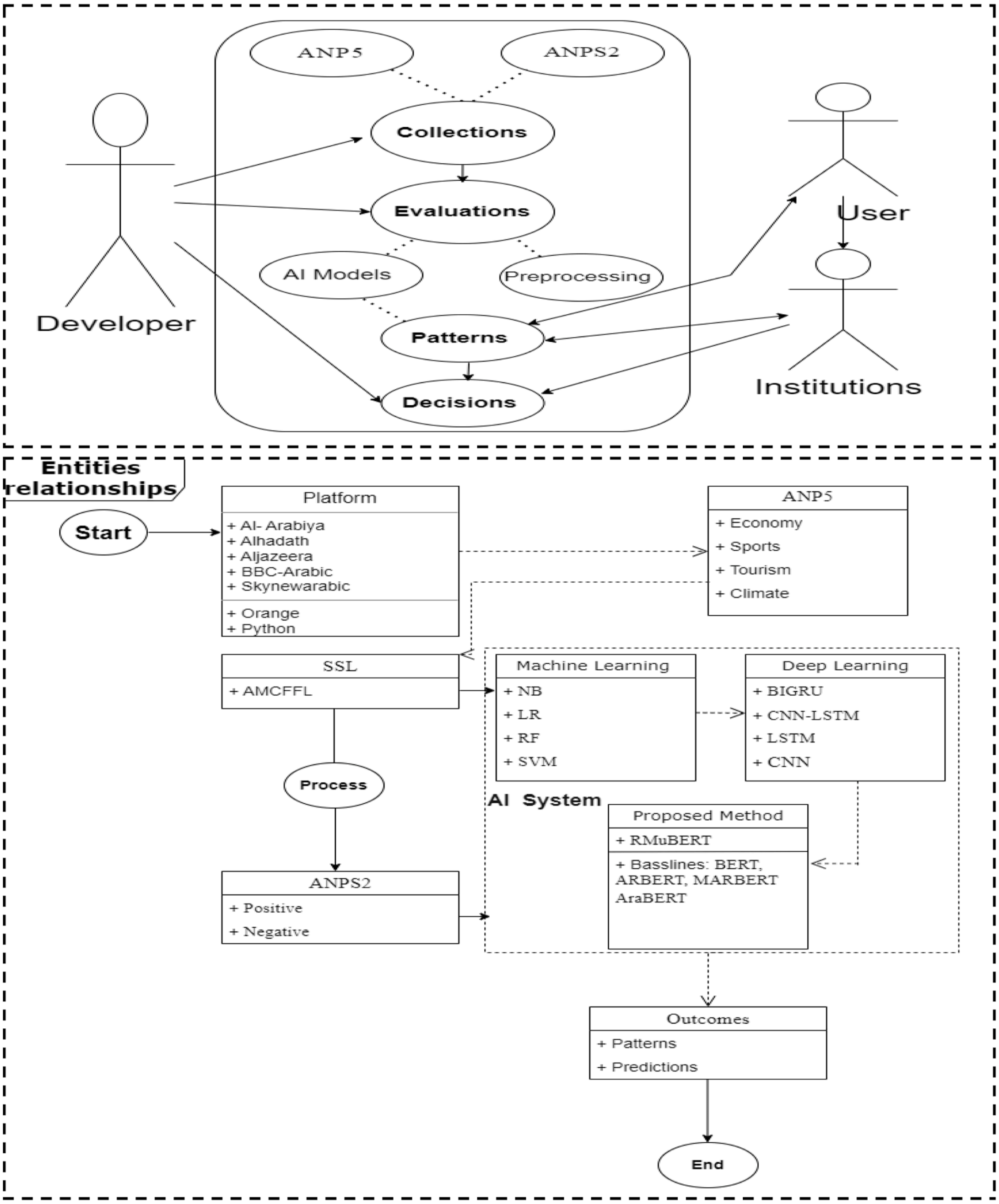


Fig. 4. Displays a flowchart illustrating the overall procedures, the primary operating between actors, processes, and the relationships among entities.

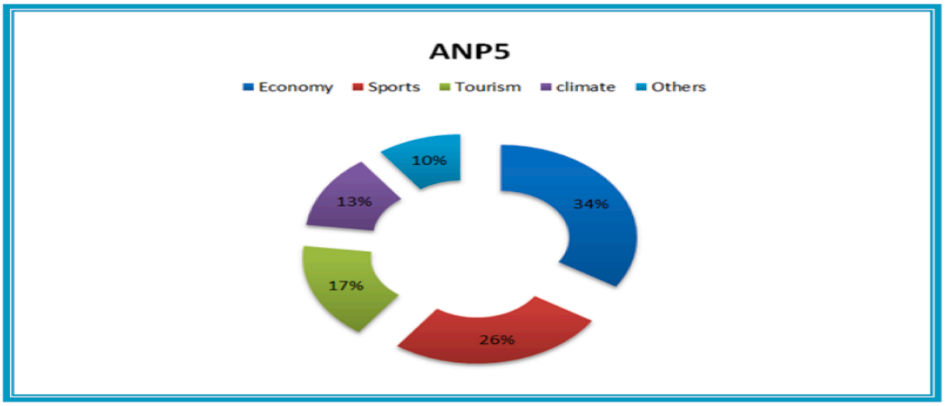


Fig. 5. Corpus Data.

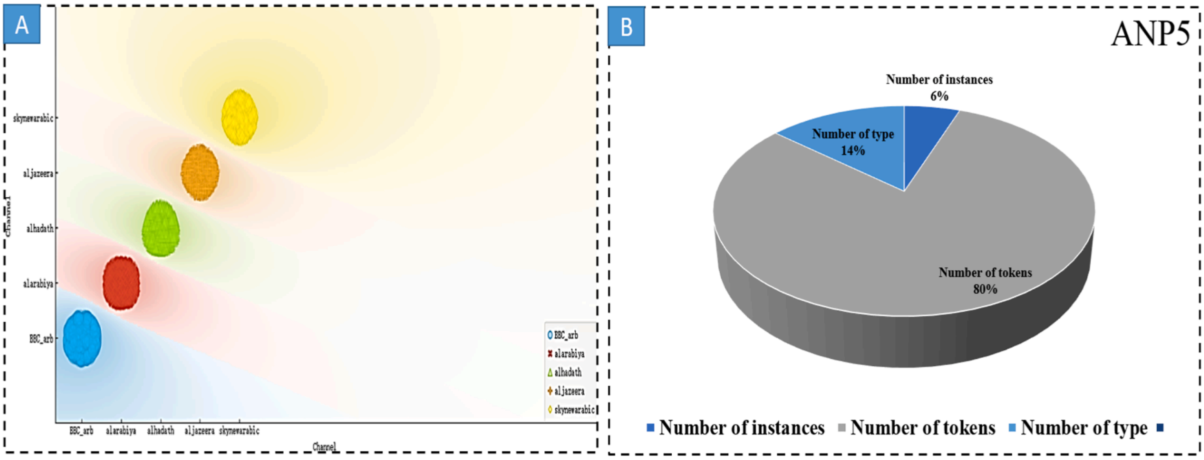


Fig. 6. Statistics of ANP5 datasets (A) ANP5 corpus based on the channel collection, and (B) the distribution regarding instances, types, and tokens.

Table 2
Statistics of ANP5 datasets.

	ANP5
Number of instances	10,000
Number of tokens	142,366
Number of Type	24,594
Size	10 k

actual-time volume of social texts and postings, including messages, created by subscribers. Walaa et al. [24] provided a comprehensive review that included SA domains, applications, and resources. In the study by George et al. [25], three multilingual deep learning classifiers were applied and compared in addition to a zero-shot categorization strategy. The results demonstrate that the zero-shot classification method cannot achieve high levels of accuracy when applied to monolingual data. Furthermore, several techniques were applied to assess the data quality, especially in healthcare, such as ASA for COVID-19 [26], rating medical

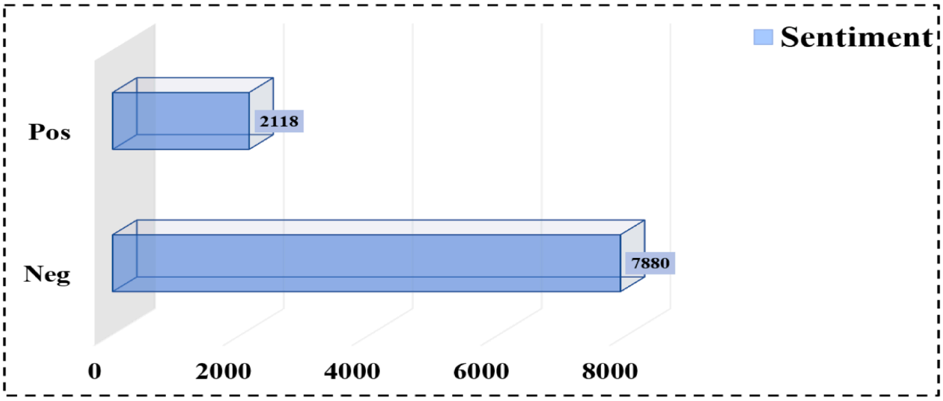


Fig. 7. ANPS2 corpus.

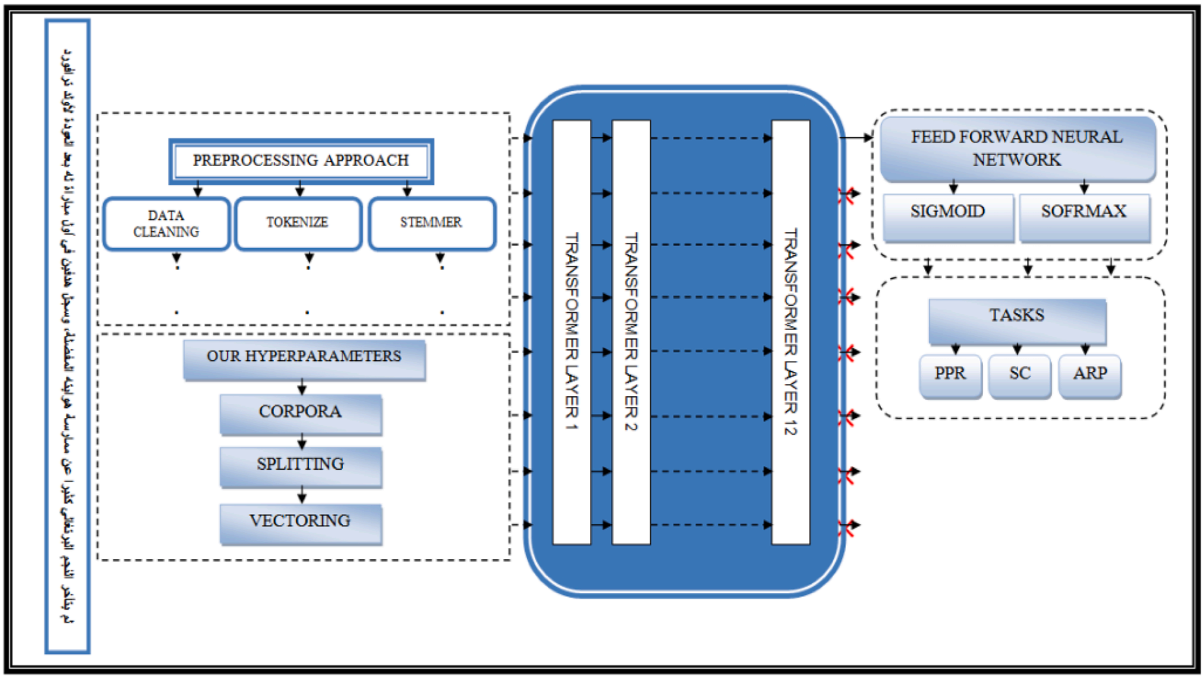


Fig. 8. The proposed RMUBERT-based architecture.

issues [27], Twitter Arabic health services [28], Facebook [29], Holistic Intelligent Healthcare Theory (HIHT) [30], impacts of data preprocessing [31], data cleaning mechanisms [32], and Herbal Treatments for Diabetes [33].

Numerous contemporary methodologies have evolved due to the growth and expansion of information flow through social networking sites, helping to address issues with sentiment analysis in general and Arab sentiments in particular. Table 1 (the article, dataset, language, model, and outcome) is the prior state-of-the-art computational procedures for sentiment analysis in various languages. We will begin sentiment analysis with non-Arabic and later on with Arabic.

On Amazon [44] datasets, Geetha et al. [34] upgraded BERT and compared it to LSTM, SVM, and Naive Bayes. With two ways for feature extraction, Bag of Words and TF-IDF, accuracy was 88.48%, 83.97%, 81.33%, and 80.12%, respectively. Basarslan et al. [35] utilized word embedding techniques such as Global Vector (GloVe), Word2Vec (W2V), and Bidirectional Encoder Representation (BERT). On the IMDB and Yelp datasets, DL classifiers (CNN, LSTM, and RNN) and ML classifiers (SVN and NB), with Bag-of-Words (BOW) and (TF-IDF). The BERT word embedding approach outperforms other techniques with all classifiers. The highest accuracy was reported with BERT and LSTM, with an IMDB accuracy of 94.00% and a Yelp accuracy of up to 89.00%. Li et al. [36] developed BERT[2] on the Chinese corpus,[3] which contains 9204 positive and negative reviews; their BERT+FC gave the best performance up to 92.65%. Xie et al. [37] modified the BERT model with the best fine-tuning on the Chinese sentiment corpus and compared it with RoBERTa-wwm-ext, RoBERT, ERNIE, and BERT transformers, and their approach was the best with (86.4%) accuracy.

Catelli et al. [39] compared the capacity of a language model based on deep neural networks, such as BERT Base Italian XXL,[4] with a system based on lexicons, such as NooJ, to evaluate sentiment in the Italian language using an ad hoc collection with performance up to 93.32%. Vasquez et al. [40] used two Bert-based techniques on five classes of Spanish datasets.[5] The first method involved perfecting Beto, a Bert-like figure pre-trained in Spanish. The second method relies on fusing TF-IDF-weighted feature vectors with Bert embeddings. The top outcome was up 56.72%.

In [41], fine-tune several BERT-based models pre-trained on English, Italian, and French datasets in two subtasks: classification and regression. For classification, F1-macro English was 0.873, Italian was 0.874, and French was 0.823; for regression, MSE was 2.71, 3.86, and 3.65, respectively. In two tasks vs. the baselines, the model performed better.

A zero-shot classification strategy and four multilingual BERT-based classifiers [46] are applied and evaluated for their efficiency and usefulness in classifying multilingual data [47]. Manias et al. [48] presented a comprehensive analysis of applying an Entity-Level Sentiment Analysis (ELSA) methodology to Twitter Data. It improved the information that may be extracted from tweets, aiming to upgrade the policy-making processes used by modern businesses and enterprises.

Nevertheless, when we come to sentiment analysis of Arabic, Al-Twairesh [38] investigated the development of language models from the classic TF-IDF to the more advanced word2vec and, most recently, the state-of-the-art pertained language model BERT. They discovered that the ArabicBERT-Large model delivers the best results on SemEval 2018 [45]; the rise in F1-score was significant at +7–21%. For two tasks, Sarcasm Detection (SD) and Sentiment Analysis (SA), Abuzayed et al. [42] employed data augmentation to fine-tune seven BERT-based models. MAR-BERT reports the best performance on the two tasks, with an F1-sarcastic performance of up to 0.86. We conclude by Farha et al. [43] by applying several transforms on two Arabic datasets with different tasks on Sarcasm, and sentiment (3C) datasets. On sentiment (3C), AraBERT was the highest accuracy (70.00%), Fpn (0.724), and on Sarcasm Detection, MARBERT was the best with recall (0.714) and F1 (0.584).

We see that the BERT-based model performed well with binary corpora and poorly with multiple categories. It is employed on smaller data sets; the large model size and the sophisticated design of BERT

[2] https://github.com/ymcui/Chinese-BERT-wwm.

[3] https://github.com/algosenses/Stock_Market_Sentiment_Analysis/tree/master/data.

[4] https://github.com/huggingface/transformers.

[5] https://github.com/juanmvsa/Sentiment-Analysis-TripAdvisor-Spanish.

Table 3
Hyperparameter value.

Setting	Value(s)
gradient − accumulation − steps	{2, 4, 6, 8}
Number-classes	{1,2, 3, 5, 6, 8}
Batch-size	{16, 32, 64, 128, 256, 512, 1024}
Max − len	{66, 80, 100, 124, 200, 250, 520}
Extra − Len	{6, 10, 12}
Rand − Seed	{42, 84, 123, 666}
adam − epsilon	{1e − 8 }
Epoch	{5, 10, 20, 50, 100, 200 }
Optimizer	Adam
Learning rate	{1.215e − 05}

might lead to overfitting, which is mainly an issue. In this way, generalization and performance on data that has not been viewed might suffer. Additional fine-tuning or domain adaptation may be required to ensure that the BERT pre-trained weights can generalize effectively to tasks with considerable domain or language changes. Second, more competitive preprocessing needs to be implemented to cover the Arabic context at three levels: characters, letters, and sentences, in addition to evaluating massive Arabic datasets. Thirdly, Arabic does poorly when compared to other languages. The Arabic-based transformer models weren't tested on a variety of Arabic data of different lengths and sizes. Finally, the transformer's Attention layer can only handle fixed-length textual data. Additionally, a significant quantity of information is lost when a context is split in two or along its center. These points addressed the necessity of cleaning and preparing Arabic data and choosing the optimum hyperparameters-compatible sequence layers for the architecture of the final selected tasks.

Here, we establish a new collection called (ANP5). Following data cleaning and processing, we benchmark the datasets using ML and DL models before semi-supervising using (AMCFFL) methodologies to categorize the sentiment analysis. We are developing a transformer based on Arabic language models, demonstrating its outstanding performance using varied Arabic data with varying categories and sizes (see next section).

3. Materials and methods

3.1. Datasets collection

This work (see Fig. 4) built a new collection called (ANP5) from five Arabic news platforms, including (Al- Arabiya, Alhadath, Aljazeera, BBC-Arabic, and Skynewarabic), as well as the Twitter platform. Afterward, 10,320 items were collected, including posts and comments about the economy, sports, tourism, climate, and some global policies, as shown in Fig. 5. Next, purge redundant items, dates, author followers, friends, favorites, images, locations, status URLs, and tweet sources from the dataset. Additionally, eliminate the Tweet ID, Username, and accompanying Retweet, Comment, and Favorites counts.

Following the abovementioned process, 10,000 items from the initial collection were chosen, as shown in Fig. 6. The Statistics of ANP5 are shown in Table 2, which include the number of instances for the corpora, the number of tokens and types, and the final size.

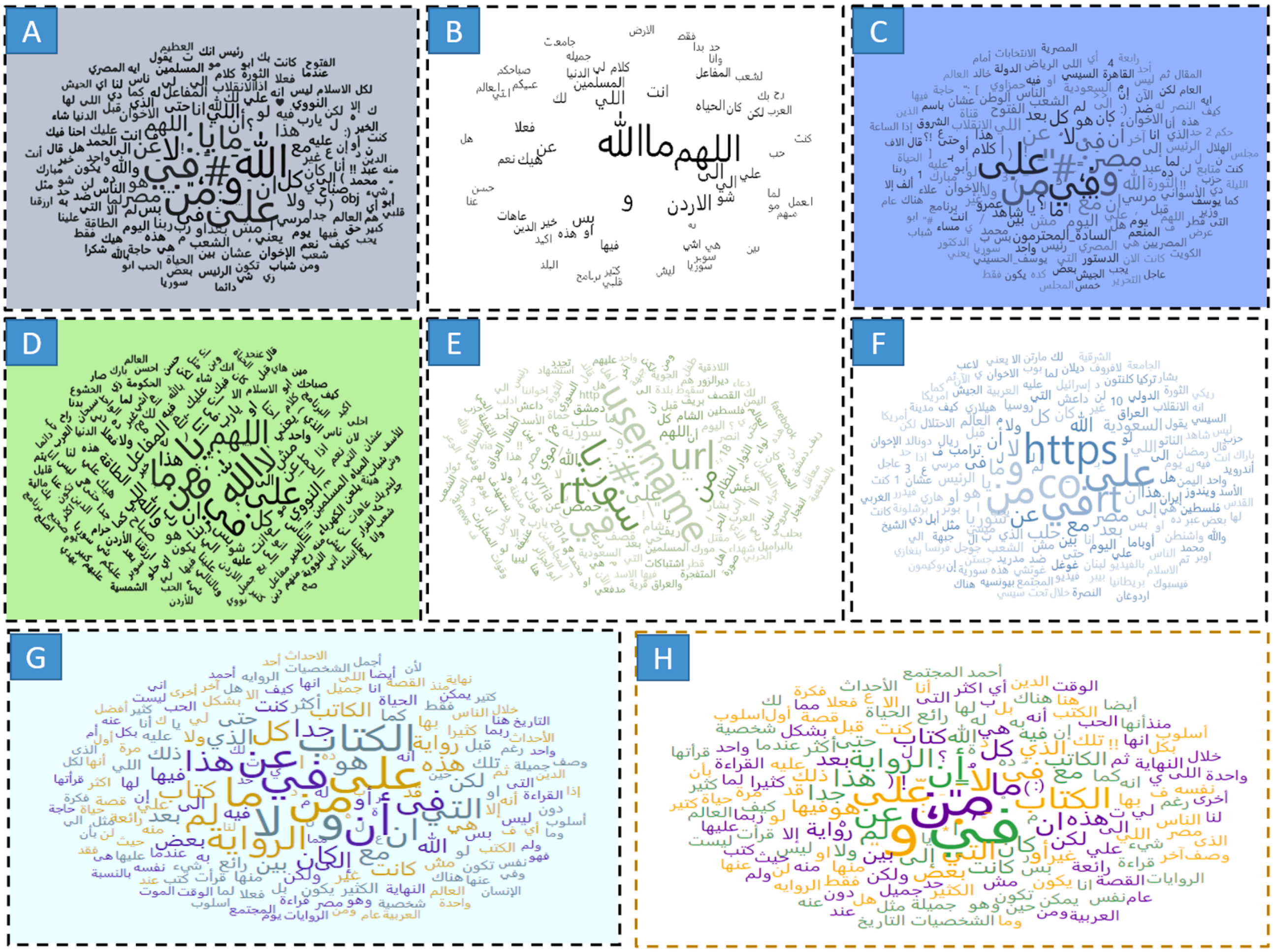


Fig. 9. Display the distribution of the datasets utilized. (A) AAQ, (B) AJGT, (C) ASTD, (D) ArTwitter, (E) STD, (F) ArSarcasm, (G) BRAD, and (H) TEAD.

Next, Semi-supervised learning analyzes the sentiment that generates (ANPS2) corpus Fig. 7 (see next section). Both corpora are shown in Table 4.

3.2. Semi-supervising learning (SSL)

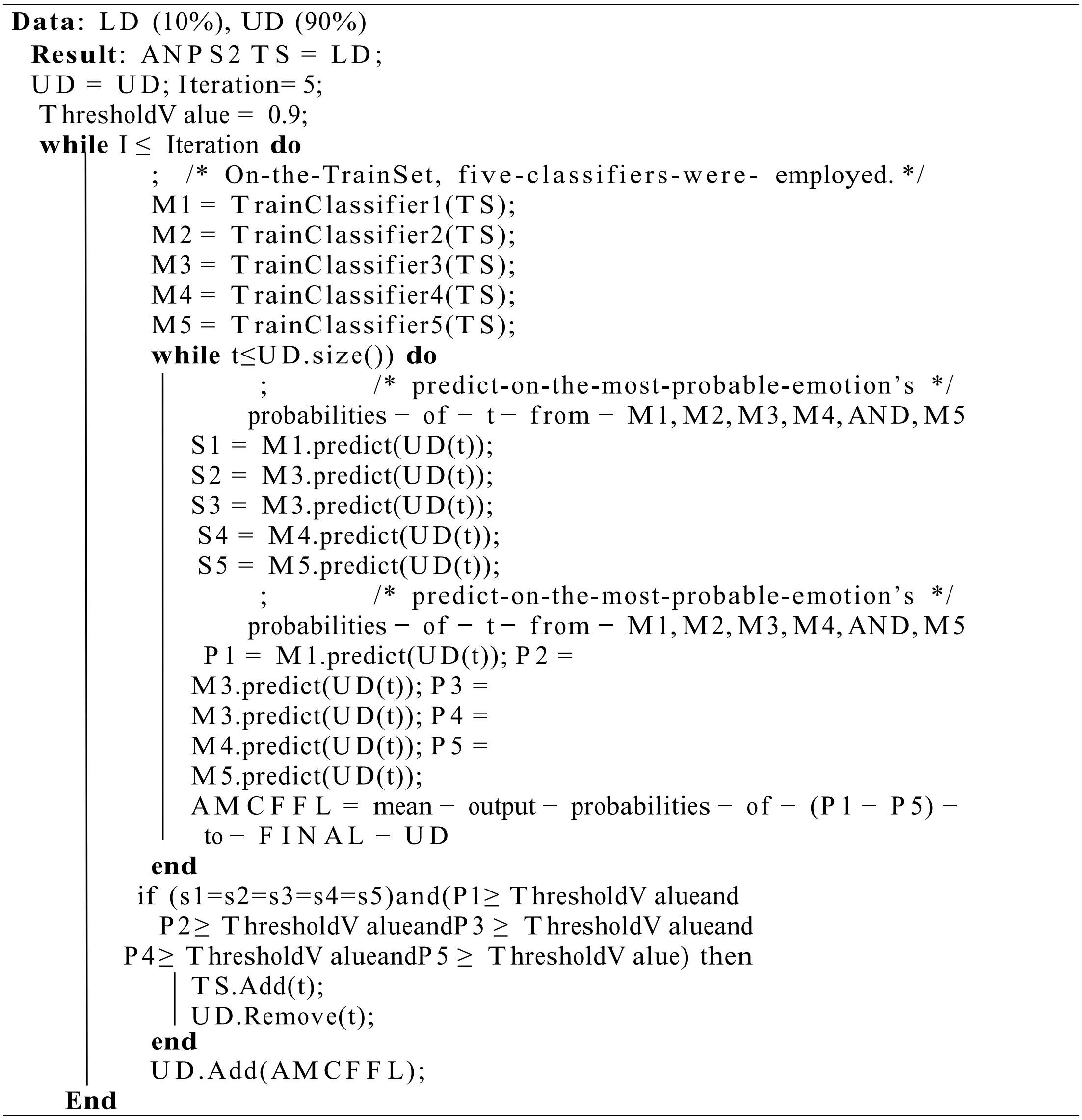
SSL is one of the critical methods in deep learning, which is learning from a limited amount of labeled data and then testing and assessing in large unlabeled datasets. SSL reduces the work and time required for data labeling. A classifier is first learned on a smaller set of labeled items (LD) in the self-training technique, a wrapper for SSL [49].

The unlabeled data collection (UD) is then classified using the newly trained classifier. The most certain examples are included in the labeled set with their expected labels based on the prediction output. Afterward, the classifiers may be re-trained using the newly 'labeled' data. This procedure may be repeated until all (UD) samples are uploaded to (LD) or a specified stopping condition is met.

3.3. Methodology

3.3.1. SSL

Semi-supervising approaches are applied with proposed new techniques, which calculate the automatic mean classifiers for the optimal final labeling (AMCFFL). Semi-supervising approaches implemented with five classifiers. Then, the proposed new techniques (AMCFFL), which are automatic mean classifiers for the most optimal final labeling, were used to determine the final annotation in the unlabeled datasets. The below design[1] represents the proposed algorithm[6] for SSL. From ANP5, 10% was (LD), and 90% was (UD); the final generated corpus named (ANPS2) includes 2118 positives and 7880 negatives, as shown in Fig. 7; afterward, we benchmark (ASA) of data by using the DL models as shown in Table 10.



```
Data: LD (10%), UD (90%)
  Result: ANPS2 TS = LD;
  UD = UD; Iteration= 5;
   ThresholdValue = 0.9;
   while I ≤ Iteration do
           ; /* On-the-TrainSet, five-classifiers-were- employed. */
           M1 = TrainClassifier1(TS);
           M2 = TrainClassifier2(TS);
           M3 = TrainClassifier3(TS);
           M4 = TrainClassifier4(TS);
           M5 = TrainClassifier5(TS);
           while t≤UD.size()) do
                  ;        /* predict-on-the-most-probable-emotion's */
                 probabilities − of − t− from − M1, M2, M3, M4, AND, M5
              S1 = M1.predict(UD(t));
              S2 = M3.predict(UD(t));
              S3 = M3.predict(UD(t));
               S4 = M4.predict(UD(t));
               S5 = M5.predict(UD(t));
                  ;        /* predict-on-the-most-probable-emotion's */
                 probabilities − of − t− from − M1, M2, M3, M4, AND, M5
               P1 = M1.predict(UD(t)); P2 =
              M3.predict(UD(t)); P3 =
              M3.predict(UD(t)); P4 =
              M4.predict(UD(t)); P5 =
              M5.predict(UD(t));
              AMCFFL = mean− output− probabilities − of − (P1 − P5) −
               to− FINAL − UD
           end
          if (s1=s2=s3=s4=s5)and(P1≥ ThresholdValueand
            P2≥ ThresholdValueandP3 ≥ ThresholdValueand
         P4≥ ThresholdValueandP5 ≥ ThresholdValue) then
              TS.Add(t);
              UD.Remove(t);
           end
           UD.Add(AMCFFL);
     End
```

Illustrate the proposed SSL algorithm with the AMCFFL technique.

[6] Model(M),LabeledDATA(LD),UnlabeledData(UD),TrainSet(TS).

Table 4
Datasets for our experiments.

Datasets	POS	NEG	NEU	Total
ANPS2 (2C)	2,118	7,880	–	9,998
ANP5 (5C)	–	–	–	10,000
STD (2C)	2,148	1,350	–	2,000
AAQ (2C)	900	2,148	–	4,296
AJGT (2C)	1,678	900	–	1,800
ArSarcasm (3C)	88,296	3,529	5,340	10,547
BRAD (2C)	3,122,615	68,198	–	156,494
TEAD (3C)	–	2,115,325	378,003	5,605,943

Table 5
Experiment 1: Accuracy of four ML classifiers on ANP5 dataset.

Models	Accuracy (%)	Macro Avg	Weighted Avg
ANP5 Dataset (5C)			
NB	79.00	0.79	0.79
LR	81.00	0.81	0.82
RF	82.00	0.82	0.83
SVM	82.00	0.82	0.82

Table 6
NB classifier: Results in terms of P, R, F, broken down by 5 Arabic platforms.

	Precision	Recall	F1
BBC – Arb	0.73	0.74	0.73
Al – Arabiya	0.86	0.67	0.76
AlHadath	0.80	0.88	0.83
Aljazeera	0.82	0.87	0.84
Skynewarabic	0.74	0.78	0.76

Table 7
LR classifier: Results in terms of P, R, F, broken down by 5 Arabic platforms.

	Precision	Recall	F1
BBC – Arb	0.75	0.77	0.76
Al – Arabiya	0.80	0.76	0.78
AlHadath	0.88	0.89	0.88
Aljazeera	0.89	0.86	0.88
Skynewarabic	0.75	0.79	0.77

Table 8
RF classifier: Results in terms of P, R, F, broken down by 5 Arabic platforms.

	Precision	Recall	F1
BBC – Arb	0.75	0.79	0.77
Al – Arabiya	0.87	0.74	0.80
AlHadath	0.86	0.90	0.88
Aljazeera	0.91	0.87	0.89
Skynewarabic	0.75	0.83	0.79

Table 9
SVM classifier: Results in terms of P, R, F, broken down by 5 Arabic platforms.

	Precision	Recall	F1
BBC – Arb	0.75	0.79	0.77
Al – Arabiya	0.86	0.74	0.80
AlHadath	0.93	0.86	0.90
Aljazeera	0.90	0.85	0.87
Skynewarabic	0.71	0.85	0.77

3.3.2. *Baselines*

3.3.2.1. BERT. The Bidirectional Encoder Representations from Transformers (BERT),[78] language representation approach was unveiled in 2018 by Google researchers Jacob Devlin and his team [50]. It has become one of the essential standard benchmarks for natural language processing technology [51].

3.3.2.2. ARBERT and MARBERT. These models were developed by [52] using a collection of books and news items and the BERT-base. Only 66GB of text from news items were used to train ARBERT. A more extensive dataset (128 GB) with tweets making up half of the dataset was used to train MARBERT. The diversity in MARBERT's training data enables it to handle Arabic dialects more effectively. Free at.[9]

3.3.2.3. AraBERT. Is a model of the Arabic pre-trained language based on the BERT architecture from Google. The BERT-Base configuration is the same for AraBERT [53]. Available at.[10]

3.3.3. *RMuBERT architecture*

A new transformer-based language model for Arabic, called RMuBERT, was designed based on the MARBERT architecture. Additionally, the earlier preprocessing methods [22] extended the by using various cleaning techniques for the Arabic context (AC). RMuBERT starts by choosing the Arabic sentence inputs $A_N = (a_1, a_2, \ldots, a_n)$ and figuring out the context length, which varies from corpus to corpus. The data cleaning process $B_M = (b_1, b_2, \ldots, b_m)$, begins from the input by removing special characters, punctuation marks, and all diacritics. We remove digits, including the date, and apply the elongation and normalization steps with AutoTokenizer pre-trained in Arabic. The stemmer follows the state-of-art concerning the Arabic root by utilizing the Arabic pystemmer to obtain the final processed Arabic context (FPC).

The outcomes of the standard fine-tuned BERT model are improved by developing customized upstream stages, Transformer Layers $T_L = (t_1, t_2, \ldots, t_l)$ as shown in Fig. 8. In addition, the suggested fine-tuning is more appropriate for varied Arabic resource settings with various classes, lengths, and sizes. The proposed hyperparameters $H_P = (h_1, h_2, \ldots, h_p)$, are shown in Table 3. The below equations illustrate proposed RMuBERT primary operations:

$$FPC = AN + BM \tag{1}$$

$$TS = PPRm + SCc1cn + APRz \tag{2}$$

$$RMuBERT = PAC + (HP + TL + OL)BERT + TS \tag{3}$$

$O_L = SG_B$, SM_M, where O_L stands for the output layer, SG_F is a sigmoid function for the binary class task, and SM_F is a softmax function for our multiple class task. T_S represents our three main tasks: PPR_m is platform post recognition, SC_{c1cn} is sentiment classification, and APR_z is Arabic sentiment prediction.

3.4. *Datasets*

The proposed method is trained using the ANPS2, ANP5, STD, AAQ, AJGT, Ar-Sarcasm, BRAD, and TEAD datasets for Arabic context sentiment categorization (see Fig. 9). The outline statistics are shown in Table 4. The following is a description of the datasets:

ANPS2 and ANP5 consist of Arabic news and posts through MSA and dialects. ANPS2 has 2,118 positive and 7,880 negative, free at.[11]

Mohammad et al. [54] offered the Syrian tweets Arabic sentiment analysis (STD) dataset, which was reconstructed based on the NRC-

[7] https://github.com/tensorflow/text/blob/master/docs/tutorials/classify_text_with_bert.ipynb.

[8] https://colab.research.google.com/drive/1PHv-IRLPCtv7oTcIGbsgZHqrB5LPvB7S.

[9] https://github.com/UBC-NLP/marbert.

[10] https://github.com/aub-mind/arabert.

[11] https://github.com/mustafa20999/-ANP5-ANPS2-datasets.

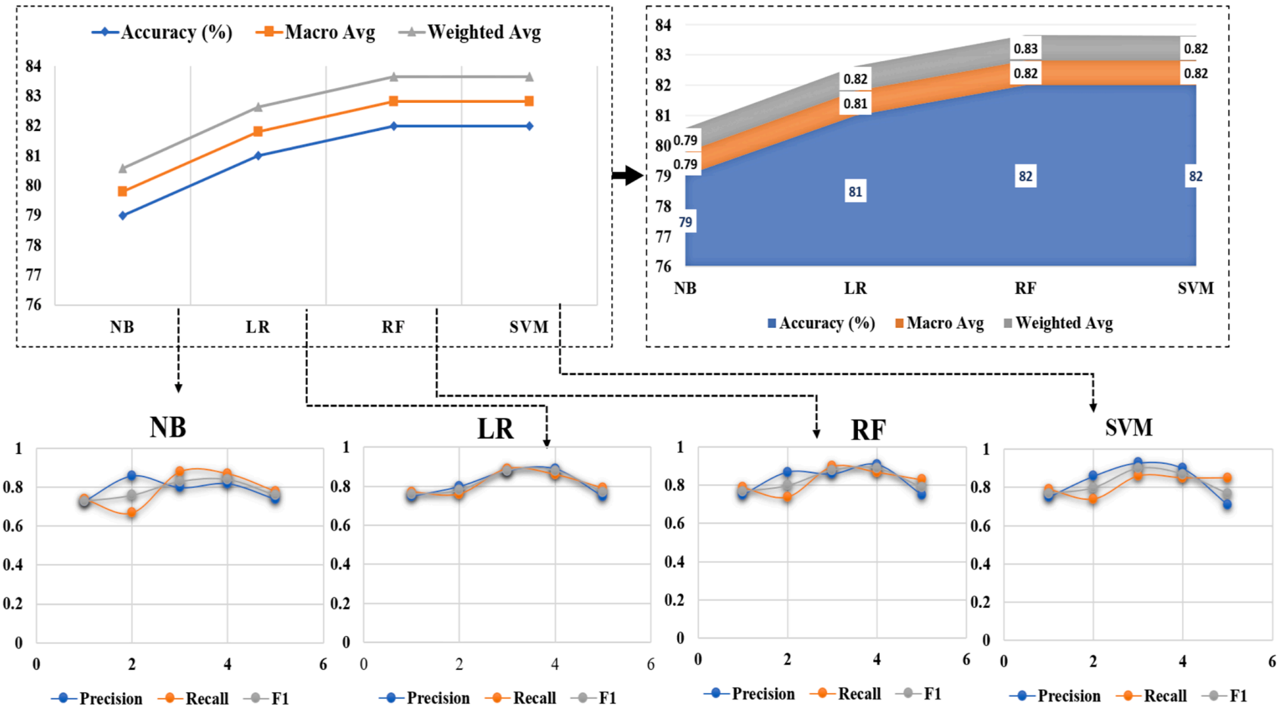


Fig. 10. Efficiency of the ANP5 dataset using four machine learning classifiers (NB, LR, RF, and SVM). P, R, and F accomplishments, split down into five Arabic platforms.

Table 10
Experiment 2: Accuracy of four DL models on ANPS2 dataset.

Models	Accuracy (%)	Weighted Avg
ANPS2 Dataset (2C)		
BIGRU	88.10	0.88
CNN-LSTM	89.30	0.89
LSTM	89.85	0.90
CNN	**90.10**	**0.90**

Canada English; then, the word-sentiment association lexicons were generated and translated. The CMU Twitter NLP tool is used for preprocessing; it contains 2,000 Syrian tweets. In May 2016, Alomari et al. [55] gathered the Arabic Jordanian general tweets (AJGT) corpus, which consists of 900 positive and 900 negative entries, authored in both current standard Arabic and the Jordanian dialect. And AAQ[12] has over 4,000 tweets.

Two human experts manually annotated tweets as favorable or harmful, and a third expert was contacted about the annotations. The Excel plugin ASAP utilities was implemented to cleanse the data. In [56] the Arabic sentiment analysis datasets SemEval 2017 and ASTD were used to build the dataset, including labels for sarcasm and dialect, which contains 10,547 tweets, 1,678 positive tweets, 3,529 negative tweets, and 5,340 neutral tweets. Available at.[13] Sarcasm is given in a CSV file; they offered an 80/20 train/test split to keep things constant in subsequent analysis.

There are 510,600 book reviews in Arabic in the Books Evaluations in Arabic Dataset (BRAD),[14] including balanced and unbalanced reviews totaling more than 156K. Sixty-eight thousand one hundred ninety-eight negatives and 88,296 positives are included. It is compressed in a CSV file with a balanced collection of positive and negative evaluations. Reviews are only included if they are either good (ratings 4 and 5) or negative (ratings 1 and 2).

There are 6 million Arabic tweets in the TEAD[15] dataset [57]; the information was collected from Twitter during the period beginning on June 1st and ending on November 30th, 2017. Emojis and emotion lexicons were employed to accomplish the task of automatically collecting and annotating the information. It has a vocabulary of 602,721 unique things. Neutral tweets 378,003, Subjective positive tweets 3,122,615, and Subjective negative tweets 2,115,325.

3.4.1. Evaluation metrics

The approach's framework was learned and assessed using the following metrics: Four categories of dimensional predictions exist: False Positives (FP) are the number of falsely classified positive classes; False Negatives (FN) are the number of falsely classified negatives; True Negatives (TN) are the number of accurately classified unfavorable classes; and True Positives (TP) are the number of correctly classified positive dimensions. After that, the efficiency metrics shown below are computed [58–65].

$$Precision = TP \div TP + FP \tag{4}$$

$$\mathrm{Recall} = TP + FN \tag{5}$$

$$F1 - score = 2{*}Precision{*}\mathrm{Recall} \div Precision + \mathrm{Recall} \tag{6}$$

$$Accuracy = TP + TN \div TP + FP + FN + TN \tag{7}$$

4. Findings

This section provides five main experiments, the utilized settings, and significant outcomes and impacts. (Experiment 1) this experiment aims first to evaluate the ANP5 dataset using four machine learning

[12] https://github.com/dahouabdelghani/DE-CNN/blob/master/datasets/AAQ.csv.

[13] https://github.com/iabufarha/ArSarcasm.

[14] https://github.com/elnagara/BRAD-Arabic-Dataset.

[15] https://github.com/HSMAabdellaoui/TEAD.

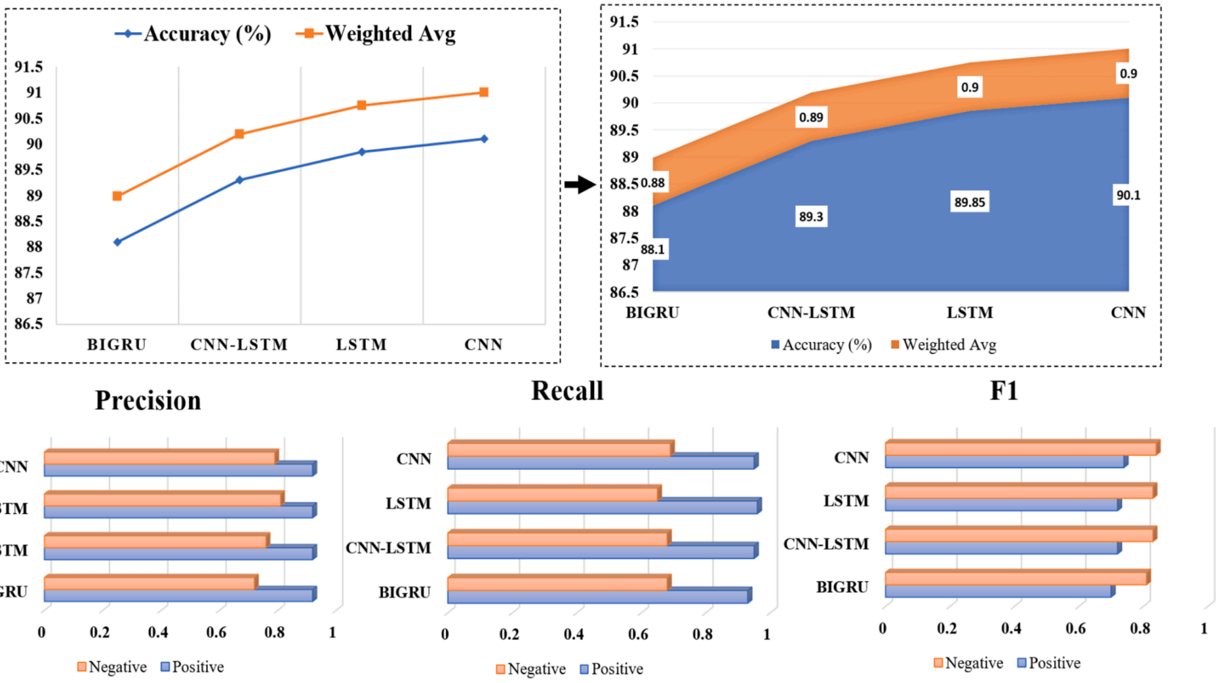


Fig. 11. DL models (BIGRU, CNN-LSTM, LSTM, and CNN) accuracy: Outcomes by Positive and Negative sentiment, broken down by P, R, F, and weighted Average.

Table 11
DL models: Results in terms of P, R, F, Macro Average, broken down by Positive and Negative sentiment.

	Precision		Recall		F1		Macro Avg
	Positive	Negative	Positive	Negative	Positive	Negative	
BIGRU	0.92	0.72	0.93	0.68	0.70	0.81	
CNN-LSTM	0.92	0.76	0.95	0.68	0.72	0.83	
LSTM	0.92	0.81	0.96	0.65	0.72	0.83	
CNN	0.92	0.79	0.95	0.69	0.74	0.84	

Table 12
A confusion matrix of [BIGRU, CNN-LSTM, LSTM, and CNN] on the ANPS2 Dataset.

ANPS2						
BIGRU				**CNN-LSTM**		
Actual	Predicted			Predicted		
		Pos	Neg		Pos	Neg
	Pos	1489	108	Pos	1512	85
	Neg	129	273	Neg	128	274
LSTM				**CNN**		
		Pos	Neg		Pos	Neg
	Pos	1536	61	Pos	1525	72
	Neg	141	261	Neg	125	277

Table 13
Experiment 3: Performance of five transformer models on our collection ANPS2 and ANP5 dataset.

Models	Accuracy (%)			
ANPS2 Dataset (2C)		Time of training	ANP5 Dataset (5C)	Time of training
BERT	78.52	1 h:17 m:43 s	69.04	1 h:20 m:17 s
ARBERT	88.75	1:13:49	87.90	1:10:25
MARBERT	88.89	1:04:57	89.00	1:06:33
AraBERT	89.40	**1:03:42**	89.20	**1:01:27**
RMUBERT	**90.87**	1:14:33	90.33	**1:17:22**

classifiers, which were Naive Bayes (Multinomial) (NB) [66], Logistic Regression (LR) [67], Random Forest (RF) [68], and Support Vector Machine (SVM) [69], accuracies were (79.00%), (81.00%), (82.00%), and (82.00%), respectively (see Table 5). (Experiment 2) aims to initial benchmark classification of Arabic sentiment by using four deep learning models (BIGRU [70], CNN-LSTM [71], LSTM [72], and CNN [73]) on the generated ANPS2 by SSL, as explained in the previous section. The configuration of DL was number-classes equal to two, split (Train/Validation/Test) (80%/10%/10%), and three or five of the kernel size, with various pooling between two and four, batch-size ['64', '128'] applied Adam optimizer with 0.001 learning rate, one hundred of an epoch. The result is shown in Table 11.

(Experiment 3) intends to assess the effectiveness of our RMuBERT compared to other methods and study the performance of the transformer models on our collection of ANPS2 and ANP5 datasets across several classes. We applied BERT and transformers based on the Arabic language (AR-BERT, MARBERT, and AraBERT) on two datasets, ANPS2 (2C) and ANP5(5C). On the ANPS2, accuracy was (78.52%, 88.75%, 88.89%, and 89.40%) respectively. On ANP5, accuracy was (69.04%, 87.90%, 89.00%, and 89.20%), all shown in Table 13.

(Experiment 4) investigate RMuBERT and AraBERT on four Arabic datasets of various lengths, classes, and sizes. As RMuBERT and AraBERT performed well in the prior trials, they chose them to be evaluated on additional Arabic datasets with varying lengths, classes, and sizes. RMuBERT and AraBERT applied on ArSarcasm (3C), STD (2C), AJGT (2C), and AAQ (2C); we used the configuration of RMuBERT in Table 3

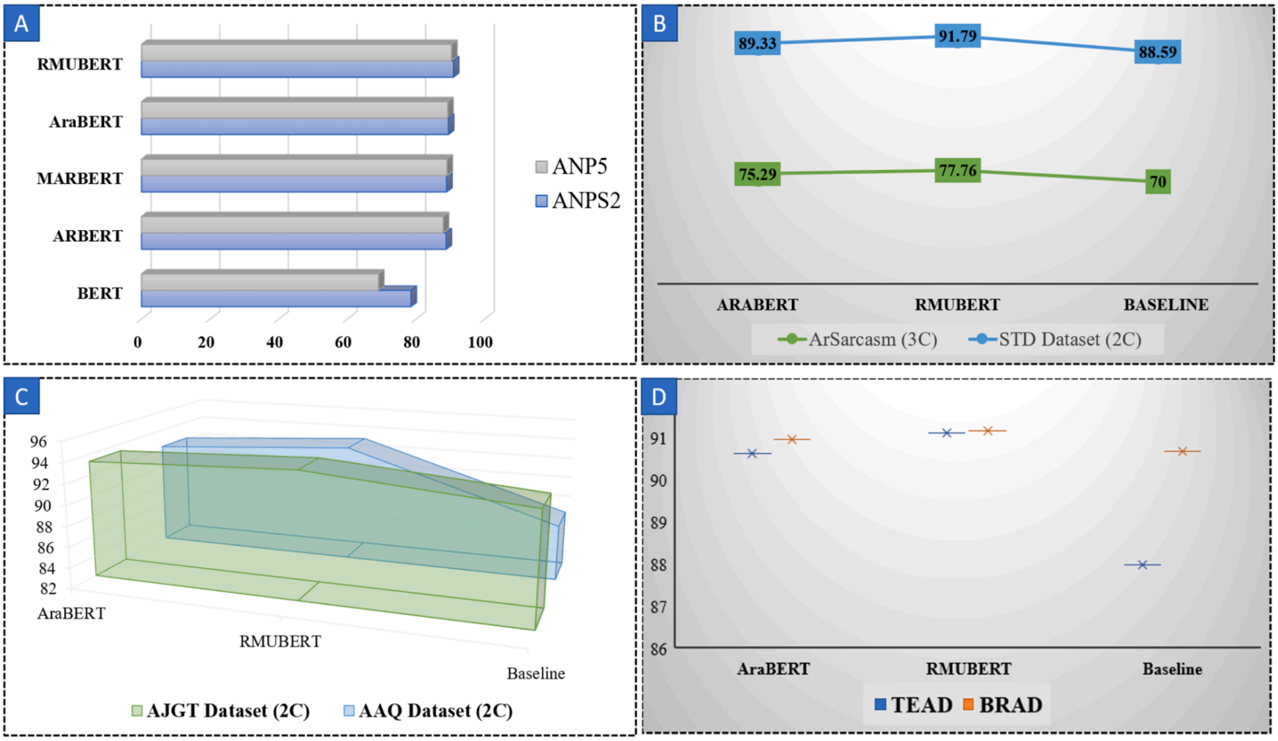


Fig. 12. (A) Five transformer models' performances on the ANPS2 and ANP5 dataset collections. (B) AraBERT and RMuBERT's performance on differnt length Arabic datasets. (C) Comparison of the baselines and the performance of RMuBERT and AraBERT across various Arabic dataset dimensions. (D) Investigations of RMuBERT and AraBERT on big data sets via the TEAD and BRAD datasets.

Table 14
Experiment 4(a): Performance of RMuBERT and AraBERT on four Arabic datasets.

Models	Accuracy (%)			
	ArSarcasm (3C)	Time of training	STD Dataset (2C)	Time of training
AraBERT	75.29	1:01:33	89.33	14:52
RMUBERT	**77.76**	45:12	**91.79**	10:12
Baseline	70.00 [56]	–	88.59 [74]	–

Table 15
Experiment 4(b): Performance of RMuBERT and AraBERT on four Arabic datasets.

Models	Accuracy (%)			
	AJGT Dataset (2C)	Time of training	AAQ Dataset (2C)	Time of training
AraBERT	93.33	5:22	92.24	25:36
RMUBERT	**94.07**	10:12	**93.48**	17:50
Baseline	92.44 [74]	–	87.29 [74]	–

Table 16
Experiment 5: Performance of RMuBERT and AraBERT on TEAD and BRAD datasets.

Models	Accuracy (%)			
	TEAD Dataset	Time of training	BRAD Dataset (2C)	Time of training
AraBERT	90.63	19:20:52	90.96	9:30:42
RMUBERT	**91.12**	17:10:30	**91.18**	8:59:22
Baseline	88.00 [5]	–	90.68 [75]	–

and epoch equal twenty for both models. Outcomes are shown in Tables 14 and 15.

(Experiment 5) testing RMuBERT and AraBERT on the massive Arabic datasets. After seeing that RMuBERT and AraBERT performed well on various datasets, the models tested on a massive Arabic dataset with the longest and largest length. With the exact configuration of RMuBERT in Table 3, the epoch equals ten for both models this time. RMuBERT and AraBERT are used on BRAD, representing the most extended length of Arabic context with 156K items. The largest Arabic sentiment TEAD[16] dataset comprises six million Arabic tweets acquired from social media (see Table 4).

5. Discussion

In (Experiment 1), first, RF and SVM performed best among the other classifiers; ten-fold cross-validation was conducted for all classifiers, and the average performance was provided. As shown in Table 5. Second, RF was the highest; RF might have been trained in a shorter time than NB and LR. It has a high degree of accuracy in its output predictions and operates well even when applied to large datasets, particularly regarding recall and precision. Even without a considerable quantity of data, it maintained its reliability consistently. With a weighted avg up to (83.00%); besides the two classes 'BBCArb' and 'AlArabiya', RF and SVM are equal with F1 (0.77) and (0.80); when it comes to the 'AlHadath' class, RF was low compare to SVM, with F1 (0.88) for RF, and (0.90) for SVM. With the last classes, 'Aljazeera' and 'Skynewarabic', RF performed well than SVM, RF F1 up to (0.89) for 'Aljazeera', and (0.79) for 'Skynewarabic', as shown in Tables 8 and 9. Also, RF F1 was the highest for each class compared to Tables 6 and 7. SVM performs well through distinct boundaries between classes; nevertheless, they are more successful in high-dimensional environments with five classes and

[16] https://github.com/HSMAabdellaoui/TEAD.

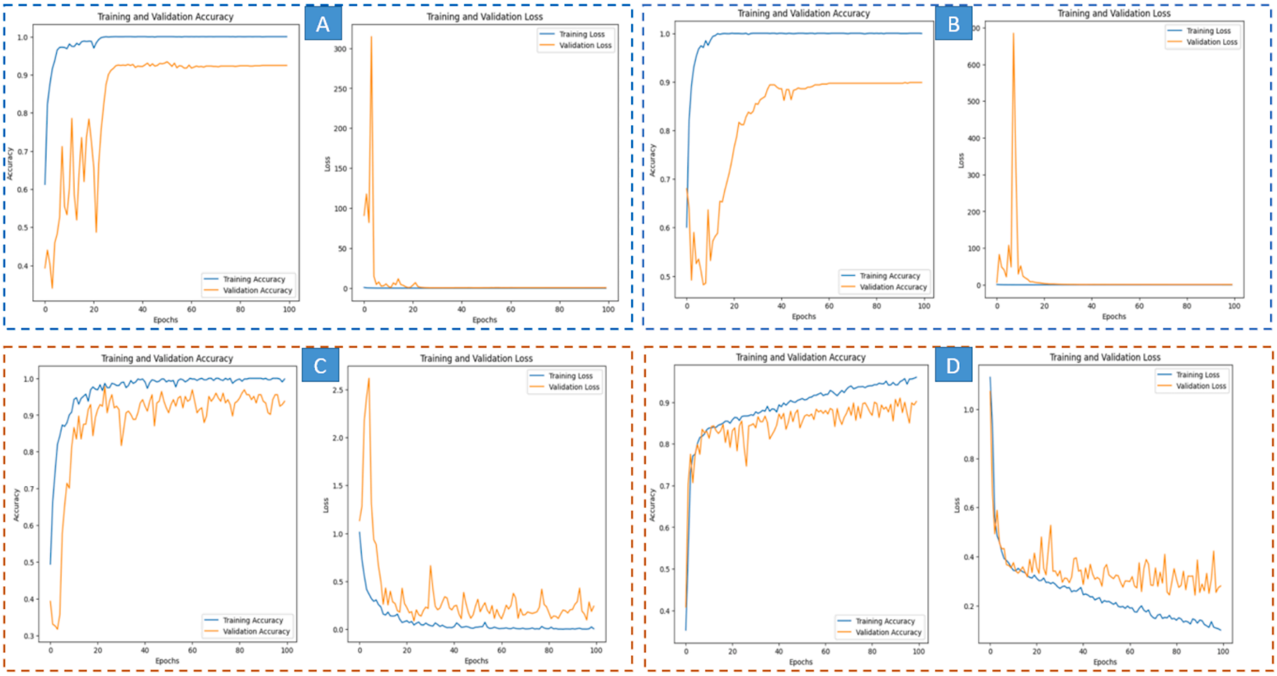


Fig. 13. (A) Displays the accuracy and loss plots of the proposed method NPS2 datasets. (B) Shows the accuracy and loss graphs for the suggested approach ANP5. (C) Displays the accuracy and loss plots of the proposed method TEAD Dataset. (D) Shows the accuracy and loss graphs for the suggested approach BRAD Dataset.

the binary. Compared to other classifiers, SVM performed much better because the total number of dimensions was more than the number of samples implemented. Fig. 10 demonstrates the ANP5 dataset's overall efficiency using four machine learning algorithms.

(Experiment 2) applied BIGRU, CNN-LSTM, LSTM, and CNN on the ANPS2 dataset, and accuracies were 88.10%, 89.30%, 89.85%, and 90.10%, respectively, as shown in Table 10. This experiment included extending the core architectures using the best hyperparameter choices and adding more filters to the improved CNN representation. It generated several feature maps using various window lengths and weights. A number of construction parts and processes were effectively optimized, and these constituted the inputs for the layer's output. Additionally, the appropriate activation function for each job was provided. CNN had the highest accuracy (Fig. 11), with precision for positive (0.92), negative (0.79), recall for positive (0.95), negative (0.69), and F1 positive (0.94), negative (0.74), with (0.84) of macro avg, as shown in Table 11. Also, according to a weighted average of 0.90, CNN outperformed other models in extracting the Arabic features. The evaluation of all is shown in Table 12.

In (Experiment 3), first, for ANPS2, AraBERT performed best compared to the BIGRU, CNN−LSTM baselines in Table 10. For ANP5, AraBERT also was the best compared to the results in Table 5. The fine-tuned AraBERT model performs better than the models in Experiment 2 by a significant margin. One possible explanation for this phenomenon is that the AraBERT models are exclusively pre-trained in Arabic. In addition, in contrast to the BERT, they use a substantial amount of data and vocabulary, which results in a wide variety of words. Second, when we applied the proposed RMuBERT, it was the highest, which enhanced performance on ANPS2 accuracy up to (90.87%) and for ANP5 up to (90.33%). Besides, AraBERT was less model time[17] than the other transformer with ANPS2 (1h, 03m, 42s), and for ANP5 was (1h, 01m, 27s) (Fig. 13A, B). All are shown in Table 13. According to the final evaluation, the transformer models based on the Arabic language performed better and required the least training time with our two datasets. In Addition, RMuBERT shows it can make a difference (1.98%) on two classes and (1.33%) on 3 class tasks, compared to the base of the MARBERT transformer result (Fig. 12A). This indicates the efficiency of RMuBERT in the Arabic context. It helped capture significant features and contextualize representation to fine-tune smaller, task-specific labeled data, ultimately gaining a grasp of the appropriate Arabic context and boosting sentiment analysis.

(Experiment 4) RMuBERT and AraBERT performed better on the four datasets than on the baseline. For ArSarcasm (3C), the accuracy of AraBERT was 75.29%, and RMuBERT was 77.76% compared to 70.00% [56]. For STD (2C), the accuracy of AraBERT was 89.33%, and RMuBERT was 91.79% compared to 88.59% [74]. For AJGT (2C), the accuracy of AraBERT was 93.33%, and RMuBERT was 94.07% compared to 92.44% [74]. For AAQ (2C), the accuracy of AraBERT was 92.24%, and RMuBERT was 93.48% compared to 87.2% [74], as shown in Tables 14 and 15, and (Fig. 12B, C). Although it required less time to train, the RMuBERT achieved maximum performance. This proves its stability via the standard datasets and assists the Arabic sentiment analysis in the various domains. Now looking forward to seeing how RMuBERT works on the massive Arabic sentiment, whether MSA or dialects (see next Experiment 5).

In (Experiment 5), RMuBERT and AraBERT had the best performance compared to the baseline. For BRAD, accuracy for AraBERT was 90:96% with training time (9:30:42), and RMuBERT was 91.18% (8:59:22) compared to 90.68% [75]. For TEAD, accuracy for AraBERT was 90.63% with training time (19:20:52), and RMuBERT was 91:12% (17:10:30) compared to 88.00% [5], (see Table 16, Fig. 12D). RMuBERT scored the highest results (Fig. 13C, D); it can evaluate the Arabic sentiment by using a variety of data subcategories and can function well on enormous Arabic datasets, whether they are comprised of MSA or dialects.

Several limitations during the experiments included Arabic preprocessing context and high efficiency in handling long Arabic sentences. For the preprocessing, we utilized a comprehensive Arabic preprocessing approach, which helped us to tackle the complicated formation of the Arabic language in different stages, whether with sentences, words, or letters.

For instance, remove @date and @time symbols, punctuation marks, diacritics, strip elongation, normalization, and stopwords. Second, the

[17] h:hours,m:minutes,s:seconds.

attention mechanism related to the proposed method helps solve long-sequence training issues and improves efficiency and accuracy.

Finally, RMuBERT design assisted with the following: Firstly, Platform posts recognition. This seeks to identify the opinion's source; it is necessary for official documents, like news stories, since they may have many viewpoints voiced by different opinion leaders, who may also be identified by name. The opinion holders are often the writers of the reviews or postings for social networking apps, blogs, and e-commerce websites, and they may be recognized using their login information. Secondly, Sentiment classification. This determines the subjective texts' sentimental inclinations. It may categorize a text into fine or coarse-grained classifications depending on the intensity of the feelings, such as positive, negative, and neutral. Thirdly, Sentiment prediction.

The RMuBERT achieved high performance on the three previous tasks and effectively handles both MSA and dialects.

6. Limitations and future research

Much information on Arabic news and publications has been gathered from several sites, customizing new approaches and applying advanced techniques for study objectives.

1. Because the events in our daily lives are changing and ongoing, necessitating continuous news. It is necessary to use larger devices and connect them to more platforms. So that requires huge real-time applications that assist in the continuous analysis, prediction, and the relationship between platforms and the link between publications and events.
2. More studies are possible to find out more about the most significant platforms in the Arab area, their locational breakdown, intriguing themes, and the greatest follow-up.
3. Examining the effects of positive and negative, rating the news's quality and engagement level, and determining what humor is real and what is not.
4. Further, the recommended framework in this article may be used in various disciplines, including Arabic translation [76], Chatbots [77], Arabic information retrieval [78], understanding the meaning of intricate music-related writings [79], and Arabic hand-written text recognition [80].

7. Conclusion

Following the presentation of a new collection (ANP5) consisting of Arabic news posts from several Arabic platforms, SSL was utilized with the AMCFFL technique to analyze the Arabic sentiment and generate a second dataset (ANPS2). Machine and deep learning models were then applied to the two datasets to perform initial benchmark categorization, which served as the first baselines. Afterward, a new RMuBERT Model was developed and compared with four transformers on the datasets: ANPS2 accuracy (90.87%) and ANP5 (90.33%). RMuBERT performed better than the baselines. Further tested RMuBERT on various Arabic corpus with different classes, lengths, and sizes; the datasets were ArSarcasm (3C), STD (2C), AJGT (2C), and AAQ (2C) with accuracies of 77.76%, 91.79%, 94.07%, and 93.48%, respectively. Also, RMuBERT performed better than baselines. Finally, the accuracy of the BRAD dataset's maximum Arabic context lengths with MSA was (91.18%). On the largest Arabic sentiment corpora with six million Arabic tweets, RMuBERT works efficiently, performing up to (91.12%) with less training time. The current analysis is based on data from various disciplines, such as economics, sports, tourism, climate, healthcare, and global policies. Institutions, government communities, companies, and academics can benefit from this research and access the resources.

The application of sentiment analysis to Arabic situations that are more particular would be an intriguing endeavor, as would the examination of the influence of preprocessing across different categories of sentiment. Tailored methods that consider the particulars of the context in which sustainability in various fields is being implemented are needed. The stable method can almost always answer the adaptive minimization query function. This is done to protect differential privacy while ensuring computational consistency. This also makes exploring feature extraction for developed frameworks with differential privacy and gaining end-to-end cryptography during the model's training.

CRediT authorship contribution statement

Mustafa Mhamed: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Richard Sutcliffe:** Writing – review & editing, Supervision, Investigation. **Jun Feng:** Writing – review & editing, Supervision, Investigation.

8. Data Availability

The ANP5 and ANPS2 datasets are publicly available.[18]

References

[1] Babu NV, Kanaga E. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Comput Sci 2022;3(1):1–20.
[2] Rocha L, Mourão F, Silveira T, Chaves R, Śa G, Teixeira F, Vieira R, Ferreira R. Saci: Sentiment analysis by collective inspection on social media content. J Web Semant 2015;34:27–39.
[3] Maynard D, Roberts I, Greenwood MA, Rout D, Bontcheva K. A framework for real-time semantic social media analysis. J Web Semant 2017;44:75–88.
[4] Zhang X, Ma Y. An albert-based textcnn-hatt hybrid model enhanced with topic knowledge for sentiment analysis of sudden-onset disasters. Eng Appl Artif Intel 2023;123:106136.
[5] Nassif AB, Elnagar A, Shahin I, Henno S. Deep learning for arabic subjective sentiment analysis: challenges and research opportunities. Appl Soft Comput 2021;98:106836.
[6] Hager Morsy MS, Aljohani N, Shoman M, Abdou S. Automatic speech attribute detection of arabic language. Int J Appl Eng Res 2018;13:5633–9.
[7] Alsayat A, Elmitwally N. A comprehensive study for Arabic sentiment analysis (challenges and applications). Egypt Informat J 2020;21(1):7–12.
[8] Retta EA, Sutcliffe R, Almekhlafi E, Enku YK, Alemu E, Gemechu TD, et al. Kiñit classification in ethiopian chants, azmaris and modern music: a new dataset and cnn benchmark. PLoS One 2023;18(4):e0284560.
[9] Elshakankery K, Ahmed MF. HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. Egypt Informat J 2019;20(3):163–71.
[10] El-Demerdash K, El-Khoribi RA, Shoman MAI, Abdou S. Deep learning based fusion strategies for personality prediction. Egypt Informat J 2022;23(1):47–53.
[11] Chan J-Y-L, Bea KT, Leow SMH, Phoong SW, Cheng WK. State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev 2022:1–32.
[12] Uymaz HA, Metin SK. Vector based sentiment and emotion analysis from text: a survey. Eng Appl Artif Intel 2022;113:104922.
[13] Salama ES, El-Khoribi RA, Shoman ME, Shalaby MAW. EEG-based emotion recognition using 3D convolutional neural networks. Int J Adv Comput Sci Appl 2018;9(8).
[14] Hamzah NA, Dhannoon BN. Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network. Egypt Informat J 2023;24(2):365–73.
[15] Sun B, Song X, Li W, Liu L, Gong G, Zhao Y. A user review data-driven supplier ranking model using aspect-based sentiment anal-ysis and fuzzy theory. Eng Appl Artif Intel 2024;127:107224.
[16] Agüero-Torales MM, Salas JIA, Ĺopez-Herrera AG. Deep learning and multilingual sentiment analysis on social media data: An overview. Appl Soft Comput 2021;107:107373.
[17] Ristoski P, Paulheim H. Semantic web in data mining and knowledge discovery: a comprehensive survey. J Web Semant 2016;36:1–2.
[18] Alharbi A, Taileb M, Kalkatawi M. Deep learning in arabic sentiment analysis: an overview. J Inf Sci 2021;47(1):129–40.
[19] Mhamed M, Sutcliffe R, Sun X, Feng J, Retta EA. Arabic sentiment analysis using gcl-based architectures and a customized regularization function. Eng Sci Technol Int J 2023;43:101433.
[20] Imran AS, Yang R, Kastrati Z, Daudpota SM, Shaikh S. The impact of synthetic text generation for sentiment analysis using GAN based models. Egypt Informat J 2022;23(3):547–57.
[21] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, E. A. Retta, A deep cnn architecture with novel pooling layer applied to two sudanese arabic sentiment datasets, arXiv preprint arXiv:2201.12664 (2022).

[18] https://github.com/mustafa20999/-ANP5-ANPS2-dataset.

[22] Mhamed M, Sutcliffe R, Sun X, Feng J, Almekhlafi E, Retta EA. Improving arabic sentiment analysis using cnn-based architectures and text preprocessing. Comput Intell Neurosci 2021;2021.

[23] Khattak A, Asghar MZ, Ishaq Z, Bangyal WH, Hameed IA. Enhanced concept-level sentiment analysis system with expanded ontological relations for efficient classification of user reviews. Egypt Informat J 2021;22(4):455–71.

[24] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 2014;5(4):1093–113.

[25] Manias G, Kiourtis A, Mavrogiorgou A, Kyriazis D. In: June. Multilingual Sentiment Analysis on Twitter Data towards Enhanced Policy Making. Cham: Springer International Publishing; 2022. p. 325–37.

[26] Alhumoud S, Al Wazrah A, Alhussain L, Alrushud L, Aldosari A, Altammami RN, et al. ASAVACT: Arabic sentiment analysis for vaccine-related COVID-19 tweets using deep learning. PeerJ Comput Sci 2023;9:e1507.

[27] Abualigah L, Alfar HE, Shehab M, Hussein AMA. Sentiment analysis in healthcare: a brief review. Recent Adv NLP: Case Arabic Language 2020:129–41.

[28] Alayba, A., 2019. Twitter Sentiment Analysis on Health Services in Arabic (Doctoral dissertation, Coventry University).

[29] Al-Bohnayyah M, Wahsheh HA. Towards developing medical sentiment analysis model. NeuroQuantology 2022;20(17):238.

[30] Ejjami, R., The Holistic Intelligent Healthcare Theory (HIHT): Integrating AI for Ethical, Transparent, and Human-Centered Healthcare Innovation.

[31] Mavrogiorgos K, Kiourtis A, Mavrogiorgou A, Kleftakis S, Kyriazis D. January. A multi-layer approach for data cleaning in the healthcare domain. In: Proceedings of the 2022 8th International Conference on Computing and Data Engineering; 2022. p. 22–8.

[32] Misra P, Yadav AS. March. Impact of preprocessing methods on healthcare predictions. Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE). 2019.

[33] Yafooz WM, Alsaeedi A. Sentimental analysis on health-related information with improving model performance using machine learning. J Comput Sci 2021;17(2): 112–22.

[34] Geetha M, Renuka DK. Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model, Interna-tional Journal of Intelligent. Networks 2021;2:64–9.

[35] Ba¸sarslan MS, Kayaalp F. Sentiment analysis on social media reviews datasets with deep learning approach. Sakarya Univ J Comput Inform Sci 2021;4(1):35–49.

[36] Li M, Chen L, Zhao J, Li Q. Sentiment analysis of Chinese stock reviews based on bert model. Appl Intell 2021;51(7):5016–24.

[37] S. Xie, J. Cao, Z. Wu, K. Liu, X. Tao, H. Xie, Sentiment analysis of chinese e-commerce reviews based on bert, in: 2020 IEEE 18th Inter-national Conference on Industrial Informatics (INDIN), Vol. 1, IEEE, 2020, pp. 713–718.

[38] Al-Twairesh N. The evolution of language models applied to emotion analysis of Arabic tweets. Information 2021;12(2):84.

[39] Catelli R, Pelosi S, Esposito M. Lexicon-based vs. bert-based senti-ment analysis: a comparative study in Italian. Electronics 2022;11(3):374.

[40] V́asquez J, Gómez-Adorno H, Bel-Enguix G. Bert-based approach for sentiment analysis of spanish reviews from tripadvisor. IberLEF@ SEPLN 2021:165–70.

[41] Markchom T, Liang H, Chen J. Uor-ncl at semeval-2022 task 3: Fine-tuning the bert-based models for validating taxonomic relations, in. In: Pro-Ceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022); 2022. p. 260–5.

[42] Abuzayed A, Al-Khalifa H. Sarcasm and sentiment detection in Arabic tweets using bert-based models and data augmentation. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop; 2021. p. 312–7.

[43] Farha IA, Magdy W. Benchmarking transformer-based language mod-els for arabic sentiment and sarcasm detection, in. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop; 2021. p. 21–31.

[44] T. U. Haque, N. N. Saber, F. M. Shah, Sentiment analysis on large scale amazon product reviews, in: 2018 IEEE international conference on innovative research and development (ICIRD), IEEE, 2018, pp. 1–6.

[45] Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S. Semeval-2018 task 1: Affect in tweets. In: Proceedings of the 12th in-Ternational Workshop on Semantic Evaluation; 2018. p. 1–17.

[46] Koroteev, M.V., 2021. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

[47] Manias G, Mavrogiorgou A, Kiourtis A, Symvoulidis C, Kyriazis D. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Comput Applic 2023; 35(29):21415–31.

[48] Manias, G., Sanguino, M.A., Salmerón, S., Mavrogiorgou, A., Kiourtis, A. and Kyriazis, D., 2023, May. Utilizing an Entity-level Semantic Analysis Approach Towards Enhanced Policy Making. In CS & IT Conference Proceedings (Vol. 13, No. 8). CS & IT Conference Proceedings.

[49] Triguero I, Garćıa S, Herrera F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inform Syst 2015;42(2): 245–84.

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

[51] Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: what we know about how bert works. Trans Assoc Comput Linguist 2020;8:842–66.

[52] Abdul-Mageed M, Elmadany A, et al. Arbert & marbert: Deep bidi-rectional transformers for Arabic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (vol-Ume 1: Long Papers); 2021. p. 7088–105.

[53] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).

[54] Mohammad SM, Salameh M, Kiritchenko S. How translation alters sentiment. J Artif Intell Res 2016;55:95–130.

[55] Alomari KM, ElSherif HM, Shaalan K. Arabic tweets sentimen-tal analysis using machine learning. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer; 2017. p. 602–10.

[56] I. A. Farha, W. Magdy, From arabic sentiment analysis to sarcasm de-tection: The arsarcasm dataset, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 32–39.

[57] Abdellaoui H, Zrigui M. Using tweets and emojis to build tead: an arabic dataset for sentiment analysis. Computación y Sistemas 2018;22(3):777–86.

[58] Retta EA, Almekhlafi E, Sutcliffe R, Mhamed M, Ali H, Feng J. A new Amharic speech emotion dataset and classification benchmark. ACM Trans Asian Low-Resour Lang Inf Process 2023;22(1):1–22.

[59] Mhamed M, Noja JA. Enhancing Arabic sentiment analysis through a hybrid deep learning approach. Int J Educ Cult Soc 2023;8(4):183–9. https://doi.org/10.11648/j.ijecs.20230804.15.

[60] Retta EA, Sutcliffe R, Mahmood J, Berwo MA, Almekhlafi E, Khan SA, et al. Cross-Corpus Multilingual Speech Emotion Recognition: Amharic vs. Other Languages Appl Sci 2023;13(23):12587.

[61] Almekhlafi E, Moeen AM, Zhang E, Wang J, Peng J. A classification benchmark for Arabic alphabet phonemes with diacritics in deep neural networks. Comput Speech Lang 2022;71:101274.

[62] Feng J, Ip HH. A multi-resolution statistical deformable model (MISTO) for soft-tissue organ reconstruction. Pattern Recogn 2009;42(7):1543–58.

[63] Sun X, Li B, Sutcliffe R, Gao Z, Kang W, Feng J. Wse-MF: a weighting-based student exercise matrix factorization model. Pattern Recogn 2023;138:109285.

[64] Mhamed M, Zhang Z, Hua W, Yang L, Huang M, Li X, et al. Apple varieties and growth prediction with time series classification based on deep learning to impact the harvesting decisions. Comput Ind 2025;164:104191.

[65] Sun X, Dong K, Ma L, Sutcliffe R, He F, Chen S, et al. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. Entropy 2019;21(1):37.

[66] Roe C, Lowe M, Williams B, Miller C. Public perception of sars-cov-2 vaccinations on social media: questionnaire and sentiment analy-sis. Int J Environ Res Public Health 2021;18(24):13028.

[67] Hidayat THJ, Ruldeviyani Y, Aditama AR, Madya GR, Nugraha AW, Adisaputra MW. Sentiment analysis of twitter data related to rinca island development using doc2vec and svm and logistic regres-sion as classifier. Procedia Comput Sci 2022;197:660–7.

[68] Munshi A, Arvindhan M, Thirunavukkarasu K. Random forest appli-cation of twitter data sentiment analysis in online social network pre-diction. Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models; 2021. p. 299–313.

[69] Styawati S, Nurkholis A, Aldino AA, Samsugi S, Suryati E, Cahyono RP, et al. International Seminar on Machine Learning, Optimization, and Data Science (ISMODE). IEEE 2021;2022:163–7.

[70] X. Yin, C. Liu, X. Fang, Sentiment analysis based on bigru information enhancement, in: Journal of Physics: Conference Series, Vol. 1748, IOP Publishing, 2021, p. 032054.

[71] Khan L, Amjad A, Afaq KM, Chang H-T. Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. Appl Sci 2022; 12(5):2694.

[72] Huang X, Zhang W, Tang X, Zhang M, Surbiryala J, Iosifidis V, et al. Lstm based sentiment analysis for cryptocurrency pre-diction. In: International Conference on Database Systems for Advanced Applications; 2021. p. 617–21.

[73] Liao S, Wang J, Yu R, Sato K, Cheng Z. Cnn for situations under-standing based on sentiment analysis of twitter data. Procedia Comput Sci 2017;111:376–81.

[74] Dahou A, Elaziz MA, Zhou J, Xiong S. Arabic sentiment classi-fication using convolutional neural network and differential evolution algorithm. Comput Intell Neurosci 2019;2019.

[75] Elnagar A, Lulu L, Einea O. An annotated huge dataset for standard and colloquial Arabic reviews for subjective sentiment analysis. Procedia Comput Sci 2018;142: 182–9.

[76] Faheem MA, Wassif KT, Bayomi H, Abdou SM. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation. Sci Rep 2024;14(1): 2265.

[77] Sutcliffe, R., 2023. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. arXiv preprint arXiv:2401.00609.

[78] Darwish K, Magdy W. Arabic information retrieval. Foundations and Trends®. Inf Retr 2014;7(4):239–342.

[79] Sutcliffe, R.F.E. and Liem, C., 2017. Capturing the meaning of complex texts about music. Proceedings of ISMIR 2017, Suzhou, China.

[80] Salourn SS. In: Arabic Hand-Written Text Recognition. IEEE; 2001 June. p. 106–9.