International Journal of Education, Culture and Society
2023; 8(4): 183-189
http://www.sciencepublishinggroup.com/j/ijecs
doi: 10.11648/j.ijecs.20230804.15
ISSN: 2575-3460 (Print); ISSN: 2575-3363 (Online)

Enhancing Arabic Sentiment Analysis Through a Hybrid Deep Learning Approach

Mustafa Mhamed[1, 2], **Jamal Ali Noja**[3]

[1]School of Information Science and Technology, Northwest University, Xi'an, China

[2]College of Information and Electrical Engineering, China Agricultural University, Beijing, China

[3]College of Agricultural Sciences, Dalanj University, Dilling, Sudan

Email address:

mustafamhamed@stumail.nwu.edu.cn (Mustafa Mhamed), mustafamhamed@cau.edu.cn (Mustafa Mhamed), mustafamhamed2099@gmail.com (Mustafa Mhamed), Jamalnoja@yahoo.com (Jamal Ali Noja)

To cite this article:

Mustafa Mhamed, Jamal Ali Noja. Enhancing Arabic Sentiment Analysis Through a Hybrid Deep Learning Approach. *International Journal of Education, Culture and Society*. Vol. 8, No. 4, 2023, pp. 183-189. doi: 10.11648/j.ijecs.20230804.15

Received: June 15, 2023; **Accepted**: July 26, 2023; **Published**: July 31, 2023

Abstract: Sentiment analysis is a key procedure in many natural language processing systems that extract emotions from textual input. In recent years, Arabic sentiment analysis has become a significant study area. With the growth of social media platforms and data flow, especially in Arabic, substantial difficulties have emerged that call for new strategies to address problems, such as the Arabic language's complicated development and the complexity of the multiple, binary, or massively imbalanced Arabic dataset categorizations. Besides, the system's limitations, whether in online analysis tools, deep or machine learning. This paper proposes a new conjunction method for Arabic sentiment analysis (ASA) called Hybrid Convolution Gate Long (HCGL). This method allows us to extract the best features, handle sequences of different lengths to capture context, address the issue of disappearing error gradients, and improve prediction performance. To match other research works, we conduct studies using a variety of data splits. Furthermore, we pay great attention to Arabic preparation by using all-encompassing procedures that help us address the Arabic language context. The proposed method performs highest in terms of 2-class way efficiency (95.88%), followed by 3-class way performance (95.92%). Additionally, we apply it to the massive Arabic sentiment dataset; it performs well, achieving 88.40% in less time.

Keywords: Deep Learning, Natural Language Processing (NLP), Hybrid Convolution Gate Long (HCGL), Arabic Sentiment Analysis (ASA)

1. Introduction

Sentiment analysis is a branch of psychology that studies people's feelings, thoughts, assessments, and behaviors about events, products, services, news, people, organizations, and problems. Opinion mining, review mining, emotion mining, and subjectivity analysis are used in the literature to describe this topic [1]. It has become a significant research field whose application is visible in various domains such as health, commerce, education, politics, and tourism [2]. Online social media are the data sources for sentiment analysis (SA), the users of which generate an ever-increasing amount of information. Thus, these data types' sources need to be considered under the Big Data approach. Additional issues must be addressed to achieve efficient data storage, access, and processing and ensure reliability [3]. Arabic is a rich, unique language and is one of the United Nation's six official languages. There are three principal forms of this Language, standard, dialect, and classical. Modern Standard Arabic (MSA) is used in formal speeches and publishing, such as magazines and journals. Dialects utilized in informal writing, mainly social media, differ from country to country. Classical Arabic is the Qur'an Language used in recitation and praying [4]. Diverse approaches were applied to the Arabic sentiment analysis (ASA). For Machine Learning (ML), such as Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). Deep Learning (DL), the same as Convolutional Neural Networks (CNN), Long short-term memory (LSTM), and Recurrent Neural Networks (RNN). Here we proposed a new architecture named (HCGL) for (ASA); they can be used for both dual and multiple classifications and employ varied window lengths and weights depending on how many feature maps need to be constructed, running faster, and tackling long Arabic contexts accurately.

The following are the work's key contributions:

We develop a domain-neutral Arabic Sentiment Analysis (ASA) model. We put the suggested method to the test using assessments from several domains with various word associations, and we evaluate the efficacy of each dataset separately.

We propose a hybrid Convolution Gate Long method (HCGL), a convolutional neural network with the gated recurrent unit through Long-short-term-memory. First, extract the best features; second, utilize the gated recurrent unit to transfer information faster; and then, use long-short-term memory to regulate the flow of information and transfer appropriate data down the long sequence chain to make predictions with different output sizes.

Our hyperparameter selections, tuning optimizations, and preprocessing tasks assist in boosting HCGL performance.

The proposed method outperforms conventional baselines in Arabic classification.

Our method is highly efficient, allowing them to be applied to massive datasets.

The rest of this paper is organized as follows. Section 2 previous work. Section 3 describes the processing method and proposed architecture. Section 4 experiments. And Section 5 conclusion and future work.

Table 1. *DL previous works on Arabic sentiment analysis.*

Paper	Datasets	Model	Result
[5]	AraSenTi	LSTM-RNN	93.5%
[6]	LABR	CNN	91.9%
[7]	TSAC	Deep-LSTM	90.00%
[8]	Twitter Tweets	Bi-GRU	78.71%
[9]	GS	HILATS	68.09%
[10]	ArTwitter	BiLSTM	91.64%
[11]	MSAC	SVM	83.45%
[12]	DT	SVM	82.1%
[13]	LB	LR	88.00%
[14]	YouTube text	SVM+LR	77.00%

2. Previous Work

Several deep learning models were trained on the Arabic sentiment datasets; Table 1 depicts the most significant previous deep learning works in Arabic sentiment.

We'll go through these works individually, beginning with the ones that use machine learning. Following that, we'll talk about neural network techniques.

Tabii et al. [11] generalizations stacking are implemented based on specific algorithms with different settings and contrasted with the majority vote. And applied Naïve Bayes [15], Maximum Entropy [16], and SVM [17] on two datasets Moroccan Sentiment Analysis Corpus (MSAC) and SemEval-2017. Performance results of ensemble classifiers show that the majority voting rule achieved the best score (83.45%). Atom and Nouman [12] have used SVM and NB on Jordanian dialect tweets (JDT). Results show the SVM was the highest accuracy in all bigrams and stemmed unigrams cases. SVM bigrams = 74.00% and SVM unigrams = 82.1%. Al Omari et al. [13] utilized Logistic regressions on the Lebanon dialect, collected from Google and Zomato, which performed 88.00%. Yafooz et al. [14] compared their Ensemble model, which consists of SVM, followed by Linear Regression, on Arabic text collected from YouTube comments. The ensemble was the best of the other classifiers' accuracy (77.00%).

Now we'll discuss deep learning methods to analyze Arabic sentiment.

Alwehaibi and Roy [5] conjunction LSTM-RNN on the AraSenTi datasets, which include tweets written in the Saudi dialect and the MSA. The model achieves a reasonable accuracy of up to 93.5%. Barhoumi et al. [6] extracted the features by CNN on the LABR dataset. The result shows that the accuracy reaches 91.9%, higher than the best previously published one (91.5%). Abdulla et al. [7] implemented LSTM, RNN, bidirectional-LSTM, and Deep-LSTM on a Tunisian Sentiment Analysis Corpus (TSAC). Deep-LSTM achieves the highest accuracy, reaching 90.00%. Al-Azani and El-Alfy [8] used machine and deep learning methods on the Twitter Tweets dataset, collected from Tweets and YouTube comments. The bidirectional Gated Recurrent Unit was the best with 78.71% accuracy. Elshakankery and Ahmed [9] combine an incremental learning approach on Mini Arabic Sentiment Tweets Dataset (MASTD). For the 3-class, accuracy was 73.67%, and for the 2-class, 83.73%. Ultimately, Elfaik et al. [10] applied RF, SVM, and Bi-LSTM on ASTD, ArTwitter, LABR, MPQA, Multi-Domain, and Main-AHS datasets. The result shows the Bi-LSTM gave the best performance with an accuracy of 76.83%, 91.64%, 79.79%, 74.74%, 91.89%, and 86.03%, respectively.

Now we summarise architectures on the Arabic sentiment analysis. For the ML, Tabii et al. [11] Atoum and Nouman [12] use SVM, Al Omari et al. [13] use Logistic regressions, and Yafooz et al. [14] apply SVM+LR. SVM was the most algorithms applied, but LR had the highest accuracy.

For the DL, Alwehaibi and Roy [5], Abdulla et al. [7], Al-Azani and El-Alfy [8], Elshakankery and Ahmed [9], and Elfaik et al. [10] applied LSTM-RNN, Deep-LSTM, Bi-GRU, HILATS, and BiLSTM, respectively. Barhoumi et al. [6] used CNN, and BiLSTM was the highest performance. Our proposed approach (see next section) is based on CNN, with GRU through LSTM on binary and multiple classifications.

3. Proposed Method

3.1. Preprocessing Steps

Natural language text data is an erratic and ambiguous. Pre-treatment of text involves placing the content in a clean, organized format with a dardized structure to make it suitable for further analysis and training.

Preprocessing techniques were comparable to those we had previously devised [18].

Applied Tokenizer from Keras [19].

Removed all diacritics, special letters, and punctuation marks.

Removed all numerals and dates.

Removed all repetitive characters.

Removed any non-Arabic characters.

3.2. Hcgl Design

Figure 1 shows the main components of the model, starting from the input and the embedding matrix, then the convolutional layers, gated recurrent unit layers, and long short-term memory layer.

3.3. Input Layer

Let's assume that the input layer gets text data in Y (y1, y2... yn), where y1, y2... yn denotes the number of words with each input term's dimension m. The dimensional space of Rm would then be defined for each word vector. As a result, the input text dimension space will be Rm×n.

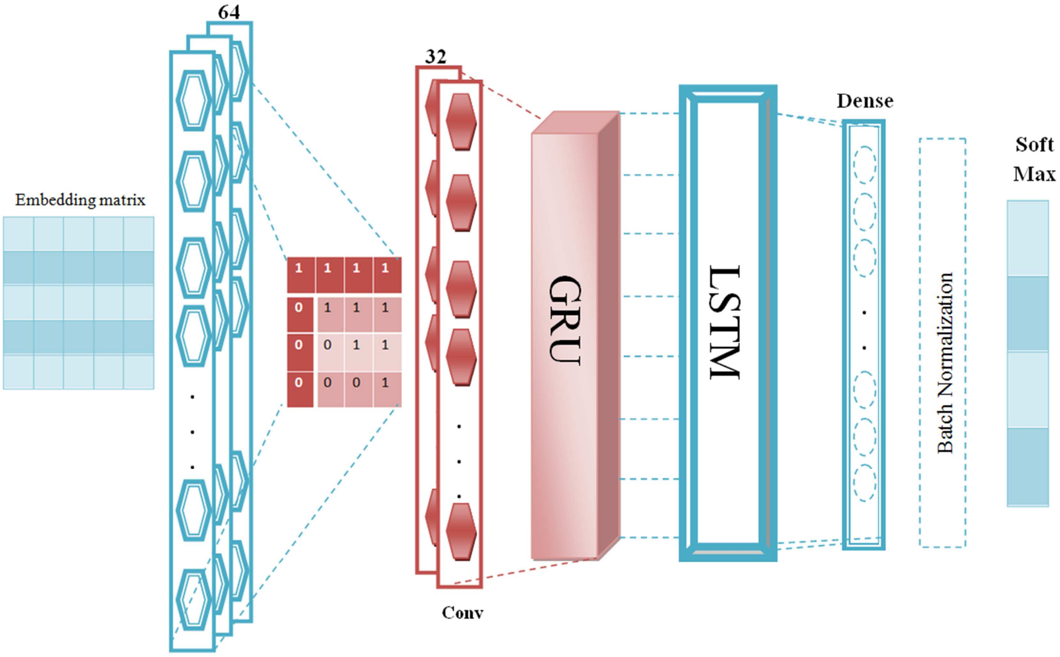

Figure 1. *Shows the HCGL architecture model.*

3.4. Word Embedding Layer

Syntactic and semantic analysis is utilized in all-natural language processing tasks to separate human Language into machine-readable components. The process of syntactic analysis, often referred to as grammatical analysis or simply syntactic analysis, establishes the grammatical structure of a text and the connections across words that appear in a parse tree. The objective of semantic analysis is to interpret Language. However, semantics is one of the most challenging topics in NLP due to Language's polysemy and ambiguity. Semantic tasks attempt to decipher word meanings and the subject matter of the text and examine sentence structure and the relationships between related words and ideas. Let's assume the vocabulary size for a text representation is d, and we will do some word embedding. Thus, the dimensions term embedding matrix will be represented as Am×d. The input text Y (yI), where I = 1, 2, 3,..., n, Y ε Rm×n, is now transferred from the input layer to the embedding layer, resulting in the text's term embedding vector. On the modern standard Arabic (MSA) from the Twitter text, we apply the AraVec [20] word embedding pre-trained by Word2vec [21] for Word representations.

3.5. Hcgl Architecture

contain from embedding layers, which include several unique words, with embedding sizes equal to 128 or 300 and max lengths 150, 50, 30; after that, convolutional neural network layers with 64 filters, kernel size similar to three, padding amounting to 'valid', activation equal ReLU, and strides equal one; followed by [global average pooling 1D, global max pooling 1D], pool size equal two, then followed by an additional layer of convolution with 32 filters, kernel size equal three, padding tantamount to 'valid', activation equal ReLU, and strides equal one. Then we applied the regularization technique on the previous two layers and the ReLU activation function. This helped us to reduce model capacity while maintaining training accuracy. After that, Dropout (0.25) and implement gated recurrent unit layer, with 128 units, again Dropout (0.25) and long-short term memory with output [90, 70, 40], and add the Flatten then batch normalization, and finally a softmax layer, i.e., a fully connected layer to predict the output of the class among three sentiment classes: Negative, Neutral, Positive, or binary classes.

Table 2. *For our experiments, we used the following datasets.*

Datasets	POS	NEG	NEU	Total
ArTwitter (2C)	1,000	1,000	-	2,000
MASC (2C)	4,476	2,257	-	6,733
LABR (2C)	6,580	6,578	-	13,158
AraSenTi (3C)	4,643	7,840	7,279	19,762
GS (3C)	559	1,232	2,400	4,191

4. Experiment

Our experiments include four aspects:

1. Preprocessing the datasets and checking the steps.
2. Utilizing the suggested approach.
3. Result analysis.

4.1. Datasets

For Arabic text sentiment classification, our model is trained on the Twitter dataset for Arabic Sentiment Analysis (ArTwitter) [7], Multi-domain Arabic Sentiment Corpus (MASC) [22], Large-scale Arabic Book Review (LABR) [23], AraSenTi [24], and Arabic Gold Standard Twitter Data for Sentiment Analysis (GS) [25]. Table 2 shows the details of the dataset, where POS represents positive tweets, NEG represents negative tweets, and NEU represents neutral tweets.

ArTwitter (2C) consists of 2,000 Arabic tweets with two classes (1,000 POS and 1,000 NEG). MASC (2C) comprises 2 classes, 4,476 POS, and 2,257 NEG. LABR (2C) is made up of two categories, 6,580 POS and 6,578 NEG. AraSenTi (3C) consists of three classes, 4,643 POS, 7,840 NEG, and 7,279 NEU. GS (3C) consists of three classes, 559 POS, 1,232 NEG, and 2,400 NEU.

Elnagar et al. [26] offered an extensive dataset with over 370,000 MSA reviews and some dialectical material in Gulf languages. The balanced subset has 94,052 reviews, 46,968 positive reviews, and 47,084 bad reviews; the unbalanced subset contains 94,052 reviews, 46,968 good reviews, and 47,084 negative reviews. Each rating in the dataset is classed as negative, positive, or neutral based on its ranking.

4.2. Experimental Settings

Our tuning and hyperparameter settings were used. The experiment settings are:

Embedding size amounting to [100, 128], Pooling [2, 4, 6], Batch-size [64, 128, 164], Kernel-size [3, 5], Number-classes [2, 3], Epoch [10, 50, 100], with Adam optimizer, and 0.001 Learning Rate. For the implementation, we used the TensorFlow framework.

Table 3. *Show accuracy with binary datasets. Bottom line shows the highest previous accuracies.*

Datasets	Model	Accuracy	F1
LABR	[6]	91.9%	-
	HCGL	95.59%	-
MASC	[11]	83.45%	-
	HCGL	88.76%	87.50
ArTwitter	[10]	91.82%	92.39
	HCGL	95.88%	95.8

Table 4. *Show accuracy with multiclass datasets. Bottom line shows highest previous accuracies.*

Datasets	Model	Accuracy	F1
AraSenTi	[5]	93.5%	-
	HCGL	95.92%	95.92%
GS	[9]	68.09%	58.29
	HCGL	72.33%	-

Table 5. *Shows the accuracy (%) of the models on HARD (2C).*

Model	Accuracy
CNN	85.12
RNN	86.70
GRU	84.07
BI-GRU	84.83
HCGL	88.40

4.3. Experiment 1: Two-Way Sentiment Classification

In this stage, we apply the proposed method (HCGL) on the binary datasets; the data was split into 80/10/10 train/validation/test, with 10-fold cross-validation. Firstly, for the big binary sentiment of the LABR (2C) Datasets, the HCGL shows the highest accuracy reached 95.59%, compared to the baseline of 91.9% [6]. Secondly, for the medium MASC (2C) Datasets, the HCGL shows the best accuracy, up to 88.76%, compared to the baseline of 83.45% [11]. Third, for the ArTwitter (2C) Datasets, HCGL was the highest, accuracy 95.88% compared to the baseline of 91.82% [10], as shown in Table 3.

4.4. Experiment 2: Three-Way Sentiment Classification

The goal was to test HCGL again on the three-way dataset, AraSenTi, and GS. 3-way classification is more difficult than 2-way classification, mainly because the Neutral class might contain instances with both positive and negative aspects, which can cause the model to become confused.

The configuration of HCGL was the same as in Experiment 1, except that there were three outputs, not two. Once again, ten-fold cross-validation was used for all models. The results are shown in Table 4.

HCGL was the best performance, for AraSenTi accuracy reached 95.92%, compared to the baseline of 93.5% [5]. GS accuracy is up to 72.33%, compared to the baseline of 68.09%[9].

4.5. Experiment 3

The aim of this experiment, therefore, was to compare HCGL with four models (CNN [27], RNN, GRU, and BI-GRU [28]) on the large Arabic datasets HARD [26].

The average performance was calculated using ten-fold cross-validation. Table 5 shows the results. For CNN, accuracy was 85.12%, RNN was 86.70%, GRU was 84.07%, BI-GRU was 84.83%, and HCGL was up to 88.40%, which is the highest performance. And also different times for the predictions; for the HCGL model, the elapsed time is 3h 55m 18s (3 hours, 55 minutes, and 18 seconds); for RNN, it was 4h 30m 19s, for GRU, it was 3h 58m 57s, For BI-GRU was 4h 5m and for CNN was 4h

20m 35s. Thus HCGL gives us the best validation accuracy and the least execution time.

4.6. Accuracy During Training

Figure 2 shows the accuracy and validation accuracy of the proposed method on LABR, MASC, and ArTwitter datasets. After 100 epochs, the HCGL model offers the highest performance, reaching 95.59%, 88.7%, and 95.88%, respectively. Figure 3 shows the same information for the AraSenTi and GS datasets (HCGL reaches 95.92% and 72.33%).

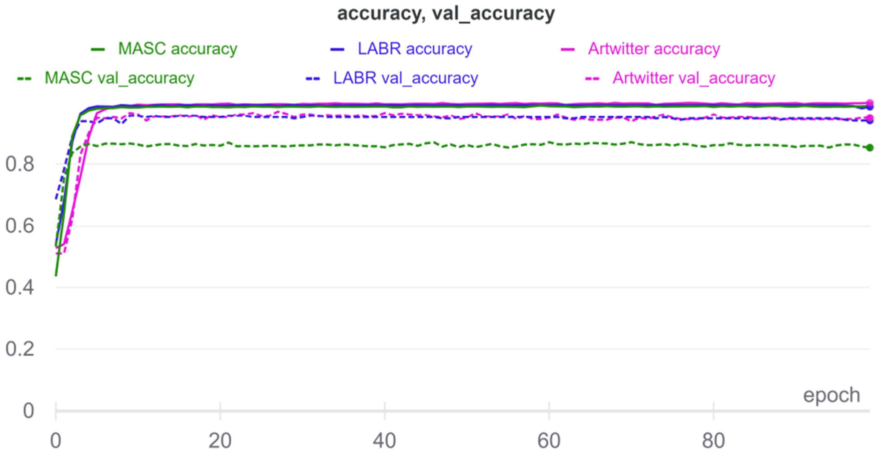

Figure 2. *Shows accuracy and validation accuracy for LABR, MASC, ArTwitter datasets.*

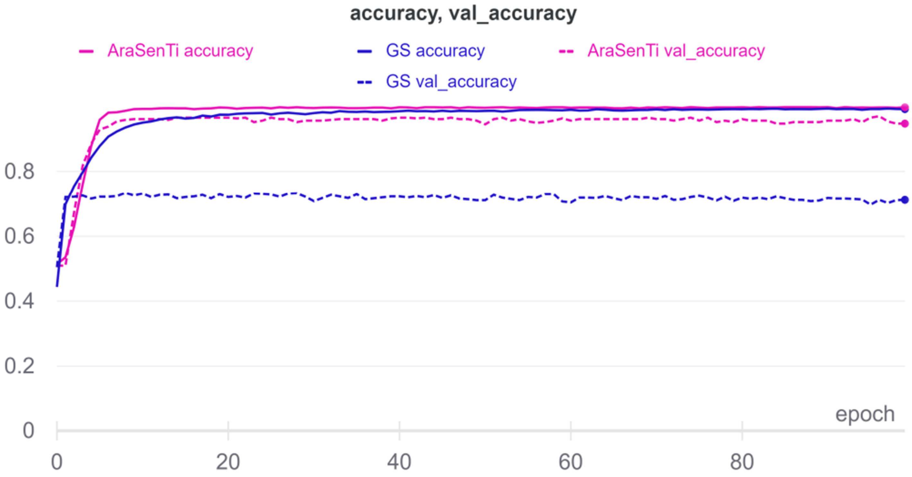

Figure 3. *Shows accuracy and validation accuracy for AraSenTi and GS datasets.*

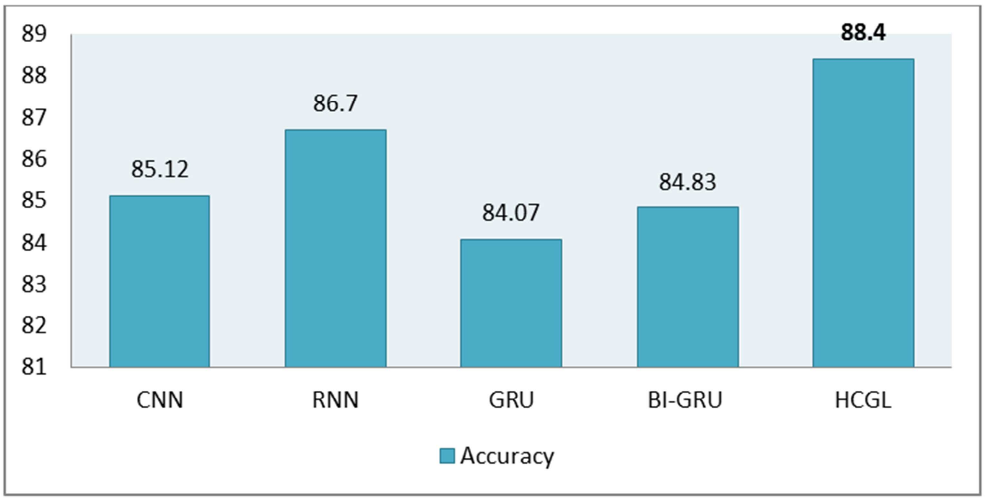

Figure 4. *Displays the HARD (2C) models' performance (%).*

5. Conclusion and Future Work

In this paper, we reviewed significant works related to Arabic sentiment analysis and then proposed a new architecture called (HCGL) on 2-class and 3-class Arabic sentiment datasets. Using this approach, we solved the problem of fading error gradients, extracted the best features, handled sequences of various lengths to capture context, and enhanced classification performance. The highest accuracy was (95.92%). Our results exceed current baselines. On 2-class, these can make a significant difference up to 3.69%, 5.31%, and 4.06%, respectively. On the 3-class, the optimum enhancements are 2.42% and 4.24%. Furthermore, it showed high efficiency in the huge Arabic sentiment with less consuming time.

Future studies will test the suggested strategies utilizing a range of historical data, including news articles, assessments of restaurants, technical evaluations, and archives from several languages.

References

[1] Alharbi, A., Kalkatawi, M., Taileb, M.: Arabic sentiment analysis using deep learning and ensemble methods. Arabian Journal for Science and Engineering, 1–11 (2021).

[2] Boudad, N., Faizi, R., Thami, R. O. H., Chiheb, R.: Sentiment analysis in arabic: A review of the literature. Ain Shams Engineering Journal 9 (4), 2479–2490 (2018).

[3] Dang, N. C., Moreno-García, M. N., De la Prieta, F.: Sentiment analysis based on deep learning: A comparative study. Electronics 9 (3), 483 (2020).

[4] Alrefai, M., Faris, H., Aljarah, I.: Sentiment analysis for arabic lan guage: A brief survey of approaches and techniques. arXiv preprint arXiv: 1809.02782 (2018).

[5] Alwehaibi, A., Roy, K.: Comparison of pre-trained word vectors for arabic text classification using deep learning approach. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1471–1474 (2018).

[6] Barhoumi, A., Camelin, N., Aloulou, C., Estève, Y., Belguith, L. H.: An empirical evaluation of arabic-specific embeddings for sentiment analysis. In: International Conference on Arabic Language Processing, pp. 34–48 (2019).

[7] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M.: Arabic sentiment analysis: Lexicon-based and corpus-based. In: 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pp. 1–6 (2013).

[8] Al-Azani, S., El-Alfy, E.-S.: Emojis-based sentiment classification of arabic microblogs using deep recurrent neural networks. In: 2018 International Conference on Computing Sciences and Engineering (ICCSE), pp. 1–6 (2018).

[9] Elshakankery, K., Ahmed, M. F.: Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis. Egyptian Informatics Journal 20 (3), 163–171 (2019).

[10] Elfaik, H., et al.: Deep bidirectional lstm network learning-based sentiment analysis for arabic text. Journal of Intelligent Systems 30 (1), 395–412 (2021).

[11] Tabii, Y., Lazaar, M., Al Achhab, M., Enneya, N.: Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers.

[12] Atoum, J. O., Nouman, M.: Sentiment analysis of arabic jordanian dialect tweets. Int. J. Adv. Comput. Sci. Appl 10 (2), 256–262 (2019).

[13] Al Omari, M., Al-Hajj, M., Hammami, N., Sabra, A.: Sentiment classifier: Logistic regression for arabic services' reviews in lebanon. In: 2019 International Conference on Computer and Information Sciences (iccis), pp. 1–5 (2019).

[14] Yafooz, W. M., Hizam, E., Alromema, W.: Arabic sentiment analysis on chewing khat leaves using machine learning and ensemble methods. Engineering, Technology & Applied Science Research 11 (2), 6845–6848 (2021).

[15] Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., Gani, A.: Hadith data mining and classification: a comparative analysis. Artificial Intelligence Review 46 (1), 113–128 (2016).

[16] El-Halees, A. M.: Arabic text classification using maximum entropy. IUG Journal of Natural Studies 15 (1) (2015).

[17] Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications 36 (3), 6527–6535 (2009).

[18] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, E. A. Retta, Improving arabic sentiment analysis using cnn-based architectures and text preprocessing, Computational Intelligence and Neuroscience 2021 (2021).

[19] Hammad, M., Al-awadi, M.: Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning, pp. 131–139. Springer, Information technology: new generations (2016).

[20] Soliman, A. B., Eissa, K., El-Beltagy, S. R.: Aravec: A set of arabic word embedding models for use in arabic nlp. Procedia Computer Science 117, 256–265 (2017).

[21] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013).

[22] Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., Abdullah, S.: Arabic sentilexicon: Constructing publicly available language resources for arabic sentiment analysis. Journal of information science 44 (3), 345–362 (2018).

[23] Aly, M., Atiya, A.: Labr: A large scale arabic book reviews dataset. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 494–498 (2013).

[24] Elmadany, A., Mubarak, H., Magdy, W.: Arsas: An arabic speech-act and sentiment corpus of tweets. OSACT 3, 20 (2018).

[25] N. Habash, O. Rambow, G. A. Kiraz, Morphological analysis and generation for arabic dialects, in: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, 2005, pp. 17–24.

[26] Elnagar, A., Khalifa, Y. S., Einea, A.: Hotel arabic-reviews dataset construction for sentiment analysis applications, 35–52 (2018).

[27] Kim, Y.: Convolutional neural networks for sentence classification. arxiv 2014. arXiv preprint arXiv: 1408. 5882 (2019).

[28] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078 (2014).