Arabic Sentiment Analysis with Noisy Deep Explainable Model

Md Atabuzzaman*
Virginia Tech, Virginia
USA

Md Shajalal*
Fraunhofer FIT & University of Siegen
Germany

Maksuda Bilkis Baby
UITS, Dhaka
Bangladesh

Alexander Boden
Fraunhofer FIT & HBRS, Sankt Augustine
Germany

ABSTRACT

Sentiment Analysis (SA) is an essential task for numerous real-world applications. However, the majority of SA research focuses on high-resource languages such as English and Chinese, while limited-resource languages like Arabic and Bengali receive less attention. Additionally, existing Arabic sentiment analysis methods based on advanced artificial intelligence (AI) approaches tend to operate as black boxes, making it challenging to comprehend the reasoning behind their predictions. This paper proposes an explainable sentiment classification framework for the Arabic language. We introduce a noise layer to different deep learning (DL) models, including BiLSTM and CNN-BiLSTM, to address the issue of over-fitting. The proposed framework enables the explanation of specific predictions by training a local surrogate explainable model, shedding light on the reasons behind each sentiment prediction (positive or negative). Experiments were conducted on publicly available benchmark Arabic SA datasets, and the results demonstrated that the inclusion of noise layers in the DL model improves performance for the Arabic language by mitigating overfitting. Our method also outperformed several state-of-the-art approaches. Moreover, the introduction of explainability with the noise layer enhances transparency and accountability, making the model suitable for practical adoption in AI-enabled systems.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Neural networks**.

KEYWORDS

Arabic Sentiment Analysis, Noise layer, BiLSTM, CNN, Explainable AI (XAI), Interpretability

ACM Reference Format:
Md Atabuzzaman, Md Shajalal, Maksuda Bilkis Baby, and Alexander Boden. 2023. Arabic Sentiment Analysis with Noisy Deep Explainable Model. In *2023 7th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2023), December 15–17, 2023, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3639233.3639241

*Both authors contributed equally

NLPIR 2023, December 15–17, 2023, Seoul, Republic of Korea

ACM ISBN 979-8-4007-0922-7/23/12.
https://doi.org/10.1145/3639233.3639241

1 INTRODUCTION

Online social media platforms have become increasingly popular, leading to the emergence of various fields dedicated to analyzing the platforms and their content to extract useful information for individuals [29]. SA is one of them. It is a branch of Natural Language Processing (NLP) which is concerned with identifying the feelings expressed in texts. SA may indeed be used to investigate this enormous volume of raw text data to provide a concise summary of what the public believes about a specific topic or a product, or even about any opinion [2].

Prior research on sentiment analysis mostly focused on high-resourced languages. Since Arabic is not a high-resourced language, still there is a lack of attention as compared to other high-resourced languages. The prior methods for Arabic Sentiment Analysis (ASA) depended on sentiment lexicons like ArSenL [8], a large-scale MSA word lexicon. Various options for analyzing Arabic-specific data were examined using recurrent and recursive neural networks [3, 9]. Dahou et al. [11] trained CNN with the semantic representation using word-embedding for ASA. Farha et al. [17] proposed a hybrid model for ASA by employing LSTMs for sequence and context interpretation and CNNs for feature extraction. Then a BERT-based model for Arabic language representation, AraBERT was proposed for many Arabic language-specific tasks including ASA [7].

However, the proposed models work like a black-box, even the developers and AI practitioners do not exactly understand what are the causes for a specific prediction (positive or negative). This lack of transparency is a drawback for an efficient ASA system for adoption in real-world applications. Though some works can be found with explanations for SA prediction using XAI tool-kits for rich languages such as English [22], to the best of our knowledge, there is no single work of Arabic sentiment where XAI is utilized to explain the reasons of prediction of the complex models. In addition, DL-based models often show over-fitting characteristics due to the lack of sufficient amount of data used to train [5] including the ASA task. This reduces the models' efficiency to determine the sentiment of low-resources languages like Arabic, Bengali, Hindi, etc.

To tackle these concerns, we propose a new interpretable[1] Arabic sentiment classification framework by adding a Gaussian noise layer to the DL-based models. We develop and train two DL models including Bidirectional LSTM (BiLSTM) and CNN-BiLSTM, CNN layer followed by BiLSTM layer for sentiment classification. The

[1]The words *Explainability* and *Interpretability* are used interchangeably throughout the paper.

experimental results indicate that adding a noise layer helps to resolve the overfitting problem of these models in SA. To explain particular prediction of our sentiment classification framework, we adopted LIME (Local Interpretable Model-agnostic Explanations), a prominent XAI method that can explain the predictions of any sentiment classifier in an interpretable and transparent manner, by learning an interpretable surrogate model locally around the prediction [26]. For experimental purposes, we employed publicly available datasets including a Large Arabic Book Review (LABR) [6] and an Arabic Hotel Review (HTL) datasets [16]. The experimental results verify our claims and mitigate the above-mentioned concerns reducing potential overfitting problems for DL-based ASA models. The contribution of this paper can be summarized as follows:

(1) We propose two different DL-based methods by introducing *noise layer* for Arabic SA to reduce over-fitting with improved performance.
(2) To the best of our knowledge, this is the initial endeavor towards enhancing the explainability of Arabic sentiment classification models.
(3) Our method consistently achieves competitive performance compared to state-of-the-art approaches in Arabic SA, and it can be applied to other regional Arabic languages.

The rest of the paper is organized as follows: We survey related literature on ASA in Section 2. Then we present our method with explainability in Section 3. In Section 4, we discuss the findings from the experiments. Finally, Section 5 concludes our methods with some future plans.

2 LITERATURE REVIEW

Recently, like other languages, Arabic SA has gained the attention of the research community [29]. Abdul-Mageed et al. [1] proposed methods to identify the subjectivity and sentiment of standard Arabic. In the following years, they proposed a corpus for sentiment analysis and a system to detect social media posts' sentiment. In these works, the authors utilized a large set of features for the experiments with machine learning algorithms [18]. Shoukry et al. [27] classified the sentiment of Egyptian Arabic tweets using SVM and NB classifiers. Later, they measured the performance of different machine learning models on preprocessed (i.e., stemming, stop words removal, and normalization) tweets [28]. Nayel et al.[23] employed a classical machine learning algorithm including SVM (Support Vector Machine) for sentiment and sarcasm detection purposes.

However, the use of DL methods is less common in Arabic SA compared to English. An LSTM-CNN model is utilized by Sarah [4] for Arabic text to classify two unbalanced classes from the ASTD dataset among four classes. Similarly, a CNN model is used by [25] with Stanford segmenter for tweets tokenization and normalization. They applied the CNN model to the ASTD dataset with a word-embedding model. Heikal et al. [20] proposed a method combining CNN and LSTM models with a pre-trained word-embedding model to predict the sentiment of the tweets. Some more prominent works on ASA also came out from the workshop and shared tasks [19].

As BERT (Bidirectional Encoder Representations from Transformers) based models show very promising performance in English SA, Oueslati et al. [15] presented an Arabic language-specific universal language model (ULM), hULMonA by fine-tuning multi-lingual BERT (mBERT). To evaluate the ULM, they collected a benchmark dataset for SA. BERT-based Arabic language model, AraBERT is developed by [7] to improve the state-of-the-art in many Arabic natural language understanding tasks. Husain et al. [21] hypothesized that tweets are more likely to contain offensive content when the tweet is positive or negative. Therefore, they fine-tuned the AraBERT using offensive language for Arabic sarcasm detection and sentiment analysis.

However, XAI is being applied for SA analysis for some languages like English and Chinese [13]. Though many works have been done on ASA, they did not employ XAI to explain the reasons for a specific prediction of the used ML or DL models. Adding a noise layer to the DL-based models also helps reduce over-fitting and eventually enhance the performance of the DL models [30]. In addition, there is no such work on ASA where a noise layer is added in the DL models to reduce over-fitting and improve the model's performance with XAI.

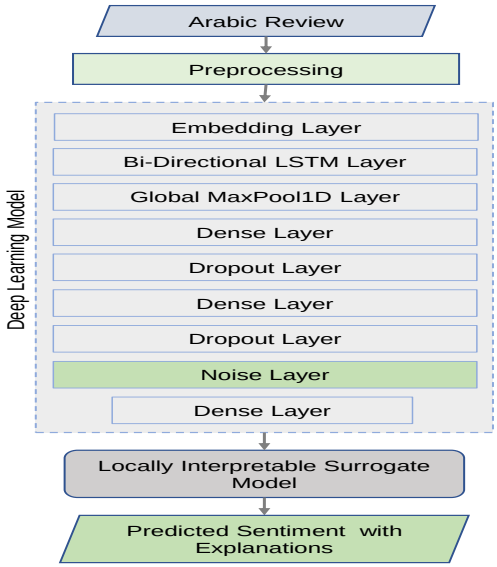

Figure 1: Overview diagram of our proposed Arabic Sentiment Analysis method

3 PROPOSED METHOD

This section presents our proposed method for Arabic Sentiment classification, which utilizes DL and provides explanations for each prediction that attempt to highlight the reasons for certain predictions. Initially, the Arabic reviews serve as input for the DL methods. Subsequently, the reviews undergo a preprocessing phase, where special characters are removed. Following this, a word tokenizer is employed to extract the list of words from the reviews then the padding technique is applied. Consequently, the padded sequences of the reviews are obtained from the preprocessing phase. BiLSTM and CNN-BiLSTM models are trained using the padded sequences

of the reviews. The performance of these trained models is evaluated using test data. Finally, to explain specific predictions, a locally trained surrogate model is employed, utilizing LIME. An overview of the overall framework is depicted in Fig. 1.

3.1 Classification Models

In the BiLSTM model, an Embedding layer is followed by a bidirectional LSTM layer and a global max pooling layer. In the CNN-BiLSTM model, an additional convolutional layer is introduced before the BiLSTM layer. To mitigate overfitting, a noise layer is incorporated just before the final output layer in both cases. The inclusion of the noise layer is motivated by the need to address overfitting in small neural networks trained with limited training data [30]. Inspired by this approach, we integrate a Gaussian noise layer into both the BiLSTM and CNN-BiLSTM models to mitigate overfitting [24, 30]. The detail for each layer is mentioned in section 4.2.

3.2 Explainable Surrogate Model with LIME

For explainability purposes, we trained a local surrogate model that mimics the classification performance of the original model. In this regard, we apply LIME (Local Interpretable Model-agnostic Explanations) introduced by [26]. LIME is a local delegate model means it is a trained model used to explain the causes of the predictions of the underlying black-box complex structure. However, it also includes generating different versions of the data for the machine learning model and testing what happens to the predictions, utilizing this perturbed data as a training set instead of the initial training data. In another sense, LIME creates a new dataset using permuted data and the associated black-box model predictions. LIME then learns an explainable model on this new dataset, which is weighted by the sampled instances' closeness to the instance of interest. LIME explains the local predictions of the model.

4 EXPERIMENTAL RESULTS WITH DISCUSSION

4.1 Datasets

For experiment and verification purposes, we apply our methods on two Arabic benchmark datasets for sentiment analysis including LABR [6] and Hotel review Dataset (HIL) [16].

LABR Dataset: Large-scale Arabic Book Review (LABR) dataset contains *63257* reviews each with a rating of 1 (one) to 5 (five) on *2131* books by *16486* users. Reviews with ratings 4 and 5 are considered as positive (1) sentiment, ratings with 1 and 2 are counted as negative (0) sentiment [6] and rating 3 is eliminated. Then *51056* reviews are considered for our experiments among them 42832 are positive (1) and *8224* are negative (0). Finally, *40844* reviews are used for training purposes and the rest *10212* are for testing.

Hotel Review dataset (HTL): A total of *15572* reviews are there in the HTL dataset [16]. The reviews are for *8100* Hotels among them *10766* reviews are positive, *2645* are Negative and the rest are Neutral. To make it a balanced and smaller dataset, we only select *2645* positive reviews randomly. Finally, *3967* reviews are used for training, and the rest *1323* reviews for testing.

Table 1: Performance of BiLSTM and CNN-BiLSTM models on imbalance LABR dataset.

model	setup	train acc.	Testing Evaluation class	pre.	rec.	f1	test acc.	O.F. (%)
BiLSTM	$Model_{ND}$	0.98	0	0.62	0.55	0.58	0.88	10.0
			1	0.92	0.94	0.93		
	$Model_{N}$	0.99	0	0.63	0.53	0.57	0.88	11.0
			1	0.92	0.94	0.93		
	$Model_{D}$	0.98	0	0.56	0.59	0.57	0.86	12.0
			1	0.92	0.91	0.92		
CNN-BiLSTM	$Model_{ND}$	0.99	0	0.63	0.50	0.56	0.88	11.0
			1	0.91	0.95	0.93		
	$Model_{N}$	0.99	0	0.62	0.50	0.55	0.88	11.0
			1	0.91	0.94	0.93		
	$Model_{D}$	0.99	0	0.60	0.55	0.57	0.87	12.0
			1	0.92	0.93	0.93		

(pre.:precision, rec.:recall, f1:$f1\_score$, O.F.:overfit.)

Table 2: Performance of BiLSTM and CNN-BiLSTM models with different settings on balance HTL dataset

Model	Setup	Tr.-Acc.(%)	Tst-Acc.(%)	O.fit. (%)
BiLSTM	$Model_{ND}$	**99.72**	**93.64**	**6.08**
	$Model_{N}$	99.86	93.51	6.35
	$Model_{D}$	99.96	92.00	7.96
CNN-BiLSTM	$Model_{ND}$	**99.85**	**94.78**	**5.07**
	$Model_{N}$	99.92	93.65	6.27
	$Model_{D}$	99.97	92.02	7.95

4.2 Experimental Setup

We conducted experiments using different settings to evaluate the performance of our models. For each proposed model, we considered three variations, each with specific characteristics:

- $Model_{ND}$: This model includes both the noise and dropout layers before the output layer or final layer.
- $Model_{N}$: This model only contains the noise layer before the output layer and does not have an immediate dropout layer before the noise layer.
- $Model_{D}$: This model does not include the noise layer but incorporates the dropout layer.

For the BiLSTM model, the architecture follows the overview depicted in Figure 1. In this model, we enumerated 10,000 unique vocabularies, representing each vocabulary with a 100-dimensional vector using the embedding layer of the Keras library.

For the CNN-BiLSTM model, a one-dimensional convolutional layer with a ReLU activation function and a kernel size of 3 was included. A dropout layer with a value of 0.5 was used to randomly drop 50% of the features during training. Additionally, a Gaussian noise layer with a value of 0.75 was employed. Each variant of the DL-based models underwent training for a total of 10 epochs.

Finally, a local interpretable surrogate model was trained to mimic the original proposed models and provide explanations for specific predictions.

Table 3: Performance comparison of our method CNN-BiLSTM with some known methods on LABR dataset.

methods	accuracy(%)
Attention-BiGRU [10]	95.6
AraBERT [7]	89.6
CNN-BiLSTM	**88.0**
Multi-chan CNN [12]	87.5
mBERT [7]	83.0
CNN [14]	81.6
BiLSTM [14]	80.7
LSTM [14]	78.5

4.3 Experimental Results

Tables 1 and 2 present the performance of our BiLSTM and CNN-BiLSTM models in determining Arabic sentiment, using different evaluation metrics on the LABR and HTL datasets, respectively. As LABR is an imbalanced dataset, in addition to accuracy, other evaluation metrics are used to assess the performance of our proposed models. Conversely, as HTL is a balanced dataset, the performance of the introduced BiLSTM and CNN-BiLSTM models is evaluated solely based on accuracy.

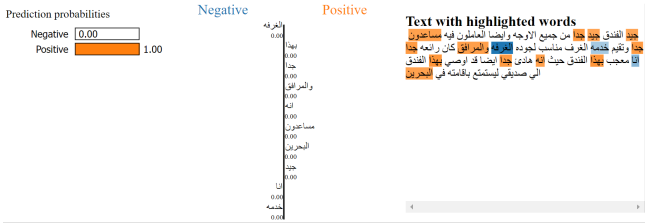

Figure 2: Explanation of a review with $Model_{ND}$ using LIME.

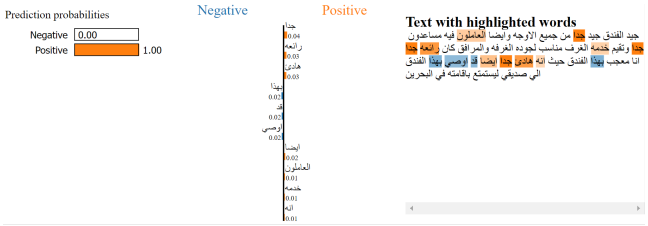

Figure 3: Explanation of a review with $Model_D$ using LIME.

When the noise layer is added along with the dropout layer before the last layer of the DL models, they exhibit improved performance and reduced overfitting. Table 1 supports this claim by showing that when the noise layer is absent before the output layer ($Model_D$), the overfitting (O.fit) is 12%. However, after adding the noise layer before the output layer ($Model_{ND}$), the overfitting decreases by 2% and 1% for the BiLSTM and CNN-BiLSTM models, respectively. The *precision* of negative reviews in the BiLSTM model's $Model_{ND}$ setup in Table 1 is 0.62, while for $Model_D$, it is 0.56. This indicates that the noise layer helps in identifying more negative reviews compared to those without the noise layer. The same trend is observed for the CNN-BiLSTM model.

Table 2 illustrates that the overfitting (O.fit.) of the BiLSTM model is 6.08% when the noise and dropout layers are included ($Model_{ND}$). Similarly, when the noise layer and dropout layer are integrated into the CNN-BiLSTM model, the overfitting is 5.07%, with training and testing accuracies of 99.85% and 94.78%, respectively. On the other hand, without the noise layer, the degree of overfitting increases. For example, in the CNN-BiLSTM model with the $Model_D$ setup, the overfitting is 7.95%, which is approximately 2.88% higher than $Model_{ND}$(5.07%). These findings demonstrate the effectiveness of adding a noise layer to the models.

4.4 XAI in Arabic Sentiment Analysis

We trained a local explainable surrogate model to get an understandable representation of the predicted sentiments of the Arabic reviews.

Let's consider a review and see how our proposed method performs on this. The review is as follows: ``جيد الفندق جيد جدا من جميع الاوجه وايضا العاملون فيه مساعدون جدا وتقيم خدمه الغرفمناسب لجوده الغرفه والمرافق كان رائعه جدا انا معجب بهذا الفندق حيث انه هادئ جدا ايضا قد اوصي بهذا الفندق الي صديقي ليستمتع باقامته في البحري'' (Translation: Good, the hotel is very good in all respects, and also the staff is very helpful, and the room service is appropriate for the quality of the room and the facilities were very wonderful. I admire this hotel as it is very quiet also. I may recommend this hotel to my friend to enjoy his stay in Bahrain.) Fig. 2 and 3 illustrate the humans' interpretable representation of the predicted positive sentiment of the same review using $Model_{ND}$ and $Model_D$, respectively. Fig. 2 which is obtained using $model_{ND}$ highlights the words (e.g. جيد (good), جدا (very), هادئ (calm/quiet), مساعدون (helpful)) which are contributing the review toward the positive sentiment. On the other hand, LIME and $model_D$ interpreted the same review in Fig 3 and it shows some major differences such as جيد (good), مساعدون (helpful) are not highlighted as positive sentiment words. However, these words have a real impact on making the review a positive sentiment. This phenomenon illustrates the effectiveness of adding a noise layer in the DL model to make it more explainable and acceptable.

4.5 Comparison with related works

Different prominent research works have been conducted on ASA on different datasets. Here, we compare the proposed method with some existing methods that are on the LABR dataset. Attention BiGRU [10], and AraBERT [7] achieved higher accuracy than our method as these are attention-based DL models. Attention BiGRU [10] employed a hybrid bidirectional gated recurrent unit (BiGRU) and bidirectional long short-term memory (BiLSTM) additive attention model with two types of embedding and achieved

SOTA results on LABR dataset. AraBERT [7] was introduced for different tasks including ASA. It is built especially for Arabic NLP. In ASA, they achieved 89.6% accuracy which is 1.6% more than our CNN-BiLSTM. Multilingual BERT (mBERT) is also employed for ASA [7] and its accuracy is 83% which is 5% lesser than our method. Moreover, LSTM, BiLSTM, and CNN, multi-channel CNN are also utilized without noise layer and their performance is quite promising but still the performance is lower than ours. These findings illustrate the significance of adding a noise layer in the DL models.

5 CONCLUSION AND FUTURE WORK

This paper proposed an explainable Arabic sentiment classification framework introducing noise layer in DL models including BiLSTM, and CNN-BiLSTM. Generally, DL-based models show overfitting characteristics when a small amount of data is used for training which makes the DL model's generalization capability poor. That is why a Gaussian noise layer is added to the proposed models to reduce overfitting and enhance performance. The experimental results also indicate that the noise layer helps to reduce the overfitting issue of the DL-based models and improve the performance. Again, these models work in a black-box manner in predicting the sentiment which is not understandable to humans. Therefore, to interpret the reasons for the particular sentiment predictions, a locally explainable surrogate model known as LIME is employed for the first time in this paper for ASA. LIME shows easy-to-understand explanations that provide an understandable representation to make users sense of a prediction.

In the future, we plan to enhance the performance of the current explainable AI algorithms for a better understanding of ASA and other Arabic NLP tasks.

ACKNOWLEDGMENTS

This paper is a collaborative effort and the second author's work as part of the *Marie Skłodowska-Curie* project funded by the *Horizon Europe 2020* research and innovation program of the *European Union* under the grant agreement no. 955422.

REFERENCES

[1] Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 587–591.

[2] Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information processing & management* 56, 2 (2019), 320–342.

[3] Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16, 4 (2017), 1–20.

[4] Sarah Alhumoud, Tarfa Albuhairi, and Wejdan Alohaideb. 2015. Hybrid sentiment analyser for Arabic tweets using R. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, Vol. 1. IEEE, 417–424.

[5] A Aziz Altowayan and Lixin Tao. 2016. Word embeddings for Arabic sentiment analysis. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 3820–3825.

[6] Mohamed Aly and Amir Atiya. 2013. 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)*. 494–498.

[7] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104* (2020).

[8] Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*. 165–173.

[9] Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2019. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv preprint arXiv:1906.01830* (2019).

[10] Mohamed Berrimi, Mourad Oussalah, Abdelouahab Moussaoui, and Mohamed Saidi. 2023. Attention mechanism architecture for arabic sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 4 (2023), 1–26.

[11] Abdelghani Dahou, Mohamed Abd Elaziz, Junwei Zhou, and Shengwu Xiong. 2019. Arabic sentiment classification using convolutional neural network and differential evolution algorithm. *Computational intelligence and neuroscience* 2019 (2019).

[12] Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, and Mohamed Abd Elaziz. 2019. Multi-channel embedding convolutional neural network model for arabic sentiment classification. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, 4 (2019), 1–23.

[13] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711* (2020).

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

[15] Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hulmona: The universal language model in arabic. In *Proceedings of the fourth arabic natural language processing workshop*. 68–77.

[16] Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *International conference on intelligent text processing and computational linguistics*. Springer, 23–34.

[17] Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. 192–198.

[18] Ibrahim Abu Farha and Walid Magdy. 2021. A comparative study of effective approaches for Arabic sentiment analysis. *Information Processing & Management* 58, 2 (2021), 102438.

[19] Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb. 2021. Proceedings of the Sixth Arabic Natural Language Processing Workshop. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

[20] Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of Arabic tweets using deep learning. *Procedia Computer Science* 142 (2018), 114–122.

[21] Fatemah Husain and Ozlem Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. 364–369.

[22] Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Intelligent computing-proceedings of the computing conference*. Springer, 1269–1292.

[23] Hamada Nayel, Eslam Amer, Aya Allam, and Hanya Abdallah. 2021. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. 386–389.

[24] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807* (2015).

[25] Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems* 112 (2020), 408–430.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[27] Amira Shoukry and Ahmed Rafea. 2012. Sentence-level Arabic sentiment analysis. In *2012 international conference on collaboration technologies and systems (CTS)*. IEEE, 546–550.

[28] Amira Shoukry and Ahmed Rafea. 2015. A hybrid approach for sentiment classification of egyptian dialect tweets. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. IEEE, 78–85.

[29] Cristian Muoz Villalobos, Leonardo Mendoza Forero, Harold De Mello, Cesar Valencia, Alvaro Orjuela, Ricardo Tanscheit, and Marco Pacheco Cavalcanti. 2022. Sentimental Analysis on Social Media Comments with Recurring Models and Pretrained Word Embeddings in Portuguese. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*. 205–209.

[30] Zhonghui You, Jinmian Ye, Kunming Li, Zenglin Xu, and Ping Wang. 2019. Adversarial noise layer: Regularize neural network by adding noise. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 909–913.