2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
17-20 October, 2021. Melbourne, Australia

2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-4207-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMC52423.2021.9659245

Sentiment Analysis in Arabic Social Media Using Deep Learning Models

Ayman Yafoz
Department of Information Systems
King Abdulaziz University
Jeddah, Saudi Arabia
ayafoz@kau.edu.sa

Malek Mouhoub
Department of Computer Science
University of Regina
Regina, Canada
Malek.Mouhoub@uregina.ca

Abstract— There are limited research contributions targeting sentiment analysis in feedback in Arabic gulf dialect, in particular, and the Arabic language in general. Furthermore, the inadequate and limited adoption of classification techniques and natural language processing is noticeable in the sentiment analysis projects addressing the Arabic language. Hence, this paper focuses on analyzing the sentiments in automobile and real estate domains through the application of the state-of-the-art word-embedding model "BERT" and a collection of deep learning models (GRU, LSTM, CNN, CNN-GRU and BiLSTM). The results of classification revealed that combining the BERT with deep learning models have shown efficiency in analyzing sentiments and yielded outstanding results.

Keywords— Sentiment Analysis, Deep Learning, BERT Word Embedding, Arabic Language

I. INTRODUCTION

The topic of sentiment analysis has been widely covered in recent publications. There are numerous contributions in sentiment analysis claiming to have a positive impact on different areas like health, education, politics, marketing among other fields. Currently, social networks play a vital role in the lives of millions of users, where they can share their opinions, emotions and sentiments [1]. On the other hand, the Arabic language is considered, by millions of Arabic states citizens, to be the official language and, by Muslims, a religion-based language. Arabic is classified as a Semitic language which consists of 28 letters. It is estimated that approximately 5% of the global human population are native Arabic speakers, and there are 22 Arab countries with a total population of approximately 359 million people, estimated in 2020. Based on research published by the United Nations, this number is expected to reach around 589 million by 2050 [2]. Furthermore, the Arabic language is rapidly increasing in popularity on the web and is considered the fourth proliferated language [3].

However, analyzing texts on social media is not appropriately covered in the framework of the Arabic language, whereas many contributions are addressing sentiment analysis in other languages such as English. Moreover, lack of available annotated datasets in Arabic, which are focusing on specific domains (such as automobiles and real estate), and the limited sentiment analysis research areas focusing on different Arabic dialects are noticeable.

Arabic language has many features which complicate performing language processing and analyses. These features include complicated semantics, grammars and morphologies due to the complex derivational and inflectional characteristics of the language. Hence, natural language processing systems designed for English are unable to handle the processing requirements of the Arabic language [4]. For instance, performing part-of-speech (POS) tagging on Arabic text is more complicated than on English text. Arabic words could be categorized into several classes. For example, the word "علم" could have various meanings such as "flag", "prominent" or "taught". Therefore, this word could be a noun, an adjective or a verb. Furthermore, Arabic dialects contain negations that are not present in MSA (Modern Standard Arabic) because they can be expressed in various ways. Negation words can carry different meanings which makes the process of identifying negation, error prone and complex, e.g. the term "مو" which means "not" [5].

Moreover, Arabic social media text often contains idioms and compound phrases. These phrases vary among different dialects and could inverse the meaning among several Arab countries, and new phrases are continuously appearing. For instance, the phrase "القرد في عين أمه غزال", which means "A monkey is a deer in the eye of its mother", conveys a negative sentiment [6]. Furthermore, there are no lowercase and uppercase in Arabic, which affects the identification of Arabic nouns in a text. This identification is significant in sentiment analysis because the analyzer could be mistaken in distinguishing Arabic nouns derived from adjectives. For instance, the word "سعيد", which could denote both a male name or the adjective "happy" [5]. Additionally, diacritization is a significant element to convey the meaning of a given word, such as "علم" which could have the diacritical mark as "عَلِمَ" conveying the past verb "knew", or the diacritical mark as "عِلْمٌ" conveying the word "science" [7]. However, diacritization is absent in many Arabic textual resources, which makes the parsing task more difficult and inaccurate in many sentiment analysis tasks.

These gaps and challenges have specifically encouraged the researchers to conduct this research. Consequently, the work reported in this paper followed a methodology that was customized for the Arabic language. The different between contributions addressing non-Arabic data and the work presented in this paper includes the pre-processing techniques, the POS tagging and the lemmatizing tools.

Moreover, we introduced in [8], [9], and [10] approaches based on word embedding, deep learning and machine learning techniques for classifying the sentiments in Arabic data related to real estate and automobiles. Following up on these works, we included BERT word embedding model and the state-of-the-art Stanza POS tagger and lemmatizer to extend and enhance the approaches in those papers. These new techniques have not been used before with Arabic data related to real estate and automobiles. The adopted techniques in this new work should improve the classification results and update the approaches reported in our previous papers. The sections of this paper are organized as follows. The second section presents the literature review, followed by the third section which explains the proposed solution. The fourth section reports on the analysis of the classification results. Finally, the fifth section concludes the paper, and provides future enhancements to upgrade the quality of work and the results in future contributions.

II. LITERATURE REVIEW

The sentiment analysis systems proposed in this research are compared with contributions targeting sentiment analysis for both Arabic and non-Arabic data. The researchers in [11] used the LABR dataset which contained 63,257 book reviews written in both colloquial dialects and MSA. These reviews were rated using a scale from one to five. The researchers performed a preprocessing procedure to clean and normalize the text. They also removed stop words from the data and used Bag-Of-Words to extract features from the text. Furthermore, four classifiers were adopted: Naive Bayes, Decision Tree, SVM and K Nearest Neighbor provided by Weka. The highest accuracy result achieved by the K Nearest Neighbor on a hierarchical classification structure of four levels was 57.8%.

Moreover, the researchers in [12] introduced a sentiment analysis framework that worked on a dataset containing 2,000 political Jordanian tweets written in both MSA and the Jordanian dialect. The tweets were evenly divided between positive and negative. The researchers relied on four features extracted from the text: lexicon, writing style, emotional and grammatical features. Eight traditional machine learning classifiers were adopted: MultiClass-Classifier, Random Forest, Simple Logistic, Dagging, Naïve-Bayes, Canonical Variate Regression, MultiBoost and SVM. The highest accuracy result they found was 72.83% by the Dagging classifier.

The researchers in [13] conducted research on performing a sentiment analysis on Twitter data to predict the movements in stock market. The dataset contained 250,000 tweets on Microsoft, gathered using Twitter API. After the dataset had been gathered from Twitter, the researchers performed data preprocessing through the filtering phases of tokenization, stop words and elimination of special characters. Following that, the processes of n-grams and Word2vec were performed (where the words were mapped to a 300-dimensional vector). Random Forest, Logistic Regression and SVM algorithms were selected to perform the classification process. The highest accuracy result they obtained was 71.82% by the SVM classifier.

The researchers in [14] performed a sentiment analysis on a dataset containing 1,000 tweets about an Indonesian courier service "JNE Semarang". The dataset was divided into 700 tweets for training and 300 tweets for testing. The researchers did a preprocessing procedure on the dataset to clean, convert the letters in tweets into lowercase, tokenize and perform stemming. The researchers selected TFIDF for feature selection and adopted K-Nearest Neighbor for classification. The highest accuracy result achieved was 82.7%.

The researchers in [15] started their work by performing a preprocessing and a filtration process on a large corpus containing more than 84,000 Arabic tweets. The tweets were gathered from Egyptian Twitter accounts that appeared to be determined to be the most active accounts, and based on trends in Egypt. Data were annotated using Amazon Mechanical Turk. The tweets were divided into four tags, which were “subjective negative”, “objective”, “subjective mixed” and “subjective positive”. This resulted in creating a dataset with 10,006 labelled tweets. Regarding feature selection, the researchers adopted TFIDF and n-grams approaches. The highest accuracy result they acquired was 69.10% by the SVM classifier on the unbalanced dataset with the features of unigram, bigram, trigram and TFIDF.

III. PROPOSED SOLUTION

The proposed solution was developed using Python. The training of the deep learning models with word embeddings was on Google Cloud that has NVIDIA Tesla P100 - GPU computing processor. Keras and TensorFlow deep learning libraries were adopted. The developed sentiment analysis system is comprised of three main levels which are described below.

A. Dataset Preparation

Collecting Arabic data is a challenging task as there is a limitation in the number of available resources where data could be gathered. In many cases, data are gathered from either online forums or social media outlets. The gathered data could contain various Arabizi (Latin letters used to write Arabic words) or dialects, which means that the sentiment analysis system will be trained on a small training dataset.

This could lead to lesser accuracy or inaccurate results [5]. The procedures of assembling, cleaning, processing, and annotating information could cost extra time, generate incorrect results, and require much effort and attention to the language details especially whenever the sentiment analysis approaches are based on those procedures [16].

In this level, the automobile and real estate datasets created by us in [8], [9], and [10] were also adopted in this research to test our approach, which enabled the comparison of our results with the previous results. The data in these datasets were collected from three famous online forums in the Arabic gulf countries, which are: (1) Haraj online forum for automobile data, and (2) Hawamer and Aqarcity online forums for real estate data. Moreover, the automobile dataset contained around 6,585 reviews grouped into three sentimental classes (negative, positive, or mixed), and were focused on almost 27 topics about automobiles. Moreover, the real estate dataset had around 6,434 reviews (3,203 reviews from Hawamer and 3,231 reviews from Aqarcity) grouped into three sentimental classes: negative, positive, and mixed. For instance, Table I illustrates a sentiment analysis operation on reviews extracted from the automobile dataset. On the other hand, Table II illustrates the results of a sentiment analysis process on reviews extracted from the real estate dataset. The datasets are uploaded on Github and can be freely accessed through the link in [17].

To prepare the datasets, the records were randomly shuffled to fairly distribute them. After that, the datasets were split into 70% and 30% for the training and the testing sets, respectively, to ensure achieving an adequate training and testing. The 10-K cross validation method was adopted to randomly shuffle the records in the datasets and divide them in K identical-sized folds. Moreover, SMOTE-NC was adopted to adequately represent minority examples in the datasets.

B. Feature Selection

The feature selection process was carried out to minimize the dimensionality of the datasets, relieve the memory load, and minimize the processing costs [18]. In this level, two feature selection methods were adopted: lemmatization and the POS tagging. As a component of feature selection, the state-of-the-art Stanza toolkit (a Python library for natural language analysis tasks) was adopted to provide the lemmatizer and POS tagger for our research. It supports the Arabic texts and showed more accurate results when compared with two other famous Arabic POS taggers, which are Farasa and MADAMIRA. For instance, the following sentence was POS tagged by Stanza, Farasa, and MADAMIRA POS taggers:

"بالنسبة للاوبتيما سيارة عملية و ممتازة و مريحة"

Which translates to:

"With regard to the Optima, it is a practical, excellent and comfortable car"

The Stanza POS tagger was more precise in tagging the above sentence as follows: (1) the coordinating conjunction "و", translating to "and", was not tagged by Farasa or MADAMIRA, (2) the adpositional phrases "ب", translating to "with", and "ل", translating to "to", were not tagged by Farasa or MADAMIRA, and (3) the adjective "عملية", which translates to "practical", was tagged by Farasa as a noun. On the other hand, MADAMIRA tagged it as a nominal, which could be confusing when performing the processing (is it a noun or adjective?).

TABLE I. THE RESULTS OF A SENTIMENT ANALYSIS OPERATION ON THE AUTOMOBILE DATASET

Sentence	Translation	Sentiment
فعلا قطع الكامري الأصلية غالية بس لها ميزة إنها متوفرة.	Indeed, the original spare parts of the Camry are expensive, but they have the benefit of being available.	Mixed
المورانو فيه عيب كبير دايما الجربوكس يروح فيه.	The Murano has a big defect, its gearbox always wears out.	Negative
الافضل اكسنت بالنسبة لكل شيء مرغوبة، قطع غيار متوفرة ورخيصة وعملية.	For me, Accent is the best in everything, it is a popular car, its spare parts are available, it is cheap, and it is practical.	Positive

TABLE II. THE RESULTS OF A SENTIMENT ANALYSIS PROCESS ON THE REAL ESTATE DATASET

Sentence	Translation	Sentiment
حي اليرموك يعتبر من اسوأ احياء الرياض. حي منتشر به جرائم السرقة و اطلاق النار و تجمعات الدرباوية و المفحطين.	The Yarmouk District is considered one of the worst neighborhoods in Riyadh. A neighborhood where the crimes of robbery and shootings are common, as well as the communities of neds and hoons.	Negative
الموقع جميل كإستراحات بعيدة عن ضجيج المدينة، لكنه لا يستحق الأسعار الحالية.	The location is beautiful as lounges and it is away from the city noise, but it does not deserve the current prices.	Mixed
رح حي عرقة الإيجارات معقولة وحي زين.	Go to Irqah District; the rentals are reasonable, and the district is nice.	Positive

On the other hand, Stanza lemmatizer was adopted to generate lemmas. It was adopted because it is compatible with the adopted Stanza POS tagger, and generated accurate lemmas to a great extent. For instance, the following sentence was lemmatized by Stanza lemmatizer:

" سيكون مشروع الصرف الصحي في أبحر أطول مشروع في العالم صراحة مهزلة"

Which translates to:

"The sewage project in Obhor will be the most time-consuming project in the world, frankly, it is a mockery"

Is lemmatized as:

" سيكون مَشرُوع صَرف صِحِّيّ فِي أَبحَر أَطوَل مَشرُوع فِي عَالَم صَرَاحَة مهزَلَة"

C. Data Processing

The BERT word embedding model is developed by Google in 2018 [19]. BERT is a state-of-the-art word embedding model which proved efficiency when applied to many NLP projects. The multilingual BERT (M-BERT) is pre-trained on a sequence of corpora extracted from Wikipedia pages in 104 languages [20] (including the Arabic language) and Google Books [21].

The training of BERT is based on predicting the masked words related to input sentences. This process is accomplished through learning self-attention with the purpose of valuing the association between a single word and the words in the same input sentence. Following that, a vector is assigned to each word to represent its association with the words in the same input sentence. The assigned vectors will be used to create word embedding, where the created embeddings for every word are based on the words in the same sentence. On the other hand, Wor2vec and Glove represent each word with fixed embeddings, irrespective of the word's context [22].

In this level, BERT and a collection of deep learning models (GRU, LSTM, CNN, CNN-GRU and BiLSTM) had classified the datasets. The epoch had the value of 50. The same value was also chosen by the authors in [23] for classifying their large-scale multi-language datasets. Furthermore, the batch size had the value of 32. This value was nominated as a decent default value by the authors in [24]. For the learning rate, the value 5e-e was chosen, which is the same value adopted by the authors in [25]. In addition, the stride had the value of 1 because it is the most common stride value [26]. The same stride value was also used by the authors in [27]. Finally, the adopted activation function was ReLU (Rectified Linear Unit), which was adopted as well by the authors in [28].

TABLE III. THE CLASSIFICATION RESULTS FOR THE AUTOMOBILE DATASET

BERT				
	Accuracy	***Precision***	***Recall***	***F1-Score***
CNN	98.47	98.47	97.99	98.20
LSTM	98.94	98.94	98.61	98.71
GRU	98.43	98.43	98.26	98.16
BiLSTM	95.32	95.32	94.11	93.92
CNN+GRU	98.16	98.16	97.52	97.71
BERT	95.98	95.98	94.57	95.07

TABLE IV. THE CLASSIFICATION RESULTS FOR THE REAL ESTATE DATASET

BERT				
	Accuracy	***Precision***	***Recall***	***F1-Score***
CNN	99.10	99.10	98.45	98.67
LSTM	98.71	98.71	97.64	97.79
GRU	98.91	98.91	98.01	98.28
BiLSTM	98.80	98.80	98.15	98.05
CNN+GRU	98.66	98.66	97.72	98.02
BERT	95.45	95.45	93.26	92.53

IV. ANALYSIS OF THE CLASSIFICATION RESULTS

In this level, the comparison is twofold. The first aspect shows the performance of the models with the automobile dataset, whereas the second illustrates their performance with the real estate dataset. The comparisons focus on four metrics, which are accuracy, precision, recall, and F1-score as shown in Table III and Table IV.

Based on Table III, for the automobile dataset, the LSTM with the BERT had the highest accuracy result of 98.94%. In terms of the precision metric, the LSTM with the BERT had the highest result of 98.94%. In terms of the recall metric, the LSTM with the BERT had the highest result of 98.61%. As for the F1-score metric, the LSTM with the BERT had the highest result of 98.71%. These results show that the LSTM with the BERT generated the highest results in the four measured criteria (accuracy, precision, recall, and F1score). Moreover, in terms of the F1-score metric, the LSTM with the BERT produced a higher result than that achieved by us in [8] using the GRU with the CBOW (the result was 84.90%).

On the other hand, Table IV illustrates the classification results for the real estate dataset. Based on Table IV, for the real estate dataset, the CNN with the BERT had the highest accuracy result of 99.10%. In terms of the precision metric, the CNN with the BERT had the highest result of 99.10%. As for the recall metric, the CNN with the BERT had the highest result of 98.45%. For the F1-score metric, the CNN with the BERT had the highest result of 98.67%. These results show that the CNN with the BERT generated the highest results in the four measured criteria (accuracy, precision, recall, and F1-score). Moreover, in terms of the F1-score metric, the CNN with the BERT produced a higher result than that achieved by us in [8] and [9] using the Ridge CV classifier (the result was 76%).

For both datasets, the results of classification revealed that combining deep learning models with the BERT produced higher F1-scores than those of the other models adopted by us in [8], [9], and [10]. Our findings also conform to the findings of the researchers in [29] who reported the superiority of BERT when adopted to analyze the data in Arabic social media. The outstanding results could be attributed to the BERT model mechanism of utilizing word pieces "subwords" instead of whole words, which should give better results when used with Arabic text. Moreover, the researchers in [30] also claim that BERT model would be more compatible with Arabic texts as Arabic language has a rich morphology, which justify to a large extent the high results achieved in our research.

Furthermore, the researchers in [31] showed that combining M-BERT with a deep learning model (more specifically CNN) for Arabic class classification (six, three, and two classes) produced higher F-Macro scores than those achieved by the adopted machine and deep learning models in their paper. The scores were 75.51% for the six classes, 78.99% for the three classes, and 87.03% for the two classes. These factors drive us to promote adopting M-BERT in future Arabic sentiment analysis projects.

V. CONCLUSION AND FUTURE WORK

In conclusion, there is a limited number of studies addressing the application of deep learning and word embedding models to analyze the sentiments in online Arabic reviews related to the industry of real estate and automobiles, particularly in the Arabic gulf dialect. In order to fill this gap, this paper proposed a framework customized to classify three categories of sentiments (positive, negative, mixed) in online Arabic reviews associated with real estate and automobiles. The framework is based on deep learning models and the BERT word embedding model. The paper also explained the levels of building the framework. Based on the high results achieved in all of the measured criteria, the analysis has shown that for both automobile and real estate datasets, adopting deep learning models with the BERT model to analyze the sentiments is more efficient than adopting other models. Moreover, regardless of the enhancements outlined in the future work below, the results of the proposed framework are, to a great extent, satisfactory when compared to other works addressing Arabic sentiment analysis as shown in this paper.

In future work, adopting advanced techniques to analyze sentiments could also enrich the work presented in this paper. These techniques include attention-based models, the transformers, ELMo, ERNIE, and reinforcement learning models. Furthermore, adopting a semi supervised approach that needs few labelled examples to learn could enlarge the scope of our work [32]. Finally, the work could also be enriched by performing a finer-grained sentiment analysis process on an aspect-based level. This approach focuses on analyzing sentiments in specific aspects (such as engine, color, speed, among others) as opposed to the whole sentence [33] —an aspect that should also be considered in future contributions.

REFERENCES

[1] B. A. Olivares-Zepahua, F. J. Ramírez-Tinoco, G. AlorHernández, J. L. Sánchez-Cervantes, and L. Rodríguez-Mazahua, "A Brief Review on the Use of Sentiment Analysis Approaches in Social Networks," *In Proceedings of the International Conference on Software Process Improvement (CIMPS 2017)*, 2017, 1st ed., pp. 263–273.

[2] B. Mirkin, "*Population Levels, Trends and Policies in the Arab Region: Challenges and Opportunities*," Technical Report, United Nations Development Programme, 2010.

[3] I. Awajan and M. Mohamad, "A Review on Sentiment Analysis in Arabic Using Document Level*," Int. J. Eng. Technol.*, vol. 7, no. (3.13) (2018), pp. 128–132, 2018.

[4] M. Nagi, N. Adly, and S. Alansary, "A Suite of Tools for Arabic Natural Language Processing: A UNL Approach*," In Proceedings of the 1st IEEE International Conference on Communications, Signal Processing, and their Applications (ICCSPA), IEEE Press*, 2013, pp. 1–6.

[5] H. Mansour, M. El-Masri, and N. Altrabsheh, "Successes and Challenges of Arabic Sentiment Analysis Research: A Literature Review," *Soc. Netw. Anal. Min. J.*, vol. 7, no. 1, 2017, pp. 1–22.

[6] N. Habash, Introduction to Arabic Natural Language Processing, 1st ed., Morgan and Claypool, 2010.

[7] A. Fahmy, M. Magdy, M. Elarnaoty, and S. AbdelRahman, "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," *International Journal of Computer Science Issues (IJCSI)*, vol. 7, no. 4, 2010, pp. 27-36.

[8] A. Yafoz and M. Mouhoub, "Analyzing Machine Learning Algorithms for Sentiments in Arabic Text,". *In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2020)*, 2020, pp. 1–7.

[9] A. Yafoz, "Towards Analyzing the Sentiments in the Fields of Automobiles and Real-Estates with Specific Focus on Arabic Online Reviews," In Canadian Conference on Artificial Intelligence (Canadian AI 2020), 2020, pp. 566–570.

[10] A. Yafoz and M. Mouhoub, "Towards Analysing the Sentiments in the Field of Automobile with Specific Focus on Arabic Language Text," *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, 2020, pp. 403–412.

[11] A. Nuseir, G. Kanaan, M. Al-Ayyoub, and R. Al-Shalabi, "Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews*," International Journal of Advanced Computer Science and Applications (IJACSA)*, vol 7, no. 2, 2016, pp. 531– 539.

[12] E.E. Abdallah and S. Abo-Suaileek, "Feature-based Sentiment Analysis for Slang Arabic Text," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol 10, no. 4, 2019, pp. 298–304.

[13] B. Majhi, G. Panda, K. Challa, and V. Pagolu, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," *In Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016, pp. 1345–1350.

[14] A. Rahman, E. Udayanti, and E. Kartikadharma, "Sentiment Analysis Towards Courier Service: Case Study on JNE Semarang," *In Proceedings of the 1st International Conference on Computer Science and Engineering Technology Universitas Muria Kudus (ICCSET)*, 2018, pp. 857–863.

[15] A. Atiya, M. Nabil, M. Aly, "ASTD: Arabic Sentiment Tweets Dataset," *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519.

[16] A. ElKorany and H. Alsayadi, "Integrating Semantic Features for Enhancing Arabic Named Entity Recognition," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 3, 2016, pp. 128–136.

[17] "Arabic-Sentiment-Analysis-Datasets" https://github.com/aymanya/Arabic-Sentiment-Analysis-Datasets (accessed May. 15, 2021).

[18] A. Alonso-Betanzos, N. Sánchez-Maroño, and V. BolónCanedo, *Feature Selection for High-Dimensional Data (Artificial Intelligence: Foundations, Theory, and Algorithms)*, 1st ed., Springer, 2015.

[19] R. Zhu, X. Huang, and Xinhui Tu. Deep Learning on Information Retrieval and its Applications. In C. Pradhan, H. Das, and N. Dey, editors, *Deep Learning for Data Analytics: Foundations, Biomedical Applications, and Challenges*, 2020, pp. 125–153.

[20] D. Garrette, E. Schlinger, and T. Pires, "*How Multilingual is Multilingual BERT?,*" ArXiv, 2019, pp. 1–6.

[21] A. Abdelali, A. Rashed, H. Mubarak, K. Darwish, and Y. Samih, "*Arabic Offensive Language on Twitter: Analysis and Experiments,*" ArXiv, 2020, pp. 1–10.

[22] F. A. Maghraby, M. El-Razzaz, and M. W. Fakhr*, "Arabic Gloss WSD Using BERT," Applied Sciences Journal*, vol. 11, no. 6 :2567, 2021, pp. 1–15.

[23] J. Lehečka, J. Švec, L. Šmídl, and P. Ircing, "Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification," *International Conference on Text, Speech, and Dialogue (TSD 2020)*, 2020, pp. 214-221.

[24] A. Garcia-Silva, J. Gomez-Perez, and R. Denaux, "*A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP"*, Springer, 2020.

[25] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," *China National Conference on Chinese Computational Linguistics (CCL 2019)*, 2019, pp. 194206.

[26] K. Togashi, M. Nishio, R. Do, and R. Yamashita, "Convolutional Neural Networks: An Overview and Application in Radiology," *Insights into Imaging Journal*, vol. 9, no. 4, 2018, pp. 611–629.

[27] H. Srivastava, S. Srivastava, S. Kumari, and V. Varshney, "A Novel Hierarchical BERT Architecture for Sarcasm Detection," *Proceedings of the Second Workshop on Figurative Language Processing*, 2020, pp. 93–97.

[28] J. Nie, P. Du, and Z. Lu, "VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification," *European Conference on Information Retrieval (ECIR 2020)*, 2020, pp. 369-382.

[29] C. Zhang and M. Abdul-Mageed, "BERT-Based Arabic Social MediaAuthor Profiling," *The 11th meeting of the Forum for Information Retrieval Evaluation (FIRE 2019)*, 2019, pp. 1-8.

[30] M. Alrabiah, and N. Alsaaran, "Arabic Named Entity Recognition: A BERT-BGRU Approach," *Computers, Materials and Continua Journal*, vol. 680, no. 1, 2021, pp. 471-485.

[31] M. Mouhoub, S. Alsafari, and S. Sadaoui, "Hate and Offensive Speech Detection on Arabic Social Media," *Online Social Networks and Media Journal*, vol. 19, 2020, pp. 1-15.

[32] M. Mouhoub, S. Alsafari, and S. Sadaoui, "Deep Learning Ensembles for Hate Speech Detection," *IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020, pp. 526-531.

[33] L. Yue, M. Yin, W. Zuo, W. Chen, and X. Li, "A Survey of Sentiment Analysis in Social Media," *Knowledge and Information Systems Journal*, vol. 60, 2019, pp. 617–663.