Sentiment analysis of Arabic Reviews

Team Members:

- **Ahmed Essam Abdel-aleem - 900193476**
- **Iman Ahmed Attia - 900192510**

The proposed problem statement

Arabic sentimental analysis using deep learning is a field of research that aims to develop algorithms and models to automatically classify Arabic text based on its sentiment, i.e., whether the text expresses a positive, negative, or neutral sentiment. It is a very crucial problem to solve as it has a variety of applications, such as monitoring social media sentiment, analyzing customer reviews, and more. We aim to focus on the Arabic Customer Reviews Analysis and attempt to solve it.

Motivations to work on sentiment analysis

Arabic sentimental analysis can be used to automatically analyze customer reviews of products and services. This can help businesses identify common complaints and issues, as well as areas for improvement.

Further, there is a growing need for Arabic language processing tools and resources, as Arabic is the fifth most spoken language in the world. Developing effective Arabic sentimental analysis models can help improve the overall quality of Arabic language processing tools and resources.

In Academic research, Arabic sentimental analysis is a growing field, and implementing a project in this area can contribute to the development of new techniques and methods for Arabic natural language processing and sentiment analysis.

Overall, implementing an Arabic sentimental analysis project using deep learning can have practical applications in several domains, as well as contribute to the advancement of the field of Arabic natural language processing.

Input/Output examples To explain the problem statement

The input is mainly a string that presents a review of a specific book or hotel written in the Arabic language. **The output** will be the predicted rating of this review based on the sentiment analysis for the words in the review. This rating will range from 1 to 5 such that rate-1 means that it’s really bad while rate-5 means that it’s really good.

Example:

Input: “هذا من أسوأ الكتب التي قرأتها في حياتي. لم أستطع تكملة الكتاب من شدة الملل”

Predicted output: 1

Explanation: since the model can detect some negative words like “أسوأ”, “لم أستطع”, “الملل“. These words can indicate that this review is strongly negative. Of course, it’s not that easy when it comes to implementation, and there are many interleaved features that will be explained later on with the selected model.

A survey of available evaluation metrics/tools for this problem.

In the following lines, we introduce some common evaluation metrics/tools that are already used in our project under Arabic sentiment analysis using deep learning with their related papers that used them.

- In their paper "Arabic Sentiment Analysis: A Survey" published in 2019, Sawalha and Saraireh reviewed various evaluation metrics that have been used in Arabic sentiment analysis research (Sawalha & Saraireh, 2019). They noted that the most commonly used metrics for binary classification tasks like sentiment analysis are **accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).**

- In their paper "Deep Learning for Arabic Sentiment Analysis: A Comparative Study" published in 2020, Al-Dahoud and Al-Smadi used several evaluation metrics to compare the performance of different deep learning models for Arabic sentiment analysis on the HARD dataset (Al-Dahoud & Al-Smadi, 2020). The evaluation metrics used in the paper included the same metrics for binary classification in the previous paper . Moreover, they used the **mean squared error (MSE)** for regression-based models. The authors used a **10-fold cross-validation** technique to evaluate the models and reported the average values of the evaluation metrics across the folds.

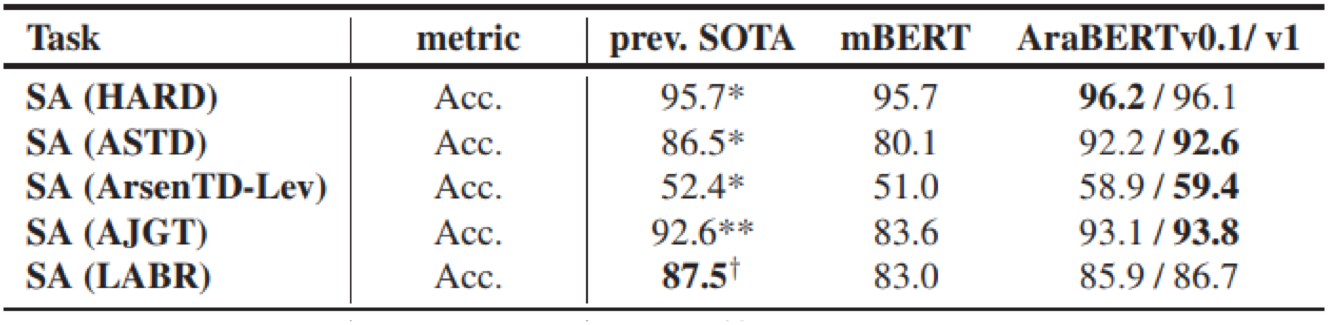
These evaluation metrics/tools can be used to assess the performance of different deep learning models for Arabic sentiment analysis, and to compare their effectiveness.

Current state-of-the-art (SOTA) results

Paper Title: AraBERT: Transformer-based Model for Arabic Language Understanding

Description: AraBERT is a pre-trained transformer-based model that was specifically designed for Arabic natural language processing tasks, including sentimental analysis. AraBERT was able to perform well in both Modern Standard Arabic (MSA) and different Arabic dialects. In 2020, a team of researchers achieved state-of-the-art results on several Arabic sentimental analysis benchmarks using AraBERT.

The table shown below highlights the performance of AraBERT on Arabic downstream tasks or different datasets compared to the multilingual BERT model (mBERT) and previous state-of-the-art systems. [1]



Task	metric	prev. SOTA	mBERT	AraBERTv0.1/ v1
SA (HARD)	Acc.	95.7*	95.7	**96.2** / 96.1
SA (ASTD)	Acc.	86.5*	80.1	92.2 / **92.6**
SA (ArsenTD-Lev)	Acc.	52.4*	51.0	58.9 / **59.4**
SA (AJGT)	Acc.	92.6**	83.6	93.1 / **93.8**
SA (LABR)	Acc.	**87.5**†	83.0	85.9 / 86.7

(*) represents the results of the paper "Arabic sentiment classification using convolutional neural network and differential evolution algorithm" [2]

(** and †) represent the results of the paper "Multi-channel embedding convolutional neural network model for Arabic sentiment classification" [3]

A short survey of available datasets

In the following lines, we will introduce a brief overview of some available Arabic sentiment analysis datasets and their characteristics:

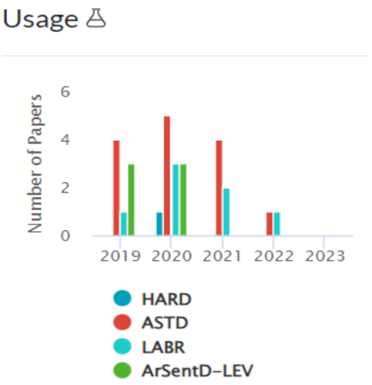
1. **The Hotel Arabic-Reviews Dataset (HARD)** contains 93700 hotel reviews in the Arabic language. The hotel reviews were collected from the Booking.com website during June/July 2016. The reviews are expressed in Modern Standard Arabic as well as dialectal Arabic. [4]

2. **Arabic Sentiment Tweets Dataset (ASTD)** is an Arabic social sentiment analysis dataset gathered from Twitter. It consists of about 10,000 tweets which are classified as objective, subjective positive, subjective negative, and subjective mixed. [5]

3. **The Arabic Sentiment Twitter Dataset (ArSenTD-LEV)** is a dataset of 4,000 tweets with the following annotations: the overall sentiment of the tweet, the target to which the sentiment was expressed, how the sentiment was expressed, and the topic of the tweet. [6]

4. **Large-Scale Arabic Book Reviews (LABR)** is a large sentiment analysis dataset to date for the Arabic language. It consists of over 63,000 book reviews, each rated on a scale of 1 to 5 stars. [7]

The attached chart indicates an approximate number of open-access papers monitoring the dataset in the last five years.



A DETAILED DESCRIPTION OF THE DATASET THAT WE SELECTED TO BE USED.

Hotel Arabic-Reviews Dataset (HARD) & Large-Scale Arabic Book Reviews (LABR) are chosen to be the perfect choices to train our project under Arabic sentiment analysis for several reasons:

- **Domain-specific:** The HARD dataset consists of Arabic hotel reviews while the LABR consists of Arabic book reviews. This is useful for developing sentiment analysis models for hotels, travel-related applications, and reading monitoring.

- **Large and diverse:** The HARD dataset consists of more than **90,000 hotel reviews in 41.4 MB** while the LABR dataset consists of more than **60,000 book reviews in 20.6 MB** written in Arabic, which provides a large and diverse set of data for training and testing sentiment models.

- **High quality:** Both LABR & HARD dataset is manually annotated by human annotators, which ensures the high quality and accuracy of the labels.

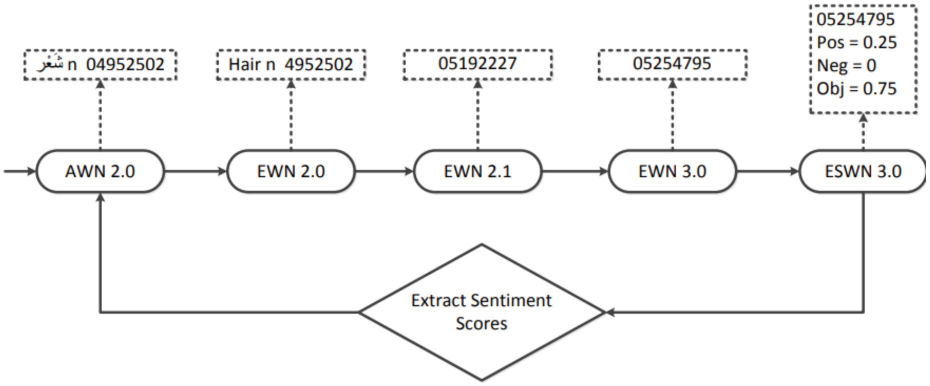
- **Widely used:** Both LABR & HARD dataset has been used in academic research and industry, which makes them well-established and well-tested benchmark datasets. This also means that there are many existing models and techniques that have been trained and tested on those datasets, which can serve as a reference for comparison and evaluation.

Overall, HARD & LABR datasets provide domain-specific, large, diverse, high-quality, and widely used sets of labeled data for training and testing sentiment analysis models in Arabic. Hence, these were strong reasons to decide to use these two datasets.

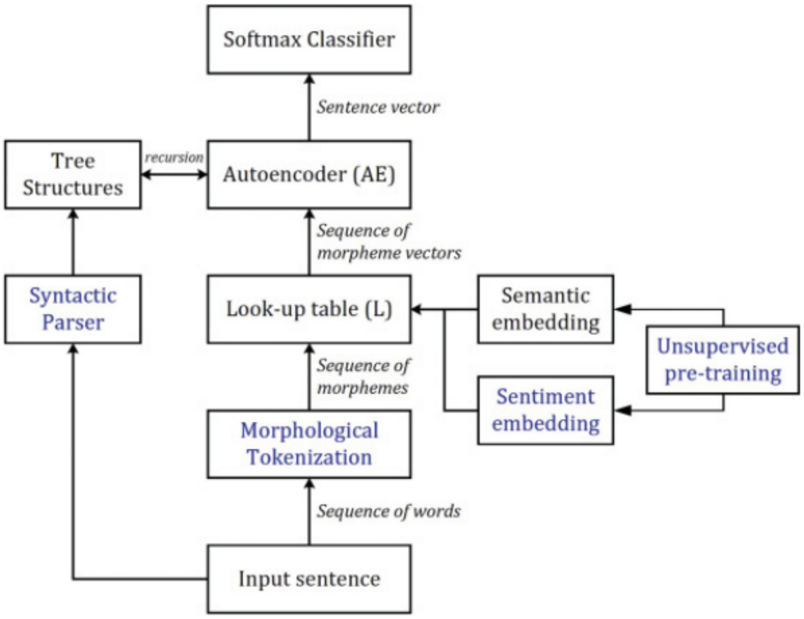
A SHORT SURVEY OF AVAILABLE MODELS AND SOLUTIONS

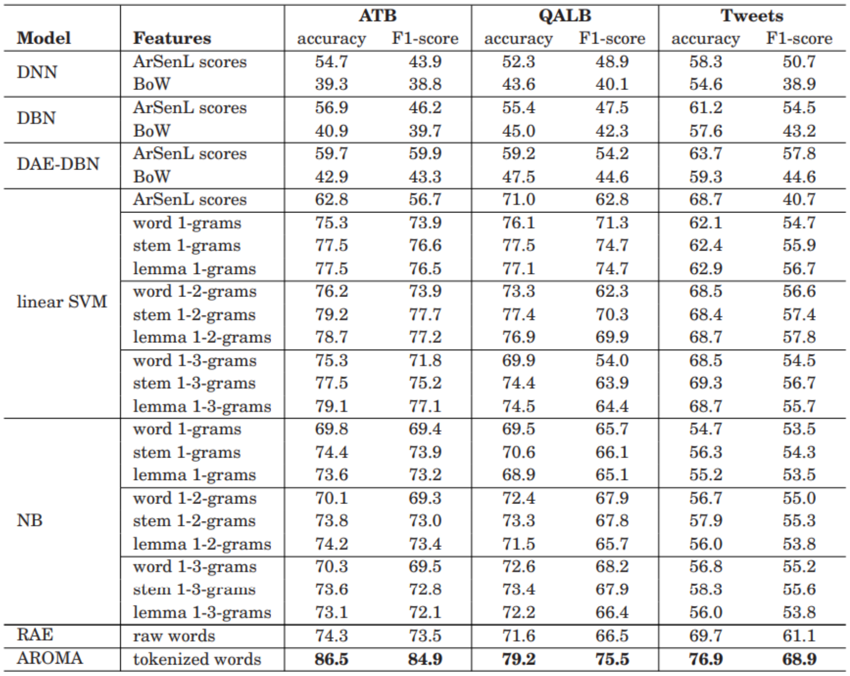
Sentiment Analysis is a popular Arabic NLP task, so here is some of the baselines:

- Previous approaches relied on **sentiment lexicons** such as ArSenL (Badaro et al., 2014), which is a largescale lexicon of MSA words that is developed using the Arabic WordNet (AWN) in combination with the English SentiWordNet (ESWN) [1]. In the figure below, the architecture used to map AWN to the ESWN in order to analyze the sentiment of the word “شعر” meaning hair.



- **Recurrent and recursive neural networks** were explored with different choices of Arabic-specific processing (Al Sallab et al., 2015; Al-Sallab et al., 2017; Baly et al., 2017). In the figure below is the Architecture of the AROMA: A Recursive Deep Learning Model for Opinion Mining propsed by Al-Sallab et al., 2017, followed by a table that contrasts its performance against other Opinion Models from the Literature.





Model	Features	ATB		QALB		Tweets	
		accuracy	F1-score	accuracy	F1-score	accuracy	F1-score
DNN	ArSenL scores	54.7	43.9	52.3	48.9	58.3	50.7
	BoW	39.3	38.8	43.6	40.1	54.6	38.9
DBN	ArSenL scores	56.9	46.2	55.4	47.5	61.2	54.5
	BoW	40.9	39.7	45.0	42.3	57.6	43.2
DAE-DBN	ArSenL scores	59.7	59.9	59.2	54.2	63.7	57.8
	BoW	42.9	43.3	47.5	44.6	59.3	44.6
linear SVM	ArSenL scores	62.8	56.7	71.0	62.8	68.7	40.7
	word 1-grams	75.3	73.9	76.1	71.3	62.1	54.7
	stem 1-grams	77.5	76.6	77.5	74.7	62.4	55.9
	lemma 1-grams	77.5	76.5	77.1	74.7	62.9	56.7
	word 1-2-grams	76.2	73.9	73.3	62.3	68.5	56.6
	stem 1-2-grams	79.2	77.7	77.4	70.3	68.4	57.4
	lemma 1-2-grams	78.7	77.2	76.9	69.9	68.7	57.8
	word 1-3-grams	75.3	71.8	69.9	54.0	68.5	54.5
	stem 1-3-grams	77.5	75.2	74.4	63.9	69.3	56.7
	lemma 1-3-grams	79.1	77.1	74.5	64.4	68.7	55.7
NB	word 1-grams	69.8	69.4	69.5	65.7	54.7	53.5
	stem 1-grams	74.4	73.9	70.6	66.1	56.3	54.3
	lemma 1-grams	73.6	73.2	68.9	65.1	55.2	53.5
	word 1-2-grams	70.1	69.3	72.4	67.9	56.7	55.0
	stem 1-2-grams	73.8	73.0	73.3	67.8	57.9	55.3
	lemma 1-2-grams	74.2	73.4	71.5	65.7	56.0	53.8
	word 1-3-grams	70.3	69.5	72.6	68.2	56.8	55.2
	stem 1-3-grams	73.6	72.8	73.4	67.9	58.3	55.6
	lemma 1-3-grams	73.1	72.1	72.2	66.4	56.0	53.8
RAE	raw words	74.3	73.5	71.6	66.5	69.7	61.1
AROMA	tokenized words	**86.5**	**84.9**	**79.2**	**75.5**	**76.9**	**68.9**

- **Convolutional Neural Networks (CNN)** was trained with pre-trained word embeddings (Dahou et al., 2019a).

- **A hybrid model** was proposed by (Abu Farha and Magdy, 2019), where CNNs were used for feature extraction, and LSTMs were used for sequence and context understanding.
- **The hULMonA model** (ElJundi et al., 2019), which is an Arabic language model that is based on the ULMfit architecture (Howard and Ruder, 2018).

Model	Explanation Summary
Pre-trained language models (AraBERT & mBERT)	These models can be fine-tuned on specific sentiment analysis tasks, allowing for efficient training and improved performance.
Convolutional neural networks (CNNs)	CNNs can be trained to automatically learn relevant features from input, making them suitable for sentiment analysis tasks. [8]
Recurrent neural networks (RNNs)	RNNs have also been used for Arabic sentimental analysis, with some studies reporting improved performance compared to CNNs. Bidirectional RNNs (BiRNNs) and long short-term memory (LSTM) networks are popular choices for Arabic sentiment analysis. [9]

A detailed description of the model to be used from literature to build on

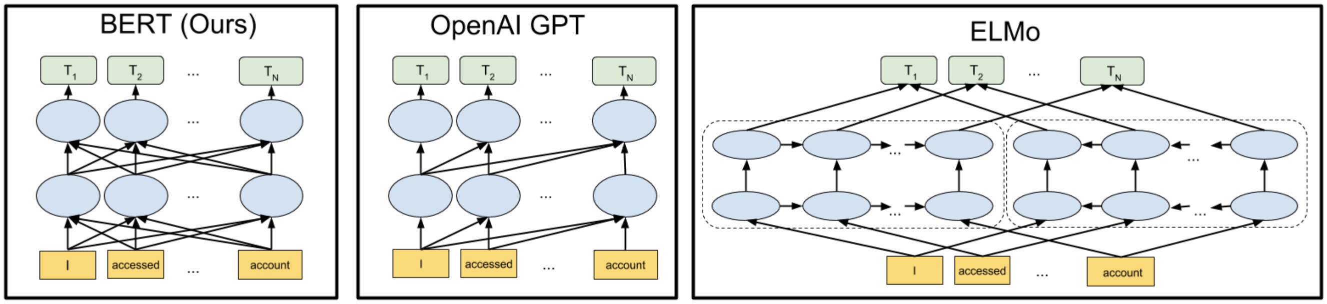
We are going to choose the Arabic Bidirectional Encoder Representations from the Transformers (AraBERT) Model. It is an Arabic pre-trained language model based on Google's BERT architecture. AraBERT uses the same BERT-Base config.

BERT builds upon recent work in pre-training contextual representations — including Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit. However, unlike these previous models, BERT is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus.

There are two versions of the model, AraBERTv0.1 and AraBERTv1, with the difference being that AraBERTv1 uses pre-segmented text where prefixes and suffixes were split using the Farasa Segmenter. We plan to experiment with both versions, however, our starting point, for now, will be to use the AraBERTv1 Base model.

Architecture of the model

Shown below is the BERT's neural network architecture is visualized in comparison to previous state-of-the-art contextual pre-training methods. The arrows indicate the information flow from one layer to the next. The green boxes at the top indicate the final contextualized representation of each input word [10].



The details of the different AraBERT Variations Architectures are displayed in the below table[11].

	AraBert	ArabicBERT Mini	ArabicBERT Medium	ArabicBERT Base	ArabicBERT Large
Hidden Layers	12	4	8	12	24
Attention heads	12	4	8	12	16
Hidden size	768	256	512	768	1024
Parameters	110M	11M	42M	110M	340M

WHY CHOOSING THIS PARTICULAR MODEL: AraBERT ?

AraBERT is considered to be one of the best models for Arabic sentiment analysis for several reasons:

1. **Pre-trained on a large Arabic corpus:** AraBERT has been pre-trained on a large corpus of Arabic text, which includes various dialects and topics. This pre-training enables the model to learn general representations of the Arabic language that can be fine-tuned on specific NLP tasks with smaller amounts of labeled data.

2. **Achieved state-of-the-art performance:** AraBERT has achieved state-of-the-art performance on several Arabic NLP benchmarks, including sentiment analysis. This indicates that the model is effective at learning and representing sentiment in Arabic text.

3. **Handles Arabic morphology and dialectal variation:** AraBERT has been specifically designed to handle the unique challenges of the Arabic language, including its rich morphology and dialectal variation. The model incorporates a modified tokenizer that can handle Arabic morphological complexity and has been trained on a diverse range of dialects, making it more robust to dialectal variation.

4. **Available and accessible:** AraBERT is open source and freely available for download (Link), making it accessible to researchers and developers who want to use it for their own Arabic NLP projects. Starting on the 17th of July 2022: The AraBERT model can be installed via `pip install arabert`.

5. **Can be easily updated/modified:** the models can be fine-tuned on a wide variety of NLP tasks in a few hours or less. The open-source release also includes code to run pre-training.

Overall, AraBERT's combination of pre-training on a large corpus of Arabic text, state-of-the-art performance on Arabic NLP benchmarks, and handling of Arabic morphology and dialectal variation makes it an excellent choice for our proposed project: Sentiment Analysis for Arabic Reviews.

How you will evaluate results, what kind of evaluation metric you will use

To evaluate the results for this specific problem, we can use a combination of different evaluation metrics, such as **F1 score, precision, recall, and ROC-AUC curve.** These metrics can help to assess the performance of the model by capturing the nuances of the sentiment expressed in Arabic text and providing a more comprehensive evaluation of the results.

For example, the F1 score is the harmonic mean of precision and recall, which can provide a good balance between precision and recall in sentiment analysis. The ROC-AUC curve is another evaluation metric that can provide insights into the trade-off between the true positive rate and the false positive rate.

To point out the comparison results, **confusion matrices** can be used to visualize the performance of the model, showing the distribution of predicted labels versus actual labels. **Heat maps and scatter plots** can also be used to provide a visual representation of the sentiment analysis results.

The model weights URL

The model weights are open-sourced and accessible for both Tensorflow and PyTorch models: (through this Link) [12]

The proposed updates to the literature model.

The updates we plan to do are:

Fine-tuning or updating the hyper parameters of the Pre-trained AraBERT model for our specific Sentiment analysis task for assessing the Arabic Reviews. This update will be beneficial as it will be customized to the two selected datasets (Hard and LABR) which may result in better accuracy and higher performance in general with the analyzing the reviews.

Your graduation project brief problem statement

Iman is currently taking a thesis I course this semester. The thesis problem statement is a holographic AI interactive assistant for AUC helpdesks. The thesis problem is completely different from this proposed problem. In the thesis project, we will tackle the following topics: chatbots, computer vision, TTS, STT, and Holography. However, in this proposal, the project will mainly tackle Arabic sentiment analysis with a whole different scope, dataset, and model.

A List of all other machine learning, deep learning, computer vision, natural language processing, pattern recognition or any related data science field past projects

Iman Machine Learning and Computer Vision Past projects:

- Music Genre Classifier using Artificial Neural Networks ANNs.
- Live Hand Tracking System using Python, OpenCV, and MediaPipe.

The past projects clearly fall under different domains than our proposed project.

Each team member contributions

Iman's contribution:

- Constructed the proposed problem statement along with the motivation.
- Current state-of-the-art results for the proposed problem.
- A survey of available models and solutions for the proposed problem.
- A detailed description of the baseline model to be used from literature to build on, also mention why you will use this particular model.
- Propsed the updates that can be made to the proposed model.

Ahmed's contribution:

- Searched for the inputs and outputs formats that will match the proposed problem statement.
- Researched the survey of used evaluation metrics for a similar problem.
- Chose suitable evaluation matrices for our project with reasons why so.
- Listed all of the available Arabic sentiment analysis datasets.
- Illustrated a detailed description of the LABR and HARD datasets.

Refernces

[1] Antoun W., Baly F.,Hajj H.,(2020). AraBERT: Transformer-based Model for Arabic Language Understanding. Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France.

[2] Dahou, A., Elaziz, M. A., Zhou, J., and Xiong, S. (2019a). Arabic sentiment classification using convolutional neural network and differential evolution algorithm. Computational intelligence and neuroscience, 2019.

[3] Dahou, A., Xiong, S., Zhou, J., and Elaziz, M. A. (2019b). Multi-channel embedding convolutional neural network model for arabic sentiment classification. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(4):1–23.

[4] Papers with code - hard dataset. HARD Dataset | Papers With Code. (n.d.). Retrieved February 22, 2023, from https://paperswithcode.com/dataset/hard

[5] Papers with code - ASTD dataset. ASTD Dataset | Papers With Code. (n.d.). Retrieved February 22, 2023, from https://paperswithcode.com/dataset/astd

[6] Papers with code - ArSentD-LEV dataset. ArSentD-LEV Dataset | Papers With Code. (n.d.). Retrieved February 22, 2023, from https://paperswithcode.com/dataset/arsentd-lev

[7] Papers with code - LABR dataset. LABR Dataset | Papers With Code. (n.d.). Retrieved February 22, 2023, from https://paperswithcode.com/dataset/labr

[8] Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., Almekhlafi, E., & Retta, E. A. (2021, September 11). Improving Arabic sentiment analysis using CNN-based architectures and text preprocessing. Computational Intelligence and Neuroscience. Retrieved February 23, 2023, from https://www.hindawi.com/journals/cin/2021/5538791/#abstract

[9] Arabic sentiment analysis using recurrent neural networks: A Review. (n.d.). Retrieved February 23, 2023, from https://www.researchgate.net/publication/351008137_Arabic_sentiment_analysis_using_recurrent_neural_networks_a_review

[10] J. Devlin and M.-W. Chang, "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing," *Google Research*, 02-Nov-2018.

[11] Al-Twairesh,N. The Evolution of Language Models Applied to Emotion Analysis of Arabic Tweets. Information 2021,12, 84.

[12] W. Antoun, F. Baly, and H. Hajj, "Arabert : Pre-training bert for Arabic language understanding," AUB MIND Lab, 28-Feb-2020. [Online]. Available: https://sites.aub.edu.lb/mindlab/2020/02/28/arabert-pre-training-bert-for-arabic-language-understanding/. [Accessed: 22-Feb-2023].

[13] Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El-Hajj, W., and Shaban, K. (2015). Deep learning models for sentiment analysis in arabic. In Proceedings of the second workshop on Arabic natural language processing, pages 9–17.

[14] Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., and Badaro, G. (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):1–20.

[15] Baly, R., Hajj, H., Habash, N., Shaban, K. B., and El-Hajj, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. ACM Transactions on Asian and LowResource Language Information Processing (TALLIP), 16(4):1–21.

[16] Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. In Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP), pages 165–173.

[17] Abu Farha, I. and Magdy, W. (2019). Mazajak: An online Arabic sentiment analyser. In Proceedings of the Fourth 14 Arabic Natural Language Processing Workshop, pages 192–198, Florence, Italy, August. Association for Computational Linguistics.

[18] ElJundi, O., Antoun, W., El Droubi, N., Hajj, H., El-Hajj, W., and Shaban, K. (2019). hulmona: The universal language model in arabic. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 68–77.

[19] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.