AraBERT and mBert: Insights from Psycholinguistic Diagnostics

Basma Sayah
Lab. d'Informatique & Mathématiques
Université Amar Telidji, Algeria
b.sayah@lagh-univ.dz

Attia Nehar
Lab. d'Informatique & Mathématiques
Ziane Achour University, Algeria
neharattia@univ-djelfa.dz

Hadda Cherroun
Lab. d'Informatique & Mathématiques
Université Amar Telidji, Algeria
hadda_cherroun@lagh-univ.dz

Slimane Bellaouar
Lab. Mathématiques & Sciences
University of Ghardaia, Algeria
bellaouar.slimane@univ-ghardaia.dz

Abstract

BERT, a groundbreaking large language model, has excelled in natural language processing tasks such as question answering. Motivated by a desire to understand BERT's knowledge and limitations across different languages, we build upon Alysson Ettinger's work by evaluating BERT Arabic versions using psycholinguistics. These diagnostics, designed to assess human brain linguistic abilities, cover aspects like common sense and pragmatic inference, which constitute fundamental knowledge for any pretrained language model. Upon translating these diagnostics into Arabic, the results of diagnostic assessments for mBERT in Arabic and AraBERT reveal linguistic deficiencies in mBERT and a moderate grasp in AraBERT. This emphasizes the need for further training on diverse texts, especially those related to everyday situations.

Keywords: AraBERT, mBert, Psycholinguistic, Linguistic evaluation, Arabic language models.

1 Introduction

Nowadays, large Language Models (LLMs) are the base of almost every Natural Language Processing (NLP) application. They are used in sentiment analysis (SA), question answering (QA), conversational agents, personal assistants, and robotics (et al., 2021).

Since the introduction of Transformers in 2017 (Vaswani et al., 2017), computers have demonstrated remarkable linguistic abilities, often comparable to those of humans. Consequently, a multitude of language models has emerged from the Transformer framework, addressing a variety of languages. Examples include: ELMo (Peters et al., 2018), BERT and mBERT (Devlin et al., 2019), GPT through all its versions (Radford et al., 2018), PaLM (et al., 2023).

Despite the popularity of these models and their impact across various fields, there is an urgent need for interdisciplinary efforts in order to understand the knowledge they infer and to discover their unknown failures. Previous studies have delved into various performance aspects, including task-specific evaluations (Jiang et al., 2021) (Wang et al., 2018), probing different layer (Conia and Navigli, 2022), and linguistics evaluations of humans on machines (Ettinger, 2020) (Lialin et al., 2022).

Unlike other languages, Arabic LLMs have not been extensively studied, despite some recent investigations (Albilali et al., 2021) (Abdelali et al., 2022). In this paper, we aim to fill this gap by investigating the capabilities of Arabic LLMs. Our initial step involves enhancing our understanding of what Arabic LLMs comprehend about the Arabic language by measuring their linguistic abilities through the discipline of psycholinguistics. Psycholinguistics, originally developed by linguists to assess the human brain's capacity to understand and produce language (Harley, 2013), serves as our guiding framework. Our investigation is specifically narrowed down to the Arabic language models araBERT and multilingual BERT

The rest of this paper is organized as follows. First, in Section 2, we introduce some preliminaries and concepts related to pre-trained LMs and psycholinguistic diagnostics. In Section 3, we review related literature that has considered the evaluation of LMs' linguistic abilities. The methodology of our investigation is presented in detail in Section 4. Finally, we report and discuss the results of the evaluation in Section˜ 5

2 Preliminaries

Driven by the purpose of this paper, this section offers a concise overview of Multilingual BERT and AraBERT. Subsequently, we delve into psycholinguistic aspects and psycholinguistic diagnostics that examine predictive human responses, all of which are relevant to the assessment of pre-trained Language Models.

2.1 Arabic LLMs and BERT

Arabic language models, especially those built on the BERT architecture, are pivotal in natural language processing. BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), constitutes a highly parallel deep neural network leveraging attention mechanisms for sequence prediction and generation (Vaswani et al., 2017).

Originally designed for language modeling and machine translation, transformers like BERT have evolved to handle more complex tasks, including computer vision (Nguyen et al., 2023). Specific BERT variations tailored for Arabic have been developed. Figure 1 provides a chronological overview of Arabic BERT models and other transformers pre-trained on diverse Arabic texts, encompassing dialects and Modern Standard Arabic (MSA) from platforms like social media, news, and academic content. For the purpose of this paper, we will focus specifically on two models: mBERT and AraBERT.

mBERT released by Devlin et al.,(2019) is a single-language model that was pre-trained using monolingual corpora in 104 languages, including Arabic. This enabled BERT to learn and generalize across multiple languages.

AraBERT developed by Antoun et al.,(2020) is a widely adopted model pre-trained on an extensive corpus of Modern Standard Arabic (MSA) texts. AraBERT is applied in various natural language processing (NLP) tasks, including text classification, named entity recognition (NER), and sentiment analysis (SA) in the Arabic language.

2.2 Psycholinguistics

Psycholinguistics, a subfield of linguistics, studies the mental processes involved in language acquisition, comprehension, and production (Harley, 2013). Within the domain of psycholinguistics, the study of human language processing incorporates fundamental metrics such as *Cloze probability* and *N400* amplitude (Kutas and Hillyard, 1984).

- Cloze probability is the likelihood or probability that individuals choose a specific word to complete a given context. It provides a quantifiable measure of how well a word fits into a particular linguistic context based on human responses.

- The N400 amplitude is a quantifiable electrical signal discerned in brain activity, particularly in electroencephalogram (EEG) recordings. The measurement of the N400 component's amplitude helps comprehend the brain's reaction to words that disrupt the contextual flow or are unexpected within a given sentence.

3 Related Work

In the literature, there is a growing effort to better understand the specific linguistic capacities achieved by neural Natural Language Processing (NLP) models. We have reviewed several studies that measured their performances and behaviors, categorizing them based on three criteria:

- Linguistic analysis: This category focuses on assessing the lexical, syntactic, and figurative skills of a language model.

- Tasks-based Analysis: This category involves evaluating the language model through specific tasks such as Sentiment Analysis (SA), Question Answering (QA), Translation, Named Entity Recognition (NER), and Dialect Identification.

- In-Depth Model Examination: This type of analysis delves into the inner workings of the model, considering aspects of explainability and probing.

In linguistic analysis, Ettinger, (2020) presents a set of diagnostics derived from human language experiments to systematically investigate the information utilized by language models during prediction generation in context. The study applies these diagnostics to assess the popular BERT model. The findings reveal that BERT demonstrates a general ability to distinguish between good and bad completions involving shared category or role reversal, though with less sensitivity compared to humans.

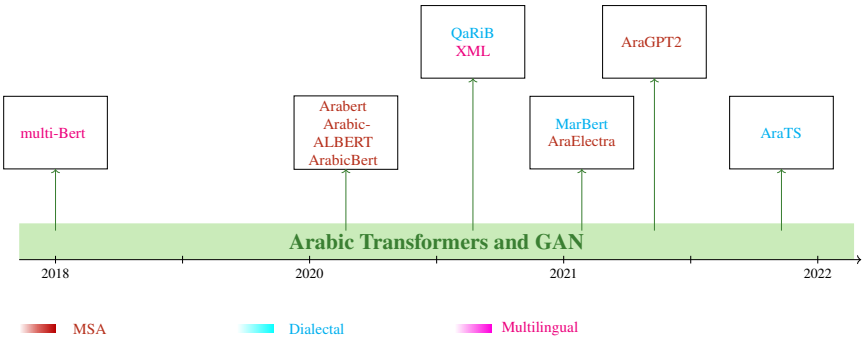


Figure 1: Some Arabic Transformers and GANs.

Additionally, BERT consistently retrieves noun hypernyms effectively. However, the model faces challenges in intricate tasks such as inference and role-based event prediction. Notably, BERT exhibits a clear insensitivity to the contextual impacts of negation.

In task-based analysis, Rönnqvist et al., () investigated mBERT's performance across languages and tasks. They found mBERT to be inferior to monolingual models, especially for Nordic languages. Chouikhi et al., (2021) addressed tokenization issues in Arabic Sentiment Analysis. Their approach, incorporating an Arabic BERT tokenizer instead of the basic BERT tokenizer, outperformed Arabic BERT and AraBERT models in classification quality and accuracy, particularly for dialect and MSA instances. Lialin et al., (2022) scrutinized 29 diverse model families, including T5, BART, and ALBERT, using the oLMpics benchmark and psycholinguistic probing datasets. Their study found that none of these models, when assessed in a zero-shot manner, could effectively address compositional questions, challenging the adequacy of current pre-training objectives for acquiring this skill.

In their in-depth model examination, Mickus et al., (2020) examined the semantic coherence of BERT's embedding space. They mention that, while showing a tendency towards coherence, BERT does not fully live up to the natural expectations for a semantic vector space. They discovered, in particular, that the position of a word in a sentence, despite having no meaning correlates, leaves an evident trace on the word embeddings and disrupts similarity relationships. Li et al., (2021) introduced a tool for probing surprisal at BERT's intermediate layers, employing density estimation with Gaussian models. They found a high correlation between surprisal and low token frequency in lower layers, decreasing in upper layers. Regarding morphosyntactic, semantic, and commonsense anomalies, the best-performing model (RoBERTa) exhibited surprisal in earlier layers for morphosyntactic anomalies, but not for semantic or commonsense anomalies. Abdelali et al., (2022) conducted a post-hoc examination of transformer models trained on diverse Arabic dialects. Using layer and neuron analysis, they found that word morphology is predominantly learned in lower and middle layers, syntactic dependencies are primarily captured in higher layers, and despite vocabulary overlap, models based on Modern Standard Arabic struggle to capture nuanced aspects of dialects. Neurons in embedding layers exhibit polysemous characteristics, while those in middle layers specialize in specific properties.

4 Methodology

In the assessment of the psycholinguistic skills of Arabic BERT models, we translated the psycholinguistic diagnostics from Ettinger's work into Arabic with the assistance of three Arabic native speakers, one of whom is a professional translator. Subsequently, we applied these diagnostics to AraBERTv2$_{base}$, AraBERTv2$_{large}$, and mBERT using the Python language in the Google Colab platform. Each diagnostic test involves sentences (contexts) with a missing word, and the task is to predict that missing word. Accurate predictions require the application of the targeted linguistic skills defined by these tests. The evaluation utilized the following metrics:

- **Word Prediction Accuracy** measures how often the language model correctly provides the expected item among its top k predictions and is designed to be the equivalent of Cloze probability in psycholinguistics (refer to Section 2).

- **Sensitivity Test** represents the percentage of items for which the probability assigned to a correct completion exceeds the probability assigned to the inappropriate one. This measure is designed to be the equivalent of the N400 in psycholinguistics (refer to Section 2).
- **Qualitative analysis** is the process of manually reviewing the results, making observations on the top k predictions, and understanding their relationships with each other and with the context, all in order to gain deeper insights into the skills of AraBERT.

All the diagnostic datasets and experiment code are shared and accessible on GitHub. [1] The following subsection provides a detailed description of the diagnostics employed in our evaluation.

4.1 CPRAG-102

This diagnostic is made up of 102 contexts. Each context comprises two consecutive sentences with a missing word (Federmeier and Kutas, 1999). In these contexts, predicting the missing word requires Common Sense to understand what is being described and Pragmatic Inference to understand how the second sentence relates to the first. Table 2 shows an example of CPRAG-102 and its Arabic translation. The 'Expected' column displays the word most likely to be predicted by humans, taking into account synonyms in our experiments. In contrast, the 'Inappropriate' column lists some incorrect word completions that fall within the same category as the expected word. The inappropriate completion is used to examine whether LMs will prioritize unsuitable completions that share a semantic category with the expected completions.

4.2 ROLE-88

It comprises 88 contexts, with one sentence per context designed to target role reversal. (Chow et al., 2016) illustrated the example in Table 3, "Completing the sentence requires semantic role identification and event knowledge, which means finding the accurate words associated with events and actions to fill in the blank". Although each completion (e.g., 'served') is suitable for only one of the noun orders and not the reverse, we use this diagnostic to test whether Arabic BERT models will face difficulty distinguishing appropriate continuations based on word order and semantic role.

[1] https://github.com/BasmaSayah/Psycholinguistic-Diagnostics-on-AraBERT

4.3 NEG-SIMP-136

This diagnostic targets understanding the meaning of negation and category membership (Fishler et al., 1983). Table 4 presents a negation example along with its corresponding translation. The affirmative sentence allows us to assess the model's capacity to associate nouns with their hypernyms. Through this diagnostic, we investigate the model's ability to distinguish between affirmative and negative sentences, specifically whether it outputs the same word as in the affirmative case, as indicated in the 'match' column, or a different word, as shown in the 'mismatch' column.

4.4 NEG-NAT-136:

This diagnostic targets naturally occurring negative sentences and was derived from a human study conducted by Nieuwland and Kuperberg (2008). Building upon the experiment conducted by Fishler et al.,1983, it involves the creation of affirmative and negative sentences chosen to be more 'natural for somebody to say,' contrasting these with the non-natural affirmative and negative sentences. Table 5 shows an example of NEG-NAT-136 and its Arabic translation.

5 Experiments and Discussion

In this section, we analyze the results of running the diagnostics on AraBERTv2$_{base}$, AraBERTv2$_{large}$, and mBERT. We compare these results with those of the English BERT as presented in the paper "What BERT Is Not". We manually reviewed the results to ensure accuracy and to avoid instances where the language models provided correct answers not present in the diagnostic dataset.

5.1 Results for Common Sense and Pragmatic Inference

Figure 2 illustrates the performance of AraBERT$_{base}$, AraBERT$_{large}$, mBERT, BERT$_{base}$, and BERT$_{large}$ on the CPRAG-102 dataset, in terms of accuracy. It represents the percentage of items for which the 'expected' completion is among the model's top k predictions, with $k \in \{1, 5\}$. For accuracy at $k = 1$, both AraBERT$_{base}$ and mBERT achieved a score of 2.94%. In contrast, AraBERT$_{large}$ achieved more higher accuracy of 8.82% on the same task. On the other hand, BERT$_{base}$ and BERT$_{large}$ performed better with accuracies of 23.5% and 35.3%, respectively. This indicates that

Table 2: Example of CPRAG-102 and its Arabic translation.

Context	Expected	Inappropriate
أرادت أن تجعل رموشها تبدو سوداء وسميكة حقًا.	الماسكارا	أحمر الشفاه ـ قلادة
لذا طلبت من صديقتها أن تعيرها ————	Maskara	lipstick | necklace
She wanted to make her eyelashes look really black and thick.		
So she asked to borrow her older friend's ——		

Table 3: Example of ROLE-88 and its Arabic translation.

Context	Completion
نسي صاحب المطعم أي زبون قام النادل ب ————	خدمته
The restaurant owner forgot which customer the waitress had ——	served
نسي صاحب المطعم أي نادل قام الزبون ب ————	خدمته
The restaurant owner forgot which waitress the customer had ——	served

Table 4: Example of NEG-SIMP-136 and its Arabic translation.

Context	Match	Mismatch
أبو الحنّاء هو ————	طائر	شجرة
A robin is a ——	bird	tree
أبو الحنّاء ليس ————	طائر	شجرة
A robin is not a ——	bird	tree

AraBERT$_{base}$ and mBERT do not perform well in common-sense and/or pragmatic inference tasks, while AraBERT$_{large}$ performs substantially better.

At $k = 5$, mBERT achieved the lowest accuracy of 5.88%, followed by AraBERT$_{base}$, which showed an improvement with an accuracy of 17.6%. AraBERT$_{large}$ achieved the highest accuracy in Arabic, reaching 23.52%. In the English part, both BERT$_{base}$ and BERT$_{large}$ achieved a 52.9% accuracy. The low accuracy scores highlight clear weaknesses in AraBERT's ability to handle common-sense and/or pragmatic inference.

Regarding completion sensitivity, Figure 2 illustrates the performance of AraBERT, mBERT, and BERT on the CPRAG-102 dataset in terms of sensitivity. This metric represents the percentage of items for which the model assigns a higher probability to the expected completion (e.g., 'Maskara,' as shown in Table2) than to any of the inappropriate completions (e.g., 'lipstick' or 'necklace'). mBERT assigns the highest probability to the expected completion only 5.88% of the time, whereas AraBERT$_{base}$ and AraBERT$_{large}$ achieve this 17.65% and 20.59% of the time, respectively. On the contrary, BERT$_{base}$ and BERT$_{large}$ exhibit a high sensitivity of 73.5% and 79.4% with English. This suggests that both versions of AraBERT and mBERT do not exhibit sensitivity in differentiating between good and bad completions within the same semantic category, with AraBERT noticeably better than the latter.

Upon introducing the threshold on the probability difference, mBERT's sensitivity remains the same, while AraBERT$_{base}$ and AraBERT$_{large}$ sensitivity drop slightly to 14.7% and 17.65%, respectively. In contrast, BERT$_{base}$ and BERT$_{large}$ sensitivity drop drastically to 44.1% and 58.8%. This still indicates that AraBERT lacks sensitivity in distinguishing between good and bad completions, whereas BERT$_{base}$ and BERT$_{large}$ exhibit some sensitivity, albeit with a small probability difference.

The qualitative analysis of the sentences where AraBERT$_{large}$ has made incorrect predictions shows that AraBERT$_{large}$ fails not only in one but in both common sense and pragmatic inference. In the phrase

أراد بابلو قطع الخشب الذي اشتراه لصنع بعض الرفوف. سأل جاره إذا كان بإمكانه أن يعيره

meaning 'Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her', AraBERT$_{large}$ predicted words related to wood but did not suggest 'saw.' This suggests that it recognized that the word to be predicted was related to the first sentence, succeeding in pragmatic inference, but failed to recognize what it was, indicating a failure in common sense understanding. In the phrase

إنها تستمر في تدويرها و تدويرها حول عنقها. يبدو أن ستيفاني سعيدة حقًا لأن دان أعطاها ذلك ————

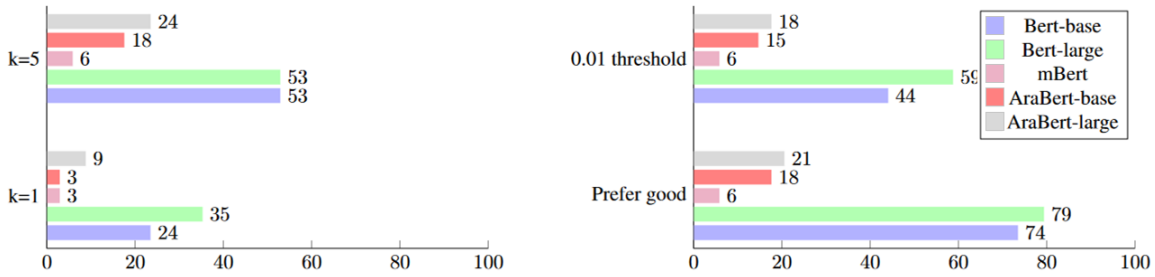
meaning "She keeps twirling it around and around under her collar. Stephanie seems really happy that Dan gave her that ——', AraBERT$_{large}$ predicts the words 'place' and 'time.' This indicates that it only used the second sentence for predictions, failing in pragmatic inference.

5.2 Results for role reversals and event knowledge

As demonstrated in Figure 3 when $k = 1$, mBERT exhibits poor performance (a 0% accuracy). This

Table 5: Example of NEG-NAT and its Arabic translation.

Context	target_aff	target_neg
مع المعدات المناسبة، يعد الغوص تحت الماء ———	آمن	خطير
With proper equipment, scuba-diving is very ———	safe	dangerous
مع المعدات المناسبة، لا يعد الغوص تحت الماء ———	آمن	خطير
With proper equipment, scuba-diving isn't very ———	safe	dangerous



(a) Accuracy. (b) Sensitivity.

Figure 2: Performances of BERT, mBERT and AraBERT on the CPRAG-102 dataset.

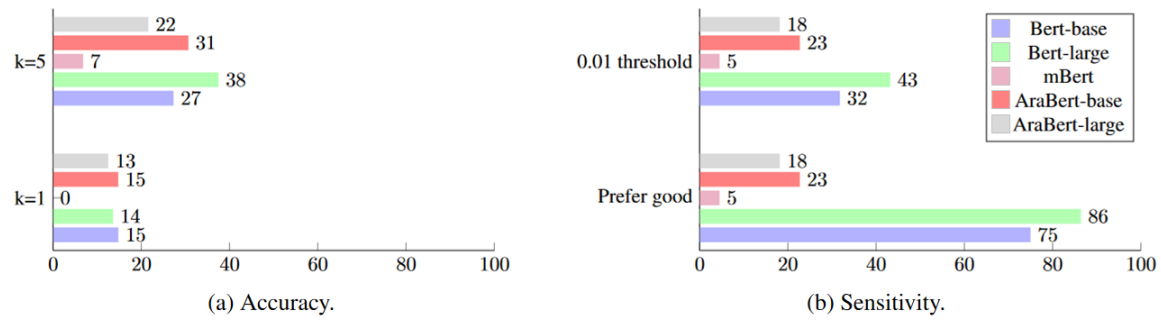
suggests that mBERT is not suitable for role reversal and event knowledge tasks. In contrast, BERT$_{base}$ and AraBERT$_{base}$ show similar accuracies, both approximately at 14.8%. However, AraBERT$_{large}$ and BERT$_{large}$, despite their larger architectures, achieve slightly lower accuracies of 13.6% and 12.5%, respectively. These results indicate that AraBERT$_{large}$ and BERT$_{large}$ may benefit from further fine-tuning tailored to these tasks, emphasizing that model size alone does not guarantee enhanced performance.

When $k = 5$, mBERT still lags with an accuracy of 6.81%. AraBERT$_{base}$ and AraBERT$_{large}$ both demonstrate improved accuracies, with AraBERT$_{base}$ surprisingly surpassing AraBERT$_{large}$. AraBERT$_{base}$ achieves an accuracy of 30.68%, while AraBERT$_{large}$ achieves 21.59%. Although AraBERT$_{base}$ and AraBERT$_{large}$ exhibit better performance than mBERT for $k = 5$, they are still outperformed by the English-language models BERT$_{base}$ and BERT$_{large}$, which achieved accuracies of 27.3% and 37.5%, respectively. Considering a larger number of predictions enhances accuracy for all models. However, English-language models BERT$_{base}$ and BERT$_{large}$ consistently outperform the multilingual and Arabic-specific models in this task.

Figure 3 illustrates the sensitivity of BERT models to role reversals. mBERT performs poorly for Arabic, exhibiting a sensitivity of only 4.54%. AraBERT$_{base}$ and AraBERT$_{large}$ show moderate sensitivity, with AraBERT$_{base}$ at 22.72% and AraBERT$_{large}$ at 18.18%. In contrast, BERT$_{base}$ and BERT$_{large}$ demonstrate high sensitivity to "good completions" with accuracies of 75% and 86.4%, respectively.

After introducing the threshold of 0.01, mBERT maintains a low sensitivity of 4.54%. AraBERT$_{base}$ and AraBERT$_{large}$ also maintain their sensitivities at 22.72% and 18.18%, respectively, while BERT$_{base}$ and BERT$_{large}$ maintain relatively higher sensitivities at 31.8% and 43.2%, respectively. Overall, the results suggest that mBERT for the Arabic language is not well-suited for role reversals and/or event knowledge tasks. The moderate sensitivity of AraBERT models indicates their ability to identify "good completions" to some extent. In contrast, the English-language models, particularly BERT$_{large}$, exhibit better performance in these tasks, highlighting potential challenges in adapting these models for Arabic language tasks or the need for further fine-tuning.

During manual analysis of sentences where mBERT, AraBERT$_{base}$ and AraBERT$_{large}$ failed, all models frequently produced the unknown token, indicating challenges in generating predictions for the given contexts. In cases where words were generated, mBERT's predictions often lacked coherence and didn't make sense whereas AraBERT$_{base}$ and AraBERT$_{large}$ produced logically consistent predictions that differed from those generated for their role-reversed versions of the sentence. This suggests that mBERT struggles with producing meaningful predictions. Conversely, the limitations



(a) Accuracy. (b) Sensitivity.

Figure 3: Performances of BERT, mBERT and AraBERT on the ROLE-88 dataset.

of the AraBERT models appear to be primarily related to event knowledge, as they generate words that are logically consistent within the sentence context but struggle to predict the accurate word related to the event or action. Furthermore, AraBERT$_{large}$ showed more uncertainty than AraBERT$_{base}$, indicating AraBERT$_{large}$ need for additional context. However, its responses were grammatically more accurate compared to AraBERT$_{base}$.

5.3 Results for negation understanding

Figure 4 illustrates the accuracy of BERT models in predicting affirmative and negative sentences. Affirmative sentences were used to evaluate BERT's ability to associate nouns with their hypernyms(categories). When we examine the accuracy scores for affirmative sentences, we see mBERT achieved the lowest accuracy, scoring 0%, indicating it is not suitable for category membership prediction. Both AraBERT$_{base}$ and AraBERT$_{large}$ achieved accuracies of 44.44% and 33.33% respectively, in predicting category membership. While English-language models BERT$_{base}$ and BERT$_{large}$ achieved a perfect accuracy of 100%. This suggests that AraBERT models are less effective in category membership prediction compared to the highly effective English models.

In the case of negative sentences, mBERT achieved an accuracy of 0%, which is evident given that it also failed with affirmative sentences. AraBERT$_{large}$ also achieved an accuracy of 0% in understanding negation, similar to BERT$_{base}$ and BERT$_{large}$. This result suggests these models' failure to understand negations. On the other hand, AraBERT$_{base}$ achieved a relatively low accuracy of 5.55% in understanding negation. The correct results it obtained may be attributed to its potential understanding of negation or pattern recognition.

After checking predictions manually, AraBERT mostly gives the same predictions for positive and negative sentences, except for one sentence which is

السلمون المرقط من مجموعة _____

meaning 'A trout is——-' AraBERT$_{base}$ predicted "fish" as a first prediction for the affirmative statement but provided a different answer, "chicken," for the negative statement. AraBERT$_{large}$, on the other hand, did not exhibit this distinction.

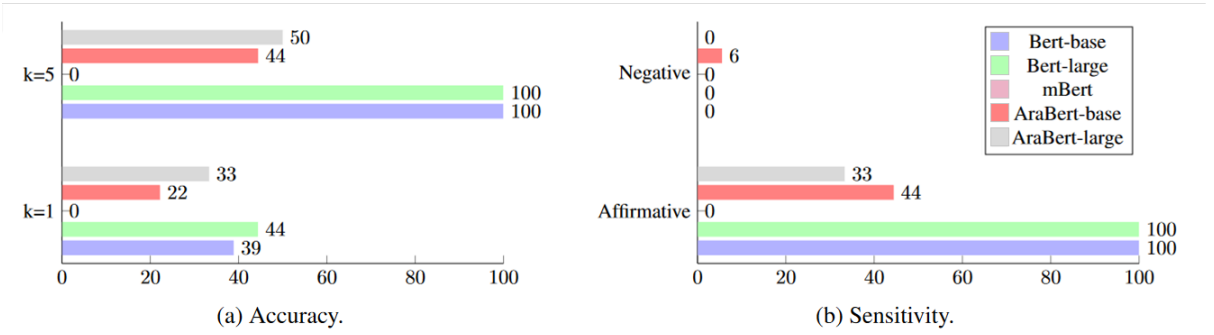
Figure 5 Illustrates the accuracies of BERT models for natural affirmative and negative sentences, with the distinction that these affirmative sentences do not test category membership. Regarding affirmative sentences, AraBERT$_{large}$ emerges as the top-performing model in this context, achieving an accuracy of 87.5%, closely followed by BERT$_{large}$ with an accuracy of 75%. BERT$_{base}$ and AraBERT$_{base}$ achieved moderate accuracies of 62.5% and 68.75%, respectively. In contrast, mBERT failed to make any correct prediction, yielding an accuracy of 0%, indicating its instability in making predictions.

Turning to negative sentences, AraBERT$_{base}$ and AraBERT$_{large}$ showed moderate performance, achieving accuracies of 43.75% and 50%, respectively, while BERT$_{base}$ and BERT$_{large}$ demonstrated strong performance with accuracies of 87.5% and 100%.

When examining the top Predictions of AraBERT$_{large}$, they all align with each other and do not contradict each other, whether for affirmative or their corresponding negative sentences, This consistency suggests that there is an opportunity to improve how the model handles negation.

6 Conclusion

In this study, we examined the capabilities of multilingual BERT for Arabic, as well as AraBERT base and large versions, using psycholinguistics. While AraBERT is better than Multilingual BERT, it has



(a) Accuracy. (b) Sensitivity.

Figure 4: Performances of BERT, mBERT, and AraBERT on the NEG-SIMP-136 dataset.

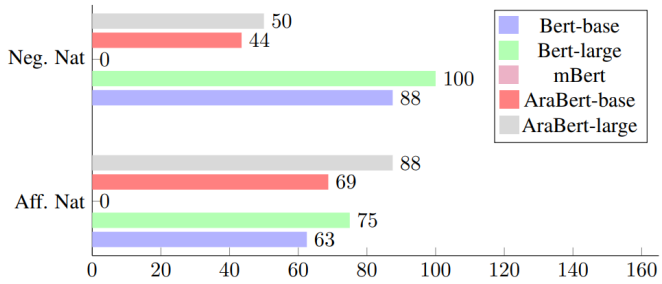


Figure 5: Performances of BERT, mBERT, and AraBERT on the NEG-NAT-136 dataset.

notable weaknesses in common sense and pragmatic inference. In this task, the large version consistently outperforms the base version of AraBERT. Additionally, AraBERT faces challenges in recognizing words related to events and actions, where the base version consistently outperforms the large version. In negation tasks, both AraBERT models often struggle to distinguish affirmative from negative sentences, except in rare cases, marking an improvement compared to English BERT models that do not make this distinction at all. All models perform well with natural negative sentences, likely relying on pattern recognition rather than a deep understanding of negation cues. This situation presents opportunities for enhancing language models' grasp of negation. Further research is needed to fully understand each model's strengths and weaknesses, facilitating more informed decisions when choosing a language model.

References

Ahmed Abdelali, Fahim Dalvi, Hassan Sajjad, and Nadir Durrani. 2022. Post-hoc analysis of Arabic transformer models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–103, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eman Albilali, Nora Altwairesh, and Manar Hosny. 2021. What does bert learn from arabic machine reading comprehension datasets? In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 32–41, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Hasna Chouikhi, Hamza Chniter, and Fethi Jarray. 2021. Arabic Sentiment Analysis Using BERT Model. In *Advances in Computational Collective Intelligence*, pages 621–632. Springer International Publishing.

Wing Yee Chow, Ellen Lau, Colin Phillips, and Cybelle Smith. 2016. A 'bag-ofarguments' mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5):577–596.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aakanksha Chowdhery et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kara Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Ira Fishler, Donald Childers, Salim Roucos, Nathan Perry, and Paul Bloom. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

Trevor Harley. 2013. *The psychology of language: From data to theory*. Psychology press.

Zhengbao Jiang, Haibo Ding, Graham Neubig, Zhengbao Jiang, and Jun Araki. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Marta Kutas and Steven Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Bai Li, Guillaume Thomas, Yang Xu, Frank Rudzicz, and Zining Zhu. 2021. How is BERT surprised? layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–28. Association for Computational Linguistics.

Vladislav Lialin, Kevin Zhao, Namrata Shivagunde, and Anna Rumshisky. 2022. Life after BERT: What do other muppets understand about language? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3180–3193, Dublin, Ireland. Association for Computational Linguistics.

Timothee Mickus, Mathieu Constant, Kees van Deemter, and Denis Paperno. 2020. What do you mean, bert? assessing bert as a distributional semantics model. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch1, Han-Seok Seo, and Khoa Luu. 2023. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492.

Mante Nieuwland and Gina Kuperberg. 2008. When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218.

Matthew Peters, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, and Mark Neumann. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the NAACL, Vol. 1*, pages 2227–2237, New Orleans, Louisiana. ACL.

Alec Radford, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. Is multilingual BERT fluent in language generation?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. page 6000–6010.

Alex Wang, Julian Michael, Felix Hill, Omer Levy, Samuel Bowman, and Amanpreet Singh. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.