منظومة لغوية للتحليل الصرفي والنحوي للجمل العربية باستخدام التعلم العميق - AraBERT Transformer

م. هديل عبد الرحمن*، أ.د. محمد فاضل سكر**

* طالبة دكتوراة، قسم الذكاء الصنعي واللغات الطبيعية، كلية الهندسة المعلوماتية، جامعة حلب

** أستاذ في قسم الذكاء الصنعي واللغات الطبيعية، كلية الهندسة المعلوماتية، جامعة حلب

الملخص

تُعد معالجة النصوص العربية آلياً من المهام المعقدة بسبب الخصائص الصرفية والنحوية الفريدة للغة، مثل الإلصاق ومرونة ترتيب الكلمات، خاصة في النصوص غير المشكولة.

في هذا البحث، نقترح منظومة لغوية تعتمد على نموذج لغوي متقدم قادر على فهم السياق اللغوي بشكل عميق، مما يساهم في استخراج 13 حقلاً نحوياً وصرفياً بدقة عالية. تعتمد المنظومة على هيكلية متعددة الرؤوس تتيح تحليل عدة خصائص لغوية في آن واحد، وقد أظهرت المنظومة أداءً عالياً واستقراراً ملحوظاً عند اختبارها على قاعدة بيانات مخصصة، متفوقة بذلك على عدة نماذج لغوية حديثة. تعكس هذه النتائج فعالية النموذج المقترح في دعم تطبيقات المعالجة الآلية للغة العربية مثل الترجمة وتصحيح الأخطاء.

الكلمات المفتاحية: AraBert v2، المحولات، إعراب الجمل العربية، التحليل الصرفي، التحليل النحوي

A Linguistic System for Morphological and Syntactic Analysis of Arabic Sentences Using Deep Learning: AraBERT – Transformer

Hadeel Abdulrahman*, Mohammed Fadel Sukkar**
*Postgraduate Student (PhD.), Dept. of Artificial Intelligence and Natural Languages, Faculty of Informatics Engineering, University of Aleppo
**Prof., Dept. of Artificial Intelligence and Natural Languages, Faculty of Informatics Engineering, University of Aleppo

Abstract

Processing Arabic text automatically is a complex task due to the language's unique morphological and syntactic characteristics, such as agglutination and flexible word order, especially in unvowelized texts.

In this research, we propose a linguistic system based on an advanced linguistic model capable of deeply understanding linguistic context, contributing to the accurate extraction of 13 syntactic and morphological fields. The system relies on a multi-headed architecture that enables the simultaneous analysis of multiple linguistic features. The system demonstrated high performance and remarkable stability when tested on a dedicated dataset, outperforming several state-of-the-art linguistic models. These results reflect the effectiveness of the proposed model in supporting Arabic language processing applications such as translation and error correction.

Keywords: AraBERT v2, Transformers, Arabic sentence parsing, Morphological analysis, Syntactic analysis

1. مقدمة

تمتاز اللغة العربية بثرائها الصرفي والاشتقاقي، حيث أن كلمة واحدة يمكن أن تأخذ عدداً كبيراً من الأشكال بناءً على الاشتقاق والتصريف، وعلى الرغم من كونها مصدر قوة للغة العربية إلا أنها

تشكل تحدياً كبيراً للأنظمة الآلية. فعلى سبيل المثال، كلمة "كتب" قد تعني "كَتَبَ" (فعل ماض)، أو "كُتِبَ" (فعل مبني للمجهول)، أو "كُتُبٌ" (اسم جمع)، وكل ذلك دون تغيير في الحروف الساكنة الأساسية [1]، حيث يكون معدل الغموض في التفسير النحوي أعلى في الكلمات غير المشكولة بمتوسط 8.7، مقارنةً بمتوسط 5.6 للكلمات المشكولة [2]. بالإضافة إلى ذلك، تمتاز اللغة العربية بخصائص فريدة تجعل عملية التحليل صعبة، كما في الإلصاق من خلال إضافة السوابق واللواحق، حيث يمكن أن تلتصق بالكلمات أدوات التعريف وحروف الجر والضمائر، مما يؤدي إلى بنى كلمات وجمل معقدة وغير تقليدية. علاوة على ذلك يمكن ترتيب الكلمات في اللغة العربية بشكل مرن، فعلى الرغم من وجود ترتيب شائع وهو (فعل– فاعل– مفعول به) على سبيل المثال، عادةً ما توضع الكلمة المراد التأكيد عليها في بداية الجملة، وفي بعض الحالات يكون الهدف من التقديم والتأخير لأغراض بلاغية أو نحوية، الأمر الذي يؤدي إلى ارتباك نحوي يصعب حله باستخدام القواعد النحوية [2].

كل العوامل السابقة تجعل مهمة تحديد مواقع الفعل والفاعل والمفعول وإسناد الحالات الإعرابية صعبة. بالتالي فإن عملية الإعراب الكامل لكلمات الجملة يتطلب تكامل الجانب الصرفي وتحديد المواضع النحوية بدقة وهو أمر صعب في النصوص غير المشكولة. لذلك جذب إعراب الجمل العربية اهتماماً متزايداً في بحوث معالجة اللغات الطبيعية، حيث تهدف عملية الإعراب إلى تحديد البنية النحوية للجملة والتي تمثل معلومات ضرورية لمجموعة واسعة من تطبيقات معالجة اللغات الطبيعية مثل الترجمة الآلية، تصحيح الأخطاء النحوية، تلخيص النصوص واستخراج المعلومات [2]

رغم التقدم الكبير الذي أحرزته أبحاث إيجاد الصيغ النحوية للجمل العربية، إلا أنه لاتزال هناك فجوة واضحة ولاسيما في:

- غياب منظومة تقدم اعراباً متكاملاً يدمج بين الصرف والنحو بجميع تفصيلاتها كزمن الفعل، نوع الاسم، التعريف والتنكير، السوابق واللواحق.
- الاعتماد على النصوص المشكولة، وتجاهل النصوص غير المشكولة

وعليه فإن مساهمة هذا البحث تتجلى في النقاط التالية:

- بناء قاعدة بيانات مخصصة لدعم هذه المهمة، تتألف من الجمل العربية المشكولة وغير المشكولة لقوالب محدودة ويترافق مع كل جملة 13 حقل صرفي ونحوي
- تحليل صرفي ونحوي للجمل العربية البسيطة غير المشكولة، من خلال الاستفادة من محول AraBERT v2 وتوظيف رؤوس تصنيف متعددة لإيجاد 13 حقلاً صرفياً ونحوياً للجملة العربية المدخلة.

بناءً على ما سبق، نقدم خلال هذا البحث إجابةً لسؤال البحث، وهو "هل يمكن تطوير منظومة لغوية قادرة على فهم السياق للنصوص العربية غير المشكولة، وتقديم تحليل صرفي ونحوي متكامل وبدقة عالية؟ "، وذلك بالاستناد إلى أن دمج نموذج لغوي عربي مدرب مسبقاً (مثل AraBERT) مع طبقات محول مخصصة ورؤوس تصنيف متعددة، سيساهم في رفع دقة الإعراب الآلي وتحسين موثوقيته في النصوص ذات البنية المعقدة.

2. هدف البحث

يهدف هذا البحث إلى تصميم وتطوير نظام آلي متكامل لإجراء تحليل صرفي ونحوي شامل لجمل عربية بسيطة غير مشكولة، من خلال بناء نموذج لغوي هجين يعتمد على بنية المحولات حيث يدمج قوة نموذج AraBERT المدرب مسبقاً لاستخلاص السمات السياقية العميقة للكلمات، مع طبقات محول إضافية لمعالجة هذه السمات، ورؤوس تصنيف متعددة لاستخراج 13 حقلاً إعرابياً وصرفياً مختلفاً بشكل متزامن. تسعى المنظومة المقترحة إلى تحقيق دقة عالية في المهام المتعددة التالية:

- تحديد الفعل وجذره وزمنه
- تحديد الاسم ونوعه النحوي وحالة التعريف والتنكير للاسم بالإضافة إلى إعرابه الكامل
- تصنيف الكلمات داخل الجملة من حيث الوظيفة (فاعل، مفعول به، صفة، مبتدأ، ...)

وذلك بالاعتماد على قاعدة بيانات مصممة خصيصاً لهذا الغرض. كما ويستهدف البحث في تطبيقاته المستقبلية دعم عدد من المجالات مثل:

- إنشاء جمل ونصوص عربية سليمة صرفياً ونحوياً
- منصات التعليم الإلكتروني للغة العربية، من خلال تقديم إعراب دقيق ومتكامل.

3. الدراسات المرجعية

تنوعت الأبحاث في مجال إيجاد الصيغ النحوية للجمل العربية بدءاً من المعربات الإحصائية التقليدية وصولاً إلى نماذج التعليم العميق الحديثة والمحولات، حيث نستعرض في هذا القسم أبرز الأبحاث والأدوات المستخدمة مع مناقشة نقاط القوة والقصور في كل مجموعة منها.

النماذج الإحصائية المعتمدة على القواعد: وهي من أقدم أساليب التحليل النحوي، وقد نُشرت العديد من الأبحاث والأدوات في هذا المجال، وكان أولها المحللات الإحصائية التقليدية مثل محلل ستانفورد Stanford Parser [3]، والذي اعتمد على نموذج القواعد النحوية الخالية من السياق(Context-Free Grammars (CGF الاحتمالي والمدرب على بيانات شجرة Penn Arabic Tree Bank (PATB). أما معرب CAMeLParser [4] فقد عمل على تقديم نظام متكامل لتحليل الاعتماديات النحوية بين الكلمات بما يتوافق مع السمات الصرفية الغامضة سياقياً، وقد استخدم أداة متطورة لإزالة الغموض الصرفي، كما وحسن نتائجه بالاعتماد على كل من MaltParser [5] مع تقنية SVM محققين دقة عالية نسبياً في تحليلات النحو العربية التقليدية. وبذلك فإن نقاط القوة في هذه النماذج كانت بالاستناد الى قواعد لغوية واضحة، في حين أن نقاط القصور تمثلت في عدم القدرة على التعميم للأنماط اللغوية غير التقليدية.

الحزم البرمجية المتكاملة: تقدم مجموعة من المهام الصرفية والنحوية، كما في [6] Farasa، والتي دمجت تقنيات تحليل الصرف والاعتمادية، حيث تقوم بمجموعة من المهام مثل تقطيع الكلمة tokenization وإيجاد أجزاء الكلام POS، وتحليل الاعتمادية باستخدام محلل MaltParser. في حين أن فريق CAMeL طور مجموعة من الأدوات الشاملة تدعى CAMeL tools والتي توفر مجموعة من الأدوات المساعدة في المعالجة المسبقة، النمذجة الصرفية، بالإضافة إلى إيجاد أجزاء الكلام [7]. كما برز نظام MADAMIRA للتحليل الصرفي وإزالة الغموض [8] حيث جمع بين مميزات الأنظمة السابقة، وحقق كل من MADAMIRA ونظام CAMeL دقة متقاربة في مهمة POS للنصوص العربية [9] بنسبة

MADAMIRA %97.1، و97.2% في CAMeL من أجل العربية الفصحى، و91.7% MADAMIRA بمقابل 91.8% في CAMeL من أجل اللهجة المصرية. وبذلك فإن تعدد الوظائف وسهولة الاستخدام تعتبر كنقاط قوة في حين أن ضعف الأداء على النصوص غير المشكولة تعتبر من نقاط القصور.

نماذج تعلم الآلة: بدأ التحليل النحوي يشهد تحسناً ملموساً، في [10]، تم استخدام مصنف يعتمد علىConditional Random Field CRF للتقطيع وإيجاد الوظيفة النحوية في آن واحد، وفي [11] تفوقت طريقة SVM-Rank المعتمدة على المميزات التقليدية على نموذج Bi-LSTM، إلا أن الجمع بين تمثيلات الكلمة embeddings وبعض المميزات المستخرجة يدوياً ساعد في تحسين أداء الشبكة العصبونية. تمتاز هذه النماذج بقدرتها على التعامل مع التسلسلات الطويلة، الا أنها تحتاج الى حجم بيانات كبير.

نماذج التعلم العميق والمحولات: قدمت الدراسة [12] أول نموذج محول يقوم على آلية الانتباه الذاتي self-attention، وتبعه طرح نموذج [13] BERT وهو نموذج لغوي عميق ثنائي الاتجاه، مدرب على كميات هائلة من النصوص لإنجاز مهمتي نمذجة اللغة المقنعة Masked Language Modeling (MLM)، ومهمة التنبؤ بالجملة التالية Next sentence prediction (NSP). ثم تلاه تطوير نموذج AraBERT [14] المدرب على عدد كبير من النصوص العربية، وقد حقق أداءً متقدماً في فهم اللغة العربية وتحليل المشاعر وتصنيف النصوص، كما في [15] حيث تفوق نموذج AraBERT في الكشف عن الأخبار الزائفة محققاً 78.90% متوسط f1-score. في حين تفوق fine-tuned AraBERT في [16] بمهمة تصنيف المواضيع المكتوبة باللغة العربية واللهجة السورية تحديداً على الخوارزميات الأخرى بمعدلات f1-score, accuracy بمقدار 99%. وجدت الدراسة [17] أن AraBERTv2 كان الأفضل بالاختبار على جميع الأشجار المعيارية treebanks في مهمة الإعراب الاعتمادي، حيث حقق دقة UAS بلغت 84.0% و 80.3% LAS على شجرةPADT، و86.4/87.5 على شجرة CATiB. في [18] طوَّر فريق CAMeL نموذج CamelParser2.0 مفتوح المصدر لتحليل الاعتمادية العربية، وقد حقق هذا النموذج دقة عالية في إيجاد الوظيفة النحوية للجمل العربية الغير مشكولة مع درجات LAS تتجاوز 91%.

وبذلك نجد بأن نقاط القوة للمحولات والنماذج المدربة مسبقاً في دعم المهام المتعددة والتفوق على النماذج التقليدية في حين أن القصور يتجلى في أن معظم النماذج تفتقر إلى تحليل إعرابي شامل يجمع بين الجوانب الصرفية والنحوية والدلالية

4. مواد وطرق البحث

4.1.قاعدة البيانات المستخدمة

يعتبر تصميم قاعدة بيانات مناسبة عنصراً مهماً في تدريب وتقييم أداء النماذج اللغوية ولاسيما في اللغة العربية، لغناها بالتراكيب الصرفية والنحوية، وقد افتقرت قواعد البيانات التقليدية بشكل عام إلى تمثيل رسمي ومنظم مناسب، ففي أغلب الأحيان يتم تخزين التحليلات النحوية بشكل نثري، الأمر الذي يتطلب مستوى عالٍ من الخبرة لدى المستخدم في قواعد اللغة العربية، مما يجعل عملية تصميم نظام الاعراب الآلي أمراً معقداً.

خلال هذا البحث تم تصميم قاعدة البيانات من خلال توليد الجمل ومكوناتها الصرفية والنحوية يدوياً من قبل الباحث، حيث اجتُزأت هذه الكلمات من قاعدة بيانات Flickr8k المعربة [19]والمخصصة لتوصيفات الصور باللغة العربية وشكلنا منها جملاً تتبع للقوالب المقترحة، مما يضمن تنوع الجمل وتوازنها ضمن القوالب النحوية. تبع هذه المرحلة، عملية تدقيق من قبل خبرة لغوية لتشكيل الجمل واعرابها، وجذور الأفعال، بالإضافة إلى سوابق الأفعال ولواحقها وكذلك الأمر بالنسبة للأسماء.

تتألف قاعدة البيانات من 3960 جملة قصيرة وغير مشكولة، أما أعمدة قاعدة البيانات فكانت عبارة عن المعلومات الصرفية والنحوية لكل جملة وبشكل مفصل، وهي: الجملة مشكولة، الجملة بدون تشكيل، الفعل، جذر الفعل، سوابق الفعل، لواحق الفعل، زمن الفعل، نوع الفعل (مفرد مذكر، مفرد مؤنث، مثنى مذكر،...)، الاسم، نوع الاسم النحوي (فاعل، مفعول به،...)، نوع الاسم (مفرد مذكر، مفرد مؤنث،....)، اعراب الاسم الكامل.

تنتمي جمل قاعدة البيانات المقترحة إلى 66 قالباً نحوياً، بحيث أننا قمنا بتوليد 60 جملة بدون وجود تكرارات بين الجمل من أجل كل قالب نحوي مقترح، كما وتنتمي القوالب النحوية

البالغ عددها 66 إلى 11 قالباً نحوياً أساسياً، مع الحرص على التنوع في الضمائر، أنماط التأنيث والتذكير، الإفراد والجمع، وتراكيب الاسم والفعل، وبذلك فإن القوالب كانت على الشكل:

- فعل ماض/ فعل مضارع + فاعل
- مع الاخذ بعين الاعتبار جميع حالات الفاعل (مفرد مذكر/ مفرد مؤنث/ مثنى مذكر/ مثنى مؤنث/ جمع مذكر/ جمع مؤنث)
- فعل ماض/ فعل مضارع/ فعل أمر+ مفعول به
- مع الاخذ بعين الاعتبار جميع حالات الفعل (مفرد مذكر/ مفرد مؤنث/ مثنى مذكر/ مثنى مؤنث/ جمع مذكر/ جمع مؤنث)
- فعل ماض/ فعل مضارع/ فعل أمر + ظرف زمان
- مع الاخذ بعين الاعتبار جميع حالات الفعل (مفرد مذكر/ مفرد مؤنث/ مثنى مذكر/ مثنى مؤنث/ جمع مذكر/ جمع مؤنث)
- فعل ماض/ فعل مضارع/ فعل أمر + حال
- مع الاخذ بعين الاعتبار جميع حالات الفعل (مفرد مذكر/ مفرد مؤنث/ مثنى مذكر/ مثنى مؤنث/ جمع مذكر/ جمع مؤنث)

ويبين الجدول (1) أمثلة عن القوالب النحوية المستخدمة:

الجدول 1- القوالب النحوية المقترحة

القالب النحوي1	فعل ماض + فاعل مفرد مذكر
القالب النحوي2	فعل ماض + فاعل مفرد مؤنث
القالب النحوي3	فعل ماض + فاعل مثنى مذكر
…	…
القالب النحوي11	فعل مضارع + فاعل جمع مذكر سالم
…	…
القالب النحوي 21	فعل مضارع مثنى مذكر + مفعول به
…	…
القالب النحوي 66	فعل أمر جمع مؤنث + حال

على سبيل المثال قمنا بتمثيل جملة "**عملن التصميم**" في قاعدة البيانات المقترحة كما في الجدول (2)

الجدول 2–مثال عن تمثيل الجمل في قاعدة البيانات المقترحة

الجملة	عملن التصميم		
الفعل	عملن	الاسم	التصميم
جذر الفعل	عمل	–	–
سوابق الفعل	Null	سوابق الاسم	ال
لواحق الفعل	ن	لواحق الاسم	Null
زمن الفعل	ماض	نوع الاسم النحوي	مفعول به
نوع الفعل	جمع مؤنث	نوع الاسم	مفرد مذكر
اعراب الفعل	فعل ماض مبني على السكون لاتصاله بنون النسوة، والنون ضمير متصل مبني على الفتحة في محل رفع فاعل	اعراب الاسم	مفعول به منصوب وعلامة نصبه الفتحة الظاهرة على اخره

الجدير بالذكر أن اعتماد القوالب النحوية ساعد على ضبط البنية اللغوية للجمل والتحكم في مكوناتها، إلا أن هذا النهج قد يقلل من تنوع التراكيب اللغوية، مما قد يؤثر على قدرة النموذج على التعامل مع النصوص الطبيعية بكفاءة. وانطلاقاً من ذلك، يقترح هذا البحث في الأعمال المستقبلية:

- توسيع قاعدة البيانات بإدخال جمل من مصادر طبيعية مثل الكتب التعليمية والمقالات
- دمج الجمل المعقدة والتراكيب البلاغية تدريجياً في مراحل لاحقة من التدريب.

4.2.تصميم المنظومة

4.2.1. الدراسة التحليلية – المحولات [12]

تُعد بنية المحولات من أبرز التطورات في مجال الشبكات العصبونية، حيث تخلت عن آليات التكرار التقليدية المستخدمة في الشبكات العصبونية التكرارية RNN والذواكر الطويلة قصيرة الأمد LSTM، واعتمدت بشكل كلي على آلية الانتباه الذاتي لمعالجة التسلسلات بكفاءة عالية[12]، سواء في اللغة أو الصور.

أولاً: البنية العامة للمحول: يتكون نموذج المحول الأساسي من جزئين رئيسيين وهما المشفّر Encoderوفاك التشفيرDecoder ، وكل منهما يتألف من عدة طبقات متماثلة.

في طبقات المشفّر، نجد وحدتين فرعيتين أساسيتين:

- آلية الانتباه الذاتي متعدد الرؤوسMulti-Head Self Attention
- شبكة تغذية أماميةFeed Forward Neural Network

أما فاك التشفير، فيحتوي على هاتين الوحدتين، بالإضافة إلى آلية انتباه Encoder-Decoder Attention

ثانياً: الانتباه الذاتي: تعمل على حساب أوزان الانتباه بين كل كلمة في الجملة وبقية الكلمات، بغض النظر عن المسافة بين الكلمات، مما يساعد النموذج على فهم العلاقات السياقية القصيرة والطويلة داخل التسلسل، وذلك من خلال تحويل كل عنصر في التسلسل الى ثلاث متجهات:

- Q(الاستعلام Query)
- K(المفتاح Key)
- V(القيمة Value) ، وفق المعادلة (1) التالية:

$$Attention(Q,K,V) = softmax\left(QK^T/\sqrt{d_k}\right)V \dots\dots (1)$$

- $\boldsymbol{QK^T}$ درجة التشابه بين كل Query وKey
- $\sqrt{\boldsymbol{d_k}}$ لتفادي القيم الكبيرة في SoftMax

ثالثاً: الانتباه متعدد الرؤوس: بدلاً من استخدام Attention واحدة، يستخدم المحول عدة رؤوس انتباه Heads تعمل بشكل موازي. كل رأس يتعلم علاقات سياقية مختلفة، ثم تُجمع نتائج الرؤوس وتُمرر عبر مصفوفة نهائية، كما في العلاقتين (2) و (3)

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O \dots\dots (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \dots\dots (3)$$

رابعاً: مزايا المحولات:

- التوازي الكامل أثناء التدريب، لعدم وجود اعتماد زمني كما في RNN
- كفاءة في التقاط العلاقات البعيدة داخل التسلسل اللغوي
- سهولة التوسع إلى نماذج ضخمة كما في BERT وGPT، وقد أدى نجاح المحولات إلى ظهور نماذج لغوية قوية مثل: BERT [13] نموذج ثنائي الاتجاه يعتمد على المحولات لفهم السياق من كلا الجانبين، وGPT-3 [20] نموذج توليدي ضخم يستخدم طبقات المحول لإنتاج نصوص شبيهة بالبشر

4.2.2. بنية النموذج المقترح:

يعتمد النموذج المقترح بشكل أساسي على بنية هجينة تجمع بين قوة نموذجAraBERT v2 المدرب مسبقاً في تمثيل السمات السياقية العميقة للنصوص العربية، مع طبقات محول إضافية لمعالجة هذه السمات، وفي النهاية رؤوس تصنيف متعددة بما أن المسألة تندرج ضمن مجال Multi Task Classification، يهدف هذا التصميم إلى تحليل الجملة العربية واستخلاص الحقول الإعرابية النحوية والصرفية لكل كلمة من كلمات الجملة.

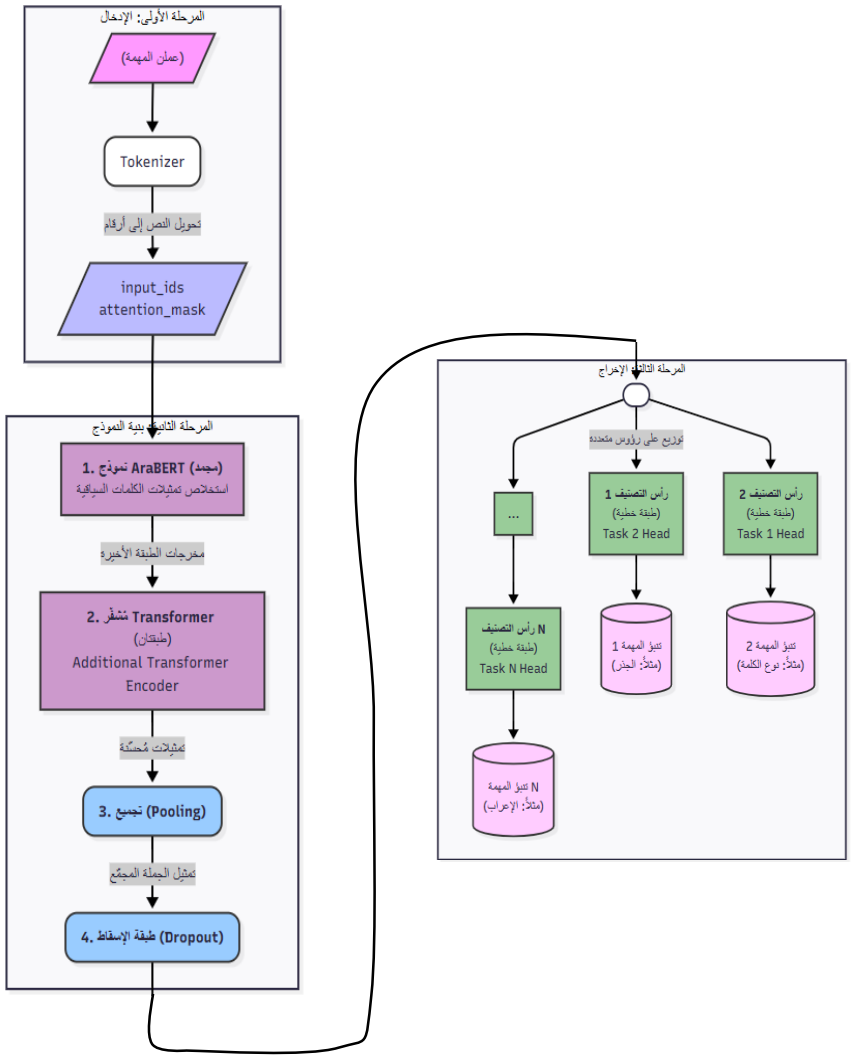
يعود سبب الاعتماد على نموذج AraBERT v2 لكونه أحدث إصدار متوفر من سلسلة نماذج AraBERT حتى تاريخ إعداد البحث، ويتميز بتحسينات ملحوظة في الأداء والاستقرار مقارنة بالإصدارات السابقة، كما أنه مخصص للغة العربية مما يضمن تمثيلاً أكثر دقة للسياق اللغوي والتراكيب النحوية مقارنة بالنماذج متعددة اللغات.

يتكون النموذج المقترح من المكونات التالية:

المرحلة الأولى: التقطيع النصي باستخدام المجزئ Tokenizer الخاص بنموذج AraBERT

المرحلة الثانية، والتي تتألف من:

1. مرحلة تمثيل الكلمات: باستخدام نموذج AraBERT v2 المدرب مسبقاً لاستخراج تمثيل الكلمات، تم تجميد أوزان النموذج الأساسي، واستخراج التمثيلات من خرج آخر طبقة خفية



للنموذج، وبذلك نحصل على التمثيلات الدلالية بطول 768 توكن.

2. محول سياقي Transformer Encoder، دخل هذا المحول هو شعاع تمثيلات كلمات الجملة الناتجة عن المرحلة السابقة، ويتألف من طبقتين، كل طبقة من نوع Transformer Encoder layer، حيث تتميز كل طبقة بما يلي:

الشكل 1- بنية النموذج المقترح

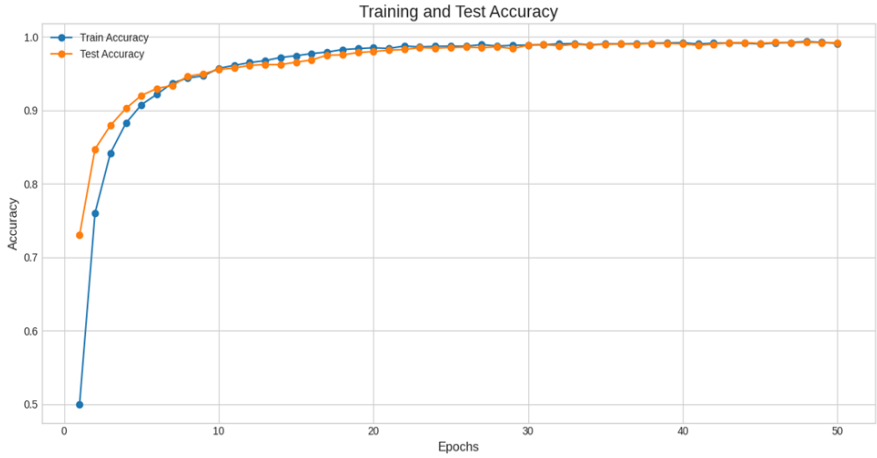
- عدد مداخل هذه الطبقة يساوي 768 وهو متوافق مع طول شعاع التضمين الناتج عن AraBERT
- آلية انتباه ذاتي متعدد الرؤوس، تتألف من 8 رؤوس انتباه ذاتي لالتقاط العلاقات السياقية المتنوعة.
- شبكة تغذية أمامية تحتوي على 512 عصبون

3. طبقة التجميع pooling layer، لدمج المعلومات السياقية الناتجة للجملة في متجه واحد بطول ثابت ويساوي 768 توكن
4. طبقة الاسقاط Dropout للحد من مشكلة overfitting وتحسين التعميم Generalization، وذلك بعد مرحلة التجميع على التمثيل الناتج، بحيث يتم اسقاط 10% من الوحدات الناتجة Units بشكل عشوائي.

المرحلة الثالثة: مرحلة الإخراج وهي عبارة عن مجموعة رؤوس مستقلة بعدد الحقول الإعرابية المستهدفة (الفعل، الجذر، الزمن، ...) والبالغ عددها 13 مخرجاً، حيث استخدمنا طبقة خطية Linear لكل حقل لتوليد التنبؤ المناسب حسب عدد فئاته. ويبين الشكل (1) بنية النظام المقترح

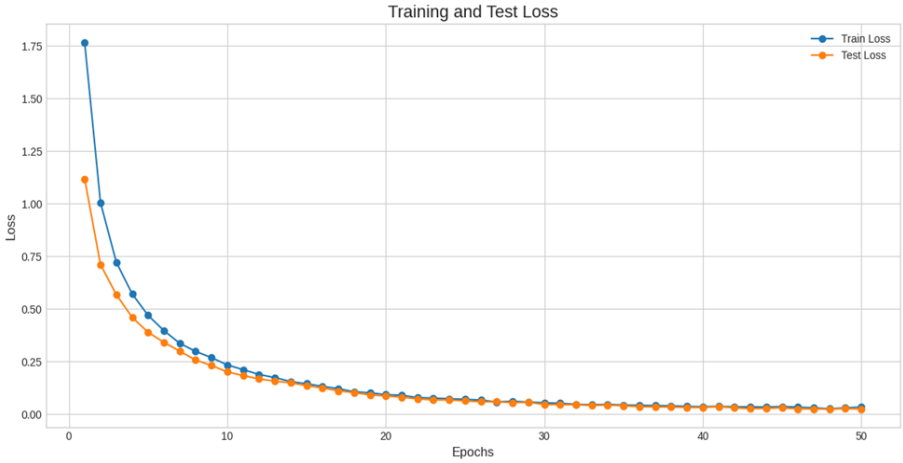
5. النتائج والمناقشة

تم تدريب النموذج باستخدام استراتيجية متعددة المهام حيث يتم توليد مخرجات متعددة متوازية من رؤوس مختلفة، يمثل كل منها تصنيفاً مستقلاً لأحد الحقول النحوية أو الصرفية وهي الفعل، جذر الفعل، سوابق الفعل، لواحق الفعل، اعراب الفعل بالإضافة الى نوع الفعل وزمنه، وكذلك الامر بالنسبة للاسم. يتم حساب خسارة loss كل مخرج بشكل مستقل باستخدام Cross Entropy، ثم يُؤخذ متوسط هذه الخسائر لتحديث النموذج. كما ويتم استخدام الدقة Accuracy للمقارنة بين التوقعات والتصنيفات الحقيقية لجميع المهام. أُنجزت عملية التدريب على دفعات بحجم ثابت batch size = 32، مع تطبيق الانتشار العكسي والتحديث باستخدام خوارزمية التحسين Adam، وكان عدد دورات التدريب المستخدمة 50 دورة، كما وتم تقسيم مجموعة البيانات إلى تدريب واختبار، بنسبة 80% تدريب و20% للاختبار، ويبين الشكل (2) مخطط الدقة لمرحلتي التدريب والاختبار،



الشكل 2- قيم Accuracy للتدريب والاختبار

حيث نلاحظ بأن قيمة الدقة التي تم الحصول عليها بنهاية الاختبار هي 99.29%، في حين أن آخر قيمة تم الحصول عليها اثناء التدريب 99.31%



الشكل 3- قيم Loss التدريب والاختبار

في حين يمثل الشكل (3) قيم الخطأ خلال عملية التدريب والاختبار، ونلاحظ بأن نسبة الخطأ التي تم التوصل إليها بنهاية عملية الاختبار كانت 0.0233، في حين أن نسبة الخطأ التي تم التوصل اليه بمرحلة التدريب 0.0280

من الشكل 2 و3 يمكن أن نستنتج أن:

- التعلم سريع: بدأ النموذج المقترح بالتعلم بسرعة كبيرة، حيث قفزت الدقة من 50% إلى 84% في 3 دورات فقط. هذا يدل على أن المهمة قابلة للتعلم وأن بنية النموذج مناسبة
- تقارب الخسارة Loss Convergence: تناقصت كل من خسارة التدريب وخسارة الاختبار بشكل مستمر ومتناسق، ولم يُلاحظ وجود تباعد بينهما، وهذا يؤكد عدم وجود Overfitting
- الاستقرار في النهاية: في آخر 15 دورة، بدأ التحسن في الدقة يتباطأ، وهذا دليل أن النموذج وصل إلى نقطة قريبة جداً من الأداء الأمثل على هذه البيانات وهي 99.29% على بيانات الاختبار

قمنا بمقارنة نتائج النموذج المقترح مع نماذج لغوية رائدة وهي Gemini 2.5 pro، GPT-4 turbo، بالإضافة إلىFelo v3.1.2 على مجموعة جمل الاختبار، حيث أعددنا أمراً واحداً prompt لإدخاله للنماذج الثلاثة، وكان على الشكل التالي:

- إعطاء ذات الأمر للنماذج الثلاثة الجاهزة بأن الدخل هو مجموعة جمل الاختبار فقط والبالغ عددها 792 جملة وسيتم إدخالها دفعة واحدة، أما المخارج المطلوبة فهي (الفعل، جذر الفعل، سوابق الفعل، لواحق الفعل، زمن الفعل، نوع الفعل، اعراب الفعل، الاسم، سوابق الاسم، لواحق الاسم، نوع الاسم النحوي، نوع الاسم، اعراب الاسم) مع إعطاء مثال مفصل عن الخرج المطلوب لكل مخرج
- تهيئة النماذج الثلاثة بخمسة جمل من قاعدة البيانات حيث تم ادخال الجملة مع الخرج المطلوب وهو ما يعرف "Few-shot"

وكانت النتائج (الدقة) كما هو موضح في الجدول (3)

الجدول 3– مقارنة نتائج المنظومة المقترحة مع كل من GPT, Felo, Gemini

	المنظومة المقترحة	GPT 4 turbo	Gemini pro	Felo
Acc%	%99.29	%59.19	%67.79	%58.48

ويبين الجدول (4) نظرة أعمق على نتائج كل من GPT، Gemini، Felo، والمنظومة المقترحة لنبين أداء كل مخرج بشكل مفصل. كما ويبين الشكل (4) دقة كل مخرج للمنظومة المقترحة بشكل منفصل خلال مرحلة الاختبار.

الجدول 4– أداء كل مخرج من مخارج المنظومة (Accuracy)

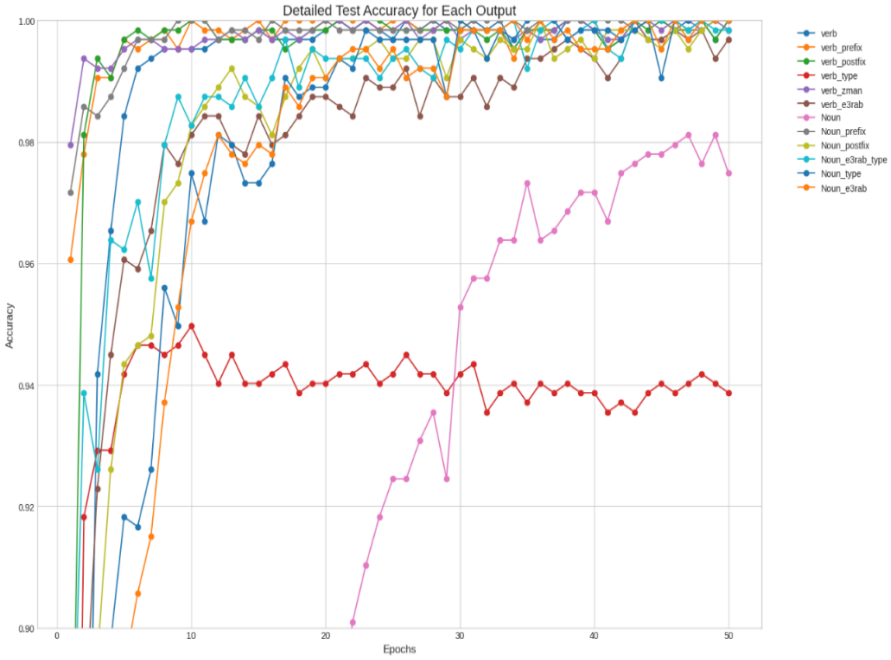
	المنظومة المقترحة	GPT 4 turbo	Gemini pro	Felo
الفعل	%100	%100	%100	%98.11
جذر الفعل	%100	%61.48	%99.69	%99.53
سوابق الفعل	%99.69	%79.09	%100	%45.13
لواحق الفعل	%100	%73.74	%61.95	%82.70
نوع الفعل	%94.03	%27.36	%0.16	%78.14
زمن الفعل	%99.84	%79.09	%100	%54.72
اعراب الفعل	%99.21	%0	%36.32	%0
الاسم	%97.80	%100	%99.69	%99.37
سوابق الاسم	%100	%100	%100	%46.54
لواحق الاسم	%99.84	%52.99	%71.70	%52.99
نوع الاسم النحوي	%99.53	%32.55	%58.49	%40.57
نوع الاسم	%100	%30.66	%46.23	%62.42
اعراب الاسم	%99.84	%32.55	%7.08	%0

ويوضح الجدول (5) مؤشرات الأداء، وهي زمن التدريب، زمن الاختبار بالإضافة الى حجم النموذج المقترح من أجل 50 دورة تدريب، وتجدر هنا الإشارة الى أن نماذج GPT، Gemini، Felo هي نماذج مبتكرة من شركات مرموقة ومدربة على ملايين النصوص التي لا يمكن حصرها وهذه النماذج غير قابلة لإعادة التهيئة والتدريب. تم تنفيذ عملية تدريب النموذج باستخدام بيئة

Google Colab، مع الاستفادة من وحدة المعالجة المخصصة للتعلم العميق (TPU V2) لتسريع عملية التدريب وتقليل الزمن الحسابي.

الجدول 5- مؤشرات الأداء

	زمن التدريب	زمن الاختبار	حجم النموذج
المنظومة المقترحة	294.17 ثانية	60.18 ثانية	541.07 MB



الشكل 4- تغيير Accuracy عبر epochs من أجل كل مخرج من مخارج المنظومة المقترحة

قمنا بدراسة المخرجات الصرفية والنحوية لكل من المنظومة المقترحة بالإضافة إلى GPT، Gemini، Felo، وكانت النتائج كما يلي:

أولاً – المنظومة المقترحة: وفقاً لنتائج الجدول (4)، كان حقل "نوع الفعل" هو الأقل دقة في المخرجات حيث بلغت 94.03%، لذلك نورد فيما يلي أمثلة عن جمل أخفق فيها النموذج في تحديد نوع الفعل من حيث التذكير أو التأنيث وكذلك العدد:

- "عملوا متفائلين"، كانت جميع الحقول صحيحة عدا حقل "نوع الفعل"

- التنبؤ الصحيح: جمع مذكر
- تنبؤ النموذج: جمع مؤنث

- "تكتبان الرسالة"، كانت جميع الحقول صحيحة عدا حقل "نوع الفعل"
  - التنبؤ الصحيح: مثنى مؤنث
  - تنبؤ النموذج: جمع مذكر

في حين أنه اعطى نتائج صحيحة من أجل باقي الجمل وعلى سبيل المثال:

- يكتبان فجرا:
- الفعل: "يكتبان"– الجذر: "كتب" – سوابق الفعل: "ي" – لواحق الفعل: "ان" – نوع الفعل: "مثنى مذكر" – زمن الفعل: "مضارع" – اعراب الفعل: " فعل مضارع مرفوع وعلامة رفعه ثبوت النون، وألف الاثنين ضمير متصل مبني في محل رفع فاعل"
- الاسم: "فجرا" – سوابق الاسم: " " – لواحق الاسم: "ا" –نوع الاسم النحوي: " ظرف زمان" – نوع الاسم: " " – اعراب الاسم: " ظرف زمان منصوب وعلامة نصبة تنوين الفتح الظاهر على آخره"

ثانياً–GPT-4 turbo:

- قام بإعراب جميع الأفعال سواء أكانت فعل ماض، مضارع أو أمر، وسواء أكانت جمع أو مفرد أو مثنى، على أنها "فعل ماض مرفوع وعلامة رفعه الضمة الظاهرة، والفاعل ضمير مستتر تقديره هو".
- في حقل زمن الفعل، لم يتعرف على فعل الأمر، وإنما اعتبره فعل ماض
- أعرب الكلمة الثانية من الجملة "مفعول به منصوب وعلامة نصبه الفتحة الظاهرة على اخره"، علماً أن اعرابات الأسماء تتنوع بين فاعل، مفعول به، ظرف زمان وحال.
- ارجع قيم NULL من أجل لواحق الاسم، علماً أن هناك أسماء لها لواحق مثل "ات"، "ان"، "ون"، ...

ثالثاً – Gemini-2.5 Pro:

- في نوع الاسم النحوي، أعرب الحال على أنه مفعول به
- أعرب ظرف الزمان بشكل صحيح، إلا أنه أخطأ بإعراب المفعول به على أنه فاعل.
- في زمن الفعل بالنسبة للأفعال الماضية فكان التصنيف لها في بعض الأحيان صحيح وفي بعض الأحيان مضارع أو أمر.
- في نوع الفعل، بعضها كان صحيح والبعض خاطئ إلا أنه أرفق جميع الأنواع بعبارة "مفرد مذكر غائب معلوم" وبالتالي لم يتطابق مع المطلوب وهو فقط مفرد مذكر (على سبيل المثال).
- فشل في إعراب تاء التأنيث المتصلة بالفعل وكذلك الأمر بالنسبة لنون النسوة
- كما فشل في إعراب المفعول به جمع المؤنث السالم (المنتهي ات) وقام بإعرابه على أنه فاعل

رابعاً – Felo v3.1.2:

- أخطأ بإعراب جميع الأفعال، فعلى سبيل المثال:
  - درس/ درست/ درسا: فعل ماضٍ مبني على الفتحة الظاهرة على آخره، والفاعل ضمير مستتر تقديره هو وبعض الجمل تقديره هما وهم وهي
  - درست: فعل ماضٍ مبني على الفتحة الظاهرة على آخره، والتاء للتأنيث الساكنة ضمير متصل مبني في محل رفع فاعل
  - يكتب: فعل مضارع مرفوع وعلامة رفعه ثبوت النون، وفاعل مرفوع ضمير مستتر تقديره هو
  - اكتب: فعل أمر مبني على حذف النون، وفاعل مستتر تقديره أنت
  - كتبن: فعل ماضٍ مبني على حذف النون، وفاعل مستتر تقديره هن
  - يكتب مساء: فعل مضارع مرفوع وعلامة رفعه ثبوت النون
- أعرب أغلب الأسماء على أنها مبتدأ عدا الحال وظرف الزمان
- اكتشف الحال وظرف الزمان بشكل صحيح، الا أنه أخطأ بالعلامة الإعرابية

6. الاستنتاجات والتوصيات:

قدمنا خلال هذا البحث مقترحاً لقاعدة بيانات للجمل العربية وقد شملت قاعدة البيانات المقترحة على أغلب المعلومات الصرفية والنحوية الخاصة بهذه الجمل، وكانت الغاية هي الوصول إلى الإعراب الكامل لمكونات الجملة، كما واقترحنا بنية لمنظومة الإعراب الكامل وإيجاد الخصائص الصرفية والنحوية للجمل العربية باستخدام محول AraBERT v2 المدرب مسبقاً مع إضافة طبقتين من المحولات، وذلك ضمن إطار Multi-Task Classification لمعالجة المهام المتعددة بشكل متوازي، وقد أثبت النتائج تفوق المنظومة المقترحة على كل من ChatGPT، Gemini و Felo، ولاسيما بعد ادخال جمل الاختبار دفعة واحدة الى هذه النماذج، وسنعمل في المستقبل على توسعة قاعدة البيانات لتشمل جمل عربية بعدد مركبات أكبر وقوالب نحوية أشمل.

7. المراجع

[1]. SHAHROUR, A., *et al*. 2015- Improving Arabic diacritization through syntactic analysis, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1309-1315). https://doi.org/10.18653/v1/D15-1152

[2]. KHOUFI, N., *et al*. 2024- Comparative Study for Text Chunking Using Deep Learning: Case of Modern Standard Arabic, Computación y Sistemas, *28*(2), 517-527. https://doi.org/10.13053/cys-28-2-4560

[3]. DARWISH, K., *et al*. 2020- A panoramic survey of natural language processing in the Arab world, Communications of the ACM, *64*(4), 72-81 https://arxiv.org/pdf/2011.12631

[4]. SHAHROUR, A., *et al*. 2016- CamelParser: A system for Arabic syntactic analysis and morphological disambiguation, in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations (pp. 228-232) https://aclanthology.org/C16-2048/

[5]. NIVRE, J., *et al*. 2006- MaltParser: A data-driven parser-generator for dependency parsing, in Proceedings of the Language Resources and Evaluation Conference, pp. 2216–2219. http://www.lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf

[6]. ABDELALI, A., *et al*. 2016- Farasa: A fast and furious segmenter for Arabic, in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations (pp. 11-16). https://aclanthology.org/N16-3003/

[7]. OBEID, O., *et al*. 2020- CAMeL tools: An open-source python toolkit for Arabic natural language processing, in Proceedings of the 12th Language Resources and Evaluation Conference, pp. 7022–7032. https://aclanthology.org/2020.lrec-1.868/

[8]. PASHA, A. *et al.* 2014- MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic, in Proceedings of the Ninth International Conference on Language Resources and Evaluation, p 1094–1101. http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

[9]. ALLUHAIBI, R., *et al.* 2021- A Comparative Study of Arabic Part of Speech Taggers Using Literary Text Samples from Saudi Novels, Information, *12(12), 523*. https://doi.org/10.3390/info12120523

[10]. GAHBICHE-BRAHAM, *et al.* 2012- Joint segmentation and POS tagging for Arabic using a CRF-based classifier, in LREC (pp. 2107-2113). https://aclanthology.org/L12-1509/

[11]. DARWISH, K. *et al.* 2017- Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet, in Proceedings of the Third Arabic Natural Language Processing Workshop, pages 130–137. https://aclanthology.org/W17-1316/

[12]. VASWANI, A. *et al.* 2017- Attention is all you need, Advances in neural information processing systems, 30. https://arxiv.org/abs/1706.03762

[13]. DEVLIN, J., *et al.* 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) (pp. 4171-4186). https://arxiv.org/abs/1810.04805

[14]. Antoun, W., *et al.* 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. https://aclanthology.org/2020.osact-1.2/

[15]. MHREZ, A., *et al.* 2024- Evaluating the Performance of Arabic Language Models on the Task of Stance Detection Toward Fake News, Journal of Homs University Vol 46. 5, P 83,106 [In Arabic]. https://journal.homs-univ.edu.sy/index.php/Engineering/article/view/5120

[16]. AHMAD, A., *et al.* 2023- Topic Detection in Arabic Social Media Text Data (Syrian Dialect), Journal of Homs University Vol 45. 14,[In Arabic]. https://journal.homs-univ.edu.sy/index.php/Engineering/article/view/3020

[17]. AL-GHAMDI, S., *et al.* 2023- Fine-Tuning BERT-Based Pre-Trained Models for Arabic Dependency Parsing, Applied Sciences, 13(7), 4225. https://doi.org/10.3390/app13074225

[18]. ELSHABRAWY, A., *et al.* 2023- CamelParser2.0: A State-of-the-Art Dependency Parser for Arabic, in Proceedings of ArabicNLP 2023 (pp. 170-180). https://aclanthology.org/2023.arabicnlp-1.15.pdf

[19]. ElJundi, O., et al. 2020- Resources and end-to-end neural network models for Arabic image captioning. VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SciTePress, pp. 233–241. https://www.researchgate.net/publication/340044948_Resources_and_End-to-End_Neural_Network_Models_for_Arabic_Image_Captioning

[20]. BROWN, T., *et al.*, 2020- Language Models are Few-Shot Learners, Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165