استخدام خوارزمية TSHD والتعلم العميق لاستخراج المعلومات النصية (حالة دراسية: السير الذاتية)

م. مجد طنوس[1] د. مهند رجب[2] د. وسيم رمضان[3]

ملخّص البحث

يُعد استخراج المعلومات النصية أحد مهام معالجة اللغات الطبيعية الهامة، نظراً لدوره البارز في معالجة البيانات النصية غير المهيكلة واستخراج معلومات مفيدة منها وهيكلتها، مما يتيح المعالجة والتحليل الحاسوبي لها. تُعتبر نماذج الإجابة عن الاسئلة القائمة على الاستخراج النصي أحد تقنيات استخراج المعلومات الحديثة التي أثبتت فعاليتها.

يقدّم هذا البحث طريقة جديدة لتطوير آلية عمل نماذج استخراج المعلومات باالاستفادة من التقطيع باستخدام خوارزمية TSHD. حيث يتم تطبيق الخوارزمية لاستخراج محتويات مقاطع الوثيقة وموضوع كل مقطع، ومن ثم استخراج المعلومات بواسطة النماذج من المقاطع عوضاً عن الوثيقة كاملة.

حققت الطريقة المقترحة تحسين في نتائج تقييم مجموعة من نماذج استخراج المعلومات على هيكلية مجموعة البيانات squad 1.1 في مجال السير الذاتية، حيث ارتفعت قيمة مقياس Exact match بنسبة زيادة وصلت إلى 7.4%، كما ارتفعت قيمة مقياس F1 score بنسبة زيادة وصلت إلى 7.8%.

الكلمات المفتاحية: معالجة اللغات الطبيعية، التعلم العميق، استخراج المعلومات النصية، نماذج المحولات، التقطيع.

[1] طالب دكتوراه في كلية الهندسة المعلوماتية جامعة البعث

[2] أستاذ في كلية الهندسة المعلوماتية جامعة البعث

[3] مدرس في كلية الزراعة جامعة البعث

Using TSHD Algorithm and Deep Learning for Textual Information Extraction (case study: Resumes)

Eng. Majd Tannous[1] Dr. Mohanad Rajab[2] Dr. Wassim Ramadan[3]

Abstract

Textual information extraction is an important task of natural language processing, due to its prominent role in processing unstructured textual data to extract useful information and structure it, thus enabling computer processing and analysis. Extractive question answering models are one of the modern information extraction techniques that have demonstrated their effectiveness.

This research presents a novel approach to enhance information extraction models using TSHD segmentation algorithm. The algorithm is applied to extract the contents of document segments and their topics. Information extraction is then performed using models focused on individual segments rather than the entire document.

The proposed method achieves an improvement in the evaluation results of various information extraction models on squad 1.1 data set in the resume domain. The exact match metric is increased by 7.4%, and F1 score is increased by 7.8%.

Keywords: Natural Language Processing, Deep Learning, Textual Information Extraction, Transformer models, Topic Segmentation.

1. مقدّمة

مع ازدياد حجم البيانات النصية غير المهيكلة بشكل هائل في عصر المعلومات، أصبح استخراج المعلومات من النصوص عملية هامة في مختلف المجالات، مثل البحث العلمي، والطب، والأعمال التجارية، والتعليم.. ومع ذلك، فإن استخراج المعلومات النصية مهمة صعبة، حيث تتطلب فهماً دقيقاً للغة.

تشكل معالجة اللغة الطبيعية واستخدام التقنيات الحديثة في تحليل النصوص نقلة نوعية في تفاعل الإنسان مع البيانات اللغوية وخاصة مع انتشار كميات هائلة من البيانات النصية غير المهيكلة. تعد نماذج المحولات "Transformer Models"، وهي نماذج عصبونية عميقة، من أبرز الطرائق القادرة على تحقيق أداء ممتاز في مجموعة متنوعة من المهام، بما في ذلك الترجمة الآلية والإجابة على الأسئلة واستخراج المعلومات، وغيرها.

تُستخدم تقنيات استخراج المعلومات النصية لتحويل البيانات النصية غير المهيكلة إلى معلومات مهيكلة سهلة الفهم وقابلة للمعالجة والتحليل من قِبل البرامج الحاسوبية، مما يوفر الوقت والجهد. تعد نماذج استخراج المعلومات، وخاصةً تلك التي تعتمد على الإجابة عن الأسئلة القائمة على الاستخراج النصي (Extractive Question Answering Models)، أدوات قوية لتحليل النصوص واستخلاص المعلومات منها تبعا للسؤال المطروح. ومع ذلك، تواجه هذه النماذج تحديات هامة في تحديد السياق المناسب لاستخراج المعلومات المطلوبة من الوثائق بدقة. فمثلاً، في مجال السير الذاتية، لاستخراج اسم الجامعة التي تخرج منها صاحب السيرة الذاتية، قد يؤدي البحث ضمن كامل الوثيقة إلى استخراج معلومات خاطئة، مثل الجامعة التي يعمل بها، أو مجلة جامعة نشر بحثاً فيها.

إنّ الطريقة التقليدية لاستخراج المعلومات من الوثائق هي البحث في الوثيقة بأكملها. وعلى الرغم من انتشار هذه الطريقة في أغلب طرائق استخراج المعلومات فإنها تتسم بعدم الدقة الكافية. حيث أن البحث في الوثيقة بأكملها قد يؤدي إلى استخراج معلومات غير ضرورية أو غير مناسبة. كما أنها تتسم بالبطء حيث قد تستغرق عملية البحث في الوثيقة بأكملها وقتاً طويلاً، خاصةً مع الوثائق الكبيرة.

وهنا يأتي دور الحل المقترح في هذه الدراسة: البحث في أجزاء من الوثيقة بدلاً من البحث في الوثيقة بأكملها. أي يمكننا البحث في أجزاء محددة من الوثيقة، مثل المقاطع أو الفقرات. تتميز هذه الطريقة بزيادة الدقة والسرعة. حيث أنه من خلال التركيز على أجزاء محددة من الوثيقة، يمكننا زيادة دقة استخراج المعلومات، ومن خلال تقليل حجم البيانات التي يتم البحث فيها، يمكننا تسريع عملية استخراج المعلومات.

الحل المقترح يرتكز على استخدام خوارزمية TSHD (Topic Segmentation based on Headings Detection) [1] إحدى الطرائق الفعالة في استخراج المقاطع من الوثائق واكتشاف موضوعاتها بالاعتماد على تحديد عناوين المقاطع.

كيف تساعد TSHD في استخراج المعلومات؟ يتم ذلك من خلال تحديد المقاطع ذات الصلة: يمكن استخدام TSHD لتحديد المقاطع ذات الصلة بالاستعلام تبعاً لموضوع المقطع، مما يسمح لنا بالتركيز على هذه المقاطع فقط عند استخراج المعلومات.

تقدم هذه الورقة استخدام خوارزمية TSHD كحل لتحسين دقة النماذج في استخراج المعلومات من الوثائق وتسريع العملية.

تم تطبيق الحل المقترح على السير الذاتية كحالة دراسية. تعد السير الذاتية من أهم الوثائق النصية التي تُستخدم في عملية التوظيف لاختيار المرشح المناسب. تحتوي هذه الوثائق على عدة مقاطع مثل (المعلومات الشخصية، والتعليم، والخبرات، وغيرها) وتتضمن المقاطع معلومات توصف جوانب متعددة لأصحابها. ونظراً لأهمية استخراج هذه المعلومات بفعالية ودقة، تم اختيار السير الذاتية كحالة دراسية في هذا البحث.

2. هدف البحث

انطلاقاً من أهمية الاستفادة من امكانيات نماذج المحولات في استخراج المعلومات النصية من جهة، بالإضافة للدور الهام الذي تقدمه خوارزمية TSHD في استخراج المقاطع وتحديد موضوعاتها من جهة أخرى. يهدف هذا البحث إلى تطوير آلية جديدة لتحسين عملية استخراج المعلومات من الوثائق ذات العناوين، من خلال تركيز اهتمام نماذج استخراج المعلومات على الجزء المهم من النص بدلاً من كامل النص. تعتمد هذه الآلية على الاستفادة من خوارزمية TSHD في استخراج مقاطع الوثيقة وتحديد موضوعاتها، مما يسمح للنموذج بالوصول المباشر للمقطع المناسب واستخراج المعلومات منه بدقة وكفاءة، ثم تخزين المعلومات المستخرجة بصيغة مهيكلة (جدول)، بحيث تصبح قابلة للمعالجة والتحليل حاسوبياً. تساهم الآلية المقترحة أيضاً في تخفيض الموارد الحاسوبية اللازمة نتيجة تصغير حجم السياق النصي. كما أن هذا التعديل لا يتطلب إعادة تدريب النماذج على مقاطع الوثائق بعد التقطيع.

3. دراسة مرجعية

خلال السنوات الأخيرة، حققت نماذج المحولات قفزة كبيرة في مجالات ومهام متعددة باعتمادها على تقنية الانتباه "Attention" [2] والتعلم المنقول "transfer learning". تختلف هيكلية النموذج تبعاً للمهام المطلوبة، ومن أبرز هيكليات النماذج التي يمكن

توظيفها من أجل استخراج المعلومات النصية: BERT [3] وRoBERTa [4] وLongformer [5] وXLNet [6] وغيرها. يتم تخصيص النماذج المدربة مسبقاً من خلال التعلم المنقول باستخدام آلية الضبط الدقيق "fine tunning"، عن طريق القيام بتدريب إضافي لتصبح هده النماذج أكثر ملاءمةً ودقةً من أجل تأدية مهام محددة بكفاءة أعلى.

تعد نماذج الإجابة عن الاسئلة القائمة على الاستخراج النصي (Extractive Question Answering Models) أحد أبرز تقنيات استخراج المعلومات، حيث تعمل على استخراج المعلومة (الإجابة) من سياق نصي "context" تبعا للسؤال المطروح. اهتمت العديد من الأبحاث خلال السنوات الأخيرة بدراسة ومقارنة فعالية نماذج استخراج المعلومات بعد تخصيصها وتدريبها على مجموعات بيانات نصية مختلفة.

تعد هيكلية مجموعة البيانات squad إحدى أبرز الهيكليات المستخدمة لتدريب هذا النوع من النماذج وتقييمها، وتضم نوعين squad 1.1 [7] و squad 2 [8]. تحتوي مجموعة البيانات squad 1.1 على 107,785 سؤال تم طرحها على 536 نص، حيث أن جميع الاسئلة لها إجابات ضمن السياق النصي المعطى. بينما أن مجموعة البيانات squad 2 تحتوي أيضاً على 53,775 سؤال إضافي ليس له إجابة ضمن السياق النصي. يشكل استخدام squad 2 من قبل النماذج تحدياً اضافياً هاماً علاوةً عن استخراج الإجابة المطلوبة، يتمثل بمقدرة النموذج على اكتشاف حالات عدم وجود إجابة للسؤال، حيث أنّ نموذجاً ما يحقق نتيجة تقييم f1-score=85.8% على squad 1.1، يحقق فقط f1-score=66.3% على squad 2 [8].

يوجد أيضاً مجموعات بيانات نصية أخرى، مثل NewsQA [9] وهي مجموعة بيانات تتكون من 119,633 سؤالاً تم طرحها على 12,744 مقالة إخبارية على شبكة

CNN. QuAC [10] وهي مجموعة بيانات تحتوي على 14 ألف حوار، و100 ألف سؤال. تتضمن مربعات الحوار: (1) طالب يطرح سلسلة من الأسئلة عن نصوص ويكيبيديا، و(2) مدرس يجيب على الأسئلة من خلال تقديم مقتطفات قصيرة من النص. CovidQA [11] وهي عبارة عن مجموعة بيانات للإجابة على الأسئلة تتكون من 2019 زوجاً مختلفاً من الأسئلة والأجوبة حول 147 مقالة ذات صلة بـCOVID-19.

يمثل نتائج التقييم باستخدام مقياس F1 score لمجموعة من النماذج في استخراج الإجابات بعد تدريبهم على مجموعات بيانات مختلفة [12]

الجدول 1: قيم F1 score لمجموعة من النماذج على مجموعات بيانات مختلفة [12]

مجموعة البيانات / النموذج	NewsQA	SQuAD 2	QuAC	CovidQA
$XLNet_{BASE}$	53.2	64.9	30.1	44.9
$BERT_{BASE}$	52.1	64.7	28.6	44.8
$RoBERTa_{BASE}$	57.0	68.2	31.3	44.5
$ALBERT_{BASE}$	51.8	64.8	19.5	42.4
$ConvBert_{BASE}$	55.7	67.4	31.5	44.9
$BART_{BASE}$	56.2	67.6	29.1	45.3
$BERT_{BASE}$-BiLSTM	52.6	65.0	28.9	45.6

نلاحظ من اختلاف نتائج تقييم النماذج من مجموعة بيانات لأخرى، تبعاً لحجم مجموعة التدريب، ودرجة وضوح صياغة الأسئلة وإجاباتها، بالإضافة لطول السياق النصي. حققت كافة النماذج أفضل النتائج على مجموعة بيانات SQuAD 2 نظرا

لوضوح صياغة الأسئلة وإجاباتها ضمن النصوص. كما نلاحظ أن نموذج RoBERTa قد حقق أفضل النتائج نسبياً، حيث يعد RoBERTa نسخة محسنة من نموذج BERT.

تم تطوير بعض النماذج الضخمة، والتي تتطلب موارد حاسوبية كبيرة، مثل نموذج Longformer. يعتمد النموذج على تطبيق آلية انتباه "attention" تتوسع خطياً مع ازدياد طول سلسلة الدخل. تم تقييم النموذج في استخراج الإجابات النصية، حيث حقق دقة نتائج أفضل من نموذج RoBERTa على مجموعات بيانات مختلفة [5].

تم تصميم بعض النماذج بهدف تحسين الأداء على حساب الدقة لملاءمة الموارد الحاسوبية الضعيفة، منها نموذج DistilBERT الذي يعد نسخة مخففة (مصغرة) عن نموذج BERT، حيث أنه أصغر حجماً وأسرع ويتطلب موارد حاسوبية أقل من نموذج BERT، إلا أن دقة نتائجه أقل [13].

تعتمد معظم الأبحاث عند تطبيق النماذج على استخراج المعلومات من الوثيقة كاملة، إلا أن بعض الوثائق مثل المنشورات البحثية والنشرات الدوائية ومقالات ويكيبيديا والسير الذاتية وغيرها، تحتوي على مقاطع ذات مواضيع معينة تتضمن معلومات مرتبطة بموضوع المقطع. نظراً لذلك، يمكن تطبيق تقنيات التقطيع "Topic segmentation" واكتشاف المواضيع "Topic identification" لاستخراج المقاطع من الوثائق واكتشاف موضوعاتها. تكمن أهمية تحقيق دقة نتائج تقطيع مرتفعة في تحسين جودة العديد من مهام معالجة اللغات الطبيعة، مثل استخراج المعلومات، وتلخيص الوثائق، وغيرها.

تعد خوارزمية TSHD (Topic Segmentation based on Headings Detection) إحدى الطرائق الفعالة في استخراج المقاطع وتحديد موضوعاتها بالاعتماد على تحديد عناوين المقاطع ضمن الوثيقة، ثم هيكلة المقاطع المستخرجة. يمكن مواءمة الخوارزمية لاستخراج المقاطع من عدة مجالات نصية، منها السير الذاتية. حققت

الخوارزمية نتائج تقييم مرتفعة تقارب F1-score = 96%، ونسبة خطأ تقطيع منخفضة جداً تقارب 2% [1]. وبالتالي، يمكن الاعتماد على خرج خوارزمية TSHD في تطوير أداء ودقة نماذج استخراج المعلومات.

4. أدوات وطرائق البحث

نقدم ضمن هذا القسم توصيف مجموعة البيانات المستخدمة، ومقاييس التقييم، وآلية عمل خوارزمية TSHD، وتوصيف نماذج استخراج المعلومات التي تم استخدامها لتقييم الطريقة المقترحة.

1.4. توصيف مجموعة البيانات "Dataset"

تم تجميع عدد من السير الذاتية مجموعها 105 مكتوبة باللغة الإنكليزية من مصادر ومواقع متعددة، كما أنها تتبع لأشخاص ذوي اختصاصات مختلفة. تعتبر هذه السير الذاتية ملفات نصية غير مهيكلة، وتتبع لقوالب وتنسيقات مختلفة، كما تختلف مقاطعها من حيث العدد والترتيب، ولا يتبع نمط وحجم ولون الخط فيها لأي مقاييس.

تم التقييم باستخدام هيكلية مجموعة البيانات squad 1.1 من أجل استخراج مجموعة من المعلومات من خلال طرح الأسئلة التالية على السير الذاتية:

- استخراج اسم صاحب السير الذاتية

السؤال: what is my full name?

- استخراج عنوان البريد الالكتروني

السؤال: what is the email address?

- استخراج رقم الهاتف

السؤال: what is mobile phone?

- استخراج اختصاص البكالوريوس

السؤال: in which major is the bachelor?

- استخراج اسم الجامعة المانحة للبكالوريوس

السؤال: from which university is the bachelor?

حيث أن حجم مجموعة بيانات الاختبار هو 444 سؤال، جميعها تحوي إجابات للأسئلة المطروحة. حيث أن هيكلية مجموعة البيانات squad بالشكل:

['id', 'title', 'context', 'question', 'answers']

id: رقم معرف فريد للسؤال

title: اسم ملف الوثيقة

context: السياق النصي الذي يجب استخراج الإجابة منه

question: السؤال المطروح، الذي يمثل المعلومة المراد استخراجها

answers: الإجابة الفعلية (المعلومة الصحيحة)، حيث يمكن تحديد ثلاث إجابات صحيحة لكل سؤال.

2.4. مقاييس التقييم

يتم تقييم استخراج المعلومات باستخدام مقياسي Exact match و F1 score.

- **مقياس Exact match (EM)**

يقوم هذا المقياس بتقييم المطابقة التامة للإجابة المتوقعة الناتجة عن النموذج مع الإجابة الفعلية، وبذلك، تكون نتيجته من أجل كل سؤال، إما 100 في حال المطابقة التامة، أو 0 فيما عدا ذلك. ثم يتم حساب متوسط قيم الـ Exact match من أجل جميع الأسئلة.

- **مقياس F1 score**

يعتمد على قياس درجة تقاطع الكلمات المشتركة بين كلمات الإجابة المتوقعة الناتجة عن النموذج مع كلمات الإجابة الفعلية، بالاعتماد على

حساب precision و recall، حيث أن الـ precision تمثل نسبة عدد الكلمات المشتركة إلى عدد كلمات الإجابة المتوقعة الناتجة عن النموذج. بينما الـ recall تمثل نسبة عدد الكلمات المشتركة إلى عدد كلمات الإجابة الفعلية. وبالتالي يتم حساب F1 score وفق العلاقة:

$$F1\ score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ثم يتم حساب متوسط قيم الـ F1 score من أجل جميع الأسئلة.

3.4. الآلية العامة لخوارزمية TSHD ومراحل عملها

تتألف خوارزمية TSHD من ثلاث مراحل رئيسية، وهي:

- **المرحلة الأولى: المعالجة المسبقة "Pre-processing"**

تُعتبر هذه المرحلة أول مرحلة ضمن الخوارزمية، حيث يتم فيها معالجة أولية للبيانات النصية الخام "raw data". تتألف هذه المرحلة من سلسلة من الخطوات المتتابعة، هي: تحويل البيانات "Data Transformation"، وتجزئة الأسطر "Lines Tokenization"، وتنظيف البيانات " Data Cleaning"، والتقييس "Normalization"، وترقيم الأسطر " Lines Enumeration"، وتنقيح الأسطر "Lines Refinement"، بهدف تجهيز البيانات للمرحلة التالية.

- **المرحلة الثانية: تحديد عناوين المقاطع "Headers Detection"**

يتم في هذه المرحلة تحديد مواقع عناوين المقاطع، بالإضافة لتوحيد تسميات عناوين المقاطع المتشابهة. وذلك عن طريق القيام بمسحين خطيين متتاليين، هما: "Cue Phrases Scan" ومن ثم "Cue Words Scan".

- **المرحلة الثالثة: التقطيع "Segmentation"**

تُعتبر هذه المرحلة آخر مرحلة ضمن الخوارزمية، حيث يتم فيها استخراج مواضيع المقاطع ومحتوياتها، وهيكلتها كأزواج بصيغة JSON.

4.4. نماذج استخراج المعلومات المستخدمة

تم تطبيق وتقييم آلية العمل المقدمة بالاستفادة من مكتبة PyTorch على ثلاثة نماذج استخراج معلومات تم إجراء ضبط دقيق لها "fine tunning" على السير الذاتية. نستعرض فيما يلي هذه النماذج:

1.4.4. نموذج autotrain-resume_parser [14]

يتبع هذا النموذج هيكلية النموذج Longformer [5]. يعتمد نموذج Longformer على تطبيق آلية انتباه "attention" تتوسع خطياً مع ازدياد طول سلسلة الدخل.

يبين **خطأ! لم يتم العثور على مصدر المرجع.** أبرز إعدادات "configuration" النموذج المطبقة

الجدول 2: أبرز إعدادات نموذج autotrain-resume_parser

اسم الخاصية	القيمة	التوصيف
hidden_act	gelu	تابع تنشيط الطبقة المخفية
hidden_size	768	حجم الطبقة المخفية (أبعاد التضمين)
max_position_embeddings	4098	حد التضمينات الأقصى
model_type	longformer	نوع النموذج
num_attention_heads	12	عدد رؤوس الانتباه
num_hidden_layers	12	عدد الطبقات المخفية
vocab_size	50265	حجم القاموس
max_length	384	الطول الأعظمي لدخل للنموذج

2.4.4. نموذج CV_Custom_DS [15]

يتبع هذا النموذج هيكلية النموذج Roberta [4]. يعد نموذج Roberta نسخة مطورة من نموذج BERT، حيث يتضمن التعديلات التالية على نموذج BERT، وهي حذف مهمة التنبؤ بالجملة التالية (Next Sentence Prediction (NSP، وتعديل طريقة الـ masking إلى Dynamic masking، بالإضافة إلى تدريب النموذج لفترة أطول على مجموعات بيانات أكبر ذات سلاسل نصية أطول. يبين الجدول 3 أبرز إعدادات النموذج المطبقة.

الجدول 3: أبرز إعدادات نموذج CV_Custom_DS

اسم الخاصية	القيمة	التوصيف
hidden_act	gelu	تابع تنشيط الطبقة المخفية
hidden_size	768	حجم الطبقة المخفية (أبعاد التضمين)
max_position_embeddings	514	حد التضمينات الأقصى
model_type	Roberta	نوع النموذج
num_attention_heads	12	عدد رؤوس الانتباه
num_hidden_layers	12	عدد الطبقات المخفية
vocab_size	50265	حجم القاموس

3.4.4. نموذج AQG_CV_Squad [16]

يتبع هذا النموذج هيكلية النموذج DistilBERT [13]. يعد نموذج DistilBERT نسخة مخففة (مصغرة) عن نموذج BERT، حيث أنه أصغر حجماً وأسرع ويتطلب موارد حاسوبية أقل من نموذج BERT، إلا أن دقة نتائجه أقل. يبين الجدول 4 أبرز إعدادات النموذج المطبقة.

الجدول 4: أبرز إعدادات نموذج *AQG_CV_Squad*

اسم الخاصية	القيمة	التوصيف
Activation	gelu	تابع تنشيط الطبقة المخفية
Dim	768	حجم الطبقة المخفية (أبعاد التضمين)
max_position_embeddings	512	حد التضمينات الأقصى
model_type	distilbert	نوع النموذج
n_heads	12	عدد رؤوس الانتباه
n_layers	6	عدد الطبقات المخفية
vocab_size	30522	حجم القاموس

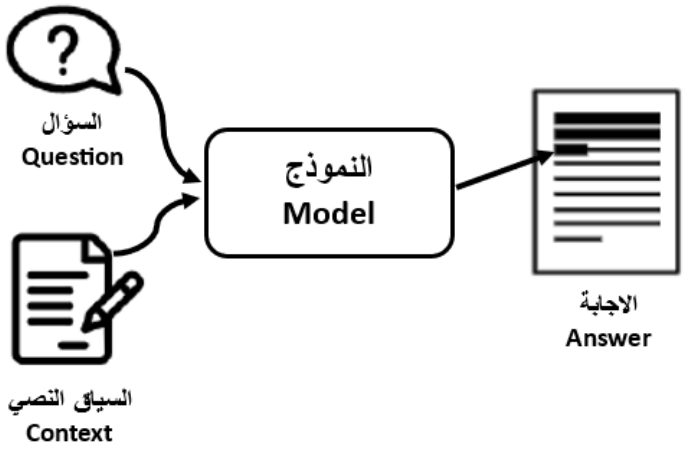
5. استخراج المعلومات باستخدام خوارزمية TSHD ونماذج المحولات

حققت نماذج المحولات قفزة كبيرة في مجالات معالجة اللغات الطبيعة، ومنها استخراج المعلومات. تسعى نماذج الإجابة عن الاسئلة القائمة على الاستخراج النصي إلى تحقيق أفضل النتائج.

يقدم هذا البحث طريقة جديدة تعمل على الاستفادة من المقاطع المستخرجة باستخدام خوارزمية TSHD، واستغلال هذه المقاطع ذات المواضيع المعروفة من قِبل نماذج استخراج المعلومات، من خلال اعتبارها السياق النصي "context" الذي يمرر للنموذج، عوضا عن الوثيقة كاملة.

سنبدأ بشرح الآلية العامة لاستخراج المعلومات باستخدام النماذج، ثم نقدم طريقة توظيف خوارزمية TSHD في تطوير آلية استخراج المعلومات بواسطة النماذج.

يوضح الشكل الآتي الشكل العام لمدخلات ومخرجات نماذج الإجابة عن الأسئلة القائمة على الاستخراج النصي.



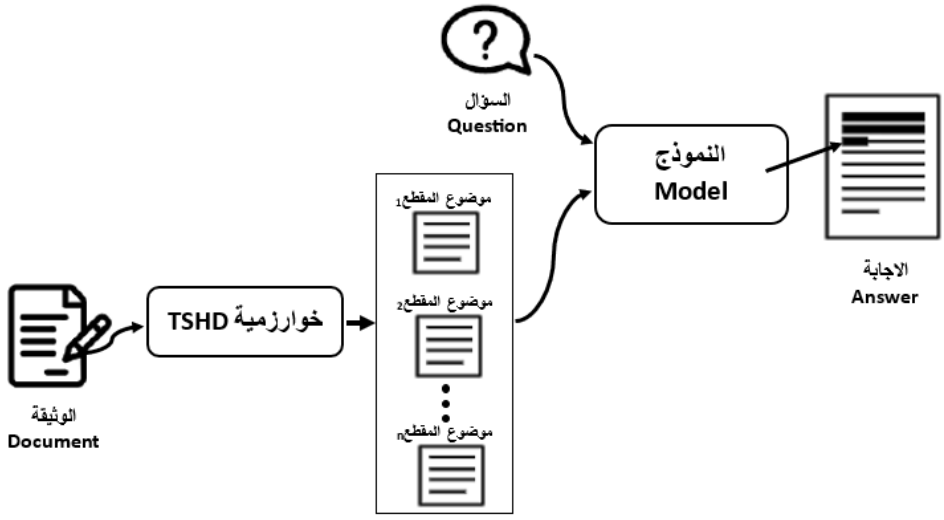
الشكل 1: مدخلات ومخرجات نموذج الإجابة عن الأسئلة القائم على الاستخراج النصي

نلاحظ من الشكل 1 طريقة استخدام النموذج في استخراج المعلومات، حيث يتم تمرير السؤال، الذي يوضح المعلومة المطلوب استخراجها، بالإضافة للسياق النصي، ويقوم النموذج بتحديد واستخراج إجابة السؤال من النص. يتم ذلك عبر المراحل الآتية:

- معالجة مسبقة لبيانات الدخل "pre-processing" لتلائم دخل النموذج من خلال تطبيق التجزئة "tokenization" وتحويل وحدات الدخل النصية إلى الأرقام "IDs" المقابلة لهم ضمن معجم المفردات "vocabulary"، ثم يتم تمريرها إلى النموذج.
- يقوم النموذج بالتنبؤ بـ start logit و end logit من أجل كل وحدة لغوية "token"، تحدد أرجحية كونها تمثل بداية أو نهاية الإجابة ضمن السياق النصي المعطى.

- يتم تطبيق معالجة لاحقة "post-processing" لتحديد مواقع الوحدات اللغوية ذوي الأرجحية الأعلى، من أجل تحديد بداية ونهاية الإجابة ضمن السياق النصي واستخراجها.

نأتي الآن إلى آلية استخراج المعلومات باستخدام النماذج وخوارزمية TSHD: يتم تمرير محتويات المقاطع المستخرجة ومواضيعها إلى النموذج عوضاً عن الوثيقة كاملة (الشكل 2).



الشكل 2: آلية توظيف خوارزمية TSHD في تطوير استخراج المعلومات من قِبل النماذج

يتم بدايةً تمرير الوثيقة كاملة إلى خوارزمية TSHD التي تقوم باستخراج محتويات مقاطع الوثيقة وتحديد موضوع كل مقطع. ثم يتم مطابقة السؤال مع موضوع المقطع المناسب، وتمرير السؤال ومحتوى المقطع المناسب إلى النموذج عوضاً عن تمرير الوثيقة كاملة، حيث يقوم النموذج باستخراج المعلومة المطلوبة (الإجابة) من المقطع، وتخزينها بصيغة مهيكلة.

6. النتائج ومناقشتها

نستعرض فيما يلي نتائج تقييم الطريقة المقترحة، حيث تم اختبارها على مجموعة من نماذج استخراج المعلومات، التي تم ذكرها سابقاً في قسم أدوات البحثً، تبعاً لهيكلية squad 1.1 في مجال السير الذاتية، لنبين تأثير تطبيق خوارزمية TSHD على دقة نتائج استخراج المعلومات من قِبل النماذج.

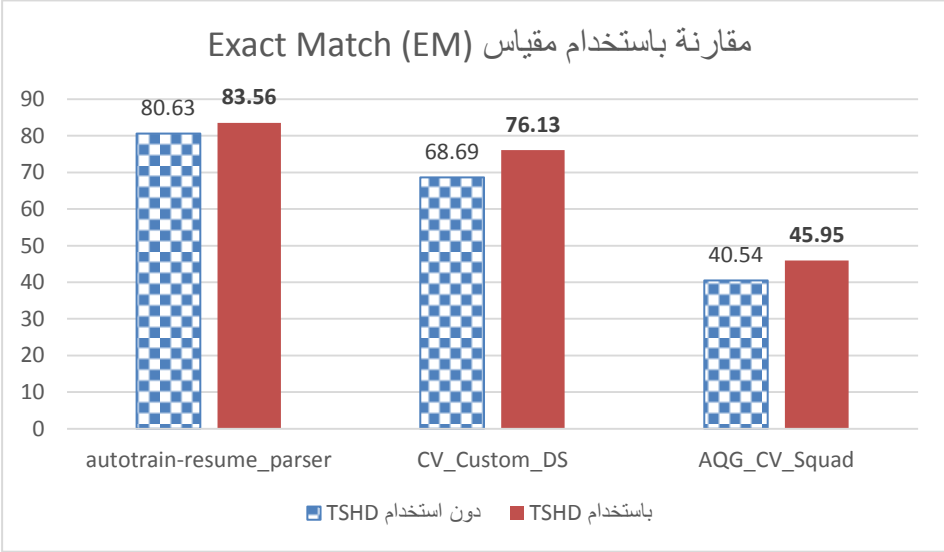
تم تقييم استخراج المعلومات على مجموعة بيانات الاختبار (444 سؤال) باستخدام ثلاثة نماذج، وهي autotrain-resume_parser و CV_Custom_DS و AQG_CV_Squad. حيث تم إجراء التقييم مرتين من أجل كل نموذج:

- الأولى دون استخدام خوارزمية TSHD، حيث يعمل النموذج على استخراج المعلومات المطلوبة من كامل الوثيقة.
- الثانية باستخدام خوارزمية TSHD، حيث يعمل النموذج على استخراج المعلومات المطلوبة من المقطع الملائم تبعاً لموضوعه.

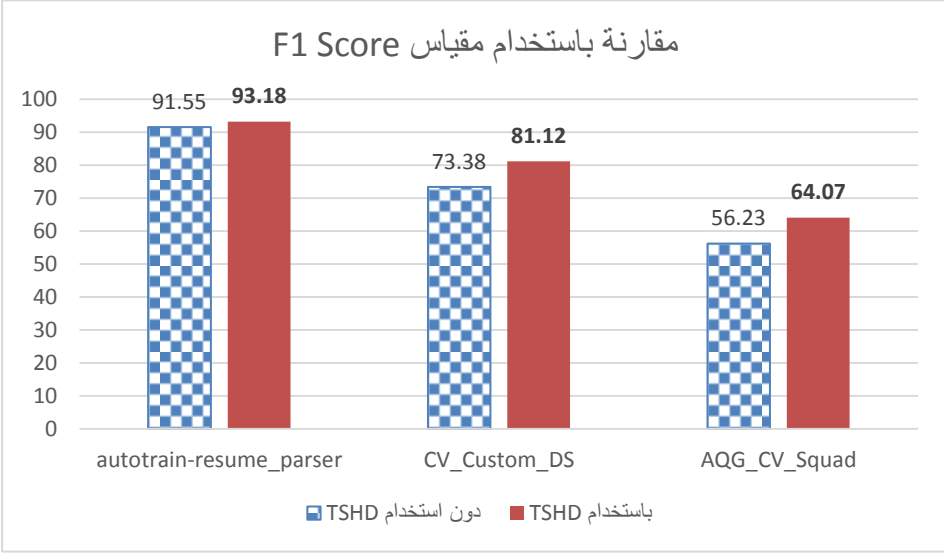
يبين كل من **خطأ! لم يتم العثور على مصدر المرجع.** والشكل 3 والشكل 4 نتائج تقييم النماذج الثلاثة في استخراج المعلومات على مجموعة بيانات الاختبار مع وبدون استخدام خوارزمية TSHD.

الجدول 5: مقارنة نتائج تقييم ثلاثة نماذج مع ودون استخدام خوارزمية TSHD

	دون استخدام خوارزمية TSHD			باستخدام خوارزمية TSHD		
	#	EM(%)	F1-score(%)	#	EM(%)	F1-score(%)
نموذج autotrain-resume_parser	358	80.63	91.55	371	83.56	93.18
نموذج CV_Custom_DS	305	68.69	73.38	338	76.13	81.12
نموذج AQG_CV_Squad	180	40.54	56.23	204	45.95	64.07
# : عدد الإجابات المتوقعة الصحيحة (مطابقة تامة) EM: Exact Match						



الشكل *3*: مقارنة تبعاً لمقياس Exact match لثلاثة نماذج مع وبدون TSHD



الشكل *4*: مقارنة تبعاً لمقياس F1 score لثلاثة نماذج مع وبدون TSHD

نلاحظ تحسن نتائج تقييم النماذج الثلاثة بمختلف مقاييس التقييم عند استخدام خوارزمية TSHD، حيث:

- من أجل نموذج autotrain-resume_parser:

❖ ارتفعت نسبة مقياس EM نتيجة استخدام خوارزمية TSHD من 80.63% (358 إجابة) إلى 83.56% (371 إجابة)، بزيادة قدرها 2.93% (13 إجابة).

❖ كما ارتفعت نسبة مقياس F1 score نتيجة استخدام خوارزمية TSHD من 91.55% إلى 93.18%، بزيادة قدرها 1.63%.

- من أجل نموذج CV_Custom_DS:

❖ ارتفعت نسبة مقياس EM نتيجة استخدام خوارزمية TSHD من 68.69% (305 إجابة) إلى 76.13% (338 إجابة)، بزيادة قدرها 7.44% (33 إجابة).

❖ كما ارتفعت نسبة مقياس F1 score نتيجة استخدام خوارزمية TSHD من 73.38% إلى 81.12%، بزيادة قدرها 7.74%.

- من أجل نموذج AQG_CV_Squad:

❖ ارتفعت نسبة مقياس EM نتيجة استخدام خوارزمية TSHD من 40.54% (180 إجابة) إلى 45.95% (204 إجابة)، بزيادة قدرها 5.41% (24 إجابة).

- ❖ كما ارتفعت نسبة مقياس F1 score نتيجة استخدام خوارزمية TSHD من 56.23% إلى 64.07%، بزيادة قدرها 7.84%.

كنتيجة لما سبق، نلاحظ التأثير الإيجابي الواضح لاستخدام خوارزمية TSHD في تحسين دقة نتائج استخراج المعلومات من قِبل نماذج المحولات، حيث ازدادت قيمة مقياس Exact match بنسبة تتراوح بين 2.93% و 7.44%، كما ازدادت قيمة مقياس F1 score بنسبة تتراوح بين 1.63% و 7.84%.

كما أن سبب تفاوت نتائج تقييم النماذج يعود لاختلاف قوة النموذج، فمثلاً: نموذج AQG_CV_Squad، الذي أعطى أقل نتائج تقييم، هو من نوع DistilBERT، حيث يعتبر أسرع وأخف مقارنةً بالنموذجين الآخرين، إلّا أنّ عدد طبقاته وبارامتراته أقل. حيث أن هدفه الرئيسي السرعة وملاءمة الموارد الحاسوبية الضعيفة، ولكن على حساب الدقة.

7. الاستنتاجات والتوصيات

قدمنا في هذا البحث طريقة جديدة لتطوير آلية عمل استخراج المعلومات من قِبل نماذج المحولات بالاستفادة من التقطيع باستخدام خوارزمية TSHD. وذلك من خلال تطبيق الخوارزمية لاستخراج محتويات مقاطع الوثيقة وموضوع كل مقطع، ومن ثم استخراج المعلومات بواسطة النماذج من المقاطع عوضاً عن الوثيقة كاملة.

تم اختبار وتقييم نتائج استخراج المعلومات من قِبل ثلاثة نماذج على مجموعة من السير الذاتية مع وبدون استخدام خوارزمية TSHD بالاعتماد على هيكلية مجموعة البيانات squad 1.1. تم تحسين نتائج التقييم باستخدام خوارزمية TSHD، حيث ارتفعت قيمة مقياس Exact match بنسبة زيادة وصلت إلى 7.44%، كما ارتفعت

قيمة مقياس F1 score بنسبة زيادة وصلت إلى 7.84%. علاوة عن ذلك، يساهم استخدام خوارزمية TSHD في تخفيض الموارد الحاسوبية اللازمة نتيجة تصغير حجم السياق النصي. كما أن استخدام الخوارزمية لا يتطلب إعادة تدريب النماذج على مقاطع الوثائق بعد التقطيع.

يمكننا أيضاً اختبار وتقييم آلية العمل المقترحة على هيكلية مجموعة البيانات squad 2، من خلال إضافة مجموعة من الأسئلة التي لا يوجد إجابة لها ضمن السياق النصي.

نرى أن خوارزميات التقطيع واكتشاف المواضيع مثل خوارزمية TSHD، لديها القدرة على إحداث تأثير كبير في العديد من مجالات استخراج المعلومات النصية، ومهام مختلفة لمعالجة اللغات الطبيعية.

8. المراجع

1. TANNOUS, M.E., RAMADAN, W.H., and RAJAB, M.A. 2023- TSHD: Topic Segmentation Based on Headings Detection (Case Study: Resumes). **Advances in Human-Computer Interaction**, Vol. 2023, 6044007.

2. VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, \Lukasz, and POLOSUKHIN, I. 2017- Attention is all you need. **Advances in neural information processing systems**, Vol. 30, .

3. DEVLIN, J., CHANG, M.-W., LEE, K., and TOUTANOVA, K. 2018- Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, .

4. LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., and STOYANOV, V. 2019- Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, .

5. BELTAGY, I., PETERS, M.E., and COHAN, A. 2020- Longformer: The long-document transformer. **arXiv preprint arXiv:2004.05150**, .

6. YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R.R., and LE, Q.V. 2019- Xlnet: Generalized

autoregressive pretraining for language understanding. **Advances in neural information processing systems**, Vol. 32, .

7. RAJPURKAR, P., ZHANG, J., LOPYREV, K., and LIANG, P. 2016- Squad: 100,000+ questions for machine comprehension of text. **arXiv preprint arXiv:1606.05250**, .

8. RAJPURKAR, P., JIA, R., and LIANG, P. 2018- Know what you don't know: Unanswerable questions for SQuAD. **arXiv preprint arXiv:1806.03822**, .

9. TRISCHLER, A., WANG, T., YUAN, X., HARRIS, J., SORDONI, A., BACHMAN, P., and SULEMAN, K. 2016- Newsqa: A machine comprehension dataset. **arXiv preprint arXiv:1611.09830**, .

10. CHOI, E., HE, H., IYYER, M., YATSKAR, M., YIH, W., CHOI, Y., LIANG, P., and ZETTLEMOYER, L. 2018- QuAC: Question answering in context. **arXiv preprint arXiv:1808.07036**, .

11. MÖLLER, T., REINA, A., JAYAKUMAR, R., and PIETSCH, M. 2020- **COVID-QA: A question answering dataset for COVID-19**. Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.

12. PEARCE, K., ZHAN, T., KOMANDURI, A., and ZHAN, J. 2021- A comparative study of transformer-based language

models on extractive question answering. **arXiv preprint arXiv:2110.03142**, .

13. SANH, V., DEBUT, L., CHAUMOND, J., and WOLF, T. 2019- DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, .

14. Kiet/autotrain-resume_parser-1159242747 · Hugging Face. https://huggingface.co/Kiet/autotrain-resume_parser-1159242747.

15. sunitha/CV_Custom_DS · Hugging Face. https://huggingface.co/sunitha/CV_Custom_DS.

16. sunitha/AQG_CV_Squad · Hugging Face. https://huggingface.co/sunitha/AQG_CV_Squad.