مراجعة في ضرورات وتحديات، التحليل الدلالي للبيانات الحكومية الموزعة

عبد الرزاق ياسين المحمد المحمود* د. محمد بسام كردي**

(الإيداع: 4 آيلول 2024، القبول: 22 2024)

الملخص:

في هذه الورقة، نقدم مراجعة لضرورة تحليل البيانات الحكومية والتحديات التي تواجهها سواءً على الصعيد المحلي أو الدولي، استخدمنا في ذلك، مقالات راجعها النظراء، تقارير من منظمات دولية، معلومات مباشرة من موظفين حكوميين، مصادر الكترونية لمبادارات حكومية، الهدف من هذه المراجعة هو لفت انتباه اصحاب القرار إلى أهمية تحليل البيانات الحكومية، باعتباره أحد المواضيع الناشئة والضرورية لتحقيق التنمية المستدامة والحكم الرشيد، استعرضنا في المراجعة بعض الحالات التي تبين أهمية تحليل البيانات في الحوكمة والتنمية والاستجابة للمتغيرات الطارئة، ناقشنا التحديات التقنية واللوجستية، راجعنا واقع تحليل البيانات في حكومات العالم، ركزنا على واقع التجربة الوطنية، ثم انتلقنا لمراجعة افضل تقنيات الذكاء الصنعي التي يمكن مواءمتها مع تحليل البيانات الحكومية، وأخيراً، ختمنا المراجعة بعدد من التوصيات.

الكلمات المفتاحية: البيانات الكبيرة، التحليل الدلالي، معالجة اللغات الطبيعية، الذكاء الاصطناعي.

*طالب دكتوراة بالمعلوماتية–الجامعة الافتراضية السورية. .

**استاذ بالجامعة الافتراضية السورية.

Semantic analysis of distributed government data in Syria

Abdul Razzaq Yassin Al-MohammedAl-Mahmoud* Dr. Mohammed Bassam Kurdi**

(Received: 4 September 2024, Accepted: 14 October 2024)

Abstract:

In this paper, we provide a review of the necessity of analyzing government data and the challenges it faces, whether at the local or international levels. In doing so, we used peer-reviewed articles, reports from international organizations, direct information from government employees, and electronic sources for government initiatives. The goal of this review is to draw the attention of decision makers to the importance of analyzing government data, as it is one of the emerging topics necessary to achieve sustainable development and good governance. In this article, we reviewed some cases that demonstrate the importance of data analysis in governance, development, and responding to emergency cases. We discussed the technical and logistical challenges. We reviewed the reality of data analysis in world governments. We focused on the reality of the national experience, then moved on to review the best artificial intelligence techniques that can be adapted to analyzing government data, and finally, we concluded the review with a number of recommendations.

Keywords: Big data, Semantic analysis, Natural language processing, Artificial intelligence.

* PhD student in Informatics - Syrian Virtual University. .
**Professor at the Syrian Virtual University.

1. المقدمة

يترافق الاستخدام المتزايد للأنظمة المعلوماتية مع ارتفاع كبير في معدل تخزين البيانات، فبحسب احصائيات الأمم المتحدة [1]، ارتفع حجم البيانات المخزنة في عام 2020 إلى 64 زيتا بايت، هذا الكم الهائل من البيانات يخفي في طياته نماذج معرفية ضمنية يمكن إظهارها باستخدام تقنيات التحليل الدلالي، يمكن الاستفادة منها في مجالات مختلفة، سواءً على المستوى الحكومي أو الخاص، ولعل القطاع الخاص كان الأسبق في استثمار مؤشرات تحليل البيانات في تعزيز الكفاءات الإدارية والتشغيلية والتنبؤ بالاتجاهات المستقبلية لسلوك السوق وفهم احتياجات العملاء، بينما تردد القطاع الحكومي في استثمار القيمة الناشئة لتحليل البيانات بسبب الطبيعة الموزعة للبيانات الحكومية وتنسيقاتها غير المتجانسة والسياسات التنظيمية والإجرائية، يضاف إلى ذلك أسباب متعلقة بجاهزية البنية التحتية والكوادر التشغيلية [2].

يعتبر تحليل البيانات الحكومية من المواضيع الناشئة والضرورية لتقييم نجاح الاستراتيجيات التنموية [3]، ومن الضرورة بمكان ملاحظة أهميته في استراتيجيّات التحوّل الرقمي، فبحسب منظمة التعاون الاقتصادي والتنمية، فإن الدول التي أهتمت بكفاءة حكوماتها الالكترونية دون الإهتمام بحوكمة البيانات، قد تأخرت عن الدول التي انتهجت سياسة البيانات المفتوحة كوسيلة لتشجيع الابتكار [4].

تنتمي البيانات الحكومية إلى أسرة البيانات الكبيرة والتي تحتاج إلى تقنيات غير عادية لتحويلها إلى معرفة مفيدة ، حيث تتجه الأنظار إلى تسخير تقنيات الذكاء الصنعي لهذا الغرض، اذ تبدو علاقة الذكاء الصنعي بالبيانات الكبيرة علاقة وثيقة، فالأول يحتاج كمية كبيرة من البيانات حتى يتعلم، لتعود البيانات الكبيرة من جهة أخرى وتستخدم خوارزميات الذكاء الصنعي في تعزيز مخرجات التحليل بنماذج معرفة ضمنية يمكن الاستفادة منها في صناعة القرار.

من خلال هذه المقدمة، حاولنا الإضاءة بشكل مختصر على أهمية تحليل البيانات الحكومية وعلاقتها بتقنيات الذكاء الصنعي. في باقي أقسام الورقة، نناقش الضرورات والتحديات، ونجري سبراً لواقع تحليل البيانات في حكومات العالم، ومراجعة لتقنيات الذكاء الصنعي، حيث ركزنا على معالجة اللغة الطبيعية وأهم مخرجاتها وهو الكلمات المضمنة Word embeddings، والتي تعتبر محركاً للنماذج اللغوية الكبيرة Large Language Model (LLM).

2. الضرورة

تسعى الحكومات من خلال التحول الرقمي إلى رفع جاهزية المؤسسات الحكومية، وإلى تحقيق مستوى جيد من التنمية المستدامة، وخلق فرص لدعم الابتكار، وفتح البيانات لمزيد من الشفافية التي ترفع من مستوى ثقة المواطن بالحكومة، لذلك فإن حوكمة البيانات واتاحتها للتحليل هو إجراء ضروري لتحقيق هذه الأهداف [5]، ولتوضيح الأمر نلقي الضوء على بعض من حالات الاستخدام التي تظهر فيها أهمية وضرورة تحليل البيانات الحكومية:

2.1. الحوكمة وصنع القرار

توفر تقنيات تحليل البيانات الكبيرة الوسائل الضرورية لجمع وتحليل البيانات دون تأخير زمني، إذ يمكن للحكومة استخدام التحليلات التنبوئة لرصد المتغيرات في الاقتصاد أو الصحة أو الأمن وتكييف سياساتها وفقاً لتلك المؤشرات، ومن ناحية أخرى، تتيح البيانات المفتوحة للمواطنين مراقبة الأداء الحكومي والمشاركة في صنع القرار، مما يعزز الثقة في المؤسسات الحكومية [6].

2.2. التنمية المستدامة

اتفقت الدول الأعضاء في الأمم المتحدة على خطة التنمية المستدامة لعام 2030، ومتابعة هذه الخطة، يتطلب جمع ومعالجة وتحليل كمية كبيرة من البيانات، لذلك فإن تحليلات البيانات الكبيرة بما توفره من تقنيات مختلفة، تساهم في تحقيق أهداف التنمية المستدامة، وقد بدا واضحاً فعالية هذه التقنيات في تحقيق أهدف خطة التنمية المستدامة خصوصاً فيما يتعلق بالأهداف 3 و7 و11 (الصحة والرفاهية، استدامة المدن والمجتمعات، وفرة الطاقة النظيفة) [7] ، بالمقابل ظهر أثر نقص البيانات على الأهداف الأخرى، وخرجت عدة دول من مؤشر التنمية، لعدم وجود البيانات اللازمة [8].

2.3. الرعاية الصحية

ينتج القطاع الصحي كميات هائلة من البيانات، 80% منها بيانات غير منظمة، تحويل هذه البيانات إلى بيانات منظمة ودمجها مع تحليلات البيانات الكبيرة، سيؤدي إلى ظهور معلومات مفيدة تسهم في رصد واكتشاف حالات مرضية او علاجية وترفع من جودة الخدمات الصحية وتُخفض التكاليف المالية [9].

2.4. الاستجابة للحالات الطارئة

لفتت جائحة كورونا انتباه الحكومات إلى أهمية تحليلات البيانات الكبيرة، وبنفس الوقت كشفت قصوراً واضحاً في جاهزية البيانات الحكومية في الاستجابة لحالات الطوارئ، إذ أن معظم المساهمات الرئيسية لتحليل بيانات الجائحة لم تأتِ من خلال منصات حكومية مجهزة لهذا الغرض، وإنما جاءت من خلال شركات تقنية مختصة بالذكاء الصنعي وتحليل البيانات [10].

3. التحديات

يواجه تحليل البيانات الحكومية مجموعة من التحديات، بسبب طبيعتها الموزعة وعدم تجانسها من الناحية البنيوية أو الدلالية، وفي بعض الأحيان تكون هذه البيانات غير مهيكلة، وبالتالي فإن تكاملها يحتاج إلى إدارة سليمة ومنظمة تعتمد معايير موحدة، وإلى أساليب تخزين قوية تراعي شروط استمرارية العمل بما فيها (سرعة الإستجابة، توزيع الأحمال، وسد الثغرات الأمنية لحماية البيانات من الهجمات الخارجية)، وهذا يتطلب استخدام أنظمة حماية وتشفير Encryption، وموزانة أحمال Load balancing تتناسب مع هذا الغرض، ومن الجانب اللوجستي فإن هذه البيانات غالباً ما تتضمن معلومات تحمل طابع السرية أو الخصوصية، ومن الضروري تقليل مخاطر تهديد الخصوصية إلى الحد الأدنى، فحماية الخصوصية من العوامل الحرجة التي قد تقوض الثقة بالمؤسسات الحكومية، مالم يتم مراعاتها [13].

4. واقع تحليل البيانات الحكومية

يتفاوت الاهتمام بتحليل البيانات الحكومية من بلد إلى آخر بحسب الظروف والمقومات، لذلك أجرينا سبراً لشرائح مختلفة من المبادرات الدولية، يمكن من خلاله فهم توجه الحكومات ومدى اهتمامها بهذا الموضوع، وقياس المسافة ما بين التجربة المحلية والتجارب الدولية، لاختيار أقصر الطرق في مواكبة التطورات المتسارعة.

4.1. الدول الصناعية المتقدمة

– الولايات المتحدة، بحسب مؤشر الذكاء الصنعي [22]، فإنها تحتل المركز الأول في استثمار تقنيات الذكاء الصنعي وتحليلات البيانات الكبيرة، معظم إداراتها تستخدم تحليلات البيانات الكبيرة، على سبيل المثال، إدارة الضمان الاجتماعي

تستخدمها لكشف المطالبات المتأخرة، هيئة الإسكان الفيدرالية تستخدمها في التنبؤ بمعدلات التخلف عن السداد، هيئة الأوراق المالية تستخدمها لمراقبة نشاط السوق المالية، وزارة الأمن الداخلي تستخدمها في مكافحة الجريمة، إدارة الغذاء والدواء تستخدمها لدراسة أنماط الأمراض المنقولة بالغذاء، وزارة التعليم تستخدمها لتعزيز التدريس والتعلم، وزارة الزراعة تستخدمها لرفع الإنتاجية والابتكار وتوفير التكاليف المالية، ناسا وإدارة النظم البيئية تستخدمها في التنبؤ بأحوال الطقس والفيضانات والبراكين ومخاطر حرائق الغابات [23] ,[24] ,[25].

– حكومات الاتحاد الأوروبي، تستخدم البيانات الكبيرة لصنع السياسات، ولديها مبادرات لحوكمة البيانات، كما أنها أطلقت نظام معلومات المدن الذكية [4],[26].

– الصين، تقوم بدمج تقنيات الذكاء الصنعي وتحليلات البيانات الكبيرة في المدن الذكية وفي مجالات مختلفة مثل النقل والأمن العام والبيئة والتصنيع، وفي هذا المجال تتفوق الصين على كل دول العالم بعدد الروبوتات الصناعية المستخدمة اذ يصل عددها إلى 290300 روبوت حسب مؤشر الذكاء الصنعي [22],[27].

– اليابان، تستخدم تحليلات البيانات الكبيرة في مجالات مختلفة مثل صناعة السفن، الطب، والزراعة الذكية، وقد أقر البرلمان الياباني قانوناً لتسريع رقمَنة الإجراءات الإدارية الحكومية، ليكون بمثابة أساس مهم لتعزيز المدن الذكية ,[30] [28] ,[29].

– سنغافورة، أطلقت في عام 2014 مبادرة الأمة الذكية لرقمَنة إجراءات وخدمات الحكومة، وفي عام 2018 حصلت على جائزة المدينة الذكية في برشلونة، وتوفر لمواطنيها البيانات المفتوحة لدعم بيئة الابتكار، وتستخدم تحليلات البيانات الكبيرة لمراقبة إجراءات السلامة وكفاءة استخدام الطاقة في المباني الذكية، ورصد الحركة المرورية، وفي وسائل النقل الذكية [31].

4.2. الدول النامية

– في ماليزيا: تستخدم القطاعات الصناعية تحليل البيانات بشكل كبير في التحقق من صحة أو تزييف النماذج والنظريات العلمية، وتدرك المؤسسات الحكومية أهمية تحليل البيانات في اتخاذ القرارات الذكية، حيث أظهر استطلاع لمايكروسوفت أن 85% من صناعي السياسات في ماليزيا يعتقدون بضرورة استخدام البيانات، بينما فقط 44% منهم بدأوا بوضع استراتيجية رقمية، في عام 2024 وبعد استثمارات مايكروسوفت في البنية التحتية، أطلقت ماليزيا خارطة طريق وطنية لإدخال الذكاء الصنعي في جميع الوزارات [19].

– في إندونيسيا: لا يزال استخدام تحليلات البيانات الكبيرة في صنع القرار أمراً نادرًا، إلا أنه مستخدم وبشكل كبير في قطاع الأعمال والشركات [32]، فبحسب شركة البيانات الدولية International Data Corporation (IDC) حقق سوق تحليل البيانات الكبيرة في إندونيسيا في النصف الأول من عام 2022 نمواً بنسبة 14.7% بعد أن زادت المؤسسات الإندونيسية استثماراتها في تحليلات البيانات الكبيرة، لتحسين اتخاذ القرارات التجارية ومواكبة تغيرات السوق [33].

– في سريلانكا: أشارت الدراسات التي اجريت على عدة قطاعات (البناء، التدقيق الخارجي، صناعة الألبسة، السياحة) أن الاعتماد على تحليل البيانات لا يزال يعاني من عوائق كثيرة أهمها الدعم الحكومي، الوعي لقيمة مخرجات التحليل، والخوف من فقدان البيانات [16]، [34].

– في المغرب: يواجه تحليل البيانات الحكومية مجموعة من التحديات أهمها نقص في التمويل والموارد البشرية، عدم وجود مواقع وظيفية لمحللي البيانات في الإدارات، ضعف في البنية التحتية اللازمة لتخزين البيانات، عدم تجانس البيانات، بالاضافة لهاجس خصوصية البيانات وأمن المعلومات [14] .

– في فلسطين: يفرض الوضع السياسي قيوداً على عملية الاستيراد، مما يعيق تطور البنية التحتية اللازمة لتحليل البيانات الكبيرة، فضلاً عن قلة خريجي الجامعات المختصين بهذا المجال [17].

– السعودية: في عام 2019، أسست الهيئة السعودية للبيانات والذكاء الصنعي (سدايا) لتكون المسؤولة عن حوكمة البيانات والذكاء الصنعي، ومواءمة استخدامات البيانات والذكاء الصنعي في القطاعات الحكومية (التعليم، الصحة، الطاقة، النقل)، يبلغ حجم الاستثمار بهذا القطاع 380 مليون دولار أمريكي لعام 2024، ومن المتوقع أن يرتفع في عام 2029 إلى 2.19 مليار دولار أمريكي [38].

– الامارات العربية: أطلقت تحت مظلة الحكومة الذكية مبادرات عديدة متعلقة بالذكاء الصنعي وتحليلات البيانات الكبيرة [11]، منها على سبيل المثال: مبادرة وزارة الصحة لربط السجلات الصحية للاستفادة من تحليلات البيانات في تحسين نظام الرعاية الصحية والتنبؤ بالمشكلات الصحية المستقبلية [20]، استراتيجية البيانات الذكية، استراتيجة دبي للبيانات المفتوحة، استراتجية الذكاء الصنعي [36].

4.3. المنظمات الدولية

– منظمة الأمم المتحدة: في عام 2017، عقدت المنتدى الأول للبيانات في كيب تاون، ودعت الأعضاء إلى اتخاذ إجراءات حاسمة لتطوير آليات إنتاج البيانات لإرشاد قرارات السياسات الإنمائية، وأكدت على ضرورة دعم الحكومات والأوساط الأكاديمية لهذا الهدف [12]. وفي الجلسة الثالثة والخمسون للمجلس الاقتصادي والاجتماعي في الأمم المتحدة، أكد تقرير لجنة خبراء البيانات على ضرورة استخدام البيانات الكبيرة في المكاتب الوطنية للإحصاء لرصد خطة التنمية المستدامة [15].

– بنك التنمية الآسيوي: يستخدم تحليلات البيانات الكبيرة في تصميم برامجه التنموية، لتحديد الفئات الأكثر حاجةً، كما يستخدمها لمراقبة انتشار الأمراض الوبائية في البلدان النامية، ويساعد الدول النامية على تجهيز البنى التحتية اللازمة لتحليلات البيانات الكبيرة في قطاعات متنوعة (المرور، النقل، الزراعة الذكية، نظم المعلومات الجغرافية، الاستشعار عن بعد، رصد البيئة، كفاءة استخدام الموارد، المصادر المستدامة للمواد، كفاءة الطاقة والمياه، وإدارة التلوث) [35].

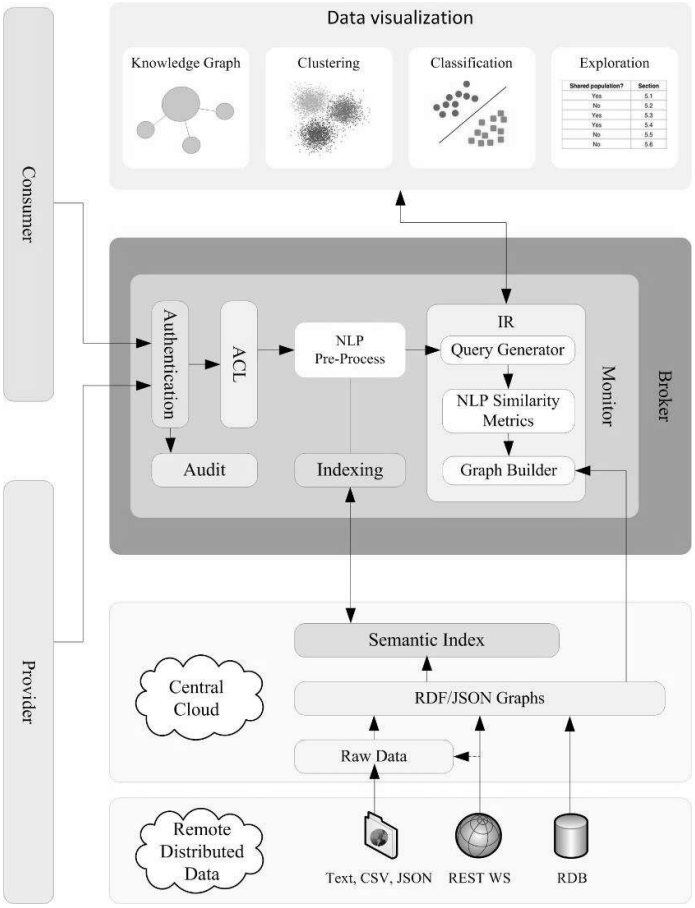
– منظمة التعاون الاقتصادي والتنمية: تعتمد حكومات المنظمة خططاً شاملة لإدارة البيانات الحكومية كوسيلة لتماسك القطاع العام وتشجيع الهيئات الحكومية على حوكمة البيانات وإدارتها على مدار دورة حياة البيانات الكاملة (تصميم، تخزين، نشر، اتلاف)، وذلك للاستفادة من البيانات الحكومية كأصل استراتيجي في صناعة القرار[4].

4.4. سوريا

تأثرت سوريا بشكل كبير بسبب الأحداث والعقوبات المفروضة عليها، وبالرغم من ذلك، فإن الحكومة تسعى لتجاوز هذه العقبات، ففي عام 2021 اعتمدت الحكومة استراتيجية التحول الرقمي للخدمات الحكومية والتي تقوم على ست ركائز (الثقة، الشفافية، المساءلة، التنمية، التكامل، والتمحور حول المواطن) [39]، وفي عام 2023، وبحسب وزارة الاتصالات، فقد تم تنفيذ عدة أعمال في اطار بناء القدرات وتأهيل الكوادر البشرية ونشر الوعي بأهمية التقنيات الناشئة، حيث تم تدريب ممثلين من الجهات العامة على مفاهيم ادارة البيانات وحوكمتها وضرورة الاهتمام بدقة البيانات، كما تم اطلاق منظومة للتوافق البيني لربط ثلاث من السجلات الوطنية (المدني، المركبات، المحروقات) وجاري العمل على ربط السجل التجاري و سجل العاملين، ومن المخطط له في عام 2025 اطلاق سياسة خاصة بالذكاء الصنعي.

في عمل سابق لنا قدمنا نموذجاً أولياً للتحليل الدلالي لبيانات الحكومة السورية الموزعة [40]، تضمن الأسس التنظيمية لإطار عمل مشترك بين القطاعات الحكومية لتكامل وتحليل البيانات الحكومية الموزعة، إبتداءً بجمع وتخزين البيانات،

مروراً بمعالجتها وتحليلها، وإنتهاءً بتصورها، تكوّنَ النموذج من خمس طبقات (مزود ومستهلك البيانات، بحيرة البيانات، الوسيط، تصوّر البيانات)، حيث يستلم النموذج البيانات من مزودي البيانات بصيغ مختلفة (منظمة، غير منظمة، مادية، افتراضية) ويخزنها بصيغة موحدة RDF/JSON في بحيرة البيانات لتحقيق تجانس البيانات، يقوم الوسيط بمعالجة البيانات وفهرستها وتقديمها للمستهلكين كخدمة من خلال طبقة تصور البيانات (Data Analysis as a Service (DAaaS، وفقاً للشروط الأمنية والحقوقية التي يحددها المزود، يوائم الوسيط ما بين المهام التي يقوم بها (معالجة، فهرسة) وتقنيات الذكاء الصنعي لتحقيق التكامل الدلالي، كما هو موضح بالشكل رقم (1).



الشكل رقم(1): النموذج الأولي للتحليل الدلالي للبيانات الحكومية الموزعة

5. المواءمة مع تقنيات الذكاء الصنعي

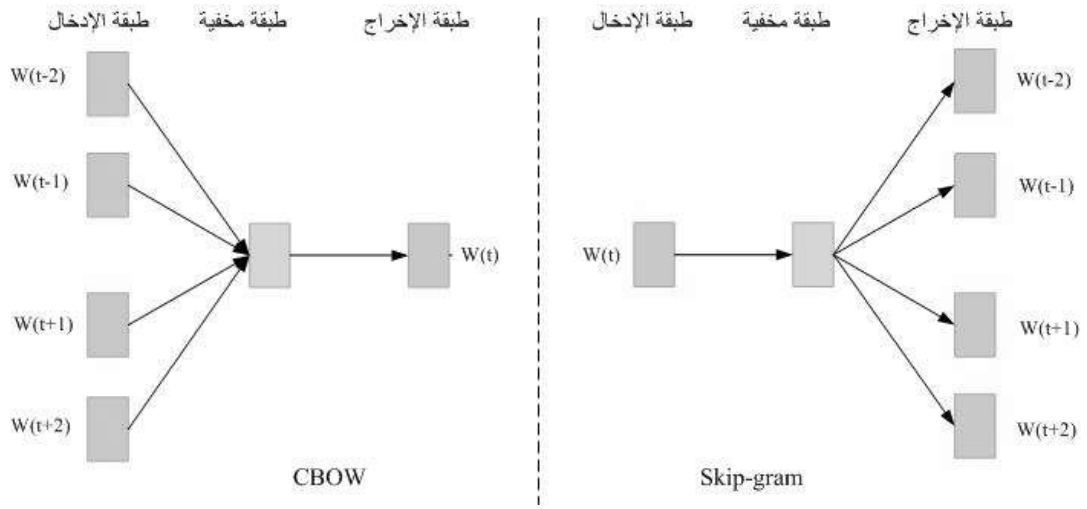
من المتوقع أن يكون الذكاء الصنعي قوةً دافعةً للنمو الاقتصادي، وهناك اعتقاد واسع النطاق أن الحكومات التي تستثمر في الذكاء الصنعي وتتبنى استراتيجيات لدمجه في الأعمال، ستتمتع بمزايا تنافسية عن الحكومات الأخرى، وقد يشكل ذلك انقساماً واضحاً بهيمنة دول على دول أخرى [37]، ولتحديد المسار الأقصر لمواكبة هذه التقنيات مع الأخذ بعين الإعتبار عدة محددات (الجودة، الكلفة، الزمن)، نجري مراجعة سريعة لتطور هذه التقنيات، بحيث تسمح لنا هذه المراجعة بالوصول إلى الهدف دون اضاعة الوقت بتجريب او تقييم ما لم يعد يتناسب مع سرعة التطورات المتلاحقة.

الشكل رقم(2): التقنيات الأساسية للذكاء الصنعي

يوضح الشكل رقم (2) تقنيات الذكاء الصنعي الأساسية: التعلم الآلي، التعلم العميق، ومعالجة اللغة الطبيعية.

يَستخدم التعلم الآلي شبكات عصبية مكونة من طبقة إدخال، وطبقة أو طبقتين مخفيتين، وطبقة إخراج، وعادة ما يكون هذا النوع من التعلم خاضع للإشراف، أي أن تصنيف البيانات يتم من خلال مختصين بالمجال المعرفي للبيانات، بينما يستخدم التعلم العميق شبكات عصبية عميقة، تختلف عن سابقتها بأن عدد طبقاتها المخفية أكثر من ثلاث طبقات وقد يصل إلى مئات الطبقات، ويكون هذا النوع من التعلم غير خاضع للإشراف حيث تتولى خوارزميات التعلم العميق استخراج الميزات من البيانات بشكل آلي.

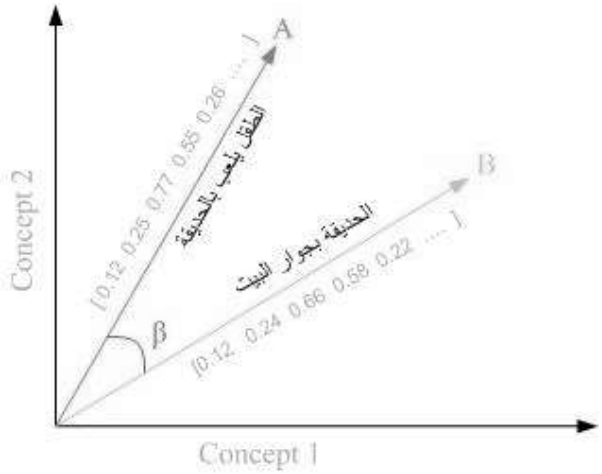
أما معالجة اللغة الطبيعية فتختص بتحليل النصوص وفهم البنية التركيبية والنحوية للجمل، وتستخدم لهذا الغرض أساليب بسيطة مثل: تجزئة الجمل Sentence segmentation، تفريق الكلمات Word tokenization، إزالة الكلمات الشائعة Stop word، تجذير الكلمات Lemmatization، تجذيع الكلمات Stemming، التعرف على الكينونات Named Entity Recognition NER، وضع إشارات على أجزاء الجملة لفهم المعاني وتميز اسماء العلم والصفات Part-Of-Speech Tagging، وتستخدم أساليب أخرى أكثر تطوراً، تم بناؤها على الأساليب السابقة مثل: حقيبة الكلمات Bag Of Word (BOW) وهي مصفوفة لتكرار الكلمات في النص، من عيوبها الأبعاد الكبيرة للمصفوفة ، وتعتبر حالة خاصة من تقنية أكثر أهمية هي N-Gram، التي تقوم بتقسيم نص من K كلمة إلى سلسلة كلمات متجاورة وفق الصيغة التالية: $N-Grams_k = K-(N-1)$، حيث N هو عدد الكلمات المتجاورة، في مرحلة تالية، تم دمج تقنيات معالجة اللغة الطبيعية مع التعلم الآلي لإنتاج نموذج شبكة عصبية تعرف باسم حقيبة الكلمات المستمرة Continuous Bag of Words (CBOW) يقوم هذا النموذج بأخذ نافذة من الكلمات المحيطة بكلمة معينة w(t) كمدخلات لنموذج تعلم آلي يتم تدريبه على التنبؤ بالكلمة المستهدفة w(t)، كما تم إنتاج نموذج آخر يدعى Skip-gram ويقوم هذا النموذج بعمل معاكس لنموذج CBOW، فيتنبأ بالكلمات المحيطة بالكلمة المستهدفة w(t)، كما هو موضح بالشكل رقم (3).



الشكل رقم (3): نموذج CBOW ونموذج Skip-gram

في عام 2013 حصلت قفزة نوعية، حيث لاحظ توماس ميكولوف [21]، أن نماذج حقيبة الكلمات تعاني من نقطتي ضعف رئيسيتين: 1) أنها تفقد ترتيب الكلمات 2) تتجاهل أيضًا دلالات الكلمات، فاقترح خوارزمية غير خاضعة للإشراف لتمثيل كلمات النصوص بمتجهات او ما يعرف باسم الكلمات المضمنة Word embedding وهي تقنية لتحويل الكلمات إلى مصفوفات رقمية تُمثل كمتجهات يمكن حساب المسافة بينها باستخدام المسافة الإقليدية Euclidean Distance أو حساب تشابه جيب التمام Cosine Similarity، وفق الصيغة التالية: $Cosine\ (\beta) = \frac{A.B}{|A|.|B|} = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}.\sqrt{\sum_{i=1}^{n} B_i^2}}$

، حيث A ,B متجهين نبحث عن التشابه بينهما، و n هو عدد الكلمات في كلّ متجه كما هو موضح بالشكل رقم (4).



الشكل رقم (4): حساب المسافة بين المتجهين A,B

قام ميكولوف بتدريب الخوارزمية باسلوبين مختلفين، مرة باستخدام نموذج CBOW، ومرة أخرى باستخدم نموذج Skip-gram، ثم قامت غوغل بتدريب الخوارزمية على 100 مليار كلمة من النصوص الإخبارية التي بحوزتها، وعلى 1.5 مليار كلمة من نصوص ويكيبيديا، وأنتجت نموذجاً لغوياً لتمثيل الكلمات بات يعرف باسم Word2Vec، في عام 2014 طورت جامعة ستانفورد خوارزمية لتمثيل الكلمات بمتجهات، ولكن باسلوب أحصائي قائم على حساب احتمال تجاور الكلمات في النصوص، ودربت الخوارزمية على 840 مليار كلمة، وأنتجت نموذجاً لغوياً لتمثيل الكلمات، يعرف باسم GloVe، في عام 2016 أجرت فيسبوك تحسينات على نموج Word2Vec وأنتجت نموذجها FastText، في عام

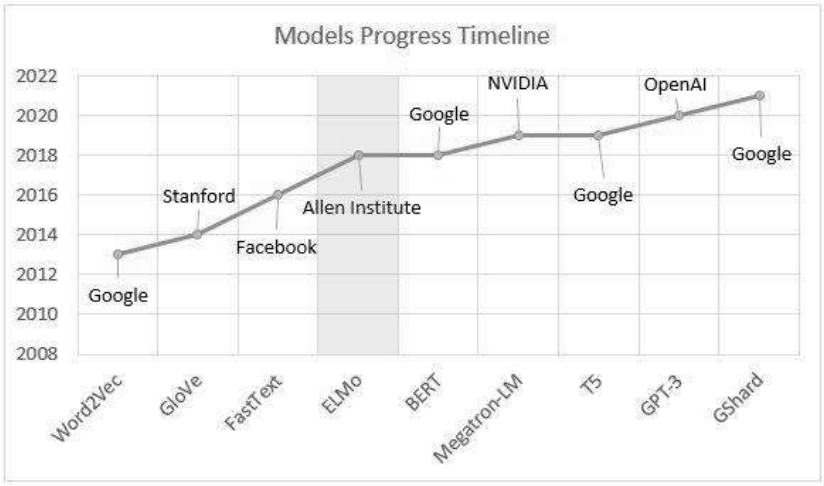
2017 نشر فريق الأبحاث في شركة غوغل ورقة بعنوان "الانتباه هو كل ما تحتاجه" [18]، فأحدثت هذه الورقة قفزة نوعية جديدة في معالجة اللغات الطبيعية والذكاء الصنعي، حيث اقترحوا استخدام المحول Transformer، وهو عبارة عن بنية نموذجية تعتمد على آلية انتباه متعددة الرؤوس، كل رأس انتباه يركز على جزء معين من النص ويُقيّم علاقته مع باقي أجزاء النص، مما يمكّن المحول من فهم العلاقات طويلة المدى بين كلمات الجملة، في عام 2018 ظهرت ثلاث نماذج لغوية، حيث تمكن معهد ألين للذكاء الاصطناعي من انتاج نموذجه ELMo والذي يتيح تضمين الكلمات بتمثيلات شعاعية ديناميكية اعتمادًا على السياق الذي استُخدمت فيه، أما غوغل فاستَخدمت المحولات في انتاج نموذجها BERT، وكذلك استخدم مختبر أبحاث الذكاء الاصطناعي OpenAI المحولات في نموذجهم اللغوي الأول GPT والذي أظهر قدرات لافتة في إنشاء النصوص. في عام 2022 أطلقت OpenAI النسخة الرابعة من نموذج GPT والذي لفت انتباه العالم لأهمية الذكاء الصنعي والقدرات المذهلة التي يوفرها، بهذه الطريقة بدأ عصر جديد من نماذج اللغات الكبيرة (LLM) التي تستخدم مئات مليارات المعاملات أو البرامترات Parameters لفهم اللغة ودلالاتها، ولم يعد ممكناً تجاهل مزايا النماذج اللغوية، والتي بطبيعة الحال توفر إمكانية الاستفادة من مقدراتها، بعدة أساليب كالتخاطب مع واجهات التطبيقات APIs، وهذا الأسلوب مقيد بمزايا النموذج والوظائف التي توفرها واجهة التطبيقات، او من خلال الضبط الدقيق للنموذج Fine-Tuning، أي تحسين أداء النموذج باعادة ضبط برامتراته وتدريبه على البيانات المستهدفة، والخيار الثالث هو بناء وتدريب نموذج لغوي جديد، وهذا يتطلب حجوماً هائلةً من البيانات وشهوراً من التدريب، مما يعني الحاجة لمصادر حوسبة عالية المواصفات وتكاليف مالية تقدر بملايين الدولارات، ولهذا السبب يطرح الباحثون سؤالاً: هل الحجوم والمصادر والتكلفة هي الأساس، أم أنه يمكن بناء نماذج رشيقة، تحقق الدقة والجودة في الإستدلال، باستخدام بيانات مصنفة وبعدد قليل من البرامترات؟.

الإجابة على هذا التساؤل، يتطلب عملاً بحثياً دقيقاً للمفاضلة بين الخيارات المشار إليها من الناحية التطبيقية وتقييم النتائج، وهو محور عملنا المستقبلي لتطوير النموذج الأولي المشار إليه في [40]، في هذه المراجعة نجري مقارنة مقتضبة بين النماذج اللغوية من خلال الجدول رقم (1) لفهم امكانيات ومزايا النماذج وتحديد الخيار الأمثل الذي يلبي متطلبات تحليل البيانات الحكومية الموزعة.

الجدول رقم (1): نماذج تمثيل الكلمات والنماذج اللغوية الكبيرة

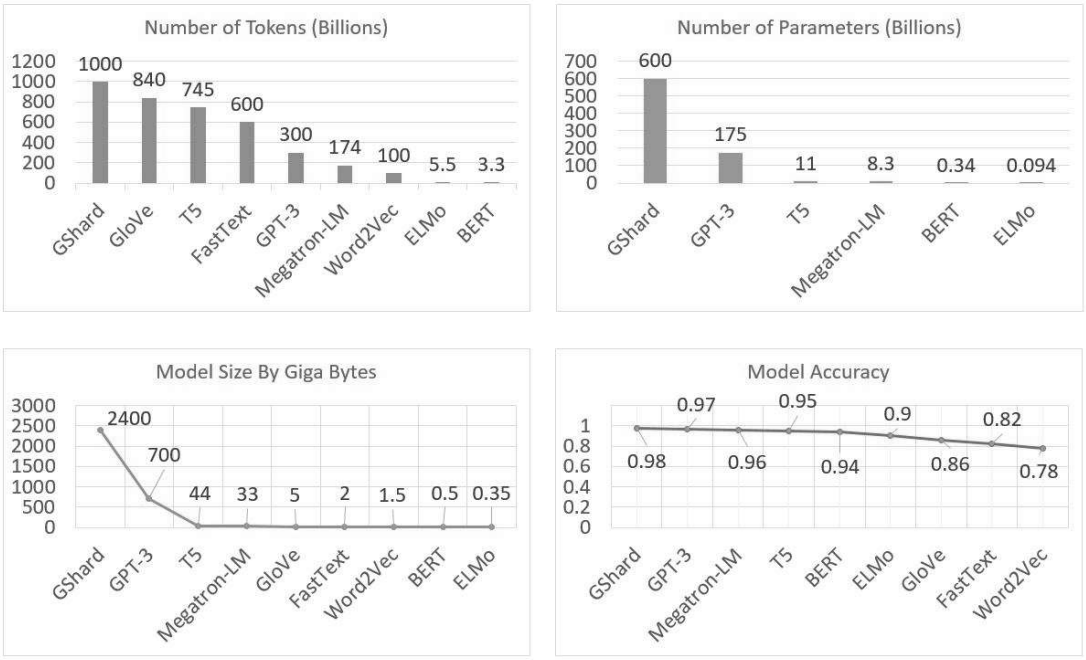
النموذج	السنة	المطور	Tokens (مليار)	Parameters (مليار)	Size (غيغا)	الدقة	التصنيف	الميزات
Word2Vec	2013	Google	100	حسب عدد المفردات	1.5	0.78	تمثيل الكلمات	سريع
GloVe	2014	Stanford	840	حسب عدد المفردات	5	0.86	تمثيل الكلمات	بطيء لكن دقيق
FastText	2016	Facebook	600	حسب عدد المفردات	2	0.82	تمثيل الكلمات	فعّال للكلمات النادرة
ELMo	2018	Allen Institute	5.5	0.094	0.35	0.9	ما بين التمثيل والنماذج	تمثيل ديناميكي للكلمات
BERT	2018	Google	3.3	0.34	0.5	0.94	نموذج لغوي	فهم عميق للنصوص
Megatron-LM	2019	NVIDIA	174	8.3	33	0.96	نموذج لغوي	تحسين النماذج الكبيرة باستخدام البنية التحتية
T5	2019	Google	745	11	44	0.95	نموذج لغوي	الترجمة، التلخيص
GPT-3	2020	Open AI	300	175	700	0.97	نموذج لغوي	توليد النصوص، الإجابة على الأسئلة
GShard	2021	Google	1000	600	2400	0.98	نموذج لغوي موزع	تسريع التدريب، تحسين الترجمة

الشكل رقم (5) يوضح التدرج الزمني لظهور نماذج التمثيل الثابت للكلمات FastText, GloVe, Word2Vec ثم التمثيل الديناميكي للكلمات ELMo والذي يعتبر مرحلة وسيطة ما بين نماذج تمثيل الكلمات والنماذج اللغوية الكبيرةGPT-3, T5, Megatorn-LM, BERT وأخيراً النماذج اللغوية الموزعة GShard.



الشكل رقم (5) : التدرج الزمني لتطور النماذج والجهة المطورة

الشكل رقم (6) يوضح مؤشرات دلالية لدقة النماذج بالمقارنة مع كلفة التدريب (عدد المقاطع Tokens، عدد البرامترات Parameters، حجوم النماذج المدربة).



الشكل رقم (6): مؤشر دقة النماذج بالمقارنة مع كلفة التدريب

6. النتائج

من الواضح أن النماذج اللغوية تتجه وبشكل مضطرد نحو استخدام كميات أكبر من المقاطع والبرامترات في مرحلة التدريب لزيادة الدقة وإضافة مهام جديدة، حيث لوحظ ارتفاع عدد المقاطع المستخدمة لتدريب النماذج من 3.3 مليار مقطع في النموذج BERT إلى تريليون مقطع في النموذج GShard، وارتفع عدد البرامترات من 94 مليون برامتر في النموذج ELMo إلى 600 مليار برامتر في النموذج Gshard، وهذا التوجه تطلب موارد حوسبة كبيرة، حيث تم توزيع النموذج GShard على آلاف الحواسيب لتقليل زمن التدريب، وكذلك الأمر ازدادت أحجام النماذج، فارتفع الحجم من 350 ميغا بايت في النموذج ELMo إلى 2.4 تيرا بايت في النموذج GShard، ولكن من جهة أخرى، فإن دقة النماذج –والتي هي بطبيعة الحال دقة مرتفعة– لم تسجل تحسناً كبيراً، فالفارق بين دقة BERT و GShard هو 0.04 وهذا تحسن طفيف اذا ما قورن بموارد الحوسبة المستهلكة، إلا أن التَحسُن الواضح كان في المهام المضافة إلى النماذج كتوليد النصوص، الإجابة على الأسئلة، الترجمة، ... وهذه الوظائف على أهميتها، إلا أنها أقل أهمية من دقة النموذج بالنسبة لتحليل البيانات.

7. المناقشة

في القسم الثاني من هذه الورقة، أوضحنا أهمية تحليل البيانات الحكومية بما توفره من إمكانيات استدلالية تسمح بالكشف عن أنماط معلومات مفيدة ورؤية مبنية على أحدث المعلومات، تساهم في رفع كفاءة الخدمات، سرعة اتخاذ القرارات، تطوير الخطط الوقائية لمواجهة الكوارث الصحية مما يعزز بشكل عام مقومات الحكم الرشيد، ناقشنا في القسم الثالث التحديات التي تعيق تكامل وتحليل البيانات الحكومية الموزعة ذات الطبيعة غير المتجانسة من الناحية البنيوية والدلالية. وفي القسم الرابع، استعرضنا واقع تحليل البيانات في الحكومات والمنظمات الدولية، حيث بدا أن كثيراً من الحكومات تتجه نحو تبني تقنيات الذكاء الصنعي في تحليل البيانات، وأن المسافة تتسع ما بين الحكومات التي نفذت استراتيجات التحول الرقمي وحوكمة البيانات والحكومات التي مازالت مترددة باتخاذ هذه الخطوة، في القسم الخامس أجرينا مراجعة لمراحل تطور تقنيات الذكاء الصنعي ومعالجة اللغة الطبيعية، مع التركيز على التقنيات التي يمكن تسخيرها في التحليل الدلالي للبيانات الحكومية (نماذج تمثيل الكلمات والنماذج اللغوية الكبيرة)، أجرينا مقارنة ما بين النماذج من حيث الكلفة والدقة وتبين أن حجم النموذج أو ضخامة الموارد الحاسوبية ليست شرطاً لتحقيق الدقة، فاستخدام نموذج رشيق مثل BERT مع اجراء بعض التحسينات Fine-Tuning يمكن أن يحقق الدقة والجودة اللازمة لتكامل وتحليل البيانات الحكومية الموزعة، دلالياً.

8. التوصيات

- إطلاق مبادرة لحوكمة البيانات، تتضمن التوجيهات التنظيمية والتشريعية التي تسمح بتبادل البيانات بين الجهات المعنية.
- وضع معايير لضمان جودة ودقة البيانات الحكومية.
- وضع استراتيجية لدمج الذكاء الصنعي في الخدمات الحكومية وتحليل البيانات.
- استخدام النموذج BERT لتحقيق التكامل الدلالي للبيانات الحكومية الموزعة.
- تطبيق التدابير الأمنية لحماية البيانات من أي تهديد خارجي أو اختراق غير مرخص.
- مراعاة سرية البيانات وحماية الخصوصية.
- تعزيز البنية التحتية اللازمة لجمع وتحليل البيانات، واعتماد الحوسبة السحابية لتخزين البيانات.
- إعداد الكوادر البشرية وتوعيتهم لأهمية تحليل البيانات باستخدام التقنيات الذكية.

9. المراجع

1. United Nations, Big Data for Sustainable Development, Retrieved from: https://www.un.org/en/global-issues/big-data-for-sustainable-development.
2. Almahmoud, A. A. (2020, September). E-Services Integration Framework Based on SOA. In Proceedings of the 2020 12th International Conference on Information Management and Engineering (pp. 1-6).
3. Abouchabaka, J., & Bentaleb, A. (2024). The impact of Big Data on the Sustainable Development Goals water and energy. In E3S Web of Conferences (Vol. 477, p. 00062). EDP Sciences.
4. OECD (2019), The Path to Becoming a Data-Driven Public Sector, OECD Digital Government Studies, OECD Publishing, Paris, https://doi.org/10.1787/059814a7-en.
5. Duggar, D., Bang, S. M., & Tripathy, B. K. (2023). Big Data Analytics in E-Governance and Other Aspects of Society. In Encyclopedia of Data Science and Machine Learning (pp. 116-128). IGI Global.
6. Hochstetter-Diez, J., Negrier-Seguel, M., Diéguez-Rebolledo, M., Vásquez-Morales, F., & Sancho-Chavarría, L. (2023). Governance Democratic and Big Data: A Systematic Mapping Review. Sustainability, 15(16), 12630.
7. Ullrich, A. (2022). Opportunities and Challenges of Big Data and Predictive Analytics for Achieving the UN's SDGs.
8. Sachs, J.D., Lafortune, G., Fuller, G., Drumm, E. (2023). Implementing the SDG Stimulus. Sustainable Development Report 2023. Paris.
9. Muhunzi, D., Kitambala, L., & Mashauri, H. (2023). Big Data Analytics in the Healthcare Sector: Opportunities and Challenges in Developing Countries. A Literature Review.
10. Mehta, N., & Shukla, S. (2022). Pandemic analytics: how countries are leveraging big data analytics and artificial intelligence to fight COVID-19?. SN Computer Science, 3(1), 54.
11. Kshetri, N. (2020). Artificial Intelligence in Developing Countries. IT Prof., 22(4), 63-68.
12. UN Statistical Commission. (2017). Cape Town global action plan for sustainable development data. Adopted by the UN Statistical Commission at its 48th Session.
13. Rahaman, M. M., Haque, A., & Hasan, M. F. (2021). Challenges and Opportunities of Big Data for Managing the E-Governance. International Journal of Business, Technology and Organizational Behavior (IJBTOB), 1(6), 490-499.

14. Khtira, R., Elasri, B., & Rhanoui, M. (2017, March). From data to big data: Moroccan public sector. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (pp. 1-6).
15. UN DESA, Report of the Committee of Experts on Big Data and Data Science for Official Statistics, 53rdrd Session of the UN Statistical Commission, 1–4 March 2022,E/CN.3/2022/25.
16. Atapattu, A. M. D. S., Wattuhewa, R. M., Waidyasekara, K. G. A. S., & Dilakshan, R. (2023). Big data analytics in the Sri Lankan construction industry: an assessment of the challenges and strategies.
17. Abu Afifa, A., & Abu-Assab, S. (2023). Big Data in the Telecommunication Sector in Palestine: Challenges and Opportunities. Artificial Intelligence (AI) and Finance, 934-944.
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
19. Sani, M. K. J. A., Zaini, M. K., Sahid, N. Z., Shaifuddin, N., Salim, T. A., & Noor, N. M. (2021). Factors influencing intent to adopt big data analytics in Malaysian government agencies. International Journal of Business and Society, 22(3), 1315-1345.
20. Alzaabi, O., Al Mahri, K., El Khatib, M., & Alkindi, N. (2023). How big data analytics supports project manager in project risk management–cases from UAE health sector. International Journal of Business Analytics and Security (IJBAS), 3(1), 11-26.
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
22. Maslej, Fattorini, et al. (April 2024) The AI Index 2024 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.
23. Helms, J. (2015). Five examples of how federal agencies use big data. Business of Government Blog, (February 25). Retrieved from: https://www.businessofgovernment.org/blog/five-examples-how-federal-agencies-use-big-data
24. Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From big data to knowledge in the social sciences. The Annals of the American Academy of Political and Social Science, 659(1), 16-32.

25. Mockshell, J., Nielsen Ritter, T., Asante–Addo, C., Ankamah–Yeboah, I., & Wong, K. (2021). CGIAR Platform for Big Data in Agriculture–Inspire Challenge Review (2017–2020).
26. European Commission. (2020). Smart Cities. The European Union. EU Regional and Urban Development. Retrieved from: https://commission.europa.eu/eu–regional–and–urban–development/topics/cities–and–urban–development_en
27. Wang, K., Zhao, Y., Gangadhari, R. K., & Li, Z. (2021). Analyzing the adoption challenges of the Internet of things (Iot) and artificial intelligence (ai) for smart cities in china. Sustainability, 13(19), 10983.
28. Tonegawa–Kuji, R., Kanaoka, K., & Iwanaga, Y. (2023). Current status of real–world big data research in the cardiovascular field in Japan. Journal of Cardiology, 81(3), 307–315.
29. Lim, J. H., Kim, J. H., & Huh, J. H. (2023). Recent trends and proposed response strategies of international standards related to shipbuilding equipment big data integration platform. Quality & Quantity, 57(1), 863–884.
30. Li, D., Nanseki, T., Chomei, Y., & Kuang, J. (2023). A review of smart agriculture and production practices in Japanese large-scale rice farming. Journal of the Science of Food and Agriculture, 103(4), 1609–1620.
31. Shamsuzzoha, A., Nieminen, J., Piya, S., & Rutledge, K. (2021). Smart city for sustainable environment: A comparison of participatory strategies from Helsinki, Singapore and London. Cities, 114, 103194.
32. Supriyanto, E. E., Warsono, H., & Herawati, A. R. (2021). Literature Study on the Use of Big Data and Artificial Intelligence in Policy Making in Indonesia. Administratio, 12(2), 139–153.
33. International Data Corporation (IDC). (2023). Indonesia Big Data and Analytics Software Market. Retrieved from: https://www.idc.com/getdoc.jsp?containerId=prAP50219423
34. Perera, M. C. D., & Abeygunasekera, A. W. J. C. (2021). Big data and big data analytics in external auditing: motivations and challenges. International Journal of Accounting & Business Finance, 7, 1–16.
35. Asian Development Bank (ADB), (2022). Strategy 2030 Digital Technology Directional Guide. Retrieved from: https://www.adb.org/documents/strategy–2030–digital–technology–directional–guide
36. The United Arab Emirates' Government portal, Retrieved from https://u.ae/en/about–the–uae/digital–uae.

37. Ozkaya, G., & Demirhan, A. (2023). Analysis of countries in terms of artificial intelligence technologies: PROMETHEE and GAIA method approach. Sustainability, 15(5), 4604.
38. Saudi Data & AI Authority (SDAIA), Retrieved from: https://ai.sa
39. F. Sulaimn, (2021). Digital transformation strategy in Government services in Syria, Retrieved from: https://www.unescwa.org/sites/default/files/event/materials/06-Sy-DT-strategy-MOCT.pdf
40. Alamhmoud, A. A., & Kurdy, M. B. (2024, January). A Semantic Analysis Prototype for Distributed Syrian Government Data. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 692-697). IEEE.