المجلة
العربية
مجلة الثقافة العربية
585

لويس بلان: معاداة الإسلام كره لفرنسا

حازم صاغية: لا أريد أن أصيب القارئ بالضجر!

تحولات الحج عبر التاريخ

الموسوعات ودوائر المعلومات

9 771319 025251

البيانات الضخمة.. مصدر البيانات المتجددة

حنان جرجس: مصر

زاد الطلب على البيانات بصورة كبيرة خلال العقود القليلة الماضية، ومع إطلاق الأمم المتحدة لأهداف التنمية المستدامة برزت الحاجة إلى توفير المزيد من البيانات للتخطيط لتنفيذ هذه الأهداف على مستوى الدول، وكذلك على المستويات الإدارية الأدنى في كل دولة. وعلى الرغم من تنوع أساليب ومنهجيات جمع البيانات سواء من خلال المسوح الميدانية أو التليفونية، إلا أن التكلفة المرتفعة وطول الوقت المطلوب لتنفيذ مثل هذه المسوح أصبحا عائقين أمام توفير البيانات من خلال هذه المنهجيات. ومن هنا ظهرت الحاجة إلى توفير البيانات من خلال أساليب غير تقليدية، تعتمد بصورة كبيرة على توظيف الجمهور العام في توليد البيانات المطلوبة، وجاءت البيانات الضخمة كنتيجة لهذه الجهود.

ما هي البيانات الضخمة؟

بدأ استخدام مصطلح البيانات الضخمة (Big Data) في تسعينات القرن الماضي، وكان جون ر. ماشي وراء انتشار استخدام هذا المصطلح. ويشير هذا المصطلح ببساطة إلى البيانات ذات الكميات الكبيرة التي يتم إنشاؤها/جمعها من التفاعلات اليومية مع المنتجات أو الخدمات الرقمية، وتتمتع هذه البيانات بصفة أو أكثر من الخصائص التالية:

- أن تكون كبيرة الحجم (Volume) حيث قد تحتوي هذه البيانات على الملايين بل المليارات من خلايا البيانات.

- أن تكون سريعة التواتر (Velocity) حيث يتم إضافة كميات كبيرة من البيانات على فترات قصيرة قد تكون بصورة يومية أو حتى كل ساعة.

- أن تكون عالية التنوع (Varity) حيث قد تشمل هذه البيانات أرقاماً أو نصوصاً أو صوراً أو فيديوهات أو ملفات صوتية.

- أن تكون لها قيمة كبيرة (Value) بحيث يمكن أن يساعد تحليلها وتحويلها إلى معلومات في عمليات التخطيط واتخاذ القرار.

وتعرف هذه الخصائص بالإنجليزية بـ Big Data 4 Vs.

وتشكل البيانات الضخمة فرصة كبيرة لتوفير البيانات من مصادر غير تقليدية

وبتكلفة قد تكون أقل، وعادةً يتم إنتاج هذه البيانات لأغراض معينة غير الأغراض الإحصائية، لكن حالياً يمكن توظيف البيانات الضخمة المتاحة من المصادر المختلفة في توفير الإحصائيات والمؤشرات التي يمكن استخدامها في مجالات شتى تشمل التخطيط والمتابعة والتقييم واتخاذ القرار المستند على الأدلة. وإذا كانت المسوح توفر لنا إحصائيات ومؤشرات عن ظاهرة ما في لحظة معينة، ويحتاج تحديثها إلى إعادة إجراء المسح، فإن البيانات الضخمة تمثل مصدراً للبيانات المتجددة التي يتم تحديثها بصورة دورية.

وتتنوع مصادر البيانات الضخمة لتشمل وسائل التواصل الاجتماعي، تطبيقات التليفون المحمول، شركات الاتصالات، تطبيقات تحديد المواقع GPS، المتاجر، البنوك، التسوق عبر الإنترنت وغيرها من المصادر التي تتيح بيانات منظمة أو غير منظمة.

وتتميز البيانات الضخمة بأنها قد تكون أرخص وأسرع من أساليب توفير البيانات الأخرى، كما أنها قد تكون مصدراً يسهل الحصول على بيانات منه في المناطق أو الظروف التي يصعب جمع البيانات فيها بالأساليب التقليدية كما كان الحال في بداية انتشار كوفيد19-. بالإضافة إلى ذلك يمكن أن تستخدم البيانات الكبيرة من المصادر المختلفة ليس فقط لقياس انتشار الظاهرة، بل أيضاً لتفسير الظاهرة ومعرفة أسبابها والتنبؤ بتطورها.

تحليل البيانات الضخمة

البيانات التي يتم توفيرها من المصادر التقليدية كالمسوح أو قواعد البيانات الإدارية تكون بيانات منظمة في شكل أعمدة وصفوف، وتستخدم العديد من الحزم الإحصائية لتحليل هذه البيانات. أما بالنسبة للبيانات الضخمة فعادةً تكون شبه منظمة أو غير منظمة مما يتطلب أساليب مختلفة لتحليل هذه البيانات خصوصاً مع أحجامها الضخمة، وتعتمد هذه الأساليب عل طرق النمذجة المختلفة وكذلك أدوات تحليل النص والتعلم الآلي machine learning.

أمثلة لاستخدام البيانات الضخمة: تحليل أنماط إنفاق السائحين في إسبانيا

لوحظ في إسبانيا أن نسبة من السائحين عند زيارتهم لها يقيمون لمدد تزيد عن 3 شهور، وعادةً ما يقيم هؤلاء السائحون في مساكن يمتلكها المواطنون أو منازل للإيجار الموسمي أو أماكن إقامة مخصصة للطلاب أو يشاركون الإقامة مع المواطنين. ومن أجل تحديد أنماط إنفاق السائحين على الإقامة والمناطق التي يميلون إلى الإقامة بها وكذلك خصائص السائحين الذين يقيمون بإسبانيا لمدد طويلة تم تحليل بيانات الإنفاق باستخدام كروت الدفع الخاصة بالأجانب الذين زاروا إسبانيا في 2023 وبقوا بها لمدة تتراوح بين 3 شهور و11 شهراً باستخدام أساليب تحليل البيانات الضخمة. بالإضافة إلى ذلك تم استخدام بيانات الإنفاق باستخدام كروت الدفع الخاصة بالأجانب لتحليل نمط الإنفاق على السلع المختلفة ومعرفة أكثر السلع وأماكن الترفيه التي ينفق عليها السائحون حسب جنسياتهم المختلفة مما يساعد على وضع خطط التسويق للسياحة في إسبانيا في الأعوام المقبلة.

هذه التطبيقات يمكن أن تستخدم بصورة أوسع للتعرف على نمط إنفاق المواطنين على السلع والخدمات المختلفة مما يساعد على توفير بيانات الإنفاق التي كان يتم عادةً توفيرها من خلال بحوث الميدانية لقياس الإنفاق التي تتطلب وقتاً وتكلفة كبيرة.

استخدام بيانات شركات الاتصالات لأغراض تنموية

تتوافر لدى شركات الاتصالات بيانات وقتية عن برج الاتصال الذي تلقى المستخدم المكالمات منه لحساب تكلفة كل مكالمة. وتعد هذه البيانات فرصة كبيرة لتوفير بيانات عن تنقل السكان يومياً خصوصاً في أيام العمل خلال الأسبوع مما يعطي فرصة للتعرف على المسافات التي يقطعها المواطنون يومياً، وتقدير الوقت والتكلفة الاقتصادية والاجتماعية للتنقل. كذلك يمكن من خلال هذه البيانات معرفة خطوط السير التي تستخدمها السيدات بكثافة مما يساعد على تطوير منظومة النقل وتوفير وسائل الانتقال الآمنة للمرأة.

كما يمكن أن تستخدم هذه البيانات بالإضافة إلى بيانات نظام تحديد المواقع العالمي (GPS) في التعرف على تردد المواطنين على أماكن تلقي الخدمات المختلفة كالمستشفيات وغيرها، مما يساعد على حساب معدلات التردد وتقدير الطلب على هذه الأماكن وبالتالي تحديد حجم الخدمات المطلوب توفيرها.

استخدام وسائل التواصل الاجتماعي في الإنذار المبكر

يمكن من خلال تحليل محتوى وسائل التواصل الاجتماعي في الحصول على إنذار مبكر بالأزمات الجارية أو التي يمكن أن تحدث في المستقبل القريب بصورة أسرع من الوسائل التقليدية. على سبيل المثال أثناء جائحة كوفيد19- تمكن الباحثون من اكتشاف تفشي مرض كوفيد19- في أوروبا خلال فصل الشتاء 2019-2020 في بعض المناطق، قبل الإعلان العام الأول عن مصادر العدوى المحلية من خلال تحليل التغريدات على تويتر حسب المناطق الجغرافية. كما أن هناك العديد من التطبيقات في دول العالم التي تساعد على الاكتشاف المبكر للمناطق التي تضربها الزلازل أو البراكين أو غيرها من الكوارث الطبيعية.

مثل هذه التطبيقات توفر بيانات سريعة لمتخذي القرار يمكن توظيفها في تقديم الإغاثة والرعاية للمناطق المنكوبة بسرعة مما يساعد على تقليل الخسائر في نهاية الأمر.

محددات استخدام البيانات الضخمة في حساب المؤشرات

على الرغم من الأهمية المتزايدة لتوظيف البيانات الضخمة في توفير الإحصائيات وحساب المؤشرات، إلا أنه يجب ألا نغفل عدداً من التحديات الأساسية التي تشمل:

اعتبارات الخصوصية: في حالة البيانات الضخمة، فإنه رغم أن الشخص قد يكون أدلى بالبيانات أو المعلومات بصورة طوعية إلا أنه في الغالب قد أدلى بها لأسباب محددة ربما للتعبير عن رأيه أو الشكوى من أمر ما أو غيرها من الأسباب، وليس لاستخدامها في أغراض الدراسة التي سيتم استخدام البيانات الكبيرة في إطارها، وبالتالي استخدام هذه البيانات يشوبه مخالفة للأعراف والقواعد العلمية التي تتطلب موافقة مدلي البيانات على استخدام البيانات التي أدلى بها في الدراسة التي يتم إجراؤها، وهو المتبع مثلاً عند إجراء المسوح الميدانية أو الهاتفية.

- تمثيل المجتمع: عند إجراء المسوح الميدانية أو الهاتفية لجمع بيانات عن ظاهرة معينة يتم جمع البيانات من عينة ممثلة لمجتمع الدراسة بمعنى أنها عينة تشمل أفراداً يمثلون كافة طبقات المجتمع وأطيافه، ونتيجة لذلك يمكن تعميم نتائج الدراسة على المجتمع ككل، وهو ما لا تسمح به البيانات الضخمة، إذ إن المدلين بالبيانات على وسائل التواصل الاجتماعي مثلاً لا يشملون الأميين أومن ليس لديهم إمكانية الوصول للإنترنت أو من لا يمتلكون حسابات على هذه المواقع، وبالتالي فهي ليست عينة ممثلة، كما أنه يصعب تجميع كل الآراء التي كتبت عن الموضوع محل الدراسة وليس هناك ضمان أن العينة التي تم تجميعها من هذه التدوينات هي عينة غير متحيزة لرأي معين. بالإضافة إلى ذلك هذه التدوينات هي تدوينات من تطوعوا للكتابة عن الموضوع وبالتالي لا تعكس رأي من لم يقوموا بالتدوين حول الموضوع والذين في كثير من الأحيان يكون لهم آراء مخالفة.

- الموارد البشرية والمالية: يتطلب تجميع البيانات الضخمة وتحليلها موارد بشرية متخصصة في البرمجة والتحليل الإحصائي والتعلم الآلي، وهي موارد بشرية نادرة ومكلفة، كما أن تكلفة إعداد البرامج والحزم التي تستخدم في تحليل هذه البيانات عند بدء العمل مرتفعة، إلا أنها ومع تكرار الاستخدام تصبح أرخص نسبياً من جمع البيانات بالوسائل التقليدية.

- اللغات المحلية والعامية: عند تجميع وتحليل ما يتم تدوينه على وسائل التواصل الاجتماعي يصادف الباحثين تحد كبير متعلق باستخدام المدونين للغات المحلية واللهجات العامية وهو ما يصعب عملية تصنيف وتحليل هذه التدوينات.

- صعوبة حساب المؤشرات المتعارف عليها: هناك مؤشرات متعارف عليها لقياس وتحليل كل ظاهرة، فعلى سبيل المثال عند متابعة تنفيذ أهداف التنمية المستدامة هناك مؤشرات لقياس التقدم المحرز في كل هدف. في حالة البيانات الضخمة قد لا يمكن توفير البيانات المطلوبة لحساب المؤشرات المتعارف عليها بصورة مباشرة، وهو ما قد يتطلب في هذه الحالة وضع منهجيات لمؤشرات بديلة يمكن حسابها من خلال البيانات المتاحة.

وعلى الرغم من هذه التحديات إلا أن دور البيانات الضخمة في توفير الإحصائيات والمؤشرات يبقى دوراً محورياً خصوصاً مع التطور التكنولوجي السريع والبحث عن المنهجيات المناسبة، مما سيساعد في التغلب على معظم هذه التحديات.