الحلقة العلمية حول
"البيانات الضخمة من أجل سياسات أفضل"

سالى عاشور*

أدى التطور التكنولوجى المستمر والمتسارع خلال سنوات الألفية الجديدة إلى ظهور مصطلحات جديدة فى مجالات البحث العلمى عمومًا والعلوم الاجتماعية خصوصًا. ومن بين تلك المصطلحات مصطلح "البيانات الضخمة" Big Data وهو مصطلح يستخدم للإشارة إلى البيانات كبيرة الحجم التى يصعب تخزينها ومعالجتها بالأدوات والبرامج والتطبيقات القديمة المتعارف عليها، ويرجع السبب فى ذلك إلى ضخامتها وتعقيدها وعدم جاهزية برامج معالجة البيانات التقليدية للتعامل معها.

وبالإشارة إلى الاهتمام المتزايد باستخدام البيانات كأساس فى إعداد البحوث العلمية والتقارير وكذلك السياسيات، وما ترتب على ذلك من حدوث تواصل وتعاون بين عدد من العلوم مثل علم الإحصاء والعلوم التكنولوجية ونظم المعلومات والبرمجيات، من أجل استخدام التخصصات المختلفة لتلك العلوم فى تحليل البيانات الكبيرة والضخمة بهدف إتاحتها لمتخذى القرار وواضعى السياسات.

وقد نظم قسم العلوم الاجتماعية السياسية فى جامعة ميلانو خلال الفترة من ٢٣ – ٢٨ يونيو ٢٠١٩ حلقة علمية بعنوان "البيانات الضخمة من

* مدرس العلوم السياسية، المركز القومى للبحوث الاجتماعية والجنائية.

أجل سياسيات أفضل" للباحثين فى مجال العلوم الاجتماعية من ستة عشر بلدًا ممثلة عن قارات العالم المختلفة بهدف مناقشة عدد من المحاور.

محاور الحلقة العلمية

المحور الأول: التقاط وتخزين البيانات الكبيرة: مقدمة فى تقنيات تجريف الويب Web Scraping Techniques

المحور الثانى: تحليل المحتوى الآلى للباحثين الاجتماعيين Automated Content Analysis

المحور الثالث: تحليل محتوى الشبكات باستخدام برامج بايثون Python وجيفى Gephi.

المحور الرابع: التعلم الآلى Machine Learning، الفرص والتحديات لصانعى القرار.

المحور الأول: التقاط وتخزين البيانات الكبيرة: مقدمة فى تقنيات تجريف الويب Web Scraping Techniques

تناول هذا المحور تعريفًا لمصطلح Web Scraping والتقنيات المختلفة التى يتم استخدامها فى عملية تجريف الويب أو تجريف الإنترنت، والمقصود بها هنا هى تقنيات استخراج البيانات من صفحات الويب/الإنترنت باستخدام برامج ولغات برمجية متخصصة من أجل حفظها ومعالجتها طبقًا لاحتياجات استخدامها.

الإنترنت أو الويب هى شبكة المعلومات، الشبكة العالمية، الشبكة العنكبوتية والمقصود بها شبكة الاتصالات العالمية والتى تسمح بتبادل المعلومات بين شبكات أصغر تتصل من خلالها أجهزة الحاسب الآلى. ويتناول هذا المحور أسباب احتياج دارسى العلوم المختلفة وصانعى القرار إلى تقنية تجريف الإنترنت أو الويب والتى ترجع بالأساس إلى الانتشار الواسع لشبكة المعلومات العالمية العنكبوتية والزيادة الكبيرة فى حجم المعلومات والبيانات

الموجودة على صفحات الويب والتى يتم إنشاؤها كل ساعة، مع عدم وجود خدمة حفظ نسخة البيانات على صفحات الويب نفسها، وقد أدى ذلك إلى الاحتياج لوجود تقنيات وبرامج تجريف الويب والتى تساعد على القيام بعمليات الحفظ بشكل دقيق وسريع، خاصة وأن البديل عن تلك العملية هو نسخ البيانات بشكل يدوى وحفظها فى ملفات، وهذا من شأنه استهلاك الكثير من الوقت والموارد خاصة فى حالة وجود حجم هائل من البيانات. كما أنها تتيح استخدام تلك البيانات فى مراحل زمنية مختلفة حتى وإن تم رفعها أو حذفها من على شبكة المعلومات العالمية (الويب).

المحور الثانى: تحليل المحتوى الآلى للباحثين الاجتماعيين

Automated Content Analysis

تناول هذا المحور التقنيات الرقمية الحديثة والبرمجة التى يتم استخدامها لتحليل محتوى وسائل الإعلام – فى الأغلب البيانات النصية– بشكل ذاتى أو تلقائى باستخدام الآلة والبرمجيات الحديثة والبرامج المستحدثة فى تحليل المحتوى. وتم الإشارة إلى أن التزايد المستمر فى الاتصالات عبر شبكة المعلومات العالمية العنكبوتية والمتاح أساسًا بتقنيات التنسيق الرقمى، هو ما دفع علماء البرمجيات إلى استحداث برامج تكون لديها القدرة على تحليل كميات ضخمة من البيانات بطريقة آلية.

وبغرض تسهيل استخدام تلك البرامج من قبل الباحثين الاجتماعيين فى إطار ما يطلق عليه Computational Social Science أو"العلوم الاجتماعية الحاسوبية"، يتم الاستعانة بتقنيات من تخصصات مختلفة ومتعددة مثل علوم الكمبيوتر واللغويات الحاسوبية والبرمجة، متضمنة للتقنيات والطرق القائمة على استخدام القاموس والكلمة، ومعالجة اللغات الطبيعية Natural Language Processing (NLP)، والتعلم الآلى machine learning.

المحور الثالث: **تحليل محتوى الشبكات باستخدام برامج بايثون Python وجيفى Gephi**

تناول هذا المحور إمكانيات بعض البرامج فى تحليل محتوى الشبكات وتم إعطاء مثال ببرنامجى بايثون Python وجيفى Gephi. حيث عادة ما يحتاج مستخدمو البيانات الضخمة والكبيرة Big Data عن طريق أجهزة الحاسب الآلى والشبكات إلى تأدية بعض المهام بصورة أوتوماتيكية. وقد تكون عمليات بسيطة مثل "البحث" Search أو استبدال Replace أو"إعادة تسمية" Rename وترتيب Arrange بعدد كبير من الملفات، أو عمليات أكثر تعقيدًا. وعادة ما كان يستخدم المحترفون فى مجال تطوير البرمجيات العديد من مكتبات اللغات C/C++/Java للقيام بتلك العمليات البسيطة وغيرها من العمليات المعقدة، ولكن لوحظ أن عمليات الكتابة ثم التصريف ثم الاختبار ثم إعادة التصريف قد صارت مؤخرًا سلسلة من العمليات المتكررة البطيئة للغاية.

وتتيح برامج مثل بايثون Python وجيفى Gephi كتابة البرامج بأسلوب مختصر ومقروء، فالبرامج المكتوبة فى بايثون تكون عادة أقصر بكثير من مكافئاتها فى لغات C ، أو C++ ، أو Java.

المحور الرابع: **التعلم الآلى** Machine Learning **الفرص والتحديات لصانعى القرار**

تناول هذا المحور مفهوم التعلم الآلى Machine Learning والذى يركز بالأساس على بناء معادلات باستخدام علوم البرمجة قادرة على استقبال بيانات كبيرة وضخمة Big Data يتم إدخالها بطرق متعددة ومختلفة، ثم يتم استخدام التحليل الإحصائى (statistical analysis)؛ للتنبؤ بالمخرجات عن طريق استخدام نماذج قد تكون حقيقية مستخدمة معلومات موجودة أساسًا أو حديثة تم إنشاؤها من الصفر.

وبشكل عام يوجد مستويان من التعلم: الاستقرائى والاستنتاجى. يقوم التعلم الاستقرائى باستنتاج قواعد وأحكام عامة من البيانات المدخلة إليه، بينما الاستنتاجى ينطلق من أحكام عامة ويطبقها فى أمثلة خاصة.

وقد تم استخدام تقنيات العلم الآلى لتحليل ومعالجة البيانات الضخمة فى مجالات عديدة مثل البنوك والمؤسسات المالية المختلفة، وكذلك الشركات التجارية والاستثمارية –على نطاق واسع خلال الأعوام السابقة– بهدف المساهمة فى التنبوء بنتائج عمل تلك المؤسسات والمساهمة فى وضع الخطط والسياسات المستقبلية لها. ويتم حاليًا دراسة فرص وتحديات استخدام تلك التقنية فى وضع السياسات العامة أو حتى بعض مجالاتها وما إذا كانت مثل تلك التقنيات قابلة للاستخدام بشكل كفء ومفيد فى مجالات السياسات العامة.