استخدام تطبيقات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية وعناصرها المقروءة والمصورة

دراسة تحليلية من المستوى الثاني

د.هبه محمود مصطفى عباس*

ملخص الدراسة:

تسعى الدراسة؛ إلى تحليل تطبيقات الذكاء الاصطناعي في رقمنة الصحف التاريخية وتحليلها، بما تتضمنه من عناصر مقروءة ومصورة، وتحديد وتحليل التحديات والإشكاليات الناتجة عن ذلك في الصحافة التاريخية اللاتينية أو غير العربية، وكذلك الصحافة التاريخية العربية؛ وذلك باستخدام أسلوب التحليل الكيفي؛ من خلال التحليل من المستوى الثاني Meta-analysis

تهدف هذه الدراسة؛ إلى رصد الدراسات التي تناولت تطبيقات تقنيات الذكاء الاصطناعي في مجال الرقمنة وتحليل المحتوى رقميًا على الصحف؛ من خلال تحليل و أرشفة وتصنيف وحفظ النصوص والمواد الصحفية المصورة، لإتاحتها للاسترجاع، وتحليل هذه الدراسات، في إطار التحليل الكيفي؛ بأسلوب التحليل من المستوي الثاني، كما تهدف الدراسة؛ إلى رصد الدراسات التي تمت كذلك على المشروعات التي تم فيها معالجة الصحف رقميًّا، بتقنيات الذكاء الاصطناعي؛ من خلال أرشفة، وتصنيف الصحف، وأرشفة النصوص التي تضمنتها، والصور والرسوم التي نشرتها في أثناء المعالجات الصحفية، لأرشفة وتصنيف وحفظ التراث الصحفي، وإتاحته للاسترجاع ، ودراستها وتحليلها؛ من خلال أسلوب التحليل من المستوى الثاني، وذلك في إطار رصد دور الذكاء الاصطناعي، وتقنياته في حفظ التراث الصحفي.

الكلمات الدالة:

الذكاء الاصطناعي- رقمنة الصحف- الصحف التاريخية – المواد المقروءة- المواد المصورة .

* مدرس الصحافة – كلية الإعلام – جامعة المنوفية

Using Artificial Intelligence Applications in Digitizing and Analyzing Historical Newspapers and their Text and Visual Elements

Meta-analysis Study

Dr.Heba Mahmoud Mustafa Abbas [**]

Abstract

This study covers the applications of artificial intelligence in digitizing and analyzing newspapers, including their written and visual elements. It also explores the many challenges and problems presented by the scientific heritage to academics at various levels regarding artificial intelligence in the field of digitizing and analyzing historical journalism.

The study then seeks to identify and analyze these challenges and problems in Latin American or non-Arabic historical journalism, as well as Arabic historical journalism, using the qualitative analysis method through second-level analysis (Meta-analysis).

Keywords:

Artificial intelligence - digitizing newspapers- historical newspapers- Text- Visual elements.

[**] **Journalism Lecture _ Faculty of Mass Communication _ Menofia University**

مقدمة

تمثل الصحف التاريخية، مصدرًا للأحداث التاريخية، ومواردًا ثقافيًا عبر حقب زمنية طويلة، ومصدرًا لغويًّا ومجالا لدراسة الخطاب، وتطوره، ودراسة قضايا عديدة، وتحليل تطورها، ومجالا لدراسة العناصر المرئية وتطوراتها عبر حقب التاريخ؛ فالصحف التاريخية- على مستوى البحث العلمي الأكاديمي- تمثل مجالا لا يمكن الاستغناء عنه.

لقد كان حفظ أعداد الصحف، يتم من خلال وضعها في الأرشيف في مجلدات ورقية مرتبة ترتيبًا زمنيًا، ثم انتقلت إلى وسائط الميكروفيلم والميكروفيش، التي تتضمن صور صفحات الصحيفة على أفلام، ويتم الاطلاع عليها ودراساتها؛ من خلال أجهزة الميكروفيلم والميكروفيش، ثم انتقلت الصحف التاريخية أو الصحف التي صدرت في فترات زمنية سابقة إلى الوسائط الرقمية، ممثلة تراثًا صحفيًا وتاريخيًا وثقافيًا _في صيغة رقمية- لا يمكن الاستغناء عنه، ولكن ماذا عن وضوح تلك الصفحات التاريخية- الصادرة في فترات زمنية سابقة- في الوسائط الرقمية، ما الإشكاليات التي ارتبطت بانتقال الصحف التاريخية، أو أعداد الصحف الصادرة في فترات زمنية سابقة، من الأرشيف الورقي والميكروفيلم والميكروفيش، إلى الأرشيف الرقمي.

لا تنتهي إشكالية البحث العلمي الأكاديمي، في مجال الصحف والصحافة التاريخية، في ظل الوسائط الرقمية عند هذا الحد، إنما ثمة تعقيدات، تتعلق بدراسة الصحف والصحافة التاريخية المحفوظة والمؤرشفة في وسائط رقمية، تتعلق بتقنيات البحث وأدوات البحث العلمي، خاصة في ظل تنامي استخدامات الذكاء الاصطناعي في مجال الدراسات الإعلامية والصحفية.

إن التاريخ الصحفي ووثائقه الصحفية التاريخية، تواجه التطور التقني للذكاء الاصطناعي بوصفه وسيطًا متاحًا لحفظ الوثائق بتقنياته المتعددة، وبوصفه وسيلة من وسائل البحث العلمي، وهو ما يحتم ضرورة إلقاء الضوء على إشكاليات بحوث تاريخ الصحافة، في ظل وسائط الأرشفة للصحف التاريخية، والبحث العلمي المعتمد على الذكاء الاصطناعي، في تحليل الصحف، وخاصة الصحف التاريخية أو الصحافة التاريخية .

وتواجه أرشفة الصحف التاريخية ورقمنتها، العديد من التحديات والمشكلات في العصر الرقمي، وتحتاج دراسة الصحف الصحافة التاريخية، في ظل التطور الرقمي والوسائط الرقمية، إلى تطوير أدوات البحث المستخدمة في تحليل، ودراسة الصحف التاريخية؛ لتتلاءم مع وسائط أرشفة تلك الصحف ورقمنتها، ولتواكب تطورات عصر البحث العلمي بأدوات الذكاء الاصطناعي، وما تتيحها من إمكانات، وتواكب التخصصات الأخرى؛ للاستفادة من هذه الثروة التاريخية الثقافية، وتحقيق إمكانية دراساتها، بما يتوافق مع أهميتها، وقيمتها، وبما يواكب تطورات أدوات البحث العلمي في العصر الرقمي للصحافة ووسائل الإعلام.

ثمة تعقيدات مرتبطة باستخدام الصحف الرقمية، ويواجه الباحثون عوائق، عند التصدي لمحتوى الصحف التاريخية بأدوات الذكاء الاصطناعي، هذه العوائق التي قد تؤثر على تفاعلهم مع المعلومات، ومن ثم فإن معرفة تقنيات و إشكاليات تطبيقات الذكاء الاصطناعي، في رقمنة

وتحليل الصحف التاريخية، وعناصرها المقروءة والمصورة، قد يسهم في إمكانية الوصول إلى الصحف، والاستفادة منها، ومن توظيف أدوات الذكاء الاصطناعي في التحليل، بناءً على معرفة تطبيقات وتقنيات الذكاء الاصطناعي، في أرشفة الصحف التاريخية وإشكالياته من جانب، ومدى فعالية البحث في الصحف التاريخية المؤرشفة بالذكاء الاصطناعي بأدوات ومناهج الذكاء الاصطناعي من جانب آخر.

ومن هنا، تأتي أهمية تناول الدراسة بشكل تحليلي ونقدي، لتطبيقات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية، وعناصرها المقروءة والمصورة، وتحديات البحث في الصحف والصحافة التاريخية، في ظل تطبيقات الذكاء الاصطناعي .

وتركز الدراسة في هذا الصدد، على رصد الدراسات التي تناولت تطبيقات تقنيات الذكاء الاصطناعي، في مجال الرقمنة، وتحليل المحتوى رقميًّا، على الصحف التاريخية؛ من خلال تحليل و أرشفة، وتصنيف، وحفظ الصحف، والنصوص، والمواد الصحفية المصورة؛ لإتاحتها للاسترجاع، وتحليل هذه الدراسات في إطار التحليل الكيفي، باستخدام أسلوب تحليل المستوي الثاني، كما تركز الدراسة، على رصد الدراسات، التي تمت على المشروعات، التي تم فيها معالجة الصحف رقميًّا بتقنيات الذكاء الاصطناعي؛ من خلال أرشفة، وتصنيف الصحف، وأرشفة النصوص التي تضمنتها، والصور والرسوم التي نشرتها في أثناء المعالجات الصحفية، لأرشفة، وتصنيف، وحفظ التراث الصحفي، وإتاحته للاسترجاع، ودراستها وتحليلها، وذلك في إطار رصد دور الذكاء الاصطناعي وتقنياته، في حفظ التراث الصحفي.

مشكلة الدراسة :

يطرح التصدي لدراسة وتحليل تطبيقات الذكاء الاصطناعي، في رقمنة وتحليل الصحف التاريخية، بما تتضمنه من عناصر مقروءة ومصورة، العديد من التحديات والإشكاليات التي تناولها التراث العلمي الأكاديمي، على مستويات متعددة في دراسته لتطبيقات الذكاء الاصطناعي، في مجال رقمنة وأرشفة وتحليل الصحافة التاريخية، ومن ثم؛ تسعى الدراسة؛ إلى تحليل تطبيقات الذكاء الاصطناعي، في رقمنة وتحليل الصحف التاريخية والصحافة التاريخية، بما تتضمنه من عناصر مقروءة ومصورة، وتحديد وتحليل التحديات، والإشكاليات الناتجة عن ذلك، في الصحافة التاريخية باللغات الأجنبية، أو غير العربية، وكذلك الصحافة التاريخية العربية؛ وذلك من خلال التحليل الكيفي؛ من باستخدام التحليل من المستوى الثاني-Meta analysis.

أهمية الدراسة:

تأتي أهمية من الدراسة مما يلي:

١. أهمية الصحف والصحافة التاريخية بوصفها وثائق تاريخية وثقافية.
٢. تقديمها تعريفًا دقيقًا بتقنيات الذكاء الاصطناعي، المستخدمة في الصحف التاريخية غير العربية، وإشكاليات استخدامها في الصحف العربية التاريخية .

٣. تصاعد استخدام تقنيات الذكاء الاصطناعي، في مجال البحث العلمي، في دراسات الصحافة والإعلام، وما يتضمنه من إشكاليات.
٤. أن الدراسة، تحلل التحديات التقنية والعلمية، التي نتجت عن أرشفة ورقمنة الصحف التاريخية، في ظل الوسائط الرقمية .
٥. أن الدراسة، تحلل إشكاليات البحث العلمي، في مجال الصحافة التاريخية، في ظل تطبيقات تقنيات الذكاء الاصطناعي .
٦. تُمثل الدراسة تحفيزًا علي تسهيل إجراء البحوث الصحفية التاريخية.

أهداف الدراسة:

تتضمن الدراسة هدفًا رئيسًا، يتفرع منه عددٌ من الأهداف الفرعية، جاءت على النحو التالي:

يتمثل الهدف الرئيس للدراسة؛ في تحليل الدراسات التي تعالج استخدام تطبيقات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية وعناصرها المقروءة والمصورة.

وتتمثل الأهداف الفرعية للدراسة في :

- تحديد التطورات التقنية للذكاء الاصطناعي المستخدمة في رقمنة الصحف التاريخية، في ضوء ما توصلت إليه عينة الدراسات موضع التحليل في الدراسة .
- رصد القضايا التي عالجتها عينة الدراسات موضع التحليل في الدراسة.
- رصد تجارب الصحف التي تم تطبيق الذكاء الاصطناعي في رقمنتها وتحليلها في عينة الدراسات موضع الدراسة.
- تحليل ومناقشة أهداف وفوائد رقمنة الصحف التاريخية وأرشفتها في ضوء ما طرحته عينة الدراسات موضع التحليل في الدراسة .
- تحليل مشكلات رقمنة الصحف التاريخية وأرشفتها، في ضوء ما طرحته الدراسات موضع التحليل في عينة الدراسة .
- مناقشة مشكلات البحث العلمي في الصحافة التاريخية، في ظل تطبيق تقنيات الذكاء الاصطناعي، في الأرشفة والتحليل، في ضوء ما طرحته الدراسات موضع التحليل في عينة الدراسة .

تساؤلات الدراسة:

١. ما الدراسات التي تعالج تطبيقات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية وعناصرها المقروءة والمصورة؟
٢. ما التطورات التقنية للذكاء الاصطناعي المستخدمة في رقمنة الصحف التاريخية؟
٣. ما القضايا التي عالجتها الدراسات موضع التحليل في دراستها لتطبيقات تقنيات الذكاء الاصطناعي في أرشفة الصحف التاريخية وتحليلها؟
٤. كيف وظفت الدراسات موضع التحليل، تطبيقات الذكاء الاصطناعي في رقمنة الصحف التاريخية وتحليلها؟

٥. ما أهداف وفوائد رقمنة الصحف التاريخية وأرشفتها، كما تعكسها عينة الدراسات موضع الدراسة؟
٦. كيف عالجت الدراسات موضع التحليل، مشكلات رقمنة الصحف التاريخية العربية والأجنبية وأرشفتها؟
٧. كيف تناولت الدراسات موضع التحليل، مشكلات البحث العلمي في الصحافة التاريخية، في ظل تطبيق تقنيات الذكاء الاصطناعي في الأرشفة والتحليل؟

نوع الدراسة:

تنتمي هذه الدراسة؛ إلى نوعية الدراسات الوصفية التحليلية؛ حيث تسعى؛ إلى الوصف والتحليل الكيفي، للدراسات العلمية الأكاديمية، التي تتناول تطبيقات الذكاء الاصطناعي، في رقمنة الصحف التاريخية والنصوص الصحفية والعناصر المصورة الصحفية التاريخية وتحليلها.

مناهج الدراسة:

تعتمد الدارسة، على منهج المسح الإعلامي؛ وذلك لمسح الدراسات المتاحة خلال الفترة الزمنية للدراسة التي تدرس استخدام تطبيقات الذكاء الاصطناعي، في رقمنة الصحف التاريخية والنصوص الصحفية والمواد الصحفية المصورة التاريخية وتحليلها .

أدوات الدراسة :

تعتمد الدراسة؛ على التحليل الكيفي، باستخدام أسلوب تحليل المستوى الثاني-Meta Analysis، وذلك لتحليل تطبيقات الذكاء الاصطناعي في رقمنة الصحف والصحافة التاريخية وتحليلها، بما تتضمنه من عناصر مقروءة ومصورة، وتحديد وتحليل التحديات والإشكاليات الناتجة عن ذلك في الصحف التاريخية اللاتينية أو غير العربية، والصحافة العربية التاريخية.

مجتمع الدراسة:

عينة الدراسات العلمية الأكاديمية :

تتمثل عينة التحليل في الدراسة، في عينة بالمسح الشامل لكل الدراسات المتاحة في السنوات الخمس الأخيرة، من ٢٠٢٠ حتى ٢٠٢٥، التي تناولت دراسة تطبيقات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية، والمواد المقروءة في الصحف، بما تتضمنه من النصوص، والخطابات، والمقالات الصحفية، والمواد المصورة، بما تتضمنه من صور ورسوم .

الفترة الزمنية للدراسة:

تمتد الفترة الزمنية للدراسة، من عام ٢٠٢٠، حتى عام ٢٠٢٥؛ وذلك نظرًا لأن الدراسة، تدرس تطبيقات الذكاء الاصطناعي، التي تتسم بالتطور السريع الدائم والتحديث المستمر؛ ومن ثم تم التركيز على الدراسات، التي تم إجراؤها خلال الخمس سنوات الماضية فقط؛ لعرض ومناقشة أحدث تقنيات رقمنة الصحف التاريخية، والنصوص الصحفية، والمواد المصورة، وتحليلها .

تم تقسيم محاور الدراسات العلمية التي تم تحليلها في الدراسة، إلى ثلاثة محاور؛ تمثلت في :

أولًا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة الصحف التاريخية وتحليلها .

ثانيًا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المقروءة (النصوص) في الصحف التاريخية وتحليلها .

ثالثًا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المصورة في الصحف التاريخية وتحليلها .

م	محاور الدراسة	عدد الدراسات باللغات الأجنبية	النسبة المئوية
١	**الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة الصحف التاريخية وتحليلها.**	٢٠	%٢٨,٢
٢	**الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المقروءة (النصوص) في الصحف التاريخية وتحليلها.**	٣٥	%٤٩,٣
٣	**الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المصورة في الصحف التاريخية وتحليلها.**	١٦	%٢٢,٥
	الإجمالي	٧١	%١٠٠

كلمات البحث

Computer vision and newspapers archives

Computer vision and visual historical newspapers

Historical newspaper archives

ai and historical newspapers

ai and historical newspapers navigator

Convolutional neural networks and historical newspaper photo

Convolutional neural networks historical newspaper

Pre-trained word and newspapers archives

Machine learning and historical newspapers

Pre-trained word embedding deep neural networks and newspapers archives

Object detection and newspaper archives

Semantic enrichment historical newspapers

Large Language Models and historical newspapers

محركات البحث

Springer nature link

Emerald insight

ACM Digital library

American University Library

arxiv.org

Google scholar

مجلات علمية أكاديمية

Journal of Open Humanities Data

أولا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة الصحف التاريخية وتحليلها .

تُحلل دراسة Bruno Frutuoso Costa et all (2025)[1] ، التقنيات الحالية، والتحديات التي ينطوي عليها، استخراج وتصنيف المقالات، وعناوين الأخبار من المجلات التاريخية، بالإضافة إلى تحويل الصور إلى صيغة نصية، وقد تولى فريق متعدد التخصصات من الخبراء، في دراسات الإعلام، والذكاء الاصطناعي ومعالجة الصور وحفظ التراث الثقافي تطوير أداة، تُركز على رقمنة المجلات التاريخية، وتُستمد البيانات المستخدمة من مجلتين برتغاليتين تاريخيتين، هما "دياريو دي نوتيسياس" و"جورنال دي نوتيسياس"، اللتين أُنشئتا في منتصف القرن التاسع عشر، ويعتمد هذا المشروع؛ على مزيج من أساليب الاستدلال، والرؤية الحاسوبية، وتعرف الأنماط، وغيرها من تقنيات الذكاء الاصطناعي والتعلم الآلي، وشملت التحديات الرئيسة، التباين في تصميم المجلات التاريخية، والحفاظ على جودة الصور بمرور الوقت، والتحسين المستمر لمعالجة الصور، وتقنيات التعرف الضوئي على الحروف، للتكيف مع أنماط وفترات الصحف المختلفة.

تناقش دراسة Johanna Mauermann and Sarah Oberbichler,(2025)[2]، إشكالية أنه عندما يتعلق الأمر بتحليل مجموعات ضخمة من الوثائق التاريخية - مثل الصحف التاريخية الرقمية - تُعدّ نماذج اللغة أدوات بالغة الفعالية، ولكن مع تزايد شيوع هذه التقنيات في أبحاث العلوم الإنسانية؛ يتعين على الباحثين- أيضًا- التفكير بشكل نقدي، في كيفية استخدامها، أحد أهم الأسئلة التي يواجهها الباحثون، عند استخدام نماذج اللغة في تحليلاتهم؛ هو التحيز، وصف روب

كيتشن في كتابه "علم البيانات النقدي" التحيز؛ بأنه "نمط ثابت من الأخطاء داخل مجموعة بيانات، أو داخل طريقة معالجة البيانات وتحليلها، يُشوّه النتائج والتفسير، تتجلى التحيزات في نماذج الذكاء الاصطناعي على مستويات مختلفة: في تصميم النموذج"، وفي البيانات نفسها، وفي تطبيقها على سياقات لم يُدرّب عليها النموذج، وفي المحفزات نفسها، لذلك؛ تُعد معالجة مسائل التحيز، أمرًا بالغ الأهمية؛ لضمان صمود الأبحاث التاريخية أمام التدقيق، وتتطلب الممارسات العلمية الجيدة؛ أن يتم أخذ هذه التحديات على محمل الجد، ولكن كيف يمكن التأكد، من أن النتائج، لن تُشوّه- بشكل كبير- بسبب أدوات الذكاء الاصطناعي التي نستخدمها؟، في هذه الدراسة، يتم تناول هذا السؤال الجوهري؛ من خلال تقييم منهجي لنماذج اللغة الكبيرة (LLMs) ، المستخدمة لاستخراج المقالات من الصحف التاريخية متعددة اللغات، تركز دراسة الحالة في الدراسة؛ على تغطية زلزال ميسينا الكارثي عام ١٩٠٨ في جنوب إيطاليا.

تهدف دراسة Sun, Shaodan,(2025)[3]؛ إلى تعزيز استرجاع الصحف التاريخية، والاستفادة منها؛ من خلال تطبيق التنظيم الدلالي، وذلك انطلاقًا من منظور دقيق لعناصر المعرفة، ويسعى هذا المسعى؛ إلى إبراز القيمة الكامنة في محتويات الصحف، مع تقديم إرشادات قيّمة، في إطار النماذج المنهجية للبحث في مجال العلوم الإنسانية.

وفقًا لعملية التنظيم الدلالي، ومفهوم عناصر المعرفة، تقترح هذه الدراسة، إطارًا شاملًا يتضمن أربع مراحل محورية؛ وصف عناصر المعرفة، واستخراجها، وربطها، وتطبيقها، في البداية، تم تصميم نموذج وصف دلالي مخصص لعناصر المعرفة، ثم- باستخدام تقنيات التعلم العميق المتقدمة- تتعمق الدراسة، في مجال تعرف الكيانات واستخراج العلاقات، تُعدّ هذه التقنيات؛ فعّالة في تحديد الكيانات داخل محتويات الصحف التاريخية، ورصد الترابطات القائمة بينها، وأخيرًا، تم تطوير منصة إلكترونية؛ تعتمد على Flask ؛ لتمكين تعرف الكيانات والعلاقات داخل الصحف التاريخية.

وقد استخدمت هذه الدراسة، مجموعة بيانات Shengjing Times Changchun ؛ لوصف محتويات الصحف، واستخراجها، وربطها، وتطبيقها؛ فيما يتعلق باستخراج عناصر المعرفة، يتفوق نموذج BERT + BS باستمرار على Bi-LSTM و++CRF ، وحتى BERT من حيث درجتي التذكر وF1؛ مما يجعله خيارًا مثاليًا؛ لتعرف الكيانات في هذا السياق، ومن الجدير بالذكر بشكل خاص نموذجBi-LSTM-Pro ، الذي يتميز بأعلى الدرجات في جميع المقاييس، لا سيما محققًا درجة F1 استثنائية، في تعرف علاقات عناصر المعرفة.

تتجاوز الصحف التاريخية، كونها مجرد تحف فنية؛ لتتحول إلى مستودعات ثمينة، تصون الذاكرة المجتمعية والتاريخية، ومن خلال التنظيم الدلالي، من منظور دقيق لعناصر المعرفة؛ يُمكنها تسهيل استرجاع الدلالة، والترابط الدلالي، وتصور المعلومات، وخدمات اكتشاف المعرفة للصحف التاريخية، عمليًا، يُمكنها تمكين الباحثين؛ من استخلاص رؤى عميقة في السياق التاريخي والثقافي؛ مما يُوسّع آفاق أبحاث العلوم الإنسانية الرقمية، وتطبيقاتها العملية.

تشير دراسة Dilawar Ali et all (2024)[4] ، إلى أنه وعلى الرغم من أن رقمنة مجموعات الصحف التاريخية، تُحسّن إمكانية الوصول إليها بشكل ملحوظ، فإن جزءًا كبيرًا من مجموعات الصحف التاريخية الرقمية، مثل مجموعات المكتبة الملكية البلجيكية(KBR) ، لا يزال غير قابل للبحث على مستوى المقالات، ولى الرغم من ذلك، فإن التطورات الحديثة في أساليب البحث القائمة على الذكاء الاصطناعي، مثل تحليل تخطيط الوثائق؛ تُتيح إمكانية إثراء البيانات الوصفية بشكل أكبر؛ لتحسين إمكانية البحث في مجموعات الصحف التاريخية، تهدف هذه الدراسة؛ إلى مناقشة ذلك، في هذه الورقة البحثية، يستكشف الباحثون كيفية استخدام مناهج الرؤية الحاسوبية والتعلم الآلي الحالية؛ لتحسين الوصول إلى الصحف التاريخية الرقمية، ولتحقيق ذلك؛ يقترحون سير عمل، باستخدام مناهج الرؤية الحاسوبية والتعلم الآلي، يهدف إلى؛ أولًا: توفير الوصول إلى مجموعات الصحف التاريخية الرقمية على مستوى المقالات؛ باستخدام تحليل تخطيط المستندات، ثانيًا: استخراج أنواع محددة من المقالات (مثل: الملاحق الأدبية من مجلة "لو بيبول" عام ١٩٣٨)، ثالثًا: إجراء تحليل تشابه الصور؛ باستخدام أساليب التصنيف الخاضعة للإشراف (أو غير الخاضعة للإشراف)، ورابعًا: إجراء تعرف الكيانات المسماة (NER)؛ لربط المعلومات المستخرجة بالبيانات المفتوحة.

تُظهر نتائج الدراسة؛ أن سير العمل المُقترح، يُحسّن إمكانية الوصول إلى الصحف التاريخية الرقمية، وإمكانية البحث فيها، كما يُسهم؛ في بناء مجموعات بحوث العلوم الإنسانية الرقمية، تُمكّن الأساليب القائمة على الذكاء الاصطناعي، من الاستخراج التلقائي للملفات، وتجميع الصور المتشابهة، والربط الديناميكي للمقالات ذات الصلة، ويتيح سير العمل المقترح؛ استخراج المقالات تلقائيًا، بما في ذلك الكشف عن نوع محدد منها، مثل مجموعة من المقالات، أو ملحق أدبي، يُعد هذا مفيدًا- بشكل خاص- لباحثي العلوم الإنسانية؛ إذ يُحسّن إمكانية البحث في هذه المجموعات، ويُمكّن من بناء مجموعات مقالات حول مواضيع محددة.

تشير دراسة Atul Kumar and Gurpreet Singh Lehal,2024)[5]، إلى أنه لسنوات عديدة، كانت الصحف مصدرًا ممتازًا للمعلومات، ولاستخراج معلومات مفيدة، و لابد من رقمنة هذه الصحف، وينطبق الأمر نفسه على الصحف البنجابية، في هذه الدراسة، تم استخدام تجزئة صور الصحف؛ لفصل المقالات المختلفة عن الصحف البنجابية، وتمت مناقشة مختلف العوائق، التي واجهت استخراج الصحف البنجابية، كما تم جمع مجموعة بيانات من ٤٠٠ صورة صحفية، وتدريب هذه الصور؛ باستخدام أحدث نماذج كشف الأجسام Faster RCNN تُظهر النتائج التجريبية، أن هذه النماذج لاستخراج المقالات؛ حققت نتائج جيدة، بمتوسط دقة ٨١,٨% لجميع الفئات.

تشير دراسة (Abdullah Almutairi,2024)[6] إلى أن رقمنة الصحف، اكتسبت اهتمامًا واسعًا حول العالم، تحتوي أرشيفات الصحف والمجلات الرقمية، على ثروة من المعلومات تمتد لعقود، لاستخراج هذه الوفرة من المعلومات، تم استخدام تقنيات التعرف الضوئي على الحروف (OCR)مع شرح يدوي مكثف للصفحات، تستخرج تقنيات التعرف الضوئي على الحروف، النص من الصورة الخام للصفحة، بينما يضيف شرح الصفحة، بيانات وصفية، حول محتوى

الصفحة، مثل فئة الصفحة وموقع المقالات، وعناصر أخرى، لأتمتة هذه العملية، اقترح إطار عمل؛ للكشف عن عناصر صفحات الصحف وتصنيفها، يستخدم الإطار، الميزات المرئية للصحيفة الرقمية؛ لتصنيف نوع الوسائط المطبوعة، وتفكيك صفحة الصحيفة إلى عناصرها الرئيسية (المقالات الإخبارية والإعلانات وعناوين الصفحات)، باستخدام اكتشاف الكائنات، بعد ذلك، يستخدم كل من الميزات المرئية والنصية؛ لتصنيف صفحة الصحيفة إلى أقسامها الرئيسية (الصفحة الأولى، والسياسة، والاقتصاد، والرياضة، والإعلان)، تُستغل مرحلة كشف العناصر، باستخدام المقالات الإخبارية فقط، لتصنيف الصفحات؛ لأن عناصر أخرى- مثل الإعلانات- قد تحتوي على خصائص بصرية ونصية غير مرتبطة بقسم الصفحة، يُهيئ هذا الإطار، صفحة الصحيفة، لتقنيات التعرف الضوئي على الحروف (OCR) لاستخراج معلومات مفيدة، مرحلة كشف عناصر الصفحات في الإطار مستقلة عن اللغة؛ مما يسمح لها باستخراج المقالات من الصحف بلغات مختلفة (مثل العربية، الإنجليزية، الفرنسية، الألمانية، إلخ)، يستخدم الإطار بنيتين للشبكات العصبية العميقة Faster R-CNN :المستندة إلى بنية الشبكة العصبية التلافيفية (CNN)، و Transformers لتصنيف العناصر، وكشفها في الوسائط المطبوعة.

تؤكد دراسة (Justyna Sikora and Chris Haffenden,2024)[7] أن التوافر المتزايد للمواد والأدوات الرقمية للتحليل الحاسوبي واسع النطاق؛ أدى إلى اهتمام متزايد، بمناهج البيانات الضخمة في العلوم الإنسانية والاجتماعية، وعلى الرغم من ذلك، لا يزال الدور الحيوي لتنظيم البيانات- بوصفه شرطًا أساسيًّا لمثل هذه المشاريع- غير مقدّر حق قدره، تُفصّل هذه الورقة البحثية، عمل مختبر KBLab في المكتبة الوطنية السويدية، في اختبار أدوات الذكاء الاصطناعي؛ للمساعدة في تنظيم أرشيف الصحف الرقمي، وجعله أكثر ملاءمةً للبحوث الكمية القائمة على التعلم الآلي، وتقدم وصفًا لبيانات الصحف في المكتبة؛ لتوجيه الباحثين المهتمين بالمادة، قبل أن تستعرض نتائج الاستكشافات، باستخدام تنظيم البيانات الآلي، وتختتم الورقة البحثية، برسم الخطوات التالية المحتملة لهذه الجهود الاستكشافية، بالإضافة إلى وضع هذا المشروع، ضمن توجه حديث أوسع نطاقًا، لوضع تصور لمفهوم جاهزية البيانات، وإعطائه الأولوية، وتتمثل حجتها الرئيسة؛ في لفت الانتباه إلى تنظيم البيانات بوصفه جزءًا أساسيًّا من أي مشروع بحثي رقمي، وليس أمرًا سابقًا أو خارجيًا عن عملية البحث.

قدمت دراسة (Laura Manrique-Gómez et all,2024)[8] ؛ إسهامين مهمين: أولاً، تُقدّم مجموعة بيانات جديدة، لنصوص صحف أمريكا اللاتينية، في القرن التاسع عشر، مُعالجةً بذلك فجوةً جوهريةً، في مجموعات النصوص المتخصصة للتحليل التاريخي واللغوي في هذه المنطقة، ثانيًا، تُطوّر إطارًا مرنًا، يستخدم نموذجًا لغويًّا واسعًا، لتصحيح أخطاء التعرف الضوئي على الحروف (OCR) ، وكشف الأشكال اللغوية السطحية، في مجموعات النصوص الرقمية، هذا الإطار شبه الآلي، قابل للتكيف مع سياقات ومجموعات بيانات مُختلفة، ويُطبّق على مجموعة البيانات المُنشأة حديثًا.

جُمعت مجموعة البيانات، في البداية من أرشيفات الصحف الرقمية الكولومبية، وركزت- بشكل أساسي- على المنشورات، التي تضمنت رسومًا كاريكاتورية أو توضيحية، والتي كانت

مخصصة للنمذجة متعددة الوسائط لاحقًا، كما امتدت هذه المراجعة، لتشمل المجموعات الورقية في الموقع، حيث لم تتم رقمنة سوى حوالي ٥٠% منها، ومن خلال هذه العملية؛ تم تحديد ٦٤ عنوانًا صحفيًا، تمثل ٧% من إجمالي ١٦٥٥ منشورًا في المجموعات، أسفر هذا التكرار الأولي، عن مجموعة بيانات تتكون من ٤٠٣٢ صفحة ممسوحة ضوئيًا من الصحف، معظمها من غرناطة الجديدة - وهي دولة كانت تضم كولومبيا، وبنما، وفنزويلا، والإكوادور.

أكملت عملية التكرار الثانية، مراجعة ٣٠٣٨ صحيفة رقمية، من ٥٨ مجموعة رقمية في جميع أنحاء المكسيك، والأرجنتين، وكولومبيا، وبيرو، وتشيلي، وبنما، وفنزويلا، وأوروغواي، وبوليفيا، وكوبا، والإكوادور، وبعض البلدان، مثل بوليفيا، وكوبا، وفنزويلا، لديها مجموعات ويب محدودة للغاية أو معدومة؛ مما يؤدي إلى نقص تمثيلها، أو غيابها عن مجموعة البيانات النهائية، بالإضافة إلى ذلك، تمت طباعة بعض الصحف في أوروبا؛ بسبب انخفاض التكاليف؛ وفي بعض الحالات، تم استخدام الاستعانة، بمصادر خارجية للطباعة، تتكون مجموعة البيانات النهائية من ١٩٧ عنوان صحيفة و٢٣٥٢٢ صفحة من الصور الممسوحة ضوئيًا، معظمها من مدينة مكسيكو (المكسيك هي الدولة الوحيدة التي رقمنت مجموعتها بالكامل)، ولكنها تشمل- أيضًا- منشورات من مدن أخرى في أمريكا اللاتينية، مثل بوينس آيرس، وليما، وبوغوتا، وسانتياغو دي تشيلي .

أشارت دراسة Nancy Girdhar et all(2024)[9] ، إلى أن الصحف التاريخية تُعدّ موارد قيّمة لفهم المجتمعات السابقة والحفاظ على التراث الثقافي، ولي الرغم من ذلك، تُشكّل رقمنة هذه الصحف تحديات؛ نظرًا لتعقيد تصميماتها، وكثرة محتواها، تُعدُّ تجزئة المقالات، التي تتضمن تحديد واستخراج المقالات الفردية من صور الصحف الممسوحة ضوئيًا؛ أمرًا بالغ الأهمية، لاستخراج المعلومات واسترجاعها بكفاءة، ولى الرغم من اقتراح بعض الخوارزميات، القائمة على القواعد، فإن قابلية تطبيق الشبكات العصبية العميقة (DNNs) لهذه المهمة، قد حظيت- مؤخرًا- باهتمام. تستكشف هذه الدراسة، قابلية تطبيق التعلم الانتقالي، لتجزئة المقالات من صور الصحف التاريخية، ولهذا الغرض، تم استخدام تسعة هياكل أساسية مُدرّبة مسبقًا، مُختارة خصيصًا من عائلة شبكاتResNet ، وتم اقتراح وحدة تجزئة مقالات، قائمة على تقريب الصندوق المُحدّد، مُصمّمة خصيصًا لهذه المهمة، علاوة على ذلك، تم تقديم مقياس لمتوسط تغطية المقالات المُقدّرة، والذي يحسب قدرة النموذج على التجزئة على مستوى المقالة، أُجريت تجارب على مجموعة بيانات) NAS فصل المقالات في(NewsEye ؛ مما يضمن ملاءمة نهج الدراسة للبيانات التاريخية، تُقيّم هذه الدراسة، أداء نماذج مُدرَّبة مُسبقًا، محققةً متوسط تغطية مُقدَّرة للمقالات بلغ ٠,٩٥٦ و٠,٩٦٩ و٠,٩٩٥ على مجموعات بيانات ONB و NLF وBNF، على التوالي، تُؤكد هذه النتائج، فعالية تعلم النقل في التكيف مع مهام تحليل التصميم التاريخي، وخاصةً تجزئة المقالات، علاوة على ذلك، تُؤكد هذه النتائج، أهمية تعلم النقل والنماذج المُدرَّبة مُسبقًا، بوصفها أدوات فعّالة، للتعامل مع تصميمات الصحف التاريخية المُعقدة.

تؤكد دراسة Bhawna Piryani et all (2024)[10] ، أن مهام الإجابة عن الأسئلة (QA) وفهم القراءة الآلية (MRC)، شهدت تطورًا ملحوظًا في السنوات الأخيرة؛ بفضل التطور السريع لتقنيات التعلم العميق، ومؤخرًا نماذج اللغات الكبيرة، في الوقت نفسه، أصبح العديد من مجموعات البيانات المرجعية، متاحة لمهام ضمان الجودة، وفهم القراءة الآلية، وعلى الرغم من ذلك، فإن معظم مجموعات البيانات المرجعية واسعة النطاق الحالية، قد تم إنشاؤها- بشكل أساسي- باستخدام مجموعات مستندات متزامنة، مثل ويكيبيديا، أو الويب، تحتوي مجموعات المستندات الأرشيفية، مثل الصحف التاريخية، على معلومات قيّمة من الماضي، لا تزال غير مستخدمة على نطاق واسع لتدريب نماذج اللغات الكبيرة، للإسهام بشكل أكبر، في تطوير مهام ضمان الجودة، وفهم القراءة الآلية والتغلب على قيود مجموعات البيانات السابقة، تقدم هذه الدراسةChronicling America QA ، وهي مجموعة بيانات ضمان جودة زمنية واسعة النطاق، تحتوي على ٤٨٧ ألف زوج من الأسئلة والأجوبة؛ تم إنشاؤها؛ بناءً على مجموعة الصحف التاريخية Chronicling America، تم إنشاء مجموعة البيانات الخاصة بالدراسة، من مجموعة صحف Chronicling America، التي تمتد على مدار ١٢٠ عامًا، أحد التحديات الرئيسية لاستخدام مجموعات الصحف التاريخية الرقمية؛ هو انخفاض جودة نصوص التعرف الضوئي على الحروف؛ لذلك، ولتمكين اختبار نماذج ضمان الجودة بشكل واقعي؛ يمكن استخدام مجموعة بيانات هذه الدراسة، بثلاث طرق مختلفة: الإجابة عن أسئلة من محتوى خام ومشوش، والإجابة عن أسئلة من نسخة أكثر وضوحًا وتصحيحًا من المحتوى، بالإضافة إلى الإجابة عن أسئلة من صور ممسوحة ضوئيًا لصفحات الصحف، هذا، بالإضافة إلى أن Chronicling America QAتمتد لأطول فترة زمنية بين مجموعات بيانات ضمان الجودة المتاحة؛ يجعلها موردًا فريدًا ومفيدًا للغاية.

اعتمدت دراسة Sarah Oberbichler (2024)[11]، على دراسات حالة الكوارث الطبيعية/البيئية والهجرة (العودة) الفترة الزمنية: ١٨٥٠-١٩٥٠، الدول: ألمانيا، النمسا، سويسرا، إيطاليا، فرنسا، المملكة المتحدة، مجموعة صحف المكتبة الوطنية السويسرية - أرشيف الصحف الإلكترونية، صحيفةBündner Nachrichten ، ومجموعة صحف ضمن Annolyzerمنصة عرض NewsEye التي تم دمجها في مختبراتÖNB ، صحيفة Illustrierte Kronen-Zeitung، وأركانوم؛ واجهة برمجة تطبيقات لتجزئة الصحف؛ مثال "صحيفة نيوز- فيلت- بلات"، واعتمدت على فصل تلقائي للمقالات، وتوصلت الدراسة؛ إلى أنه يتطلب العمل مع النماذج متعددة الوسائط أو الضبط الدقيق، موارد حاسوبية، وطاقة أكبر بكثير، ولا تُبرر درجات F1 التي تزيد على ٩٠% هذا الاستهلاك الإضافي للموارد.

تناقش دراسة Filip Dobranić and Matevž Pesek, (2024)[12]، أنه في الآونة الأخيرة، أظهرت أساليب تخطيط الوثائق القائمة على الشبكات العصبية نتائج واعدة، إلا أن النماذج المُدرَّبة مُسبقًا، لا تزال تُمثِّل تحديًا، في تحقيق أداء جيد، عند تطبيقها على مصادر رقمية جديدة تمامًا، دون أي ضبط دقيق، أو في الحالات التي يكون فيها الانحراف عن قواعد تصنيف النموذج الأصلي متطلبًا صعبًا، في هذه الدراسة، نُقدِّم مجموعة بيانات جديدة مُعلَّقة،

تضم ٤٢٣ صفحة من الصحف، والأداء الأساسي لنموذج مُعَدَّل بدقة، على مجموعة البيانات، التي تمتد على مدار ٤٦ عامًا، نُقيِّم الحد الأدنى من البيانات اللازمة لضبط نموذج YOLOv8n بدقة لتصنيف الإعلانات ونعي الوفيات في صحيفة"Slovenski narod" ، الصادرة في مطلع القرن العشرين، نُقارن أداء مجموعات بيانات مُعلَّقة، أصغر حجمًا تدريجيًّا؛ لتحديد منطقة تناقص العائدات، تُظهِر الدراسة؛ أن الزيادة في الأداء لكل صورة توضيحية إضافية، يتم ضبطها بدقة بواسطة نموذج YOLOv8n ؛ تتضاءل بعد حوالي ٢٥٠ تسمية توضيحية، بغض النظر عن عدد الفئات المدربة.

يحتاج العمل المستقبلي؛ إلى مقارنة كمية التعليقات التوضيحية المطلوبة عبر العديد من الصحف التاريخية، وليس صحيفة واحدة فقط، بالإضافة إلى التجربة مع فئات أخرى مميزة بصريًا (مثل العناوين والإشعارات العامة) ؛ من أجل اكتساب المزيد من الثقة، في عملية تحسين الأداء الملحوظة التي تم قياسها في مئات التعليقات التوضيحية، تقدم هذه الدراسة، الخطوة الأولى في كل من إنشاء مجموعة البيانات، وكذلك تجارب معدل التعلم في الدراسة، يجب أن تقارن التجارب الإضافية، الأداء في مهام مماثلة على مجموعات بيانات أكبر بكثير، ومتنوعة بلغات أخرى.

تقدم دراسة (Vincent Wai-Yip LUM and Michael Kin-Fu YIP,2023)[13] منهجيات استخراج العناوين والرسوم التوضيحية، من صحيفة تاريخية لسرد القصص، بما يدعم البحث العلمي الرقمي، استكشفت الورقة البحثية، كيف تُسهّل الأدوات الرقمية الجديدة فهم محتوى الصحيفة في سياق الزمان والمكان، وقد اختيرت صحيفة "هونغ كونغ نيوز" من صحيفة هونغ كونغ الشعبية المبكرة لدراسة الحالة؛ نظرًا لقيمتها التاريخية الفريدة لدى الباحثين، قُيّمت المنهجيات المقترحة بتقنية التعرف الضوئي على الحروف (OCR) باستخدام نماذج الكشط، والتعلم العميق لاكتشاف الكائنات، وطُوّر منتجان للتصور؛ لعرض جدوى الأساليب المقترحة في الدراسة؛ لخدمة غرض سرد القصص.

تشير دراسة(Nancy Girdhar et all,2023)[14] إلى أن رقمنة الوثائق التاريخية، مهمة بالغة الأهمية للحفاظ على التراث الثقافي، وجعل كميات كبيرة من المعلومات، متاحة لعامة الناس، أحد التحديات في هذه العملية؛ هو فصل المقالات الفردية عن صور الصحف القديمة، وهو أمر مهم لتحليل النصوص واسترجاع المعلومات، تقدم هذه الدراسة، نهجًا جديدًا، وهو نهج قائم على القواعد لفصل المقالات (STRAS) في الصحف التاريخية، يتضمن النهج المقدم STRAS،استخدام المعلومات النصية؛ عن طريق استخراج تضمينات مناطق النص؛ باستخدام صور الإدخال الممسوحة ضوئيًّا، وملفات تنسيق PAGE المقابلة لها، ثم يتم تصنيف مناطق النص ذات التضمينات السياقية المتشابهة، ويتم فصل المقالات؛ بناءً على مجموعة قواعد محددة، تم اختبار النهج المقدم، على الصحف الفرنسية والفنلندية، في القرنين التاسع عشر، وأوائل القرن العشرين، بالإضافة إلى ذلك، تم تقديم مقاييس جديدة؛ خصيصًا لمهمة فصل المقالات معدل خطأ المقالة ،(AER)، ودرجة تغطية المقالة(ACS) ، ودرجة التنبؤ الصحيحة للمقالة (PPA) تُقيّم هذه الدراسة، أداء نماذج مختلفة، بما في ذلك نموذج التخطي

(sgSTRAS)، ونموذج حقيبة الكلمات المستمرة(cbowSTRAS) ، ونموذج FastText (ftSTRAS)، ونموذج SpaCy المُدرّب مسبقًا (preSTRAS) ، وتُظهر النتائج؛ أن نموذج sgSTRASيُحقق أعلى متوسط درجاتACS، وهو ٠,٨٣٤٣ و٠,٨٦١١ في مجموعتي البيانات الفرنسية والفنلندية، على التوالي، متفوقًا على جميع النماذج الأخرى، وتُظهر نتائج الدراسة؛ أن السمات النصية الدلالية، تحتوي على معلومات قيّمة، وأن اختيار طريقة تضمين مناسبة؛ يُؤثر -بشكل كبير- على الأداء العام للنهج المُقترح لتجزئة المقالات، على حد علم الباحثين، هذه هي الدراسة الأولى، التي تطبق نهجًا قائمًا على قواعد التشابه النصي الدلالي، لفصل المقالات في الصحف التاريخية؛ مما يسد فجوة في الأدبيات الموجودة، ويفتح آفاقًا جديدة، لمزيد من البحث في هذا المجال.

تقدم دراسة Cyrille Suire and Nicolas Sidère (2023)[15] ، موردًا تعليميًا مفتوحًا (OER)، حول البحث التاريخي الرقمي مع الصحف التاريخية، يهدف هذا المورد؛ إلى تزويد الطلاب بالوسائل اللازمة لفهم المخاطر الناجمة عن التعامل مع مجموعات كبيرة من الوثائق الرقمية، بالإضافة إلى سبل الاستفادة، من تطورات معالجة اللغة الطبيعية على المجموعات الكبيرة متعددة اللغات من الصحف التاريخية الأوروبية، يستغل هذا المورد نتائج مشروع "نيوز آي هورايزون ٢٠٢٠" للبحث والابتكار، وهو جزء من مجموعة من ٧ موارد تعليمية مفتوحة، طُوّرت، وتشاركت ضمن برنامج إيراسموس، و مشروع منصة الأساليب الرقمية للفنون والعلوم الإنسانية (DiMPAH)، وأشارت الدراسة، إلى أن تحليل الوثائق وفهمها في سياق الصحف التاريخية، يتضمن :

- التعرف الضوئي على الحروف.
- فصل المقالات .
- إثراء النصوص الدلالية، ويتضمن:

تحليل الوثائق، وتعزيز محتواها بالبيانات الوصفية الدلالية، في حالة الصحف التاريخية، يتضمن هذا -عادةً- التركيز على الكيانات المُسمّاة، وهي أسماء علم، تُشير إلى كيانات مُحددة، مثل الأشخاص والمنظمات، والمواقع، والتواريخ، ولكي نكون أكثر تحديدًا، يهدف إثراء النص الدلالي؛ إلى تحديد هذه الكيانات داخل المستندات، وتوضيحها لقاعدة المعرفة، وربطها بالموقف الذي تم ذكرها فيه، يُعدّ تعرف الكيانات المُسمّاة؛ من أهمّ المهام وأكثرها دراسةً في مجال معالجة اللغة الطبيعية (NLP) وله تطبيقات صناعية وعلمية عديدة، ويُسهّل (OER) استرجاع المعلومات في الصحف التاريخية بشكلٍ واضح، وهو من أهمّ المواضيع التي طُوّرت في موارد التعليم المفتوح .

-تُعرَّف الكيانات المُسمّاة:

الكيانات المُسمّاة، عادةً ما تكون أسماء أعلام، تُشير إلى كيانات مُحدّدة، مثل الأشخاص، والمؤسسات، والمواقع، والتواريخ، وما إلى ذلك.

-لاستكشاف الوثائق التاريخية، تُعدّ أدوات مثل NER ، قيّمة للغاية للباحثين، والمؤرخين، وأمناء المكتبات؛ إذ تُضفي هيكليةً، على كميات البيانات غير المنظمة، وتُحسّن الوصول إلى المجموعات الرقمية التاريخية، يُمكن اعتبار اكتشاف الكيانات، خطوةً أولى في استكشاف مجموعات البيانات، هذه المهمة، مفيدة- بالطبع- لتحليل الصحف التاريخية الرقمية، ولكنها حاسمة أيضًا في العديد من الحالات الأخرى، مثل:

- تصنيف المحتوى لمقدمي الأخبار.
- أتمتة دعم العملاء.
- استخراج المعلومات القيمة من الوثائق الطبية.
- مساعدة تقييم المخاطر للمؤسسات المالية.
- تصنيف النص.
- استخراج الحدث.

استخراج الأحداث، هو نوع من استخراج المعلومات(IE) ، ويتضمن استخراج معلومات محددة من نصوص، تتعلق بأحداث معينة، تتضمن هذه المهمة مهمتين فرعيتين: كشف الأحداث (ED)، واستخراج حجج الأحداث، يتضمن كشف الأحداث، استخراج معلومات مهمة، مثل كلمة مفتاحية، أو جملة، أو عبارة، أو امتداد نصي، يُشير إلى الحدث محل النقاش في النص، على سبيل المثال، قد يذكر مقال إخباري، تفشي وباء حديث، أو تعيين رئيس جديد، في مثل هذه الحالات، عادةً ما يُشار إلى الأحداث المراد تحديدها، باسم الوباء، أو مصطلح "انتخابات"، يركز استخراج حجج الأحداث، على استخراج معلومات أكثر تفصيلًا حول الحدث، مثل الموقع.

- **كشف الموقف:**

اكتشاف الموقف، هو مجال فرعي من معالجة اللغة الطبيعية؛ يهدف إلى تحديد ما إذا كان مؤلف جزء من النص، مؤيدًا أو معارضًا أو محايدًا، تجاه كيان مستهدف (مثل شخص أو منظمة) مذكور أو ضمنيًا في النص.

لتطبيق المفاهيم، والأساليب، والأدوات، التي طُوّرت في موارد التعليم المفتوح، تم إعداد دراستي حالة حول: الخطاب الإعلامي والسياسي، لجائحة الإنفلونزا الإسبانية من جهة، وظهور أنماط ملابس جديدة للنساء في القرنين التاسع عشر والعشرين، من جهة أخرى، تُتيح هاتان الدراستان، تجربة مفاهيم إنشاء النصوص، والتحقق من صحة البيانات، وتصور البيانات، وتحليل النصوص، بأهداف ومنهجيات بحثية مختلفة، في كلتا الحالتين، وفي كل مرحلة من مراحل البحث، يُقترح منظور نقدي للطلاب، حول البحث عن التحيزات التقنية التي قد يواجهونها

هدفت هذه الموارد التعليمية المفتوحة؛ إلى تزويد الطلاب، بوسائل لفهم مجموعات ضخمة من الوثائق التاريخية بأية لغة، واكتشاف كيفية استخدام أدوات التحليل؛ لتبسيط مهامهم، وإيجاد طرق جديدة؛ لاكتساب المعرفة وتطوير البحث، كما تهدف الدورة؛ إلى تجاوز تأثير الصندوق

الأسود، الذي غالبًا ما يكون متأصلًا، في أتمتة المهام في المجموعات الكبيرة؛ مما يمنح الطلاب الأدوات اللازمة لفهم ما يحدث من المدخلات إلى المخرجات بشكل أفضل.

الصحف التاريخية مرآة للماضي، تصف الأحداث التاريخية من منظور معاصر، وعلى الرغم من ذلك، وخلافًا للأخبار الحالية، تُحلَّل وتُعلَّق عليها يدويًا بأثر رجعي، من خلال تدقيق المؤرخين، مبدئيًا، يعني هذا أن جودة هذه البيانات المُعلَّقة، المُقدَّمة لتدريب أنظمة الذكاء الاصطناعي، لها قيمة أكبر من البيانات المُستقاة من البيانات المعاصرة، هذا يعني- أيضًا- إمكانية تطبيق دروس الماضي في الحاضر بشكل منهجي؛ من خلال استخدام البيانات الضخمة التاريخية بوصفها وسيلة للتعلم (أيضًا من خلال التعلم الآلي) ما يجب فعله في الحاضر، ومن التطبيقات النموذجية لذلك، دعم صانعي السياسات؛ من خلال مساعدتهم- على سبيل المثال- على الاستفادة من أزمات الماضي لإدارة أزمات الحاضر، واستباق أزمات المستقبل.

تعكس حالتا الاستخدام الموصوفتان في هذا المصدر التعليمي المفتوح القضايا المعاصرة، تظهر مناقشات منتظمة في مجموعات مختلفة، حول كيفية ارتداء بعض أعضائها للملابس، أو عدم ارتدائها، يتشابه الخطاب الحالي لمثل هذا النقاش العام دائمًا، مع الخطاب الذي واجهناه في "النساء في السراويل"، كما أنه مع تأثير جائحة كوفيد-١٩ علينا جميعًا مؤخرًا، فإن الطريقة التي تم بها التعامل مع جائحة كوفيد-١٩، تشبه- إلى حد كبير- الطريقة التي تم بها التعامل مع الإنفلونزا الإسبانية، قبل أكثر من قرن: الرفض الأولي لأي خطر، والسخرية من النظرة المستقبلية لأقنعة الوجه التي تعتبر عديمة الفائدة، والعلاجات المعجزة القابلة للنقاش، والوعد بأن مجتمع ما بعد الجائحة سيكون أفضل بشكل جذري.

تناقش دراسة (Nicolas Gutehrlé and Iana Atanassova,2022)[16] ، أنه في السنوات الأخيرة، قادت المكتبات والأرشيفات حملات رقمنة مهمة، أتاحت الوصول إلى مجموعات ضخمة من الوثائق التاريخية، وعلى الرغم من أن هذه الوثائق، غالبًا ما تكون متاحة بصيغة XML ALTO، فإنها تفتقر إلى معلومات حول بنيتها المنطقية، تتناول هذه الورقة البحثية، مشكلة تحليل التخطيط المنطقي المطبق على الوثائق التاريخية باللغة الفرنسية، وتقترح طريقة قائمة على القواعد، تقوم بتقييمها ومقارنتها مع نموذجين من نماذج التعلم الآلي، وهما RIPPERو Gradient Boosting، تحتوي مجموعة بيانات الدراسة، على صحف ودوريات ومجلات فرنسية، نُشرت في النصف الأول من القرن العشرين في منطقة فرانش كونتيه.

تشير نتائج الدراسة، إلى أن نظامها القائم على القواعد، يتفوق على النموذجين الآخرين، في جميع التقييمات تقريبًا، ويتمتع بنتائج استرجاع أفضل بشكل خاص؛ مما يشير إلى أن نظام الدراسة، يغطي أنواعًا أكثر من كل تسمية منطقية، مقارنةً بالنموذجين الآخرين، عند مقارنة RIPPERمع Gradient Boosting ؛ تشير الدراسة، إلى أن Gradient Boosting يتمتع بدرجات دقة أفضل، بينما يتمتع RIPPER بدرجات استرجاع أفضل، ويُظهر التقييم، أن نظام الدراسة، يتفوق على نموذجي التعلم الآلي، ويوفر استرجاعًا أعلى بكثير، كما يؤكد إمكانية استخدام نظام الدراسة لإنتاج مجموعات بيانات مُعلّقة كبيرة، بما يكفي لاستيعاب مناهج التعلم الآلي، أو التعلم العميق لمهمة تحليل التخطيط المنطقي، قد يُؤدي دمج القواعد ونماذج التعلم

الآلي في أنظمة هجينة إلى تحقيق أداء أفضل، علاوة على ذلك، ومع التطور السريع لتخطيط الوثائق التاريخية، يتمثل أحد الحلول الممكنة للتغلب على هذه المشكلة؛ في تطبيق خوارزميات تعلم القواعد، لإعادة بناء مجموعات القواعد، بما يتناسب مع فترات النشر المختلفة.

تناولت دراسة Sarah Oberbichler et all (2021)[17]، الفرص والتحديات متعددة التخصصات، للعمل مع التراث الثقافي الرقمي، مثل الصحف التاريخية الرقمية، وتقترح الدراسة، سير عمل متكامل للتأويل الرقمي، يجمع بين مناهج بحثية تخصصية بحتة من علوم الحاسوب، والعلوم الإنسانية، وأعمال المكتبات، وقد أدت الاهتمامات والدوافع المشتركة للتخصصات؛ إلى مشاريع وتعاون متعدد التخصصات، مثل مشروع "نيوز آي"، الذي يعمل على إيجاد حلول مبتكرة، لكيفية البحث (أو إعادة البحث) في بيانات التراث الرقمي، والوصول إليها، واستخدامها، وتحليلها، وترى الدراسة؛ أن التعاون بين التخصصات المختلفة، يمكن أن يستفيد من الفهم الجيد لسير العمل، والتقاليد الخاصة بكل تخصص من التخصصات المعنية، ولكن يجب إيجاد مناهج متكاملة؛ لاستغلال الإمكانات الكاملة للمصادر الرقمية بنجاح، علاوة على ذلك، تقدم الورقة، نظرة ثاقبة على الأدوات والأساليب الرقمية، والتأويلات في التطبيق، موضحةً أن البحث المتكامل متعدد التخصصات، يحتاج إلى بناء شيء ما بين التخصصات، مع احترام وفهم خبرات وتوقعات كل منها.

إن إعادة التعرف الضوئي على الحروف في المستندات، ليست واقعية دائمًا، بالنسبة للمجموعات الكبيرة جدًا؛ بسبب التكاليف الحسابية، والأهم من ذلك؛ بسبب انخفاض دقة الصورة الناتجة عن حملات الرقمنة السابقة، وبالتالي، فإن هناك نهجًا آخر؛ لتحسين جودة النص، دون إعادة التعرف الضوئي على الحروف، وهو العمل مع تصحيح التعرف الضوئي على الحروف اللاحق، والذي قدم له باحثو NewsEye دراسات شاملة، بالاعتماد على التعلم العميق، وتحليل أخطاء التعرف الضوئي على الحروف، بالإضافة إلى المعايير العامة، بالإضافة إلى ذلك، يبحث علماء الكمبيوتر من جامعتي لاروشيل وهلسنكي، عن أدوات وطرق آلية، قادرة على إزالة أخطاء التعرف الضوئي على الحروف المتبقية (المعروفة أيضًا باسم ضوضاء التعرف الضوئي على الحروف)، إما عن طريق تطبيق طرق مختلفة؛ لتصحيح الإملاء، أو باستخدام طرق عصبية متقدمة، إلى جانب تمثيل الكلمات.

غالبًا ما يقع تحليل التردد بين المتخصصين، بمعنى أنه ليس صعبًا من الناحية المنهجية على علماء الكمبيوتر، ولكن قد يكون من الصعب على خبراء مجال العلوم الإنسانية إتقانه بأنفسهم، غالبًا ما تفتقر الأدوات المقدمة في واجهات المستخدم الرسومية، إلى إمكانيات تطبيع الترددات، فيما يتعلق بخطوط الأساس المختلفة، على سبيل المثال، فإن مقارنة حدوث الكلمات المحملة سياسيًا في صحيفتين، ليست مفيدة للغاية، ما لم نتمكن من تطبيعها، مقابل عدد الرموز التي نشرتها الصحيفتان معًا، أو بعض خطوط الأساس الأخرى ذات الصلة، هذا فقط، يعطينا نتائج قابلة للمقارنة وللتفسير، يمكن أن يكون تحديد خط أساس ذي صلة مهمة غير لا يمكن تجاهلها؛ حيث قد ينتهي الأمر بالتطبيع، إلى عكس شيء آخر غير المقصود في الأصل (على سبيل المثال، الاهتمامات التحريرية بدلًا من الاختلافات في تفضيلات الكلمات)، ومن ثم، فإن سير

عمل التأويل المتكامل، يتطلب الخبرة من جانب خبراء المجال (ما هو ذو صلة، وما هو معروف بالفعل) وعلماء الكمبيوتر (ما يمكن حسابه، وما نوع البيانات الوصفية المتاحة)، في حين أن التفكير النقدي في أدوات التعداد والتعدين النصي، وكذلك في الممارسة التاريخية نفسها، أمر ضروري في عملية تحليل البيانات وتفسيرها.

تقدم دراسة (Benjamin Lee,2022)[18]، علم الآثار للبيانات لمجموعة بيانات Newspaper Navigatorالتابعة لمكتبة الكونجرس، والتي أنشأتها بوصفها جزءًا من برنامج المبتكر المقيم بمكتبة الكونجرس، تتكون مجموعة البيانات، من محتوى مرئي مستخرج من ١٦ مليون صفحة صحيفة تاريخية في قاعدة بيانات Chronicling America باستخدام تقنيات التعلم الآلي، في دراسة الحالة، أفحص الطرق المتعددة، التي يتم بها تحويل صفحة صحيفة Chronicling America وإخراجها من سياقها في أثناء رحلتها من قطعة أثرية مادية، إلى سلسلة من الصور الاحتمالية، والرسوم التوضيحية والخرائط، والقصص المصورة، والرسوم الكاريكاتورية، والعناوين الرئيسية، والإعلانات في مجموعة بيانات Newspaper Navigatorوفقًا لذلك، أستمد من مجالات الدراسة، بما في ذلك علم الآثار الإعلامية، ودراسات البيانات النقدية، ودراسات العلوم والتكنولوجيا، والإثنوغرافيا الذاتية، في جميع أنحاء العالم، ولاستخراج مجموعة بيانات Newspaper Navigator، تضع الدراسة في الاعتبار رحلات الرقمنة، لأربع صفحات مختلفة في الصحف السوداء المضمنة في Chronicling America، والتي تعيد جميعها إنتاج نفس صورة WEB Du Bois، في مقال يعلن عن إطلاق The Crisis، المجلة الرسمية لـ NAACP عند تتبع رحلات صفحات الصحف، تقوم الدراسة بتفكيك كيفية قيام كل خطوة في خطوط أنابيب Chronicling America و Newspaper Navigator، مثل عملية التصوير، وبناء بيانات التدريب، ليس فقط بطباعة التحيز على مجموعة بيانات Newspaper Navigator الناتجة، ولكن أيضًا؛ بنشر التحيز؛ من خلال خط الأنابيب، من خلال خوارزميات التعلم الآلي المستخدمة، على طول الطريق، تقوم الدراسة، بالتحقيق في قيود مجموعة بيانات Newspaper Navigator وتقنيات التعلم الآلي بشكل عام، من حيث صلتها بالتراث الثقافي، مع التركيز- بشكل خاص- على التهميش والمحو، عبر التحيز الخوارزمي، الذي يعيد كتابة الأرشيف نفسه ضمنيًا.

تؤكد الدراسة في عرض دراسة الحالة فيها، أهمية علم آثار البيانات، بوصفه آلية لوضع مجموعات بيانات التراث الثقافي في سياقها، ودراستها نقديًا، ضمن المجتمعات التي تُنشئها، وتُصدرها وتستخدمها. يُقدم هذا البحث الإثنوغرافي الذاتي لمجموعة بيانات مُستكشف الصحف (Newspaper Navigator)، آملًا أن يُؤخذ بعين الاعتبار، ليس فقط من قِبَل مستخدمي هذه المجموعة تحديدًا، بل أيضًا من قِبَل مُمارسي العلوم الإنسانية الرقمية، والمستخدمين النهائيين لمجموعات بيانات التراث الثقافي بشكل عام.

إن الأدوار المتزايدة للتعلم الآلي والذكاء الاصطناعي، في بناء مجموعات بيانات التراث الثقافي والعلوم الإنسانية، تتطلب فحصًا نقديًا للتحيزات العديدة، التي أدخلتها الآلات، والخوارزميات، والبشر الذين يبنونها وينشرونها، من تصنيف الصور، إلى التعرف الضوئي على الحروف،

تتراكم آثار القرارات التي تتخذها الآلات ظاهريًا؛ من خلال خط أنابيب الرقمنة، وتتضاعف في كل خطوة؛ مما يتوسط تفاعلات الباحثين مع القطع الأثرية المقدمة رقميًا؛ من خلال عملية البحث والاكتشاف، ونتيجة لذلك؛ بدأ العلماء داخل مجتمع العلوم الإنسانية الرقمية؛ في الدعوة إلى وضع مجموعات بيانات التراث الثقافي، في سياقها الصحيح، داخل الأنظمة الاجتماعية والتقنية، التي يتم إنشاؤها واستخدامها فيها، أحد هذه الأساليب لهذا السياق؛ هو علم آثار البيانات، وهو شكل من أشكال التنقيب الإنساني لمجموعة بيانات، يعرّفها بول فايف، بأنها "استعادة وإعادة بناء كائنات الوسائط، داخل بيئاتها المتغيرة، في دراسات البيانات الحرجة، أثبت هذا التنقيب في مجموعة البيانات - بما في ذلك بنائها ومعالجتها عبر التعلم الآلي - أنه نهج واسع النطاق، وعلى الرغم من ذلك، لم يُعتمد علم آثار البيانات بوصفه ممارسة قياسية بعد، بين ممارسي التراث الثقافي، الذين ينتجون مثل هذه المجموعات من البيانات باستخدام التعلم الآلي.

تناقش دراسة[19] Raphaël Barman et all (2021)، أن الكميات الكبيرة من الوثائق التاريخية الرقمية، التي جُمعت على مدى العقود الماضية، تُهيأ بشكل طبيعي للمعالجة والاستكشاف الآليين، وتتزايد الأعمال البحثية، التي تسعى إلى معالجة النسخ المطابقة للأصل، واستخراج المعلومات منها تلقائيًا، مع تحليل تخطيط الوثيقة؛ بوصفه خطوة أساسية أولى، وإذا كان تحديد وتصنيف الأجزاء المهمة في صور الوثائق، قد شهد تقدمًا ملحوظًا خلال السنوات الأخيرة؛ بفضل تقنيات التعلم العميق، فلا تزال هناك تحديات عديدة، من بينها استخدام أنماط تجزئة أكثر دقة، ومراعاة الوثائق المعقدة وغير المتجانسة مثل الصحف التاريخية، علاوة على ذلك، فإن معظم المناهج، تأخذ في الاعتبار السمات البصرية فقط، متجاهلة الإشارة النصية، في هذا السياق، نقدم نهجًا متعدد الوسائط، للتجزئة الدلالية للصحف التاريخية، يجمع بين السمات البصرية والنصية، واستنادًا إلى سلسلة من التجارب على الصحف السويسرية واللوكسمبورغية ذات التاريخ الزمني؛ فإننا نبحث، من بين أمور أخرى، في القدرة التنبؤية للسمات البصرية والنصية، وقدرتها على التعميم عبر الزمن والمصادر، تُظهر النتائج تحسنًا ثابتًا في النماذج متعددة الوسائط، مقارنةً بخط الأساس البصري القوي، فضلًا عن قوة التحمل في مواجهة التباين العالي للمواد.

تقدّم دراسة [20] Min Lu et all (2021)، نموذجًا جديدًا لتحليل تخطيطات الصحف التاريخية يُسمى Panoptic-DLA، وخلافًا للأعمال السابقة، التي تناولت تحليل التخطيط، بوصفه مشكلة منفصلة؛ للكشف عن الكائنات أو التجزئة الدلالية، تُعرّف الدراسة مهمة تحليل التخطيط؛ بأنها التجزئة البانوبتيكية، الخالية من الاقتراحات، والتي تُعيّن قيمة فريدة لكل بكسل في صورة المستند، مع ترميز كلٍّ من التسمية الدلالية، ومعرّف المثيل، يتكوّن النموذج من فرعين؛ فرع التجزئة الدلالية، وفرع تجزئة المثيل، أولًا، تُفصَل البكسلات إلى أشياء ومواد، بواسطة التصنيف الدلالي، مع اعتبار الخلفية مواد، وكائنات المحتوى، مثل الصور والفقرات، أشياء، بعد ذلك، تُجمّع "الأشياء" المُتنبّأة، إلى مُعرّفات مثيلاتها بواسطة تجزئة المثيل، اعتمد فرع التجزئة الدلالية DeepLabV3+ للتنبؤ بتسميات الفئات لكل بكسل، ولتقسيم المناطق المتجاورة بشكل جيد، يُنتج فرع تجزئة المثيل خريطة نقاط ناعمة، تشبه الجبل، وخريطة اتجاه المركز؛

لتمثيل كائنات المحتوى، تم تدريب الطريقة، واختبارها على مجموعة بيانات من الصحف الأوروبية التاريخية، ذات تصميم محتوى معقد، أظهرت التجربة؛ أن الطريقة المقترحة، تحقق نتائج تنافسية، مقارنةً بأساليب تحليل التصميم الشائعة، كما برهنت الدراسة، على فعالية هذه الأساليب وتفوقها، مقارنةً بالأساليب السابقة.

ثانيًا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المقروءة (النصوص) في الصحف التاريخية وتحليلها :

تشير دراسة Keerthana Murugaraj et all (2025)[21]، إلى أن تحليل الخطاب التاريخي في أرشيفات الصحف واسعة النطاق، يتطلب أساليبَ قابلةً للتوسع والتفسير؛ للكشف عن المواضيع الخفية، تُقيّم هذه الدراسة- بشكل منهجي- مناهج نمذجة المواضيع للمقالات الصحفية من عام ١٩٥٥ إلى عام ٢٠١٨، بمقارنة تحليل LDA الاحتمالي، وتحليل NMF لعوامل المصفوفة، والنماذج العصبية مثل Top2Vec و BERTopicعبر استراتيجيات معالجة مسبقة متنوعة، نقارن هذه الأساليب، من حيث تماسك الموضوع، وتنوعه، وقابليته للتوسع، وقابليته للتفسير، في حين أن تحليل LDA ،يُستخدم بشكل شائع في تحليل النصوص التاريخية، تُظهر نتائجنا؛ أنBERTopic ، الذي يستفيد من التضمينات السياقية، يتفوق- باستمرار- على النماذج الكلاسيكية، في جميع الجوانب المختبرة؛ مما يجعله خيارًا أكثر متانة للمجموعات النصية واسعة النطاق، بالإضافة إلى ذلك، نُسلّط الضوء على المفاضلات بين استراتيجيات المعالجة المسبقة وأداء النموذج، مع التركيز على أهمية تصميم مسار مُخصص، تُسهم هذه الرؤى؛ في تطوير مجال معالجة اللغة الطبيعية التاريخية، مُقدمةً إرشاداتٍ عملية للمؤرخين، وعلماء الاجتماع الحاسوبي، في اختيار أنجع نهج لنمذجة المواضيع لتحليل الأرشيفات الرقمية، سيكون هذا الكود متاحًا للعامة على منصة.GitHub

تشير دراسة Karin Stahel, Irenie How et all (2025)[22] ، إلى أن مجموعات الصحف التاريخية الرقمية، أصبحت متاحة بشكل متزايد، إلا أن حجمها ومحتواها المتنوع، لا يزالان يمثلان تحديات، للباحثين المهتمين بأنواع أو مواضيع محددة من المقالات، في خطوة نحو تطوير نماذج لمعالجة هذه التحديات، أنشأنا مجموعة بيانات من المقالات من البيانات المفتوحة لـ Papers Past في نيوزيلندا، مع شرحها باستخدام تسميات متعددة للأنواع، والموضوعات، ودرجات ثقة المعلق، يتماشى إطار التعليق لدينا، مع النهج المنظوري للتعلم الآلي، مع الاعتراف بالطبيعة الذاتية للمهمة، واحتضان التهجين، وعدم اليقين في الأنواع، تصف هذه الدراسة أساليب أخذ العينات، والتعليق فيها ومجموعة البيانات الناتجة عن ٧٠٣٦ مقالة من ١٠٦ صحيفة نيوزيلندية، تمتد في الفترة من ١٨٣٩ إلى ١٩٠٣، سيتم استخدام مجموعة البيانات هذه؛ لتطوير نماذج تصنيف، قابلة للتفسير، تتيح الاستكشاف الدقيق، واكتشاف المقالات في صحف Papers Past بناءً على الجوانب المشتركة للشكل والوظيفة والموضوع، سيتم في نهاية المطاف، إصدار مجموعة البيانات الكاملة، بما في ذلك التعليقات التوضيحية غير المجمعة والوثائق الداعمة، بشكل مفتوح؛ لتسهيل إجراء المزيد من الأبحاث.

قدمت دراسة (Michael McRae,2025)[23] ، مجموعة بيانات جديدة، واسعة النطاق، لمقالات تاريخية من صحف جنوب الولايات المتحدة، تمتدّ من عام ١٩٦٠ إلى عام ١٩٧٥، وتغطي وكالات أنباء متعددة: أسوشيتدبرس، ويونايتدبرس إنترناشونال، ورابطة مؤسسات الصحف، بخلاف الأعمال السابقة، التي ركّزت على محتوى الصفحة الأولى، تُوثّق هذه المجموعة، مقالات من جميع أنحاء الصحيفة، مُقدّمةً رؤيةً أوسع، حول تغطية جنوب الولايات المتحدة، في منتصف القرن العشرين، تتضمن مجموعة البيانات، نسخةً خضعت لعملية تنظيف نصوص، قائمة على نماذج اللغة الكبيرة (LLM) للحدّ من تشويش التعرف الضوئي على الحروف؛ مما يُعزّز ملاءمتها للتحليل الكمّي للنصوص، بالإضافة إلى ذلك، يتم الاحتفاظ بنسخ مُكرّرة من المقالات؛ لتمكين تحليل الاختلافات التحريرية في اللغة، والتأطير بين الصحف، يُوسّم كل مقال، حسب وكالة الأنباء،؛ مما يُسهّل إجراء دراسات مُقارنة للأنماط التحريرية بين الوكالات، يفتح هذا المورد، آفاقًا جديدة للبحث في العلوم الاجتماعية الحاسوبية، والعلوم الإنسانية الرقمية، واللغويات التاريخية، مُقدّمًا منظورًا مُفصّلًا حول كيفية نقل صحف جنوب الولايات المتحدة للأخبار الوطنية والدولية، خلال فترة تحوّل في التاريخ الأمريكي، ستُتاح مجموعة البيانات عند النشر، أو عند الطلب لأغراض البحث.

تُسهم دراسة Jonathan Bourne,(2025)[24]، في معالجة مشكلة، أنه غالبًا ما تفتقر الصحافة المؤرشفة رقميًا في القرن التاسع عشر، إلى تقنية التعرف الضوئي على الحروف(OCR) ؛ مما يُقلل من إمكانية الوصول إلى هذه الأرشيفات؛ ويجعلها غير قابلة للقراءة مجازيًا وحرفيًا، وذلك من خلال تطبيق تقنية التعرف الضوئي على الحروف (OCR) على طبعة الدوريات في القرن التاسع عشر(NCSE)، وهي مجموعة من ٨٤ ألف صفحة، من الصحف والدوريات الإنجليزية من القرن التاسع عشر، باستخدامPixtral 12B ، وهو نموذج لغة مُدرّب مسبقًا؛ لتحويل الصور إلى نص، قورنت قدرة Pixtral على التعرف الضوئي على الحروف، بأربعة أساليب أخرى للتعرف الضوئي على الحروف، محققةً معدل خطأ متوسط في الأحرف قدره ١%، أي أقل بخمس مرات، من أفضل نموذج تالٍ، تتميز مجموعة بيانات NCSE v2.0 الناتجة؛ بتحسين تحديد المقالات، وتقنية التعرف الضوئي على الحروف عالية الجودة، وتصنيف النصوص، إلى أربعة أنواع وسبعة عشر موضوعًا، تحتوي مجموعة البيانات على ١,٤ مليون مُدخل، و٣٢١ مليون كلمة، تُظهر أمثلة حالات الاستخدام تحليلًا لتشابه المواضيع، وسهولة القراءة، وتتبع الأحداث، يتوفر الإصدار ٢,٠ من NCSE مجانًا؛ لتشجيع البحث التاريخي والاجتماعي.

تستكشف دراسة Kevin Cohen et all (2025)[25]، استخدام نماذج اللغة الكبيرة (LLMs)لتحسين مجموعات البيانات، واكتشاف المفارقة في صحف أمريكا اللاتينية في القرن التاسع عشر، استُخدمت استراتيجيتان؛ لتقييم فعالية نموذجي BERT و GPT-4o في التقاط الفروق الدقيقة لطبيعة المفارقة، من خلال مهام التصنيف الثنائي ومتعدد الفئات، أولًا، نفّذت الدراسة، تحسينات على مجموعات البيانات، ركّزت على إثراء الإشارات العاطفية والسياقية؛ إلا أن تأثيرها على تحليل اللغة التاريخية كان محدودًا، أما الاستراتيجية الثانية، وهي عملية شرح شبه آلية، فقد عالجت بفعالية اختلال التوازن بين الفئات، وعززت مجموعة البيانات،

بشرح عالي الجودة، على الرغم من التحديات التي يفرضها تعقيد المفارقة، يُسهم هذا العمل؛ في تطوير تحليل المشاعر؛ من خلال إسهامين رئيسين: تقديم مجموعة بيانات إسبانية تاريخية جديدة مُصنّفة؛ لتحليل المشاعر واكتشاف المفارقة، واقتراح منهجية شرح شبه آلية، تُعدّ فيها الخبرة البشرية؛ أساسية لتحسين نتائج نماذج اللغة الكبيرة، مع تعزيزها بدمج السياقات التاريخية والثقافية، بوصفها سمات أساسية.

تركز دراسة Nikos Kontonasios et all (2025)[26]، على استخراج البيانات المنظمة من أرشيف صحفي كبير؛ باستخدام الذكاء الاصطناعي، ويدرس مشروع PortADa وصول السفن إلى موانئ برشلونة، ومرسيليا، وهافانا، وبوينس آيرس، في الفترة من عام ١٨٥٠ إلى عام ١٩١٤؛ بهدف إنشاء قاعدة بيانات مفتوحة المصدر، تحتوي على آلاف السجلات، المتعلقة بالتجارة البحرية في القرن التاسع عشر، وفي إطار هذا الجهد، يعمل مركز المعلوماتية الثقافية التابع لمعهد الدراسات الدولية والإقليمية في جامعة كريت، على تطوير آلية؛ لاستخراج البيانات من وصول السفن المسجلة في صحيفة "Le Sémaphore de Marseille" التاريخية (١٨٢٧-١٩٤٤)

يدمج خط الأنابيب تحليل تخطيط المستندات، والتعرف الضوئي على الحروف(OCR) ، وتقنيات توليد الذكاء الاصطناعي؛ لاستخراج البيانات ذات الصلة من صور الصحف الممسوحة ضوئيًا، وتخزينها بتنسيق جيد الهيكل، وقابل للقراءة آليًا لمزيد من التحليل، تُعدّ أرشيفات الصحف التاريخية؛ مصادر قيّمة للمعلومات؛ إذ لا توثّق التغيرات السياسية والاقتصادية فحسب، بل توثّق- أيضًا- التطورات الثقافية، والاجتماعية، والتكنولوجية.

ويُعدّ مشروع PortADA مشروعًا أوروبيًا (تبادل موظفي (MSCA في مجال التاريخ البحري، ويعتمد على الصحف البحرية التاريخية؛ لدراسة التجارة البحرية، والتاريخ الاقتصادي، وأنشطة الموانئ في القرنين التاسع عشر، وأوائل القرن العشرين، يركز المشروع، على وصول السفن إلى موانئ برشلونة، ومرسيليا، وهافانا، وبوينس آيرس من عام ١٨٥٠ إلى عام ١٩١٤، ويهدف المشروع؛ إلى استخراج بيانات وصول السفن، من أرشيفات الصحف المحلية، وتطوير قاعدة بيانات مفتوحة المصدر، تحتوي على آلاف السجلات، التي توثق التجارة البحرية في القرن التاسع عشر.

تشير دراسة Alan Thomas et all (2024)[27]، إلى أنه لا تزال جودة التعرف الضوئي على الحروف (OCR) الضعيفة، تُشكّل عائقًا رئيسًا، أمام باحثي العلوم الإنسانية، الذين يسعون إلى الاستفادة من المصادر الأولية الرقمية، مثل الصحف التاريخية، تعتمد الأساليب التقليدية؛ لتصحيح ما بعد التعرف الضوئي على الحروف (OCR) على نماذج التسلسل، التسلسل في مهام الترجمة الآلية العصبية، حيث تُطابق نصوص التعرف الضوئي على الحروف الخاطئة، مع نصوص مرجعية دقيقة، نُحوّل تركيزنا، نحو تكييف برامج التعلم الآلي التوليدية؛ لتتبنى نهجًا قائمًا على التوجيه؛ من خلال ضبط التعليمات في برنامج Llama 2 ومقارنته ببرنامج BART مُعدّل بدقة علىBLN600 ، وهو مجموعة موازية من مقالات الصحف البريطانية من القرن التاسع عشر، نُبيّن إمكانات النهج القائم على التوجيه في اكتشاف أخطاء التعرف الضوئي على

الحروف وتصحيحها، حتى مع محدودية بيانات التدريب، وقد حققنا تحسنًا ملحوظًا في جودة التعرف الضوئي على الحروف (OCR) مع تفوق Llama 2 علىBART ، محققًا انخفاضًا بنسبة ٥٤,٥١% في معدل أخطاء الأحرف مقابل ٢٣,٣٠% لـBART، وهذا يمهد الطريق أمام العمل المستقبلي؛ للاستفادة من نماذج الذكاء الاصطناعي التوليدية(LLMs) ؛ لتحسين إمكانية الوصول إلى النصوص التاريخية، وإطلاق العنان لإمكاناتها الكاملة؛ للبحث في العلوم الإنسانية.

تُقدّم دراسة Andrei-Marius Avram et all (2024)[28]، قاعدة بيانات HistNERo، وهي أول مجموعة بيانات رومانية، لتعرف الكيانات المسماة (NER) في الصحف التاريخية، تحتوي مجموعة البيانات، على ٣٢٣ ألف رمز نصي، تغطي أكثر من نصف القرن التاسع عشر (أي عام ١٨١٧)، وحتى أواخر القرن العشرين (أي عام ١٩٩٠)، قام ثمانية متحدثين أصليين للغة الرومانية، بشرح مجموعة البيانات بخمسة كيانات مسماة. تنتمي العينات، إلى إحدى المناطق التاريخية الأربع التالية في رومانيا، وهي بيسارابيا، ومولدوفا، وترانسيلفانيا، ووالاشيا. استخدمنا مجموعة البيانات المقترحة هذه؛ لإجراء عدة تجارب؛ لتعرف الكيانات المسماة؛ باستخدام نماذج لغوية رومانية مُدرّبة مسبقًا. تُظهر نتائجنا، أن أفضل نموذج، حقق درجة F1 صارمة، بلغت ٥٥,٦٩%، بالإضافة إلى ذلك، من خلال تقليل التناقضات بين المناطق، من خلال تقنية التكيف المجالية الجديدة، قمنا بتحسين الأداء على هذه المجموعة، إلى درجة F1 صارمة، تبلغ ٦٦,٨٠%، وهو ما يمثل مكسبًا مطلقًا، يزيد على ١٠.%

تُقدّم أطروحة Niels Bijl (2024)[29] ، طريقةً مبتكرةً لاستخراج عناوين الكتب من أرشيفات الصحف التاريخية الممسوحة ضوئيًا، بتقنية التعرف الضوئي على الحروف، باستخدام تقنية تعرف الكيانات المسماة(NER) ، وهي مهمةٌ لم تُستكشف على نطاقٍ واسع في الأدبيات الحالية، من خلال تحديد الكتب التي يُسلّط عليها النقاد والصحفيون الضوء، والتي تُمكن من فهم الأذواق الثقافية والأدبية المُتطورة للمجتمع، باستخدام مجموعة بيانات من صحيفة Leeuwarder Courant، تُطبّق الدراسة نماذجَ مُختلفة من نماذج تعرف الكيانات المسماة، بما في ذلك نموذج BiLSTM-CRF، والنماذج القائمة على المُحوّل، تفوّقت النماذج القائمة على المُحوّل على غيرها، محققةً درجة F1 بلغت ٨٤,٣% في مجموعة بيانات الاختبار، مُبيّنةً فعالية هذه النماذج في استخراج النصوص التي تُمثّل عناوين الكتب من أرشيفات الصحف، بالإضافة إلى تقييم الأداء، على مستوى تعرف الكيانات المسماة، أُجري تقييمٌ لقياس مدى قدرة أفضل نموذج لتعرف الكيانات المسماة، على تحديد الكتب المُناقشة فعليًا، وقد تحقّق ذلك؛ من خلال مُطابقة نص عنوان الكتاب، المُستخرج مع العناوين الموجودة في "المكتبة الهولندية الشاملة(NBT)"، وهي تجميعٌ شاملٌ لجميع الكتب الصادرة عن دور النشر الهولندية، على الرغم من الأداء العالي لـNER ، أسفرت عملية المطابقة عن نتيجة F1 دون المستوى الأمثل، بلغت ٥٩,٤%، ويعود هذا الفارق- في المقام الأول- إلى عدم تصنيف بيانات التدريب خصيصًا لأغراض NER؛ مما جعل إعادة توظيفها بوصفها مجموعة بيانات NER غير كافية، ونتيجةً لذلك؛ غالبًا ما أغفل النموذج الترجمات؛ مما أدى إلى عدم اكتمال استخراج العناوين، وأظهر

تحليل إضافي، أنه حتى مع تنبؤات NER المثالية؛ فإن مطابقة العناوين مع NBT ؛ حققت نتيجةF1 ، بلغت ٦٥,٥% فقط، وتُبرز هذه النتيجة؛ الحاجة إلى معلومات إضافية، إلى جانب العنوان الرئيسي، مثل الترجمات والمؤلفين، وربما الناشرين؛ لتحسين دقة مطابقة العناوين مع NBT.

تركز دراسة Yu Zhang et all (2024)[30]، على أن مجموعات الصحف التاريخية الضخمة، تحتوي على معلومات ثرية، حول التطور التاريخي للقضايا الاجتماعية، وتُشكل موردًا فريدًا لدراسة البناء الاجتماعي لقضايا مثل جنوح الأحداث، وعلى الرغم من ذلك، فإن التحليل اليدوي لملايين الصفحات من المقالات الصحفية، أمرٌ غير ممكن، هذه الدراسة تقترح مجموعة من الأساليب الحاسوبية، بما في ذلك التحليل المعجمي عبر السياقات، والتحليل الدلالي الديناميكي، وتحليل التكافؤ؛ لتسهيل دراسة البناء الاجتماعي التاريخي، ونطبق هذه الأساليب، على صحف بروكويست التاريخية، جمعنا بياناتٍ في الفترة من ١٧٩٠ إلى ٢٠٠٦ ؛ لدراسة البناء الاجتماعي لجنوح الأحداث خلال هذه الفترة، تُظهر نتائج الدراسة؛ فعالية الأساليب المقترحة، في الكشف عن رؤىً ثاقبة حول البناء الاجتماعي لجنوح الأحداث؛ مما يُسهم في فهمٍ أفضل لهذه القضية المُعقّدة، ووضع فرضياتٍ مُحددةٍ لمزيدٍ من الدراسة، وبشكلٍ عام، تُظهر هذه الدراسة؛ إمكاناتٍ واعدةً؛ للاستفادة من تقنيات معالجة اللغة الطبيعية، لتحليل بيانات الأخبار التاريخية لدراسة البناء الاجتماعي للقضايا المجتمعية.

تناقش دراسة Christian Schultze et all (2024)[31]، مسألة أنه لا يزال الكشف الصحيح عن تخطيط المقالات في صفحات الصحف التاريخية أمرًا صعبًا، ولكنه مهم لتطبيقات معالجة اللغة الطبيعية (NLP) والتعلم الآلي في مجال التاريخ الرقمي، عادةً ما توفر بوابات الصحف الرقمية نصوص التعرف الضوئي على الحروف(OCR) ، وإن كانت بجودة متفاوتة، للأسف، غالبًا ما تكون معلومات التخطيط مفقودة؛ مما يحد من نطاق هذا المصدر الغني، صُممت مجموعة البيانات الخاصة بهذه الدراسة؛ لمعالجة هذه المشكلة المتعلقة بالصحف التاريخية الصادرة باللغة الألمانية، تحتوي مجموعة بيانات Chronicling Germany على ٥٨١ صفحة صحفية تاريخية مُعلقة من الفترة الزمنية بين عامي ١٨٥٢ و١٩٢٤، وقد أمضى خبراء المجال التاريخي، أكثر من ١٥٠٠ ساعة، في شرح مجموعة البيانات، تقدم هذه الورقة البحثية خط أنابيب معالجة، وتحدد نتائج خط الأساس لبيانات الاختبار داخل النطاق وخارجه، باستخدام خط الأنابيب، كل من مجموعة البيانات الخاصة بهذه الدراسة، والرمز الأساسي المقابل، متاحان مجانًا على الإنترنت، يشكل هذا العمل؛ نقطة انطلاق لأبحاث مستقبلية في مجال التاريخ الرقمي، ومعالجة الصحف التاريخية باللغة الألمانية، علاوة على ذلك، يوفر فرصة لدراسة مهمة منخفضة الموارد في مجال الرؤية الحاسوبية.

تشير دراسة Brevin Franklin et all (2024)[32] ، إلى أنه غالبًا ما يُحلل علماء الاجتماع وعامة الناس الأحداث المعاصرة؛ من خلال استخلاص أوجه التشابه مع الماضي، وهي عملية تُعقّدها طبيعة النصوص التاريخية الواسعة، والمُشوّشة، وغير المُهيكلة، على سبيل المثال، نُسخت مئات الملايين من صفحات الصحف التاريخية بشكل مُشوّش، قد تكون الطرق التقليدية

المُتفرقة للبحث عن المواد ذات الصلة في هذه المجموعات الضخمة، مثل الكلمات المفتاحية، هشة؛ نظرًا لتعقيد المفردات وتشويش التعرف الضوئي على الحروف، تُقدّم هذه الدراسة News Déjà vu، وهي أداة بحث دلالي مُبتكرة، تستفيد من نماذج اللغة الكبيرة المُحوّلة، ونهج المُشفّر الثنائي؛ لتحديد المقالات الإخبارية التاريخية، الأكثر تشابهًا مع استعلامات الأخبار الحديثة، يتعرّف News Deja Vu أولًا على الكيانات ويُخفّيها، للتركيز على أوجه التشابه الأوسع، بدلًا من الكيانات المُسمّاة المُحدّدة، التي تتم مُناقشتها، بعد ذلك، يقوم مُشفّر ثنائي خفيف الوزن، مُدرّب على التباين، باسترجاع المقالات التاريخية الأكثر تشابهًا دلاليًا، مع استعلام حديث، مُوضّحًا كيف أن الظواهر التي قد تبدو فريدة في الوقت الحاضر، لها سوابق تاريخية مُختلفة، صُممت حزمة News Deja Vu سهلة الاستخدام، الموجهة لعلماء الاجتماع؛ لتكون في متناول من يفتقرون إلى إلمام واسع بالتعلم العميق، تعمل الحزمة مع مجموعات بيانات نصية ضخمة، وتوضح الدراسة، كيفية تطبيقها على مجموعة ضخمة من المقالات الإخبارية التاريخية مفتوحة المصدر، على الرغم من أن الخبرة البشرية، لا تزال مهمة لاستخلاص رؤى أعمق، فإن News Deja Vu توفر أداة فعّالة لاستكشاف أوجه التشابه، بين كيفية إدراك الماضي والحاضر.

تشير دراسة Filip Dobranić, et all (2024)[33] ، إلى أن إعداد مجموعات الصحف التاريخية يُعدّ مسعىً معقدًا؛ إذ يتألف من خطوات متعددة، يجب تكييفها بعناية مع المحتوى المعني، بما في ذلك التصوير، والتنبؤ بالتصميم، والتعرف الضوئي على الحروف، والتعليق اللغوي، ولتلبية التكاليف الباهظة المرتبطة بهذه العملية؛ تقدم الدراسة نهجًا بسيطًا، لإنتاج مجموعات نصوص عالية الجودة، ويتم تطبيقه على مجموعة ضخمة من الصحف التاريخية السلوفينية من القرن الثامن عشر والتاسع عشر والعشرين؛ مما ينتج عنه مجموعة نصوص ضخمة، تضم مليار كلمة، تبدأ الدراسة ببيانات مشوشة مُعالجة بالتعرف الضوئي على الحروف (OCR)باستخدام تقنيات مختلفة في فترات زمنية مختلفة من قِبل المكتبة الوطنية والجامعية في سلوفينيا، ولمعالجة التباين المتأصل في جودة البيانات النصية، وهو تحدٍّ شائع في المكتبات الرقمية عالميًا، نُجري عملية تصحيح مُستهدفة لما بعد الرقمنة، إلى جانب آلية معالجة فعّالة للنصوص المشوشة عبر استدلال نموذج اللغة، بعد ذلك، نُخضع الناتج المُصحّح والمُصفّى لتعليقات لغوية شاملة، مُثرين النص، بعلامات أجزاء الكلام، والمُعلّمات، وتسميات الكيانات المُسمّاة. وأخيرًا، تُجري الدراسة تحليلًا؛ من خلال نمذجة الموضوع على مستوى مُعلّمات الاسم، إلى جانب تحليل تردد الكيانات المُسمّاة؛ للتأكد من جدوى أسلوب الدراسة في إعداد النص.

تناقش دراسة Hermann Kroll et all (2023)[34] ، أنه غالبًا ما تُتيح المكتبات الرقمية الوصول إلى أرشيفات الصحف التاريخية، عبر البحث القائم على الكلمات المفتاحية، تُعدّ الشخصيات التاريخية وأدوارها، نقاط وصول معرفية مهمة للغاية في البحث التاريخي، ويتيح هيكلة المقالات الإخبارية وتجميعها للمستخدمين، وصولاً أكثر تطورًا؛ لاستكشاف هذه المعلومات، وعلى الرغم من ذلك، فإن القيود الواقعية، مثل نقص بيانات التدريب، وقيود

الترخيص، والنصوص غير الإنجليزية التي تحتوي على أخطاء التعرف الضوئي على الحروف؛ تجعل بناء مثل هذا النظام، صعبًا ومكلفًا عمليًا، في هذا العمل، نتناول هذه المشكلات؛ من خلال عرض المكتبة الوطنية الهولندية؛ من خلال تقديم واجهة قائمة على الأدوار، تُهيكل المقالات الإخبارية حول الشخصيات التاريخية، وقد أبرزت التقييمات المتعمقة لكل مكون، والمقابلات مع خبراء المجال، فعالية نموذجنا الأولي، وملاءمته لمجموعة مكتبات رقمية واقعية.

تشير دراسة Leen-Kiat Soh et all (2023) [35] ، إلى أنه مع رقمنة ملايين صفحات الصحف التاريخية، اختبر الباحثون في مختلف المجالات، متعة البحث عن الأسماء والأحداث من منازلهم ومكاتبهم، كما طبقوا مناهج حاسوبية؛ لدراسة مساحات شاسعة من هذه الصفحات ونصوصها الإلكترونية، على الرغم من ذلك، نؤكد أن معظم مجتمعات البحث، لم تستكشف سوى سطح الأسئلة، والأساليب التي يجب استكشافها في هذه الصفحات التاريخية، أما البحث النصي، والتصفح متعدد الأوجه، والمسح البصري للمحتوى، فلا تستكشف سوى سطح أنواع الأساليب، التي تتيحها صفحات الصحف الرقمية، بينما يسعى الباحثون إلى الإجابة عن أسئلتهم التاريخية، يهتم مختبر أبحاث المكتبات الرقمية في Aida بكل من تحديات توسيع نطاق البحث في الصحف الرقمية، وأساليب تحليل صفحاتها المبتكرة، ونركز بشكل خاص على تحليل الصور واستراتيجيات التعلم الآلي.

وعلى الرغم من ذلك، للغوص تحت سطح النص تحدياته، على سبيل المثال، تنتشر "تأثيرات الضوضاء" بشكل كبير عند إنشاء الصور الرقمية من نسخ مجهرية سابقة، كما هو شائع في مجموعات الصحف التاريخية، تُدخل تأثيرات الضوضاء تداخلًا على الإشارات الأولية للصفحات، سواءً للرؤية البشرية، أو رؤية الحاسوب والمعالجة. هناك أنواع مختلفة من تأثيرات الضوضاء شائعة، بما في ذلك عدم توزيع السطوع بالتساوي (أي تأثيرات النطاق)، وظهور الأحرف من الجانب الآخر للصفحة (التسييل)، ومسح المستندات المائل (الاتجاه المنحرف)، والعلامات على الصحيفة التي تُخفي النص (البقع). هناك نطاق واسع من الشدة، لكل من هذه التأثيرات، ويمكن أن تتراوح الصور من شديدة الوضوح، إلى شديدة الضوضاء، داخل مجموعات البيانات وعبرها.

لقد ابتكرت هذه الدراسة، ودرست تقنيات فعّالة في التعامل مع الضوضاء، بما في ذلك أساليب أقل اعتمادًا على النص الإلكتروني الدقيق، تتضمن إحدى هذه التقنيات، الاستفادة من الإشارات البصرية في الصفحات الرقمية، على سبيل المثال، يمكن تحديد المحتوى الشعري في صحيفة تاريخية؛ بناءً على الإشارات البصرية التي يقدمها، عند مقارنته بمحتوى صحفي آخر، تشمل هذه الإشارات، نهايات أسطر نصية غير مُبررة، وفجوات ملحوظة بين المقاطع، عادةً ما تكون هذه الإشارات البصرية، غائبة في معظم المحتويات غير الشعرية، وبالتالي تُوفر وسيلةً؛ لتحديد وجود المحتوى الشعري حاسوبيًا، وهو أمر مفيد بشكل خاص، عندما لا يكون التعرف الضوئي على الحروف (OCR) دقيقًا بما يكفي .

تشير دراسة[36] Melissa Dell et all (2023) ، إلى أنه لا تتعرف مجموعات بيانات النصوص الكاملة الحالية للصحف الأمريكية المتاحة للملكية العامة، على تخطيطات عمليات مسح الصحف المعقدة في كثير من الأحيان، ونتيجة لذلك؛ يقوم المحتوى الرقمي، بخلط النصوص من المقالات، والعناوين الرئيسية، والتعليقات التوضيحية، والإعلانات، ومناطق التخطيط الأخرى، كما يمكن أن تكون جودة التعرف الضوئي على الحروف منخفضة، تطور هذه الدراسة خط أنابيب مبتكرًا وعميقًا؛ لاستخراج نصوص المقالات الكاملة من صور الصحف، وتطبقه على ما يقرب من ٢٠ مليون عملية مسح ضوئي في مجموعة Chronicling America التابعة لمكتبة الكونغرس والمتاحة للملكية العامة، يتضمن خط الأنابيب، اكتشاف التخطيط، وتصنيف الوضوح، والتعرف الضوئي على الحروف المخصص، وربط نصوص المقالات التي تمتد عبر عدة مربعات محيطة؛ لتحقيق قابلية توسع عالية، تم تصميمه؛ باستخدام هياكل فعالة مصممة للهواتف المحمولة، توفر مجموعة بيانات American Stories الناتجة بيانات عالية الجودة، يمكن استخدامها؛ للتدريب المسبق لنموذج لغوي كبير؛ لتحقيق فهم أفضل للغة الإنجليزية التاريخية، ومعرفة العالم التاريخي، يمكن- أيضًا- إضافة مجموعة البيانات، إلى قاعدة البيانات الخارجية لنموذج لغوي مُعزَّز بالاسترجاع؛ وذلك لتسهيل الوصول إلى المعلومات التاريخية - بدءًا من تفسيرات الأحداث السياسية، ووصولًا إلى تفاصيل حياة أسلاف الناس - على نطاق أوسع، علاوة على ذلك، تُسهِّل نصوص المقالات المُهيكلة، استخدام أساليب قائمة على المُحوِّلات في تطبيقات العلوم الاجتماعية الشائعة، مثل تصنيف المواضيع، وكشف المحتوى المُعاد إنتاجه، وتجميع القصص الإخبارية، وأخيرًا، يُوفِّر موقع "قصص أمريكية" مجموعة بيانات ضخمة بجودة كبيرة؛ لتطوير نماذج تحليل تخطيط متعدد الوسائط، وغيرها من التطبيقات متعددة الوسائط.

تناقش دراسة [37] Marten Düring et all (2023) ، مسألة أن إعادة استخدام النصوص، كشفت عن تكرارات ذات معنى للنصوص في مجموعات كبيرة من النصوص، يستخدم باحثو العلوم الإنسانية، إعادة استخدام النصوص لدراسة- على سبيل المثال- الاستقبال اللاحق للنصوص المؤثرة أو للكشف عن ممارسات النشر المتطورة في وسائل الإعلام التاريخية، غالبًا ما يُدعم هذا البحث، بتصورات تفاعلية تُبرز العلاقات والاختلافات بين أجزاء النص، هذه الدراسة تبنى على أعمال سابقة في هذا المجال، وتقدم واجهة "إمبريسو Impresso "، لإعادة استخدام النصوص على نطاق واسع، وهي الواجهة الأولى من نوعها على حد علمنا، والتي تدمج بيانات إعادة استخدام النصوص، مع أشكال أخرى من الإثراء الدلالي؛ لتمكين استكشاف متعدد الاستخدامات، وقابل للتطوير للعلاقات بين النصوص في مجموعات الصحف التاريخية، طُوّرت واجهة إعادة استخدام النصوص على نطاق واسع" بوصفها جزءًا من مشروع "إمبريسو Impresso"، وتجمع بين عمليات بحث وتصفية فعّالة ومنظورات قراءة دقيقة وبعيدة، تدمج الدراسة، بيانات إعادة استخدام النصوص، مع عمليات إثراء مستمدة من نمذجة المواضيع، وتعرف الكيانات المسماة وتصنيفها، واكتشاف اللغة ونوع المستند، بالإضافة إلى مجموعة غنية من بيانات التعريف الصحفية، تتناول هذه الدراسة، أهداف البحث التاريخي،

ومهام المستخدم الشائعة لتحليل بيانات إعادة استخدام النصوص التاريخية، وتعرض النموذج الأولي للواجهة، مع نتائج تقييم المستخدم.

تعرض هذه الورقة البحثية، واجهة إعادة استخدام النصوص على نطاق واسع للاكتشاف، بمساعدة التصور، والتي ستُكمّل تطبيق إمبريسو، تُسهّل الواجهة الجديدة "القراءة القابلة للتوسع"؛ لإعادة استخدام النصوص في الوسائط التاريخية للمؤرخين وغيرهم من الباحثين في العلوم الإنسانية، تجمع القراءة القابلة للتوسع بين القراءة عن قرب، والقراءة عن بُعد، في حالة الصحف، تُشير القراءة عن قرب؛ إلى تفحص مجموعات إعادة استخدام النصوص الفردية، والمقاطع التي تحتويها، أو دراسة المقالات التي تنتمي إليها، وتشير القراءة عن بُعد؛ إلى توزيعات مقاييس إعادة استخدام النصوص، والبيانات الوصفية، والإثراءات الدلالية، نُوضح أهداف البحث التاريخي (الوسائط) العامة، ونُحدد قائمة بالمهام العامة؛ لاستكشاف بيانات إعادة استخدام النصوص.

يتبع هيكل هذه الورقة البحثية- بشكل عام- عملية إنشاء الواجهة التي تم اتباعها، ويُحدد القسم الثاني؛ العمل وفقًا لأحدث التطورات، فهو يُقدم أساليب وأدوات البحث، والتطوير البحثي، ويناقش التطورات الحديثة، والتحديات المتبقية في كشف واستكشاف إعادة استخدام النصوص في الصحف التاريخية، ويختتم بعرض موجز، لبيانات إعادة استخدام النصوص من إمبريسو، وبالانتقال إلى الاهتمامات البحثية التاريخية، في بيانات إعادة استخدام النصوص، يُركز القسم الثالث على خمسة أهداف بحثية، في تاريخ (وسائل الإعلام)، وإحدى عشرة مهمة عامة مرتبطة بها، تم تحديدها، ثم يعرض القسم الرابع النموذج الأولي للواجهة، وعلاقته بدراسات حالة، توضح مهامَّ محددة، ويعرض القسم الخامس، نتائج تقييم أجراه ثلاثة عشر مستخدمًا، وأخيرًا، يُختتم القسم السادس الورقة بنظرة مستقبلية على العمل.

مشروع " إمبريسو Impresso "رصد وسائل الإعلام للماضي (٢٠١٧-٢٠٢٠) ، تم اكتشاف إعادة استخدام نص داخل مجموعة من الصحف السويسرية واللوكسمبورجية، إلى جانب أشكال أخرى، من الإثراء الدلالي -على سبيل المثال- نمذجة الموضوع، والتعرف على الكيان المسمى، واكتشاف تشابه الصور، واكتشاف نوع المحتوى واللغة) يتكون مجموعها من ٧٦ صحيفة، نُشرت باللغات الفرنسية والألمانية واللوكسمبورغية، بين عامي ١٧٣٨ و٢٠١٨، وتحتوي على ما يقرب من ٥٠ مليون عنصر محتوى، تم اكتشاف ٣ أخطاء في ٥,٥ مليون صفحة، في تطبيق إمبريسو، يدعم الإصدار المؤرخين والباحثين في العلوم الإنسانية، بوظائف بحث، وتصفية، واكتشاف فعّالة؛ لاستكشاف البيانات المُثرية، يتميز هذا الإصدار بشموليته؛ إذ يدعم مجموعة واسعة من حالات الاستخدام المختلفة، ويشمل ذلك، على سبيل المثال، البحث المتقدم، والمقارنة- بمساعدة التصور- لمجموعات كبيرة من المقالات التي يُنشئها المستخدمون، وإنشاء مجموعات بيانات بحثية، لمزيد من المعالجة خارج التطبيق، بالإضافة إلى ذلك، تنشر إمبريسو، مجموعات البيانات المصاحبة في مستودعات بيانات مخصصة.

تشير أطروحة Lea Marxen,(2023)[38]، إلى أن وكالات الأنباء - منذ بداياتها في ثلاثينيات وأربعينيات القرن التاسع عشر- لعبت دورًا مهمًّا في سوق الأخبار الوطنية والدولية؛ بهدف

تقديم الأخبار بأسرع ما يمكن، وبأعلى قدر من الموثوقية، وبينما نعلم أن الصحف، تستخدم محتوى الوكالات منذ فترة طويلة لإنتاج قصصها، إلا أن مدى تأثير الوكالات على أخبارنا، غالبًا ما يظل غير واضح، وعلى الرغم من أن الباحثين، قد تناولوا هذا السؤال بالفعل، فإنه باستخدام الأساليب الحسابية – مؤخرًا- ؛ لتقييم تأثير وكالات الأنباء في الوقت الحاضر، لا تزال الدراسات واسعة النطاق حول دور وكالات الأنباء في الماضي نادرة، تهدف هذه الأطروحة؛ إلى سد هذه الفجوة؛ من خلال الكشف عن وكالات الأنباء في مجموعة كبيرة من المقالات الصحفية السويسرية واللوكسمبورغية (مجموعة المطبوعات) للأعوام ١٨٤٠-٢٠٠٠، باستخدام أساليب التعلم العميق، ولتحقيق ذلك؛ قامت الدراسة أولًا، ببناء وشرح مجموعة بيانات متعددة اللغات، مع ذكر وكالات الأنباء، والتي تستخدمها الدراسة بعد ذلك؛ لتدريب وتقييم العديد من نماذج الكشف والتصنيف، القائمة على BERT بناءً على هذه التجارب، تم اختيار نموذجين (للغتين الفرنسية والألمانية)؛ للاستدلال على مجموعة البيانات الصحفية، تُظهر النتائج، أن حوالي ١٠% من المقالات، تشير صراحةً إلى وكالات الأنباء، مع تسجيل أعلى نسبة من محتوى الوكالات بعد عام ١٩٤٠، على الرغم من أن الاستشهاد المنهجي بالوكالات، بدأ ببطء في العقد الثاني من القرن العشرين، تكشف الاختلافات في استخدام محتوى الوكالات، عبر الزمن والبلدان واللغات، وكذلك بين الصحف، عن شبكة معقدة من تدفقات الأخبار، والتي يُتيح استكشافها؛ فرصًا عديدة للأعمال المستقبلية.

تشير دراسة Elvys Linhares Pontes et all (2022)[39] ، إلى أن للمكتبات الرقمية دورًا محوريًّا في التراث الثقافي؛ إذ تُتيح الوصول إلى ثقافتنا وتاريخنا؛ من خلال فهرسة الكتب والوثائق التاريخية (الصحف والرسائل)، تستخدم المكتبات الرقمية أدوات معالجة اللغة الطبيعية (NLP)لمعالجة هذه الوثائق، وإثرائها بالمعلومات الوصفية، مثل الكيانات المُسمّاة، على الرغم من التطورات الحديثة في نماذج معالجة اللغة الطبيعية هذه، إلا أن معظمها مُصمم للغات مُحددة، ووثائق معاصرة غير مُحسّنة للتعامل مع المواد التاريخية، والتي قد تحتوي- على سبيل المثال- على اختلافات لغوية وأخطاء في التعرف الضوئي على الحروف(OCR) ، تركز هذه الدراسة، على مهمة ربط الكيانات(EL) ، التي تُعدُّ أساسيةً لفهرسة الوثائق في المكتبات الرقمية، وقامت بتطوير بنية ربط كيانات متعددة اللغات لمقالات تاريخية، تتكون من تحليل متعدد اللغات، وتصحيح التعرف الضوئي على الحروف، وتحليل مُرشّح؛ لتخفيف تأثير الوثائق التاريخية، في مهمة التعرف الضوئي على الحروف، شفرة المصدر متاحة للجمهور، أُجريت التجارب على مجموعتين من الوثائق التاريخية، تغطيان خمس لغات أوروبية (الإنجليزية، والفنلندية، والفرنسية، والألمانية، والسويدية)، أظهرت النتائج، أن نظام الدراسة، حسّن الأداء العالمي لجميع اللغات ومجموعات البيانات؛ من خلال تحقيق درجة F.

أشارت دراسة Sandeep Puthanveetil Satheesan,(2022)[40] ، إلى أنه قد وفّر توافر مجموعات الصحف الرقمية، وتوليدها للباحثين في مجالات متعددة، أداةً فعّالة لتطوير أبحاثهم، وبشكل أكثر تحديدًا، تُتيح لنا الصحف التاريخية الرقمية؛ رؤيةً شاملةً للماضي، تقترح هذه الورقة البحثية، نظامًا لتحليل البيانات الضخمة، قابلًا للتطوير والتخصيص، يُمكّن الباحثين، من

دراسة أسئلة مُعقّدة حول مجتمعنا، كما صوّرته وسائل الإعلام الإخبارية، على مدى القرون القليلة الماضية؛ وذلك من خلال تطبيق أدوات تحليل نصوص متطورة، على مجموعات صحف تاريخية ضخمة، تناقش الدراسة تجربتها، في بناء نسخة أولية من هذا النظام، بما في ذلك كيفية تعاملها مع التحديات التالية:

معالجة ملايين صفحات الصحف الرقمية، من منشورات مُختلفة حول العالم، والتي تُقدّر بمئات التيرابايت من البيانات؛ وتطبيق تجزئة المقالات، والتعرف الضوئي على الحروف (OCR) على الصحف التاريخية، والتي تختلف بين المنشورات، وداخلها بمرور الوقت؛ واسترجاع المعلومات ذات الصلة للإجابة عن أسئلة البحث من مجموعات البيانات هذه؛ من خلال تطبيق التعلم الآلي المُدمج؛ وتمكين المستخدمين، من تحليل تطور المواضيع والديناميكيات الدلالية؛ باستخدام مُشغّلات تحليل مُتوافقة ومتعددة، كما تقدم الدراسة- أيضًا- بعض النتائج الأولية؛ لاستخدام النظام المقترح لدراسة البناء الاجتماعي لجرائم الأحداث في الولايات المتحدة، ومناقشة التحديات المتبقية المهمة، التي يتعين معالجتها في المستقبل.

تُقدّم دراسة Filip Grali´ nski et all (2021)[41] ، تحديًا للتعلم الآلي، يُسمى "تحدي أمريكا(ChallAm) "، استنادًا إلى مقتطفات التعرف الضوئي على الحروف (OCR) من الصحف التاريخية التي جُمعت على بوابة .Chronicling America تُوفّر ChallAm مجموعة بيانات لمقتطفات التعرف الضوئي على الحروف، مُعلّمة ببيانات meta4 حول أصلها، ومُقرونة بمحتوياتها النصية، المُسترجعة بواسطة أداة التعرف الضوئي على الحروف، ويُحدّد هذا التحدي ثلاث مهام للتعلم الآلي: تحديد تاريخ المقالة، وتحديد موقع العدد، واستنتاج كلمة في فراغ نصي، نُشر التحدي على منصةGonito ، وهي بيئة تقييم لمهام التعلم الآلي، والتي تُقدّم قائمةً بالمتصدرين لجميع الحلول المُقدّمة، تُوفّر Gonito 9 خطوط الأساس لجميع المهام الثلاث للتحدي.

تناقش دراسة Marcel Broersma and Frank Harbers,(2021)[42] ، مسألة أن طبيعة التحليل اليدوي للمحتوى، التي تتطلب جهدًا مكثفًا، وصعوبة الوصول إلى المواد المصدرية؛ تجعل التحليلات الكمية لمحتوى الأخبار، نادرة في تاريخ الصحافة، وعلى الرغم من ذلك، فإن رقمنة أرشيفات الصحف، تتيح الآن أساليب رقمية مبتكرة؛ لإجراء أبحاث طولية منهجية، تتجاوز نطاق دراسات الحالة العرضية، تجادل الدراسة بأن التعلم الآلي المُشرف، يُقدم مناهج واعدة؛ لتحليل المواد المصدرية الوفيرة، والتحليلات الميدانية للبيانات الضخمة، ورسم خريطة للتحول الهيكلي للخطاب الصحفي طوليًا؛ من خلال التحليل التلقائي لأعراف الشكل والأسلوب، التي تعكس المعايير والممارسات المهنية الأساسية، يمكن دراسة بنية التغطية الإخبارية عن كثب، وعلى الرغم من ذلك، فإن التصنيف التلقائي لفئات الترميز الكامنة، والخاصة بفترة زمنية معينة، أمر معقد للغاية، ويزيد هيكل أرشيفات الصحف الرقمية (مثل التجزئة، والتعرف الضوئي على الحروف) من تعقيد هذا الأمر، في حين أن خوارزميات التعلم الآلي، غالبًا ما تكون بمثابة صندوق أسود، توضح هذه الورقة البحثية، كيف أن جعل عمليات التصنيف شفافة؛ يُمكّن باحثي الصحافة؛ من استخدام هذه الأساليب الحسابية، بطريقة موثوقة

وصحيحة، توضح الدراسة ذلك؛ بالتركيز على المشكلات التي واجهتها، عند تصنيف أنواع الأخبار تلقائيًا، وهي فئة ترميزية مُنيرة، وإن كانت مُعقدة للغاية، وتشير هذه الدراسة، إلى أنه يُمكن لمثل هذا النهج؛ أن يُعزز مراجعة تاريخ الصحافة، لاسيما التحول من التقارير القائمة على الرأي، إلى التقارير المُركزة على الحقائق، والتي غالبًا ما يُفترض، أنها لم تُدرس جيدًا.

يصف بحث Simon Hengchen and Nina Tahmasebi,(2021)[43]، إنشاء عدة نماذج لتضمين الكلمات، بالاعتماد على مجموعة كبيرة من مواد الصحف السويدية التاريخية، المتاحة عبر بنك اللغة السويدية "سبراكبانكن تكست"، أُنتجت هذه البيانات، في إطار مهمة "سبراكبانكن تكست" المستمرة، للتعاون مع الباحثين في العلوم الإنسانية، ومعالجة اللغة الطبيعية(NLP) ، ولتوفير موارد لغوية متاحة مجانًا، لتطوير أحدث أساليب، وأدوات معالجة اللغة الطبيعية.

تناقش دراسة Estelle Bunout (2021)[44] ، كيفية دمج أنواع مختلفة من التعليقات التوضيحية التلقائية للمقالات الصحفية الرقمية، في عملية الاستجواب التكراري للمواد المصدرية، وإنشاء مجموعات بحثية من مجموعة من النصوص غير المنظمة (المحفوظة ضمن مجموعة منظمة)، تقوم الدراسة بتعليق توضيحي، على مجموعة كبيرة من مقالات الصحافة السويسرية(١٨٣,٢٧٠ مقالة)، مستخرجة عبر واجهة إمبريسو، باستخدام نمذجة المواضيع(MALLET)، بالإضافة إلى مُصنّف بايز بسيط.

لمناقشة المنهجية التي تقترحها الدراسة؛ هي استكشاف كيف يمكن لاستخراج النصوص؛ أن يساعد في تحديد الخطابات التاريخية، التي يصعب البحث عنها؛ باستخدام الكلمات المفتاحية؛ نظرًا لغموضها المتأصل، وكيفية استيعابها في مجموعة كبيرة من النصوص، تجادل الدراسة، بأن التعليقات التوضيحية الآلية، يمكن أن توفر مجموعة من الأدلة الداعمة للخطاب الذي نبحث عنه، والتي تُستخدم بوصفها خطوة وسيطة ومنهجية في التحليل.

تناقش دراسة Ahmed Hamdi et all (2021)[45] ، مسألة، أنه يتزايد استخدام معالجة الكيانات المسماة على النصوص التاريخية؛ نظرًا للكم الكبير من الوثائق والمحفوظات المخزنة في المكتبات الرقمية، وعلى الرغم من ذلك، نظرًا لضعف الموارد التاريخية المُعلّقة؛ فإن أداء استخراج المعلومات، يتخلف عن أداء النصوص المعاصرة، وتُقدّم هذه الدراسة تطوير مورد NewsEye، وهو مجموعة بيانات متعددة اللغات؛ لتعرف الكيانات المسماة وربطها، مُثرية بالمواقف تجاهها، تتكون مجموعة البيانات، من مواد صحفية تاريخية مُتعاقبة، نُشرت بين عامي ١٨٥٠ و١٩٥٠ باللغات الفرنسية والألمانية والفنلندية والسويدية، يُعدّ هذا المورد التاريخي، ضروريًا في سياق تطوير وتقييم أنظمة معالجة الكيانات المسماة؛ فهو يسمح- بالتساوي- بتحسين أداء الأساليب الحالية على الوثائق التاريخية؛ مما يُمكّن من فهرسة دلالية كافية، وفعالة للوثائق التاريخية في مجموعات التراث الثقافي الرقمي.

تشير دراسة Houssem Menhour et all (2021)[46] ، إلى أن الصحيفة ظهرت بوصفها شكلًا ثقافيًا متميزًا في أوروبا في أوائل القرن السابع عشر، وهي مرتبطة ارتباطًا وثيقًا بالفترة التاريخية الحديثة المبكرة، تُعد الصحف التاريخية، ذات أهمية بالغة للأمم وشعوبها، ويعتمد

عليها الباحثون من مختلف التخصصات؛ لتعزيز فهمنا للماضي، وسعيًا لتلبية هذه الحاجة، يوفر المكتب الرئيسي لمكتبات جامعة إسطنبول والتوثيق، إمكانية الوصول إلى قاعدة بيانات ضخمة من الصحف التاريخية الممسوحة ضوئيًا، ولزيادة إمكانية الوصول إلى الوثائق؛ نحتاج إلى تشغيل مهام التعرف الضوئي على الحروف (OCR) وتعرف الكيانات المسماة (NER) على قاعدة البيانات بأكملها، وفهرسة النتائج؛ لتمكين آلية البحث عن النص الكامل، نقوم بتصميم وتنفيذ نظام شامل لجميع مراحل العمل، بدءًا من حذف مجموعة البيانات من الموقع الأصلي، وصولًا إلى توفير واجهة مستخدم رسومية لإجراء استعلامات البحث، وهو نظام ناجح في تحقيق ذلك، يوفر النظام المقترح إمكانية البحث عن الكلمات المفتاحية، المتعلقة بالأشخاص، والثقافة، والأمن، وتصويرها.

تناقش دراسة Molly Brandt Skelbye,(2021)[47]أن الشبكات العصبية الهجينة العميقة CNN–LSTM، أثبتت قدرتها على تحسين دقة نماذج التعرف الضوئي على الحروف (OCR)للغات مختلفة، تدرس هذه الورقة البحثية، مدى تحسين هذه الشبكات، لمعدلات دقة التعرف الضوئي على الحروف في الصحف التاريخية السويدية؛ من خلال تجربة محرك التعرف الضوئي على الحروف مفتوح المصدرCalamari ، تمكنت الدراسة؛ من إثبات تفوق نماذج CNN–LSTM الهجينة العميقة المختلطة على النماذج السابقة، في مهمة تعرف الحروف في الصحف التاريخية السويدية الممتدة من عام ١٨١٨ إلى عام ١٨٤٨، وقد حققت الدراسة، متوسط معدل دقة حروف (CAR) بلغ ٩٧,٤٣%، وهي نتيجة حديثة ومتطورة، على نصوص الصحف السويدية في القرن التاسع عشر، بيانات وشفرات ونماذج الدراسة، منشورة بموجب ترخيص.CC BY

تشير دراسة Nicolas Gutehrlé and Iana Atanassova (2021)[48]، إلى أنه في السنوات الأخيرة، قادت المكتبات والأرشيفات حملات رقمنة مهمة، أتاحت الوصول إلى مجموعات ضخمة من الوثائق التاريخية، وعلى الرغم من أن هذه الوثائق، غالبًا ما تكون متاحة بصيغةXML ALTO ، فإنها تفتقر إلى معلومات حول بنيتها المنطقية، في هذه الورقة البحثية، يتم تناول مشكلة تحليل التخطيط المنطقي، المطبق على الوثائق التاريخية، وتم اقتراح طريقة، تستند إلى دراسة مجموعة بيانات؛ لتحديد القواعد التي تُعيّن تسميات منطقية لكل من كتل وأسطر النص، من وثائقXML ALTO ، تحتوي مجموعة البيانات في الدراسة، على صحف باللغة الفرنسية، نُشرت في النصف الأول من القرن العشرين، يُظهر التقييم؛ أن منهجية الدراسة، تُحقق أداءً جيدًا؛ في تحديد الأسطر الأولى من الفقرات وأسطر النص؛ حيث تبلغ قيمة F1أعلى من ٠,٩، أما بالنسبة لتحديد العناوين، فتبلغ قيمة0.64 F1 ، يمكن تطبيق هذه الطريقة، على المعالجة المسبقة لوثائقXML ALTO ، استعدادًا للمهام اللاحقة، وكذلك لشرح مجموعات البيانات واسعة النطاق، لتدريب خوارزميات التعلم الآلي والتعلم العميق.

تستكشف دراسة Till Grallert (2021)[49] ، استخدام التاريخ الرقمي في الدراسة المنهجية للصحافة الدورية في شرق البحر الأبيض المتوسط، في أواخر العهد العثماني (١٩١٨-١٩٠٦) بوصفه مجالًا خطابيًّا، ويُقّيم التحديات المنهجية والعملية للتاريخ الرقمي، المتجذرة في البنى

التحتية الاجتماعية، والتقنية للشمال العالمي، عند تطبيقه على الجنوب العالمي، ويعتمد في ذلك ، على دراسة حالة لأربع دوريات عربية من بغداد، و بيروت، والقاهرة، ودمشق، بعد توضيح الحاجة إلى بناء مجموعة مقالات، والتحديات التي يطرحها هذا الجهد، وتستكشف الدراسة، مجموعة مقالات رقمية يبلغ مجموع كلماتها حوالي ٢,٦٥ مليون كلمة؛ من خلال تحليل الشبكات الاجتماعية، ونسب التأليف الأسلوبي.

وتناقش الدراسة، تحديات ووعود التاريخ الرقمي -المفهوم على نطاق واسع، بأنه التأريخ بمساعدة مناهج البحث الحاسوبية -لدراسة أسئلة المجتمعات التاريخية، واستنادًا إلى المصادر الرقمية للجنوب العالمي، من خلال تاريخ الصحافة الدورية العربية في شرق البحر الأبيض المتوسط، في أواخر العهد العثماني .

وتستكشف الدراسة، تاريخ الصحافة الدورية العربية في المنطقة؛ من خلال الاستعارة المكانية لـ"المجال الفكري" بوصفها إشارة إلى عالم الأفكار الإنسانية في مجمله؛ حيث لا يتجلى إلا بعضها، وبالتالي يمكن تتبعه في الإنتاج الفكري الملموس، إنه يجب تجاوز الدوريات الفردية، والانخراط في دراسة منهجية للصحافة الدورية، بوصفها مجالًا خطابيًّا، وعلى نطاق واسع، من أجل فهم أفضل للتاريخ الفكري لشرق البحر الأبيض المتوسط ، في مرحلة تاريخية حاسمة، وللإنتاج الدوري نفسه، من المهم ملاحظة أن معظم التحديات، إن جوانب التاريخ الرقمي التي نوقشت هنا، لا تقتصر- بأي حال من الأحوال- على دراسة الحالة المحددة، بل تقدم الدراسة، تنويعات واضحة لموضوع، سيجد صدىً لدى جميع المؤرخين، في وقتٍ أصبح فيه سؤال: ما إذا كان شيء ما رقميًا أم لا، بلا معنى على نحوٍ متزايد، وفي وقتٍ أصبحت فيه الحوسبة مهيمنةً؛ حيث كان كل شيءٍ، يُدار دائمًا من خلال التكنولوجيا.

تشير الدراسة، إلى أن المناهج الشاملة والتركيبية، تُعاني من قيود شديدة؛ بسبب الغياب شبه التام للدراسات المنهجية حول الدوريات الفردية، حتى المناهج الأكاديمية، التي تتناول الصحافة بوصفها مصدرًا لانتزال قصصيةً إلى حد كبير، مع التركيز على مقالات الرأي، والافتتاحيات، في عينة صغيرة من الدوريات المرجعية من بيروت والقاهرة، والأسباب الرئيسية، هي حجمها، وتشتت مجموعاتها، والتي – على الرغم من ذلك- لا تزال غير معترف بها، ومن ثم، فإن ما نحتاج إليه؛ هو الشفافية بشأن الأساليب والمصادر الفعلية المتاحة، فضلًا عن النهج المبني على التجربة.

تُجادل هذه الدراسة؛ بأن النموذج الرقمي المهيمن للبنى التحتية الاجتماعية والتقنية، المبني على النصوص الإنجليزية واللاتينية، يُسهم في خلق انقسام استعماري جديد، بين وفرة المواد الثقافية الرقمية في الشمال العالمي، واختفاء أي شيء يتجاوزها تقريبًا، وهذا يُجبر الباحثين، الذين يعملون على نصوص بلغات غير غربية، أو مكتوبة بنصوص غير لاتينية، على بذل جهود كبيرة؛ لبناء مجموعات نصوص، ويُحدّ -بشدة- من نطاق تدقيق الدراسة الحاسوبي؛ لذلك، تُقدّم الدراسة مشروعًا لبناء مجموعات النصوص "الإصدارات الدورية العربية المفتوحة " (OpenarabicPE)بوصفها إطار عمل لمواجهة التحديات المُحددة للرقمنة الشاملة للدوريات

العربية؛ وذلك من خلال الجمع بين نُسخ عدد صغير من دوريات أوائل القرن العشرين من المكتبات، مع نسخ رقمية من جهات مُختلفة؛ بغرض التحقق من صحة الأولى .

تجدر الإشارة إلى أنه، كما هو الحال في أية خطوة أخرى، ضمن عملية هذا المشروع البحثي، يواجه الباحثون من المجتمعات غير الغربية عيبًا كبيرًا، يُعدّ التعرف الآلي على الكيانات المسماة (NER)راسخًا للغات الغربية، ولكن لا يوجد- حاليًا- برنامج متاح بسهولة يدعم اللغة العربية، وتخضع ملفات التفويض المتاحة للجميع، مثل ملف التفويض الدولي الافتراضي (VIAF) الذي يجمع عشرات ملفات التفويض الوطنية، والتي تعتمد- بدورها- على بنى تحتية باهظة الثمن، لنفس الشروط.

تشير دراسة Ghaith Dekhili and Fatiha Sadat (2020)[50] ، إلى أن النماذج القائمة على الشبكات العصبية، أثبتت كفاءتها في التعرف على الكيانات المسماة، وهي إحدى مهام معالجة اللغة الطبيعية (NLP) المعروفة، علاوة على ذلك، أصبحت آلية الانتباه، جزءًا لا يتجزأ من نمذجة التسلسلات، ونماذج التحويل الفعّالة في مهام مختلفة، تتيح هذه التقنية، تمثيل السياق، في تسلسل مع مراعاة الكلمات المجاورة، تقترح هذه الدراسة، بنية تتضمن طبقات BiLSTM مدمجة مع طبقة CRF وطبقة انتباه بينهما، وقد عُززت هذه البنية بتضمينات كلمات سياقية، مُدرّبة مسبقًا، وطبقات إسقاط، علاوة على ذلك، بالإضافة إلى استخدام تمثيلات الكلمات، تستخدم الدراسة، تمثيلات قائمة على الأحرف، مُستخرجة بواسطة طبقاتCNN ؛ لالتقاط المعلومات الصرفية والإملائية.

تُظهر تجارب الدراسة، تحسنًا في الأداء العام، يُلاحظ أن نموذجها العصبي المُنتبه، المُعزز بتضمينات الكلمات السياقية، يُعطي درجات أعلى، مقارنةً بقيمها الأساسية، على حد علمنا، لا توجد دراسة تجمع بين تطبيق آلية الانتباه، وتضمين الكلمات السياقية في الصحف الوطنية، والصحف التاريخية.

تُقدّم دراسة Maud Ehrmann et all (2020)[51] ، لمحة عامة عن الإصدار الأول من مشروع) HIPE تحديد الأشخاص، والأماكن التاريخية، والكيانات الأخرى)، وهو مشروع مشترك، رائد، مُخصّص لتقييم معالجة الكيانات المُسمّاة في الصحف التاريخية باللغات الفرنسية، والألمانية والإنجليزية، منذ إطلاقه قبل نحو عشرين عامًا، أصبح معالجة الكيانات المُسمّاة (NE) مُكوّنًا أساسيًا في أي تطبيق تقريبًا لاستخراج النصوص، وقد شهد تغييرات كبيرة، في الآونة الأخيرة، اتّسم تطوره بتوجهين رئيسين؛ اعتماد هياكل التعلم العميق، والاهتمام بالمواد النصية المُستمدة من مجموعات التراث التاريخي والثقافي، في حين أن الأول، يُتيح فرصًا جديدة، يطرح الثاني تحديات جديدة، مع مُدخلات غير مُتجانسة وتاريخية ومُشوّشة، في هذا السياق، يهدف مشروعHIPE ، الذي يُقام بوصفه جزءًا من مؤتمرCLEF 2020 ، إلى ثلاثة أهداف: تعزيز متانة المناهج الحالية المتعلقة بالمدخلات غير القياسية، وتمكين مقارنة أداء معالجة NE للنصوص التاريخية، وعلى المدى البعيد، تعزيز الفهرسة الدلالية الفعالة للوثائق التاريخية، يُعرض في هذا المشروع مهام، ومجموعات، ونتائج ١٣ فريقًا مشاركًا.

تقترح دراسة Lorella Viola and Jaap Verheul,(2020)[52]، منهجًا تجريبيًا؛ لتتبع التطور التاريخي لخطاب وسائل الإعلام؛ بوصفه وسيلة للتحقيق في بناء المعنى الجماعي، استنادًا إلى نظرية الدلالات التوزيعية، ونظرية الخطاب النقدي، تستكشف الدراسة، قيمة دمج تقنيتين مستخدمتين على نطاق واسع؛ للتحقيق في اللغة والمعنى في مجالين منفصلين؛ تضمين الكلمات العصبية (اللغويات الحاسوبية) (والنهج التاريخي للخطاب) (DHA) اللغويات التطبيقية)، بوصفه حالة استخدام، تحقق الدراسة في التغييرات التاريخية في الفضاء الدلالي للخطاب العام للهجرة في المملكة المتحدة، وتستخدم أرشيف تايمز الرقمي (TDA) من عام ١٩٠٠ إلى عام ٢٠٠٠ كمجموعة بيانات، بالنسبة للجزء الحسابي، تستخدم نماذج TDA word2vecالمتاحة للجمهور ، تم تدريب هذه النماذج؛ وفقًا لنوافذ زمنية منزلقة؛ بهدف محدد لرسم خريطة للتغيير المفاهيمي، ثم تستخدم الدراسة DHA ؛ لتثليث النتائج الناتجة عن نماذج متجه الكلمات مع البيانات الاجتماعية والتاريخية؛ لتحديد تفسيرات معقولة للتغييرات في النقاش العام؛ من خلال جلب تركيز التحليل إلى مستوى الخطاب، باستخدام هذه الطريقة، تهدف الدراسة؛ إلى تجاوز رسم خرائط المعاني المختلفة، التي تعبر عنها الكلمات المفردة، وإضافة العمق الاجتماعي التاريخي واللغوي الاجتماعي المفقود حاليًا، إلى النتائج الحسابية، تستند الدراسة؛ إلى أساس أن التغييرات الاجتماعية، ستنعكس في التغييرات في الخطاب العام، على الرغم من أن الارتباط لا يثبت السببية المباشرة، فإن الدراسة تجادل بأن الأحداث التاريخية، واللغة، والمعنى، يجب اعتبارها دورة متبادلة التعزيز؛ حيث تشكل اللغة المستخدمة؛ لوصف الأحداث معاني صريحة، والتي- بدورها- تؤدي إلى أحداث أخرى، والتي ستنعكس- مرة أخرى- في الخطاب العام.

تناقش دراسة Bernhard Liebla and Manuel Burghardt,(2020)[53]، أنه في حين حظيت الصحف التاريخية- مؤخرًا- باهتمام كبير في مجال العلوم الإنسانية الرقمية، إلا أن تحويلها إلى بيانات قابلة للقراءة آليًا، باستخدام تقنية التعرف الضوئي على الحروف (OCR) ، يطرح بعض التحديات الرئيسية، ولمواجهة هذه التحديات؛ قامت الدراسة بتطوير خط أنابيب متكامل؛ للتعرف الضوئي على الحروف (OCR) باسم "أوريغامي"، يُعد هذا الخط جزءًا من مشروع حالي لرقمنة صحيفة "برلينر بورسن تسايتونغ (BBZ) "الألمانية وتحليلها الكمي، من عام ١٨٧٢ إلى عام ١٩٣١، يُعيد خط أنابيب "أوريغامي"، استخدام مكونات التعرف الضوئي على الحروف (OCR) مفتوحة المصدر، ويقدم- علاوة على ذلك- بنية جديدة، قابلة للتكوين؛ لاكتشاف التخطيط، وتعرفًا بسيطًا على الجداول، وقصًا ثنائي المراحل، لمحوري س- ص ؛ لاكتشاف ترتيب القراءة، وتطبيقًا جديدًا قويًا لإزالة تشويه المستندات، تصف هذه الدراسة، المراحل المختلفة لسير العمل، وتناقش كيفية مواجهتها للتحديات المذكورة، التي تفرضها التقنيات التاريخية الجديدة.

تناقش دراسة Ehrmann, Maud et all (2020)[54] ، أنه بعد عقود من الرقمنة الكبيرة؛ أصبح بالإمكان- الآن- استرجاع كمية غير مسبوقة من نسخ الوثائق التاريخية، والوصول إليها؛ عبر بوابات التراث الثقافي الإلكترونية، وإذا كان هذا يمثل خطوة كبيرة إلى الأمام من حيث

الحفظ، وإمكانية الوصول؛ فإن التحدي الأساسي التالي - والوعد الحقيقي للرقمنة - هو استغلال محتويات هذه الأصول الرقمية، وبالتالي تكييف وتطوير تقنيات اللغة، المناسبة للبحث عن المعلومات واسترجاعها من "بيانات الماضي الضخمة"، وعلى الرغم من ذلك، فإن تطبيق أدوات معالجة النصوص على الوثائق التاريخية بشكل عام، والصحف التاريخية بشكل خاص؛ يطرح تحديات جديدة، ويتطلب- بشكل حاسم- موارد لغوية مناسبة، في هذا السياق، تقدم هذه الورقة، مجموعة من مجموعات بيانات الصحف التاريخية، تتكون من موارد نصية وصور، تم تنظيمها ونشرها في سياق مشروع "إمبريسو - رصد وسائل الإعلام للماضي"، مع مجموعات النصوص والمعايير والتعليقات الدلالية، ونماذج اللغة باللغات الفرنسية، والألمانية، واللوكسمبورغية، التي تغطي حوالي ٢٠٠ عام، كان هدف مجموعة موارد الطباعة؛ هو الإسهام في موارد اللغة التاريخية؛ وبالتالي تعزيز قوة الأساليب المتبعة في التعامل مع المدخلات غير القياسية، وتعزيز المعالجة الفعالة للوثائق التاريخية.

تتناول دراسة Jani Marjanen et all (2020)[55] ، القضايا المنهجية، في تحليل البيانات التاريخية، وتتضمن نمذجة موضوعية لديناميكيات الخطاب في الصحف التاريخية، وتطبق الدراسة مجموعتين من نماذج المواضيع (LDA) و ((DTMعلى مجموعة كبيرة نسبيًا من الصحف التاريخية؛ بهدف رصد ديناميكيات الخطاب وفهمها، تركز دراسة الحالة، على الصحف والدوريات المنشورة في فنلندا، بين عامي ١٨٥٤ و١٩١٧، ولكن يمكن- بسهولة- تطبيق منهج الدراسة، على أية بيانات تاريخية، يتمثل الإسهام الرئيس للدراسة؛ في؛ أولًا: إجراء مُدمج لأخذ العينات والتدريب والاستدلال، لتطبيق نماذج المواضيع، على مجموعات نصوص تاريخية ضخمة وغير متوازنة؛ ثانيًا: مناقشة الاختلافات بين نموذجي مواضيع لهذا النوع من البيانات؛ ثالثا: تحديد أهمية الموضوع لفترة زمنية؛ وبالتالي تعميم توزيع المواضيع على مستوى الخطاب؛ رابعًا: مناقشة دور التفسير الإنساني في تحليل ديناميكيات الخطاب؛ من خلال نماذج المواضيع.

ثالثًا: محور الدراسات التي تتناول تطبيقات الذكاء الاصطناعي في رقمنة العناصر المصورة في الصحف التاريخية وتحليلها :

تُقدّم دراسة Giulia Taurino et all (2025)[56] ، مخططًا لمشروع علم آثار البيانات، الذي يستخدم الذكاء الاصطناعي؛ لتنظيم مجموعات التصوير الصحفي والبحث فيها، استنادًا إلى أرشيف صور صحيفة بوسطن غلوب، في حين أن الهدف الرئيس للمشروع؛ هو تعزيز وصول الجمهور، إلى أرشيفات الأخبار، فإن الرقمنة واسعة النطاق، واستعادة المعلومات، من مجموعات التصوير الصحفي تثير تساؤلات أخلاقية، حول الملكية الفكرية، والحق في حماية الهوية، عند إتاحة السجلات عبر الإنترنت، تُقدّم الدراسة، دليلًا عمليًّا، يتناول هذه القضايا؛ من خلال الذكاء الاصطناعي، مع توفير وصول عادل إلى أرشيفات الصحافة، التي غالبًا ما تكون غير متاحة في المؤسسات الإعلامية الخاصة، يناقش الجزء الأول من الورقة البحثية، كيف يُمكن للتعلم الآلي، حل مشكلة نقص الموارد اللازمة لتحليل البيانات على البدائل الرقمية، بعد تقديم مقدمة حول استخدام التعلم الآلي؛ لتسهيل الوصول إلى المعلومات في أرشيف صور

صحيفة بوسطن غلوب، تُحدد الدراسة مهمتين حسابيتين آليتين جزئيًا؛ الأولى؛ مجموعة أدوات ذكاء اصطناعي؛ لنسخ ملاحظات الأرشيفيين، واستعادة أسماء المصورين، وتواريخ إنشائهم، والتي يُمكن لأمناء المكتبات، والأرشيفيين، استخدامها؛ لتقييم حقوق النشر على السجلات؛ الثانية؛ خط أنابيب؛ للكشف عن الوجوه وتمويهها، والذي يكشف عن المناطق التي يوجد فيها أشخاص، يمكن تعرفهم، ويسمح بإخفاء الهوية، في ظل التحديات التي تواجه أرشفة الأخبار، الناجمة عن "الكمية الرقمية" من البيانات اليتيمة، والخصخصة، وغيرها من العوائق أمام سجلات الصحافة، تستكشف هذه الدراسة، نهجًا أخلاقيًا لهيكلة البيانات في أرشيفات الأخبار؛ لإتاحتها للجمهور؛ من خلال الحفاظ على الملكية الفكرية والخصوصية.

تركز دراسة William Mocaër et all (2025)[57] ، على أنه يُعد فهم صور الصحف، مهمةً صعبةً؛ نظرًا لهياكلها الهرمية المعقدة، المُقدمة من خلال مجموعة متنوعة من التصميمات الكثيفة، تقدم هذه الدراسة، نموذجًا مبتكرًا، قائمًا على المحول، مُصممًا خصيصًا؛ لمواجهة هذه التحديات؛ من خلال نهج شامل ومتكامل، يتميز النموذج المقترح؛ بقدرته على استخراج البنية الهرمية للصحف، بما في ذلك الأقسام والمقالات، كما يُجري تحديد مواقع الكتل وتصنيفها (العنوان، الفقرة، الصورة، الجدول)، والتنبؤ بترتيب القراءة على مستويات متعددة، يوفر النموذج، تحليلًا شاملًا ومفصلًا ومتسقًا لمحتوى الصحف.

يعتمد هذا النهج، على عملية تكرارية لاستخراج المعلومات؛ من خلال تسلسل هرمي للمستويات، حيث يُعالَج كل مستوى، واحدًا تلو الآخر، ولتعزيز الكفاءة الحسابية؛ يُنفَّذ كل مستوى؛ باستخدام آلية انتباه متوازية، وبالاعتماد على النمذجة الهيكلية عالية المستوى؛ يُحقق النموذج معالجة شاملة، دون الحاجة إلى أية معالجة مسبقة، أو لاحقة إضافية؛ مما يضمن التكيف مع مجموعة واسعة من تنسيقات الصحف.

تم تدريب النموذج؛ باستخدام وثائق تركيبية، ترصد تنوع الصحف وتعقيدها في العالم الحقيقي، تُمكّن هذه الوثائق التركيبية، النموذج من تعلم تمثيلات دقيقة لتصاميم الصحف؛ مما يضمن قدرته على التعميم، عبر مجموعة واسعة من التكوينات الهيكلية، تُبرز التقييمات الأولية، إمكانات النموذج في إعادة بناء التسلسلات الهرمية للصحف بدقة، وتوفير رؤى ثاقبة حول محتواها.

تقدم هذه الطريقة حلًّا واعدًا؛ لاستخراج البنية الدقيقة في المستندات عالية البنية مثل الصحف، ولكن يمكن تطبيقها، على مجموعة أوسع من المستندات، التي تعالج الحاجة المتزايدة إلى حلول رقمية فعالة وقابلة للتطوير، تعتمد على الذكاء الاصطناعي.

تقدم دراسة Thomas Smits et all (2025)[58] ، مجموعة بيانات مكونة من ٧٢٠٨١ صورة محفورة على الخشب ، مستخرجة من مجلة أخبار لندن المصورة١٨٤٢ إلى عام١٨٩٠، في منتصف القرن التاسع عشر، أحدثت مجلة أخبار لندن المصورة، ثورة في استهلاك الأخبار؛ من خلال الجمع بين النص، والرسوم التوضيحية عالية الجودة، المحفورة على الخشب، والمنشورة على نطاق واسع، وفي حين أن الرقمنة، قد سهَّلت التحليل النصي للدوريات

التاريخية، إلا أن المحتوى المرئي، لا يزال يمثل تحديًا للاستكشاف المنهجي، ونحن نعالج هذه الفجوة؛ من خلال توفير مجموعة بيانات واسعة النطاق، من الرسوم التوضيحية للأخبار في القرن التاسع عشر، وتضميناتها متعددة الوسائط، وتضمنت منهجيات الدراسة ست خطوات:

جمع ٥٦٦٩٩ صفحة ممسوحة ضوئيًا، من مجلة أخبار لندن المصورة من أرشيف الإنترنت؛ إضافة تعليقات توضيحية إلى ٩٠٨ صفحة CLIP (استخدام النموذج المضبوط بدقة؛ لاستخراج الرسوم التوضيحية؛ تطبيق نموذجYOLOv8 ؛ لضبط نموذج الكشف عن الكائنات ؛ لتحويل تعليقات الرسوم التوضيحية، إلى نص قابل للقراءة Tesseract OCR ؛ لتوليد تضمينات متعددة الوسائط؛ استخدام برنامج-Open لاسترجاع النصوص والصور متعدد الوسائط، تتيح مجموعة البيانات الناتجة وتطبيقها؛ تحليلًا مرنًا للتمثيلات Flask آليًا؛ تطوير تطبيق العناصر المرئية للأخبار في القرن التاسع عشر، وتقترح آفاقًا جديدة، للبحث في العلوم الإنسانية الحاسوبية، وتاريخ الإعلام، والدراسات الدورية، ودراسات الثقافة البصرية ؛ من خلال نشر مجموعة البيانات، ورمز المشروع، والتضمينات، يهدف مشروع الدراسة؛ إلى تسهيل جهود مماثلة، مع مواد تاريخية أخرى؛ مما يسهم في فهم أوسع للثقافة البصرية في الوقت نفسه، تبُرز هذه الورقة البحثية أيضًا؛ قيود تفسير الصور التاريخية باستخدام نماذج الذكاء الاصطناعي الحديثة، محددةً الآثار المحتملة للتحيز والتشويه التفسيري.

تُعدُّ دراسة Adriana Rodríguez-Alfonso (2025)[59]، جزءًا من مشروع رقمنة المجلات باللغة الإسبانية، الذي أُجري في جامعة توبنغن بألمانيا، تتناول الدراسة، صور النساء في ثلاث مجلات رائدة من أواخر القرن التاسع عشر في إسبانيا" فيدا نويفا) "مدريد، ١٨٩٨-١٩٠٠)، و"لا فيدا ليتريا ريا" (مدريد، ١٨٩٩)، و"لا فيدا جالانتي" (برشلونة، ١٨٩٨-١٩٠٥)، وقد جمعت هذه المجلات، التي صُممت بوصفها منصات لنشر الفنون والآداب، رسامين، ورسامات، ومصورين، وكتابًا بارزين من الحركة الهسبانية، كما قدمت رؤى قيّمة، حول المفاهيم الرمزية والاجتماعية السائدة للأنوثة .

نظرًا للعلاقة الوثيقة بين الطب والفن، في أواخر القرن التاسع عشر، تقاطعت التصورات الاجتماعية المحيطة بالمرأة، في هذه المجلات بشكل متكرر، مع النظريات السائدة حول الصحة النفسية للمرأة، تأثرت هذه النظريات، بنماذج التدهور الأوروبية، بالإضافة إلى التعديلات المحلية التي أجراها أطباء نفسيون إسبان بارزون، من أتباع الفلسفة الوضعية.

وتشير الدراسة إلى النتائج المستخلصة من رقمنة ومعالجة وتحليل، هذه المجموعة من المجلات الثقافية الإسبانية من مطلع القرن العشرين، وباستخدام تقنيات معالجة اللغة الطبيعية (NLP)، مثل إزالة غموض معاني الكلمات، والتحليل الدلالي، تُحدد الدراسة الأطر الدلالية المختلفة، المرتبطة بالمرأة، وأدوارها المجتمعية، وتمثيلاتها في المجتمع الإسباني، في القرن التاسع عشر، وتكشف هذه التقنيات؛ عن مدى تشابك الدلالات الطبية، والسياسية، والاجتماعية في الخطاب المحيط بهوية المرأة وأدوارها في ذلك الوقت.

تُثري الدراسة- أيضًا- نقاشًا حول تطبيق الأدوات الحاسوبية؛ في تحليل الصحافة الدورية، مُسلّطةً الضوء على إمكانات وتحديات تكييف الأساليب الرقمية عبر اللغات، وخاصةً للمواد المكتوبة باللغة الإسبانية، وتتناول قضايا، مثل التكيف بين اللغات، والفروق الدقيقة، في تطبيق هذه الأساليب على النصوص التاريخية.

تؤكد دراسة Seojoon Lee et all (2024)[60] ، في مجال الدراسات الثقافية والتاريخية، يُعد جمع المحتوى القائم على الصور من البيانات الضخمة، جانبًا أساسيًا من تحليل البيانات، وعلى الرغم من ذلك، فإن هذه العملية معقدة تمامًا، مثل استخراج الموارد من مناطق شاسعة، واستجابةً لذلك؛ هناك تقدير متزايد في الأوساط الأكاديمية لـ "الرسوم الكرتونية رباعية الألواح " (FPC)بوصفها مصدرًا قيّمًا لمحتوى الصور في الصحف الرقمية للبيانات الضخمة في جمهورية كوريا، وعلى الرغم من ذلك، فإن تحديد هذه الرسوم الكرتونية رباعية الألواح، وسط اتساع أرشيفات البيانات الضخمة، يمثل رحلة شاقة، لا سيما بالنظر إلى تنسيق بيانات الصور غير المنظم - وهي مهمة تستغرق وقتًا طويلًا ومكلفة، لمعالجة هذه المشكلة، تقدم هذه الدراسة آلية حسابية جديدة؛ للكشف عن :FPC تطوير نموذجYOLOv5_FPC؛ من خلال الضبط الدقيق لنموذج التعلم العميقYou Only Look Once Version 5 (YOLOv5)، المصمم خصيصًا للكشف عن صور FPC طبّقنا نموذج YOLOv5_FPC على أرشيف مكتبة أخبار تشوسون إلبو (١٩٢٠-١٩٤٠) لاستخراج بيانات FPC تلقائيًا، مغطيًا ٤٧,٧٧٧ ملف صورة بتنسيق .JPG حدّدنا ١٠٤٠ كائنًا FPC ضمن ١٠٣٥ ملفًا، تتضمن كائنات FPC، لم يكتشفها باحثون سابقون من قبل، تقدم الدراسة وصفًا مفصلًا لمنهجيتها، والتي تشمل جمع البيانات التي تم اكتشافها، من أرشيفات صحف البيانات الضخمة، وتصنيفها، وتدريبها، وكشفها، وتوزيعها، نتائج هذه الدراسة، المتاحة الآن بوصفها مجموعة بيانات مفتوحة المصدر، في مجلة بيانات العلوم الإنسانية المفتوحة (JOHD)؛ تدعو إلى نقاشات بين باحثي العلوم الإنسانية، الذين يركزون على ثقافة وتاريخ كوريا بين عامي ١٩٢٠ و١٩٤٠.

تستكشف دراسة Precup A (2024)[61] ، كفاءة أدوات الذكاء الاصطناعي، في إنتاج الفيلم الوثائقي "ذاكرة الطباعة"، الذي ينتجه حاليًا التلفزيون الوطني وقناة TVR Târgu Mureș هذا الإنتاج السمعي البصري، مخصص لتاريخ الصحافة الرومانية، والذي تم تقديمه إلى الوقت الحاضر، بتنسيق تلفزيوني لمراجعة الصحف اليومية، يعيد فيلم "ذاكرة الطباعة"، الواقع والأجواء، من فترة ما بين الحربين في رومانيا، باستخدام مقالات من صحف ذلك الوقت، ومقاطع فيديو قديمة، ورسوم توضيحية صوتية (من الأرشيف الوطني للأفلام، ومجموعات الصور القديمة، وقواعد بيانات الأفلام الوثائقية، والدراما)، في هذا السياق، توفر أدوات الذكاء الاصطناعي؛ إمكانية تنويع الرسوم التوضيحية القديمة؛ من خلال تحويل صور الأرشيف، إلى تسلسلات فيديو قصيرة، وهي قيمة للغاية؛ نظرًا للكمية المحدودة العامة لتسجيلات الفيديو المتاحة من القرن العشرين، تشير الدراسة؛ إلى حقيقة أنه في حالة مثل هذه الأفلام الوثائقية التلفزيونية التاريخية؛ لا تزال أدوات الذكاء الاصطناعي، في بداية تطوير التكنولوجيا، اللازمة لتوليد محتوى فيديو "قديم"، خاص بالنصف الأول من القرن العشرين، وتثبت دراسة الحالة؛ أنه

في هذه اللحظة؛ يمكن اعتبار النتائج تسلسلات خيالية افتراضية أو محاكاة فيديو تكنولوجية للحظة من الماضي، بناءً على دليل واحد على تلك اللحظة بالذات، وهو لقطة فوتوغرافية محفوظة في الأرشيف.

توثّق دراسة Marna Giardinetti et all (2024)[62]، مجموعة بياناتEyCon، التي تضم ما يقرب من ١٣٠,٠٠٠ صورة وصفحة بتنسيقJPEG، النزاعات المسلحة من تسعينيات القرن التاسع عشر إلى عام ١٩١٨، مع التركيز على السياقات خارج أوروبا، جمع فريق المشروع آلاف الصور الرقمية والبيانات الوصفية من مؤسسات مختلفة، بما في ذلك وثائق لم تكن متاحة سابقًا، ولتحسين البيانات الوصفية؛ أجرى الفريق تحليلات تشابه بصرية ومتعددة الوسائط، بالإضافة إلى الكشف عن البشر والحيوانات، عولجت التعليقات التوضيحية؛ لاستخراج كيانات مسماة لبيانات وصفية بتنسيق XML، وعولجت التحديات في تحديد ونشر الصور الرسومية، الناتجة عن قيود الأدوات الآلية في الكشف عن العنف، بالاستعانة بالخبرة البشرية؛ لضمان دقة التصنيف، وتواجه مجموعة البيانات، المتاحة عبر الإنترنت وعلى Zenodo للتنزيل، وإعادة الاستخدام؛ مشكلات في الرؤية الحاسوبية للصور التراثية، مثل التدهور الناتج عن البهتان، وتغير اللون، والخدوش، والضوضاء؛ مما يُضعف الخوارزميات المعتمدة على السمات البصرية. ويُؤدي نقص تمثيل الثقافات الفوتوغرافية المبكرة في مجموعات البيانات؛ إلى تحيز في تطبيق الحلول القياسية على المواد الأرشيفية.

كما تم رقمنة نسخ من عام ١٩١٥ إلى عام ١٩١٧ من الصحيفة الفرنسية المصورة Sur le Vif، والتي وثقت الحرب العالمية الأولى، مع ثروة نادرة من الرسوم التوضيحية الفوتوغرافية، كما أنشأ الفريق، بنية بيانات، تحتوي على العديد من المجموعات الفرعية، القائمة على التنسيق؛ لتشمل مجموعة متنوعة من المواد الفوتوغرافية المنسقة (المطبوعات الفردية، والألبومات الفوتوغرافية، والمجموعات الفوتوغرافية والدوريات والمجلات) التي تم جمعها في إطار هذا المشروع.

تشير دراسة Pack, Chulwoo, (2023)[63] ، إلى أن أحد أهداف تجزئة صور المستندات؛ هو تحليل صورة مستند رقمية، إلى مجموعة من المناطق المتجانسة؛ عن طريق تمييز المناطق النصية عن المناطق غير النصية، في حين تمت معالجة تجزئة الصفحات، على التخطيطات المقيدة، والصور النظيفة بنجاح في الماضي، إلا أن التجزئة على التخطيطات غير المقيدة والصور المشوشة، كما هو الحال في الصحف التاريخية، لا تزال مشكلة مفتوحة؛ بناءً على العوامل التالية؛ أولاً، استخدام القواعد الاستدلالية، مع تقنيات معالجة الصور التقليدية، أقل من الأمثل، لتغطية اختلافات التخطيطات، وجودة الصورة الموجودة في الصحف التاريخية، ثانيًا، يتطلب أداء التجزئة القوي للشبكة العصبية التلافيفية العميقة (DCNN)، عادةً قدرًا عظيمًا من الحقائق الأساسية المنسقة بدقة، وهي باهظة التكلفة، ثالثًا، تتطلب شبكات (DCNN) ، عادةً تصغير الحجم؛ لمعالجة صور الصحف التاريخية الكبيرة؛ لتناسب ذاكرة وحدة معالجة الرسومات؛ مما يؤدي- عادةً- إلى فقدان الدقة، في ظل هذه العوامل الصعبة؛ تسعى الدراسة؛ إلى تحسين دقة وكفاءة العملية الرقمية؛ لتحديد مجموعة من المناطق النصية، من صور صحف

تاريخية كبيرة وغير مقيدة، مليئة بالتشويش، أولًا، تبحث الدراسة، في جدوى استخدام التمثيل البصري التقليدي، القائم على السمات الهندسية؛ لتقسيم الصحف التاريخية، سواءً باستخدام شبكة DCNN أو بدونها، ثانيًا، تبحث الدراسة؛ في إمكانية توليد بيانات شبه أرضية فعالة؛ باستخدام نماذج تدهور الوثائق؛ وذلك لتلبية الحاجة إلى وضع علامات باهظة الثمن على مجموعات البيانات، ثالثًا، تبحث الدراسة؛ في إمكانية تصغير حجم الصور الكبيرة بشكل تكيفي، مع الحفاظ على الإشارات البصرية، ذات الصلة ببنية التخطيط؛ للحد من فقدان الدقة، تُسهم هذه الدراسة؛ في تجزئة صور الوثائق، وتحليلها بشكل عام، خاصةً في مجال الصحف التاريخية، على وجه التحديد، تقترح الدراسة وتقيم طرقًا جديدة؛ لاستخدام تقنيات معالجة الصور: (١) دمج النمو الإقليمي؛ استنادًا إلى ميزة Docstrum الهندسية، (٢) دمج خريطة الجاذبية التي هي تمثيل مرئي، يعتمد على فسيفساء Voronoi وDCNN، و (٣) تقليص حجم الصورة التكيفي، الذي يجمع بين نقاط القوة في الاستراتيجيات المستقلة عن المحتوى، والاستراتيجيات الواعية بالمحتوى، وتقييم جودة صورة الوثيقة؛ مما يساعد- بدوره- في توليد معلومات مكانية أكثر دقة، بكفاءة؛ من خلال طلب موارد حوسبة أقل، وتكلفة التحقق من صحة الصورة.

تعالج دراسة Eva Sánchez-Salido et all (2023)[64]، مشكلة أن التقنيات الحديثة للبحث عن المعلومات، تعتمد على تقنيات التعلم الآلي، مثل مناهج التعلم الإحصائي أو العميق، والتي تتطلب موارد حاسوبية ضخمة، بالإضافة إلى توفر مجموعات بيانات ضخمة؛ لتطوير التطبيقات التي تُسمى، في هذا المجال الفرعي الملموس من الذكاء الاصطناعي، ما يُسمى بالنماذج، في الوقت الحاضر، تُعالج إمكانية إعادة استخدام النماذج المُطورة؛ من خلال تقنيات الضبط الدقيق، ونقل البيانات. عندما تُكتب مجموعة البيانات المتاحة بلغة، أو مجال ذي موارد محدودة؛ تنخفض دقة هذه الأساليب؛ لذا من المهم البدء بالمهمة باستخدام أحدث التقنيات.

هذه المشكلة الرئيسة، التي تمت معالجتها في هذه الدراسة، والتي تأتي من اهتمام مؤرخي الفن بمجموعة رقمية قائمة على الصور من الصحف، تسمى Diario de Madrid (DM) من الصحافة الإسبانية، بين القرنين الثامن عشر والتاسع عشر، وهي متاحة مجانًا في المكتبة الوطنية الإسبانية (BNE) ، ينصب تركيزهم، على المعلومات المتعلقة بالكيانات، مثل الشخصيات التاريخية، والمواقع، وكذلك الأشياء المعروضة للبيع، أو المفقودة، وغيرها؛ للحصول على تصورات تحديد الموقع الجغرافي، وحل بعض الألغاز التاريخية، تتمثل الخطوة الأولى المطلوبة من الناحية الفنية؛ في الحصول على نسخ من الصحف الرقمية الأصلية، من مجموعة DM (1788-1825) بعد ذلك، تتمثل الخطوة الثانية؛ في تطوير نموذج تعرف الكيانات المسماة (NER)؛ لتصنيف أو شرح المجموعة المتاحة تلقائيًا، بالكيانات ذات الأهمية لأبحاثهم، لتحقيق ذلك، بمجرد إنشاء مجموعة نصوصCLARA-DM؛ يجب شرح مجموعة نصوص فرعية يدويًا، لخطوة التدريب باستخدام تقنيات معالجة اللغة الطبيعية (NLP) الحالية، باستخدام جهد بشري، مدعوم بأدوات حاسوبية مختارة، لتطوير نموذج الشرح اللازم (CLARA-AM)، تُجرى تجربة؛ باستخدام أحدث نماذج التعلم العميق (DL) ومجموعة نصوص متاحة، تُكمل المجموعة التي تم تطويرها في الدراسة.

إن من أهم إسهامات هذه الورقة البحثية المنهجية، التي طُوّرت لمعالجة مشكلات مشابهة لتلك المتعلقة بمجموعات مؤرخي الفن الرقمية، اختيار أدوات محددة عند توفرها، وإعادة استخدام نماذج التعلم العميق المُطوّرة؛ لإجراء تجارب جديدة في مجموعة متاحة، وإعادة إنتاج التجارب في مجموعة مؤرخي الفن أنفسهم، وتطبيق تقنيات التعلم الانتقالي، ضمن مجال ذي موارد محدودة، وُصفت أربعة موارد مختلفة مُطوّرة: المجموعة المُنسوخة، ونموذج النسخ القائم على التعلم العميق، ومجموعة النصوص المُعلّق عليها، ونماذج التعلم العميق المُطوّرة؛ للتعليق باستخدام مجموعة مُحدّدة من العلامات القائمة على المجال في مجموعة صغيرة، يُمكن الوصول إلى نموذج النسخ CLARA-TM المُتعلّم لـ DM ابتداءً من يناير ٢٠٢٣ على موقع READ-COOPالإلكتروني، تحت عنوان "المطبوعات الإسبانية نموذج ذكاء اصطناعي عام مجاني"؛ لتعرف النصوص باستخدام.XVIII-XIX

في دراسة [65]Arnold, Taylor, et all (2023)، تم تطبيق المشاهدة عن بعد، على مجموعة من ١٧٨٩٧ قصة مصورة من سلسلة Peanuts لتشارلز شولتز ، بوصفها دراسة حالة أساسية، تستخدم المشاهدة عن بعد تقنيات حسابية؛ لدراسة الوسائط المرئية واسعة النطاق، وتعتمد على مجالات متعددة التخصصات، بما في ذلك دراسات الوسائط المرئية، والدراسات الثقافية، وعلوم البيانات، وعلم العلامات، تركز الدراسة على القصص المصورة، وخاصة Peanuts؛ نظرًا لقراءتها الواسعة النطاق، وقيمتها التاريخية والثقافية، وتعقيدها؛ بوصفها وسيلة مبنية على التفاعل بين النص والصورة والمعنى، أولاً، تناقش الدراسة العمل السابق، الذي تم إجراؤه عند تقاطعات دراسات القصص المصورة، والرؤية الحاسوبية، بعد ذلك، تحدد الدراسة عمليات تطبيق الرؤية الحاسوبية على القصص المصورة، بعد ذلك، تقدم العديد من الأمثلة، بما في ذلك: اكتشاف اللوحة (الاختلافات في طول اللوحة على مدار مسيرة رسام الكاريكاتير)؛ اكتشاف التسمية التوضيحية (تحديد وموقع التسميات التوضيحية في اللوحات)؛ والنص الموازي للقصص المصورة (تحليلات الرؤية الحاسوبية/استبعاد نص حقوق النشر والتوقيعات والتاريخ وما إلى ذلك)، تكشف الدراسات المُدمجة لاكتشاف اللوحات، وكشف التعليقات، والنصوص المُصاحبة للقصص المصورة؛ عن رؤى جديدة، حول نجاح وديمومة وتأثير إحدى أشهر القصص المصورة الصحفية في العالم، وفي النهاية، تكشف الرؤية الحاسوبية عن ثباتٍ وتناسقٍ دقيقين في فن شولتز، وهو ما لعب دورًا مهمًا، وإن لم يُدرَس بالقدر الكافي، في شعبية القصص المصورة.

تشير دراسة [66]Dilawar Ali et all (2022)، إلى أن الصحف، تحتوي على ثروة من المعلومات التاريخية، على شكل مقالات ورسوم توضيحية، وقد دأبت المكتبات ومؤسسات التراث الثقافي؛ على رقمنة مجموعاتها لعقود؛ لتمكين الوصول إلى المعلومات، واسترجاعها عبر الإنترنت، ويواجه التعامل مع المجموعات الرقمية، مثل (مجموعات مكتبة بروكسل الملكية) (KBR) المستخدمة في هذه الدراسة، تحديات عديدة؛ إذ لا تحتوي إلا على بيانات وصفية على مستوى الصفحة؛ مما يُصعّب استخراج المعلومات من سياقات محددة، ويعتمد البحث الواعي للسياق- بشكل كبير- على تحسين البيانات الوصفية؛ لذلك، عند استخدام البيانات

الوصفية على مستوى الصفحة؛ يصبح تحديد المواقع الجغرافية للمعالم الأقل شهرة، أكثر صعوبة، وللتغلب على هذا التحدي؛ قامت الدراسة، بتطوير مسار لتحديد المواقع الجغرافية، وتصوير الصور التاريخية، وتتمثل الخطوة الأولى؛ في تحويل البيانات الوصفية على مستوى الصفحة؛ إلى بيانات وصفية على مستوى المقالة، وفي الخطوة التالية؛ تم تصنيف جميع المقالات، التي تحتوي على صور المباني؛ بناءً على خوارزميات تصنيف الصور، علاوة على ذلك، ولتحديد المواقع الجغرافية للصور التاريخية بشكل صحيح؛ تقترح الدراسة نهجًا هجينًا، يستخدم كلًّا من البيانات الوصفية النصية وخصائص الصور، تختتم هذه الورقة البحثية، بتناول تحدي تصوير المحتوى التاريخي، بطريقة تُضيف قيمةً لبحوث العلوم الإنسانية، تجدر الإشارة إلى أن عددًا من المشاهد الحضرية التاريخية، تُصوَّر باستخدام إعادة التصوير، وهو أمرٌ يصعب إتقانه، تُمثل هذه الدراسة، خطوةً مهمةً، نحو إثراء البيانات الوصفية التاريخية، وتسهيل ربط المجموعات المتقاطعة، وتحديد المواقع الجغرافية، وتصوير صور الصحف التاريخية، علاوةً على ذلك، تتميز المنهجية المقترحة، بطابعها العام، ويمكن استخدامها؛ لمعالجة الصور غير المُعلَّمة، من مواقع التواصل الاجتماعي، بما في ذلك فليكر وإنستغرام .

تشير دراسة Benjamin Charles Germain Lee (2022)[67] ، إلى أن صحف "لا فارا" و "إل تيمبو" و "لا بوز دي أورينتي"، تُمثل ثلاثًا من أبرز صحف باللغة اللادينو التاريخية المنشورة في أرجاء العالم اليهودي السفارادي، في الشتات، خلال القرن العشرين، وبينما حظي تاريخ وثقافة اليهود السفارادييين، باهتمام أكاديمي متزايد في السنوات الأخيرة، فإن الكم الكبير من صحف اللادينو، لا يزال غير مستغل إلى حد كبير، ولا يزال هذا المجال هامشيًا، من منظور الدراسات اليهودية، تُطبّق الدراسة تحليلًا حاسوبيًا للمحتوى المرئي؛ لاستكشاف صحافة لغة اللادينو على المستوى الكلي، باستخدام نموذج تعلّم آلي، طوّرته الدراسة لمشروعها "مستكشف الصحف"، قامتُ الدراسة؛ ببناء مجموعة بيانات من الصور المستخرجة، والرسوم التوضيحية، والخرائط، والقصص المصورة، والرسوم الكاريكاتورية التحريرية، والإعلانات من أكثر من ١٥,٠٠٠ صفحة رقمية من صحف اللادينو، تُمثل هذه الطريقة، نهجًا ناشئًا في مجال أبحاث العلوم الإنسانية الرقمية مع الدوريات، وتُتيح فرصًا؛ لتسهيل الوصول والبحث في مجال الدراسات اليهودية.

تركز دراسة Melvin Wevers et all (2022)[68]، على الصور، الصور عبر التعهيد الجماعي، وتدريب النماذج الحالية وتحسينها، وبالتالي، يُمكن إضافة تسميات ذات صلة بمجموعة هذه الدراسة إلى النموذج، وهي غير موجودة في النماذج الحالية، و سُتسلطّ الدراسة، الضوء على فرص تطبيق الذكاء الاصطناعي وتحدياته، على مجموعة من الصور التاريخية، التصوير الصحفي، الرؤية الحاسوبية، اكتشاف المشهد، التاريخ الرقمي، الرقمنة هذه التصنيفات غير تاريخية، وغالبًا ما تكون غير ذات صلة بمجموعات الدراسة، وستضع الدراسة تصنيفات لمجموعات فرعية.

تركز الدراسة على مكتب فوتو برسو دي بور، الحاصل على جائزة صورة الصحافة العالمية، الذي أسسه سي س عام، ١٩٤٥ ؛ حيث إن أهمية هذه المجموعة، تكمن في حجمها؛ حيث

تتضمن حوالي ٢ مليون صورة، التقطت لحوالي ٢٥٠ ألف حدث على مدار ستين عامًا، مصحوبة ببيانات وصفية موسعة، لا تقتصر المجموعة على الأحداث الوطنية الكبرى فحسب، بل تشمل- أيضًا- مواضيع ذات اهتمام إنساني محدود النطاق، مثل صاحب المتجر القريب، لا يقتصر هدف الدراسة على رقمنة ونشر جميع الصور السلبية الـ ٢ مليون صورة لمكتب فوتو برسو دي بور فحسب، بل تهدف أيضًا؛ إلى استكشاف كيف يمكن للذكاء الاصطناعي إثراء هذه المجموعة؛ مما يعود بالنفع على مستخدمي الأرشيف، والمؤرخين الثقافيين، الذين يدرسون الصور التاريخية، يركز أحد جهودنا، على اكتشاف المشهد، و هي طريقة لاكتشاف "المشهد" المُمَثَّل في الصورة، وتعتمد الدراسة ، على التعلم الانتقالي لتكييف نماذج الرؤية الحاسوبية الحالية، مع مجموعتنا، واحتياجات المستخدمين، على الرغم من ذلك، يمكن للنماذج الحالية توليد تسميات بدقة عالية.

توضح دراسة Melvin Wevers,(2021)[69]، كيفية استخدام التعلم الانتقالي؛ لتحسين كشف المشاهد، عند تطبيقه على مجموعة صور صحفية تاريخية، بعد تطبيق التعلم الانتقالي على نموذج Places-365 ResNet-50 المُدرّب مسبقًا، حققنا دقةً من الدرجة الأولى، بلغت ٠,٦٨، ودقةً من الدرجة الخامسة بلغت ٠,٨٩ على مجموعة بياناتنا، التي تتكون من ١٣٢ فئة، بالإضافة إلى وصف استراتيجية الشرح، والتدريب في الدراسة، تتناول الدراسة- أيضًا- استخدام التعلم الانتقالي، وتقييم نماذج الرؤية الحاسوبية لمعاهد التراث.

تشير دراسة Melvin Wevers, and Thomas Smits (2020)[70] ، إلى أن أبحاث العلوم الإنسانية الرقمية، ركزت- بشكل أساسي- على تحليل النصوص، وينبع هذا التركيز؛ من توافر التكنولوجيا لدراسة النصوص الرقمية، يتيح التعرف الضوئي على الحروف للباحثين، استخدام الكلمات المفتاحية؛ للبحث عن النصوص الرقمية وتحليلها، وعلى الرغم من ذلك، تحتوي أرشيفات المصادر الرقمية- أيضًا- على أعداد كبيرة من الصور، توضح هذه الدراسة، كيفية استخدام الشبكات العصبية التلافيفية (CNNs) ؛ لتصنيف المصادر البصرية التاريخية الرقمية وتحليلها، تقدم الدراسة ثلاثة مناهج مختلفة؛ لاستخدام الشبكات العصبية التلافيفية؛ لفهم أعمق للاتجاهات البصرية، في أرشيف الصحف الهولندية الرقمية، تشمل هذه المناهج؛ الكشف عن السمات الخاصة بالوسيلة (فصل الصور عن الرسوم التوضيحية)، والاستعلام عن الصور؛ بناءً على الجوانب البصرية المجردة (تجميع الإعلانات المتشابهة بصريًا)، وتدريب شبكة عصبية؛ بناءً على فئات بصرية، طورها خبراء في هذا المجال، نجادل بأن الشبكات العصبية التلافيفية، تسمح للباحثين؛ باستكشاف الجانب البصري للتحول الرقمي، فهي تسمح لأمناء الأرشيف والباحثين؛ بتصنيف وتحديد الاتجاهات، في مجموعات كبيرة من المصادر البصرية الرقمية، بطرق جديدة كليًا، ركزت هذه الدراسة، على نوعين من الصور، التي ظهرت في الصحف الهولندية؛ الصور التي كانت جزءًا من مقالات الصحف والصور في الإعلانات، واستخرجت الدراسة الصور، من مجموعة الصحف الرقميةDelpher ، والتي تمكنت الدراسة من الوصول إليها، خلال مشاريع الباحثين في الدراسة، بوصفهم باحثين مقيمين في المكتبة الوطنية الهولندية

(KB)، بالإضافة إلى كميات كبيرة من المعلومات النصية، يحتوي الأرشيف- أيضًا- على مجموعة واسعة من المحتوى المرئي.

تحتوي مجموعة البيانات الأولى (CHRONIC) صور الصحف التاريخية المصنفة، على ٤٥٢,٥٤٣ صورة للفترة من ١٨٦٠ إلى ١٩٣٠، ولأن الدراسة كانت مهتمة بشكل رئيس بصور الأخبار؛ تم تضمين الصور المتعلقة بالمقالات الصحفية فقط؛ مما أدى إلى استبعاد صور أقسام الإعلانات، واستبعاد الصور الصغيرة نسبيًا؛ عن طريق تصفية الملفات، التي يقل حجمها عن ٣٠ كيلوبايت، تحتوي (CHRONIC) على مجموعة واسعة من المواد المرئية، مثل الرسوم التوضيحية وصور الأحداث الإخبارية، والرسوم الكاريكاتورية (السياسية)، والصور المصاحبة للكتب المصورة ، بالإضافة إلى عدد كبير من مسائل الشطرنج وتقارير الطقس.

وتتكون مجموعة البيانات الثانية (SIAMESET)، من ٤٢٦,٧٧٧ إعلانًا، نُشرت في صحيفتين هولنديتين مؤثرتين Algemeen Handelsblad و NRC Handelsblad، بين عامي ١٩٤٥ و١٩٩٥، في حين أن صور مجموعة بيانات (CHRONIC) ، منفصلة عن المحتوى النصي، تتكون إعلانات SIAMESET من محتوى مرئي ونصي، في الواقع، تعتمد العديد من الإعلانات- في الغالب- على النص، قبل أن تتمكن الدراسة من تدريب CNN ؛ للبحث عن أوجه التشابه داخل الصور، كان على الباحثين لذلك؛ إنشاء مجموعة بيانات من الإعلانات، التي تحتوي على درجة عالية من المحتوى المرئي، وتم في الدراسة، تصفية المجموعة الأولية، والتي تتكون من ١,٦ مليون إعلان، على خطوتين، أولًا، تم بإزالة الصور التي يبلغ عرضها أو ارتفاعها، أقل من ٥٠٠ بكسل والإعلانات ذات الأبعاد، التي تشبه الإعلانات المبوبة (ارتفاع >٥٠٠٠ بكسل وعرض <٩٠٠ بكسل)، ثانيًا، تم بإزالة الإعلانات التي تحتوي على نسبة أحرف أعلى من ٠,٠٠٠٥، وقد أدى هذا؛ إلى الحصول على مجموعة بيانات، تضم ٤٢٦,٧٧٧ إعلانًا للفترة ١٩٤٥-١٩٩٥.

تشير دراسة Yann Soullard et all (2020)[71]، إلى أن تحليل صور الصحف التاريخية، مهمة صعبة؛ بسبب التصميم المعقد للصحف، وتنوعها بين المجموعات، في حين أن الأساليب التقليدية، هي طرق قائمة على القواعد، مع العديد من الخطوات المتتالية، تُظهر الأعمال الحديثة، أنه يمكن استخدام أساليب التعلم العميق بنجاح؛ لتوفير تسمية بكسل للحقول المختلفة، التي تحدث في الصفحة، يسمح هذا؛ بالاستخراج التلقائي لهيكل المستند، والوصول إلى الكيانات الدلالية المختلفة، قد تنطبق التحسينات الأخيرة المقترحة؛ لتعزيز قدرات الشبكة العصبية التلافيفية، مثل الآلية المبوّبة بشكل جيد أيضًا على المهمة في النهاية، في هذا الصدد، تقترح الدراسة؛ بنية شبكة عصبية ملتوية بالكامل(FCN) ، تُخرج تسمية بكسل للكيانات الدلالية المختلفة، التي تحدث في صور الصحف التاريخية، يعتمد نموذج الدراسة، على بنية كتلة بوابة متعددة المقاييس (MSGB) جديدة، مصنوعة من اتصالات كثيفة، وآليات بوابات، تتعامل مع تحليل متعدد المقاييس للصورة المدخلة، مع الاهتمام الذاتي، تظهر التقييمات التي أجريت على أربع مجموعات بيانات صحفية تاريخية، تتضمن ما يصل إلى ١١ فئة دلالية؛ أن اقتراح الدراسة، يتفوق على هياكل FCN القياسية.

جدول رقم (١) يوضح تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل الصحف التاريخية

م	تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل الصحف التاريخية	الصحف التاريخية التى تم تطبيق التقنية عليها لرقمنتها أو تحليلها	الفترة الزمنية التاريخية التي تم التطبيق عليها.
١	أساليب الاستدلال، والرؤية الحاسوبية، وتعرف الأنماط، وغيرها من تقنيات الذكاء الاصطناعي والتعلم الآلي	مجلتان برتغاليتان تاريخيتان، هما "دياريو دي نوتيسياس" و"جورنال دي نوتيسياس"	-
٢	نماذج اللغة الكبيرة LLMs	استخراج المقالات من الصحف التاريخية متعددة اللغات، تركز دراسة الحالة في الدراسة على تغطية زلزال ميسينا الكارثي في جنوب إيطاليا.	عام ١٩٠٨
٣	نموذج وصفي دلالي تقنيات التعلم العميق المتقدمة نموذجBERT	تعزيز استرجاع الصحف التاريخية والاستفادة منها من خلال تطبيق التنظيم الدلالي لوصف محتويات الصحف واستخراجها، وربطها، وتطبيقها.	-
٤	مناهج الرؤية الحاسوبية -التعلم الآلي أساليب التصنيف الخاضعة للإشراف وغير الخاضعة للإشراف. الكيانات المسماة (NER)	مجموعات الصحف التاريخية الرقمية، مثل مجموعات المكتبة الملكية البلجيكية (KBR) استخراج أنواع محددة من المقالات (مثل: الملاحق الأدبية من مجلة "لو بيبول".	- عام ١٩٣٨
٥	نماذج كشف الأجسام Faster (RCNN)	تجزئة صور الصحف لفصل المقالات المختلفة عن الصحف البنجابية.	-
٦	بنيتا الشبكات العصبية العميقة-Faster R CNNالمستندة إلى بنية الشبكة العصبية التلافيفية (CNN)، وTransformers	باستخراج المقالات من الصحف بلغات مختلفة (مثل العربية، الإنجليزية، الفرنسية، الألمانية).	-

٧	أدوات الذكاء الاصطناعي.	مختبر KBLab في المكتبة الوطنية السويدية، اختبار أدوات الذكاء الاصطناعي للمساعدة في تنظيم أرشيف الصحف الرقمي.	
٨	نموذج لغوي واسع النطاق؛ لتصحيح أخطاء تعرف الضوئي على الحروف.	أرشيفات الصحف الرقمية الكولومبية. ٤٠٣٢ صفحة ممسوحة ضوئيًا من الصحف، معظمها من غرناطة الجديدة - وهي دولة كانت تضم كولومبيا، وبنما، وفنزويلا، والإكوادور. ٣٠٣٨ صحيفة رقمية من ٥٨ مجموعة رقمية في جميع أنحاء المكسيك، والأرجنتين، وكولومبيا، وبيرو، وتشيلي، وبنما، وفنزويلا، وأوروغواي، وبوليفيا، وكوبا، والإكوادور، وبعض البلدان، مثل بوليفيا، وكوبا، وفنزويلا ١٩٧ عنوان صحيفة و٢٣٥٢٢ صفحة من الصور الممسوحة ضوئيًا، معظمها من مدينة مكسيكو (المكسيك هي الدولة الوحيدة التي رقمنت مجموعتها بالكامل)، ولكنها تشمل أيضًا منشورات من مدن أخرى في أمريكا اللاتينية، مثل بوينس آيرس، وليما، وبوغوتا، وسانتياغو دي تشيلي.	-

٩	الشبكات العصبية العميقة DNNs التعليم الانتقالي لتجزئة المقالات من صور الصحف فصل المقالات في NewsEye	تجزئة المقالات من صور الصحف التاريخية.	-
١٠	التعلم العميق. نماذج اللغات الكبيرة. ChroniclingAmericaQA	مجموعة الصحف التاريخية Chronicling America	-
١١	منصة عرض NewsEye التي تم دمجها في مختبرات (ÖNB). أركانوم: واجهة برمجة تطبيقات لتجزئة الصحف	دراسات حالة الكوارث الطبيعية/البيئية والهجرة (العودة). مجموعة صحف المكتبة الوطنية السويسرية أرشيف الصحف الإلكترونية، صحيفة Bündner Nachrichten، مجموعة صحف ضمن Annolyzer (منصة عرض NewsEye التي تم دمجها في مختبرات ÖNB)، صحيفة Illustrierte Kronen-Zeitung، وأركانوم؛ واجهة برمجة تطبيقات لتجزئة الصحف؛ مثال "صحيفة نيوز- فيلت- بلات	الفترة الزمنية من ١٨٥٠ -١٩٥٠
١٣	نموذج YOLOv8	تصنيف الإعلانات ونعي الوفيات في صحيفة "Slovenski narod"	الصادرة في مطلع القرن العشرين
١٣	التعرف الضوئي على الحروف؛ باستخدام نماذج الكشط والتعلم العميق.	صحيفة "هونغ كونغ نيوز"	-
١٤	نهج فصل المقالات SRRAS، ذلك نموذج التخطي (sgSTRAS) ، ونموذج حقيبة الكلمات المستمرة(cbowSTRAS) ،	الصحف الفرنسية والفنلندية.	في القرنين التاسع عشر وأوائل القرن العشرين.

	ونموذج(FastText (ftSTRAS ، ونموذج SpaCy المُدرّب مسبقًا(preSTRAS)		
١٥	مورد تعليمي مفتوح OER حول البحث التاريخي الرقمي في التاريخية	الصحف التاريخية الأوروبية.	-
١٦	نظام قائم على القواعد يتم تقييمها ومقارنتها مع نموذجين من نماذج التعلم الآلي.	صحف ودوريات ومجلات فرنسية، نُشرت في منطقة فرانش كونتيه.	في النصف الأول من القرن العشرين.
١٧	NewsEye التعلم العميق، وتحليل أخطاء التعرف الضوئي على الحروف.		
١٨	استخدام تقنيات التعلم الآلي على صفحاتChronicling America	١٦ مليون صفحة صحيفة تاريخية في قاعدة بيانات Chronicling America	-
١٩	نهج متعدد الوسائط لتحليل تخطيطات الصحف التاريخية	الصحف السويسرية واللوكسمبورغية ذات التاريخ الزمني.	-
٢٠	نموذج جديد لتحليل تخطيطات الصحف التاريخية يُسمى Panoptic-DLA	الصحف الأوروبية التاريخية ذات تصميم محتوى معقد.	-

يشير الجدول إلى تعدد تقنيات الذكاء الاصطناعي المُستخدمة في رقمنة وتحليل الصحف التاريخية التي تضمنتها الدراسات موضع التحليل في الدراسة، كما يوضح الجدول تناول الدراسات موضع الدراسة تطبيقات تقنيات الذكاء الاصطناعي في رقمنة وتحليل صحف تاريخية صادرة من دول عديدة حول العالم، ومعالجة بعض الدراسات لعدد كبير من أعداد الصحف التاريخية.

جدول رقم (٢) يوضح تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل النصوص الصحفية التاريخية

م	تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل النصوص الصحفية التاريخية.	الصحف التي تم تطبيق التقنية على نصوصها التاريخية.	الفترة الزمنية التاريخية التي تم التطبيق عليها.
١	– NMF – LDA النماذج العصبية Top2Vec- لBERTopic	الخطاب التاريخي لمقالات صحفية	من عام ١٩٥٥ إلى عام ٢٠١٨
٢	نموذج مجموعة بيانات من المقالات باستخدام تسميات متعددة.	١٠٦ صحيفة نيوزيلندية	تمتد في الفترة من ١٨٣٩ إلى ١٩٠٣
٣	على نماذج اللغة الكبيرة (LLM) للحدّ من	مقالات تاريخية من صحف	تمتدّ من عام

	تشويش التعرف الضوئي على الحروف.	جنوب الولايات المتحدة.	١٩٦٠ إلى عام ١٩٧٥
٤	نموذج لغة مُدرّب مسبقًا لتحويل الصور إلى نص Pixtral 12B	٨٤ ألف صفحة من الصحف والدوريات الإنجليزية.	القرن التاسع عشر
٥	نماذج اللغة الكبيرة (LLMs) - نموذجي BERT وGPT-4o	صحف أمريكا اللاتينية.	القرن التاسع عشر
٦	خط الأنابيب لتحليل تخطيط المستندات. والتعرف الضوئي على الحروف (OCR). وتقنيات توليد الذكاء الاصطناعي لاستخراج البيانات.	مشروع PortADa وصول السفن إلى موانئ برشلونة، ومرسيليا، وهافانا، وبوينس آيرس. صحيفة " Le Sémaphore de Marseille" التاريخية	في الفترة من عام ١٨٥٠ إلى عام ١٩١٤ (١٨٢٧-١٩٤٤).
٧	تكييف برامج التعلم الآلي التوليدية وتبني نهجًا قائمًا على التوجيه؛ من خلال ضبط التعليمات في برنامج Llama 2 ومقارنته ببرنامج BART مُعدّل بدقة على BLN600	مقالات الصحف البريطانية	القرن التاسع عشر
٨	قاعدة بياناتHistNERo مجموعة بيانات رومانية لتعرف الكيانات المسماة (NER) في الصحف التاريخية.	تنتمي العينات إلى إحدى المناطق التاريخية الأربع التالية في رومانيا، وهي بيسارابيا، ومولدوفا، وترانسيلفانيا، ووالاشيا.	القرن التاسع عشر (أي عام ١٨١٧) وحتى أواخر القرن العشرين (أي عام ١٩٩٠).
٩	نماذج تعرف الكيانات المسماة. نموذج BiLSTM-CRF والنماذج القائمة على المُحوّل.	استخراج عناوين الكتب من أرشيفات الصحف التاريخية الممسوحة ضوئيًا باستخدام مجموعة بيانات من صحيفة Leeuwarder Courant	-
١٠	التحليل المعجمي عبر السياقات، والتحليل الدلالي الديناميكي، وتحليل التكافؤ.	صحف بروكويست التاريخية، لدراسة البناء الاجتماعي لجنوح الأحداث خلال هذه الفترة	الفترة من ١٧٩٠ إلى ٢٠٠٦.
١١	خط أنابيب معالجة معلومات التخطيط المفقودة.	الصحف التاريخية الصادرة باللغة الألمانية، تحتوي مجموعة بيانات	من الفترة الزمنية بين عامي ١٨٥٢ و١٩٢٤.

		Chronicling Germanyعلى ٥٨١ صفحة صحفية تاريخية مُعلقة	
١٢	أداة بحث دلالي مُبتكرة News Deja Vu تستفيد من نماذج اللغة الكبيرة المُحوّلة ونهج المُشفّر الثنائي لتحديد المقالات الإخبارية التاريخية.	مجموعة ضخمة من المقالات الإخبارية التاريخية مفتوحة المصدر	-
١٣	نهج لإنتاج مجموعات نصوص عالية الجودة.	تطبيقه على مجموعة ضخمة من الصحف التاريخية السلوفينية	القرن الثامن عشر والتاسع عشر والعشرين
١٤	تقديم واجهة قائمة على الأدوار تُهيكل المقالات الإخبارية حول الشخصيات التاريخية.	عرض المكتبة الوطنية الهولندية	-
١٥	مختبر أبحاث المكتبات الرقمية Aida	-	-
١٦	خط أنابيب مبتكر وعميق؛ لاستخراج نصوص المقالات الكاملة من صور الصحف يطبق علىChronicling America	لاستخراج نصوص المقالات الكاملة من صور الصحف الأمريكية، من ٢٠ مليون عملية مسح ضوئي في مجموعة Chronicling America التابعة لمكتبة الكونغرس والمتاحة للملكية العامة	-
١٧	واجهة امبريسو Impresso- نمذجة المواضيع	الصحف السويسرية واللوكسمبورجية يتكون مجموعها من ٧٦ صحيفة نُشرت باللغات الفرنسية والألمانية واللوكسمبورغية	بين عامي ١٧٣٨ و٢٠١٨
١٨	أساليب التعلم العميق- نماذج الكشف والتصنيف القائمة على BERT نموذجين (للغتين الفرنسية والألمانية.	المقالات الصحفية السويسرية واللوكسمبورغية (مجموعة المطبوعات) للكشف عن دور وكالات الأنباء في الماضي	للأعوام ١٨٤٠- ٢٠٠٠
١٩	نهج ربط البيانات EL - بنية ربط كيانات متعددة اللغات لمقالات تاريخية.	مجموعتين من الوثائق التاريخية تغطيان خمس	-

		لغات أوروبية (الإنجليزية، والفنلندية، والفرنسية، والألمانية، والسويدية)	
٢٠	التعلم الآلي المدمج - تحليل المواضيع والديناميكيات الدلالية باستخدام مُشغّلات تحليل مُتوافقة ومتعددة.	تطبيق أدوات تحليل نصوص متطورة على مجموعات صحف تاريخية ضخمة.	-
٢١	(ChallAm)- التعلم الآلي.	الصحف التاريخية التي جُمعت على بوابة Chronicling America	-
٢٢	التعلم الآلي المُشرف.	مشكلات تصنيف أنواع الأخبار تلقائيًا.	-
٢٣	نماذج تضمين الكلمات بالاعتماد على مجموعة كبيرة من مواد الصحف السويدية التاريخية أُنتجت هذه البيانات في إطار مهمة "سبراكبانكن تكست".	الصحف السويدية التاريخية.	-
٢٤	واجهة إمبريسو Impresso باستخدام نمذجة المواضيع (MALLET) ، مُصنّف بايز.	مجموعة كبيرة من مقالات الصحافة السويسرية (١٨٣,٢٧٠ مقالة).	-
٢٥	تطوير موردNewsEye	مواد صحفية تاريخية مُتعاقبة نُشرت باللغات الفرنسية والألمانية والفنلندية والسويدية.	بين عامي ١٨٥٠ و١٩٥٠
٢٦	التعرف على الكيانات المسماة(NER)	المكتب الرئيسي لمكتبات جامعة إسطنبول، والتوثيق للوصول إلى قاعدة بيانات ضخمة من الصحف التاريخية الممسوحة ضوئيًا.	
٢٧	الشبكات العصبية الهجينة العميقة –CNN LSTM	الصحف التاريخية السويدية.	الممتدة من عام ١٨١٨ إلى عام ١٨٤٨
٢٨	مجموعة بيانات؛ لتحديد القواعد التي تُعيّن تسميات منطقية لكل من كتل وأسطر النص من وثائق XML ALTO	صحف باللغة الفرنسية.	النصف الأول من القرن العشرين
٢٩	تحديات التأريخ بمساعدة مناهج البحث الحاسوبية.	أربع دوريات عربية من بغداد، و بيروت، والقاهرة، ودمشق.	في أواخر العهد العثماني (1918-1906)

٣٠	النماذج القائمة على الشبكات العصبية، بنية تتضمن طبقات BiLSTM مدمجة مع طبقة CRF	-	-
٣١	معالجة الكيانات المُسمّاة (NE) للنصوص التاريخية.	الصحف التاريخية باللغات الفرنسية والألمانية والإنجليزية.	
٣٢	تضمين الكلمات العصبية (اللغويات الحاسوبية) والنهج التاريخي للخطاب (DHA)، نماذجTDA word2vec	الفضاء الدلالي للخطاب العام للهجرة في المملكة المتحدة، وتستخدم أرشيف تايمز الرقمي(TDA)	من عام ١٩٠٠ إلى عام ٢٠٠٠
٣٣	خط أنابيب متكامل للتعرف الضوئي على الحروف (OCR) باسم "أوريغامي".	مشروع رقمنة صحيفة "برلينر بورسن تسايتونغ " (BBZ)الألمانية وتحليلها الكمي.	من عام ١٨٧٢ إلى عام ١٩٣١
٣٤	مشروع "إمبريسو - رصد وسائل الإعلام للماضي"، مع مجموعات النصوص والمعايير والتعليقات الدلالية ونماذج اللغة.	الإسهام في موارد اللغة الفرنسية والألمانية واللوكسمبورغية.	تغطى حوالي ٢٠٠ عام
٣٥	نماذج المواضيع (LDA) و(DTM)	رصد ديناميكيات الخطاب وفهمها، الصحف والدوريات المنشورة في فنلندا.	بين عامي ١٨٥٤ و١٩١٧

يوضح الجدول كثرة وتنوع تقنيات ونماذج الذكاء الاصطناعي المستخدمة في رقمنة وتحليل النصوص الصحفية التاريخية التي رصدتها الدراسة من عينة الدراسات موضع التحليل التي تعني بدراسة استخدام تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل النصوص الصحفية التاريخية.

جدول رقم (٣) يوضح تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل العناصر الصحفية المصورة التاريخية.

م	**تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل العناصر الصحفية المصورة التاريخية .**	**الصحف التي تم تطبيق التقنية على المواد المصورة التاريخية بها.**	**الفترة الزمنية التاريخية التي تم التطبيق عليها.**
١	التعلم الآلي، مجموعة أدوات ذكاء اصطناعي لنسخ ملاحظات الأرشيفين، واستعادة أسماء المصورين وتواريخ الصور. خط أنابيب؛ للكشف عن الوجوه، وتمويهها.	صحيفة بوسطن غلوب.	-
٢	نموذج مبتكر قائم على المحول.	-	-

٣	تطبيق نموذج YOLOv8؛ لضبط نموذج الكشف عن الكائنات لتحويل تعليقات الرسوم التوضيحية، إلى نص قابل للقراءة Tesseract OCR استخدام برنامج-Open؛ لاسترجاع النصوص والصور متعدد الوسائط، تتيح مجموعة البيانات الناتجة وتطبيقها، تحليلًا مرنًا للتمثيلات Flask آليًا.	مجلة أخبار لندن المصورة.	1842إلى عام 1890 في منتصف القرن التاسع عشر.
٤	رقمنة المجلات باللغة الإسبانية.	صور النساء في ثلاث مجلات رائدة من أواخر القرن التاسع عشر في إسبانيا "فيدا نويفا "(مدريد)، و"لا فيدا ليتريباريا" (مدريد)، و"لا فيدا جالانتي" (برشلونة).	(١٨٩٨-١٩٠٠) (١٨٩٩) (١٨٩٨-١٩٠٥)
٥	آلية حسابية جديدة للكشف عن : FPC تطوير نموذج YOLOv5_FPC، من خلال الضبط الدقيق لنموذج التعلم العميق You Only Look Once Version 5 (YOLOv5)	أرشيف مكتبة أخبار تشوسون إلبو - كوريا	(١٩٢٠-١٩٤٠)
٦	أدوات الذكاء الاصطناعي	إنتاج الفيلم الوثائقي "ذاكرة الطباعة" هذا الإنتاج السمعي البصري مخصص لتاريخ الصحافة الرومانية.	النصف الأول من القرن العشرين
٧	استخراج كيانات مسماة لبيانات وصفية بتنسيقXML، وتواجه مجموعة البيانات، المتاحة عبر الإنترنت وعلى Zenodo للتنزيل وإعادة الاستخدام، مشكلات في الرؤية الحاسوبية للصور التراثية.	النزاعات المسلحة من تسعينيات القرن التاسع عشر إلى عام ١٩١٨، مع التركيز على السياقات خارج أوروبا. الصحيفة الفرنسية المصورة Sur le Vif ، والتي وثقت الحرب العالمية الأولى.	– من عام ١٩١٥ إلى عام ١٩١٧

٨	ميزة Docstrum الهندسية، تمثيل مرئي يعتمد على فسيفساءVoronoi DCNN ، وتقليص حجم الصورة التكيفي الذي يجمع بين نقاط القوة في الاستراتيجيات المستقلة عن المحتوى والاستراتيجيات الواعية بالمحتوى.	صور الصحف التاريخية.	–
٩	تطوير نموذج التعرف على الكيانات المسماة (NER)	الصور من الصحف تسمى Diario de Madrid (DM) من الصحافة الإسبانية.	بين القرنين الثامن عشر والتاسع عشر
١٠	المشاهدة عن بعد؛ تقنيات حسابية لدراسة الوسائط المرئية واسعة النطاق. الرؤية الحاسوبية	مجموعة من ١٧٨٩٧ قصة مصورة من سلسلة *Peanuts* لتشارلز شولتز.	
١١	تطوير مسار لتحديد المواقع الجغرافية، وتصوير الصور التاريخية. نهجٌ هجينٌ، يستخدم كلاً من البيانات الوصفية النصية وخصائص الصور.	(مجموعات مكتبة بروكسل الملكية (KBR)	-
١٢	الذكاء الاصطناعي والرؤية الحاسوبية.	على مجموعة من الصور التاريخية، التصوير الصحفي، مكتب فوتو برسو دي بور.	
١٣	التعلم الانتقالي؛ لتحسين كشف المشاهد نموذج Places-365 ResNet-50 المُدرّب مسبقًا.	مجموعة صور صحفية تاريخية.	-

١٤	الشبكات العصبية التلافيفية (CNNs) لتصنيف وتحليل المصادر البصرية التاريخية الرقمية. المناهج الكشف عن السمات الخاصة بالوسيلة (فصل الصور عن الرسوم التوضيحية)، والاستعلام عن الصور؛ بناءً على الجوانب البصرية المجردة (تجميع الإعلانات المتشابهة بصريًا)، وتدريب شبكة عصبية؛ بناءً على فئات بصرية طورها خبراء في هذا المجال.	أرشيف الصحف الهولندية الرقمية مجموعة الصحف الرقمية Delpher صور الصحف التاريخية المصنفة، على ٤٥٢,٥٤٣ صورة. ٤٢٦,٧٧٧ إعلانًا نُشر في صحيفتين هولنديتين مؤثرتين هما Algemeen Handelsblad و NRC Handelsblad،	– للفترة من ١٨٦٠ إلى ١٩٣٠ – بين عامي ١٩٤٥ و١٩٩٥
١٥	أساليب التعلم العميق. بنية شبكة عصبية ملتوية بالكامل (FCN) نموذج يعتمد بنية كتلة بوابة متعددة المقاييس (MSGB) جديدة.	أربع مجموعات بيانات صحفية تاريخية.	-

يتضح من الجدول تطبيق العديد من نماذج الذكاء الاصطناعي في رقمنة وتحليل المواد المصورة في صحف تاريخية عالمية صادرة في دول مختلفة بلغات متعددة.

جدول رقم (٤) يوضح القضايا الصحفية والمواد الصحفية المصورة التي تم معالجتها في رقمنة وتحليل الصحف التاريخية وعناصرها المقروءة والمصورة بتقنيات الذكاء الاصطناعي

م	القضايا الصحفية والمواد الصحفية المصورة التي تم معالجتها في رقمنة وتحليل الصحف التاريخية بتقنيات الذكاء الاصطناعي.	معدل التكرار	
		التكرار	%
١	دراسة حالة علي زلزال ميسينا الكارثي في جنوب إيطاليا.	١	١,٤
٢	دراسات حالة الكوارث الطبيعية/البيئية.	١	١,٤
٣	دراسة التجارة البحرية، والتاريخ الاقتصادي، وأنشطة الموانئ في القرنين التاسع عشر، وأوائل القرن العشرين.	١	١,٤
٤	استخراج عناوين الكتب من أرشيفات الصحف التاريخية الممسوحة ضوئيًا، باستخدام مجموعة بيانات من صحيفة Leeuwarder Couran.	١	١,٤
٥	لدراسة البناء الاجتماعي لجنوح الأحداث الفترة من ١٧٩٠ إلى ٢٠٠٦.	١	١,٤
٦	الكشف عن دور وكالات الأنباء في الماضي.	٢	٢,٨

	إجراء دراسات مُقارنة للأنماط التحريرية بين الوكالات.		
٧	الهجرة (العودة). الفضاء الدلالي للخطاب العام للهجرة في المملكة المتحدة.	٢	٢,٨
٨	استعادة أسماء المصورين، وتواريخ الصور، وتمويه الوجوه، لحماية حقوق الملكية الفكرية والخصوصية.	١	١,٤
٩	الأطر الدلالية المختلفة، المرتبطة بالمراة، وادوارها المجتمعية، وتمثيلاتها في المجتمع الإسباني، في القرن التاسع عشر، الكشف عن مدى تشابك الدلالات الطبية، والسياسية، والاجتماعية في الخطاب المحيط بهوية المرأة وأدوارها في ذلك الوقت.	١	١,٤
١٠	تتبع التطور التاريخي لخطاب وسائل الإعلام؛ بوصفه وسيلة للتحقيق في بناء المعنى الجماعي، استنادًا إلى نظرية الدلالات التوزيعية، ونظرية الخطاب النقدي.	١	١,٤
١١	إنتاج الفيلم الوثائقي "ذاكرة الطباعة" هذا الإنتاج السمعي البصري، مخصص لتاريخ الصحافة الرومانية.	١	١,٤
١٢	النزاعات المسلحة من تسعينيات القرن التاسع عشر، إلى عام ١٩١٨، مع التركيز على السياقات خارج أوروبا.	١	١,٤
١٣	مجموعة من ١٧٨٩٧ قصة مصورة من سلسلة Peanuts لتشارلز شولتز، إحدى أشهر القصص الصحفية المصورة في العالم.	١	١,٤
١٤	تحليل حاسوبي للمحتوى المرئي؛ لاستكشاف صحافة لغة اللادينو.	١	١,٤
١٥	مجموعة من الصور التاريخية، التصوير الصحفي، مكتب فوتو برسو دي بور.	١	١,٤

يُشير الجدول إلى تعدد واختلاف القضايا التي تضمنتها عينة الدراسات موضع التحليل في الدراسة والتي تتعلق باستخدام تقنيات الذكاء الاصطناعي في رقمنة وتحليل الصحف التاريخية وعناصرها المقروءة والمصورة، وجاءت القضايا ما بين قضايا تتعلق بظواهر طبيعية، وقضايا تتعلق بالبناء الاجتماعي لجنوح الأحداث، والكشف عن دور وكالات الأنباء في الماضي والتي تكررت مرتين في الدراسات موضع التحليل في الدراسة، وجاءت بنسبة ٢,٨%، وموضوعات الهجرة بنسبة ٢,٨%، والنزاعات المسلحة، وقضية تتعلق بالأطر الدلالية المختلفة، المرتبطة بالمرأة، وأدوارها المجتمعية، وتمثيلاتها في المجتمع الإسباني وقضايا تتعلق بموضوعات تاريخية مصورة.

جدول رقم (٥) يوضح المشكلات التي تواجه رقمنة، وأرشفة، وتحليل الصحف التاريخية، وعناصرها المقرورة والمصورة، وفقا للدراسات موضع التحليل ومعدل تكرارها.

المشكلات التي تواجه رقمنة وأرشفة الصحف التاريخية وفقا للدراسات موضع التحليل											
م	المشكلة التي تواجه رقمنة وأرشفة وتحليل الصحف التاريخية	مشكلات تقنية				مشكلات بحثية					
		صحف لاتينية		صحف عربية		صحف تاريخية بشكل عام		صحف لاتينيه		صحف عربية	
		التكرار	%	التكرار	%	التكرار	%	التكرار	%	التكرار	%
١	أحد الإشكاليات التي يواجهها الباحثون عند استخدام نماذج اللغة في تحليلاتهم هو التحيز.	١	١,٤			١	١,٤	٢	٢,٨		
٢	التوافر المتزايد للمواد والأدوات الرقمية للتحليل الحاسوبي واسع النطاق؛ أدى إلى اهتمام متزايد بمناهج البيانات الضخمة في العلوم الإنسانية، والاجتماعية، وعلى الرغم من ذلك، لا يزال الدور الحيوي لتنظيم البيانات، بوصفه شرطًا أساسيًا لمثل هذه المشاريع، غير مقدّر حق قدره.					١	١,٤				
٣	تُشكّل رقمنة الصحف تحديات؛ نظرًا لتعقيد تصميماتها، وكثرة محتواها.	١٠	%١٤,١								
٤	تشويش وانخفاض جودة نصوص التعرف الضوئي على الحروف.	١٥	٢١,١								
٥	النماذج المُدرَّبة مُسبقًا، لا تزال تُمثِّل تحديًا في تحقيق أداء جيد عند تطبيقها على مصادر رقمية جديدة تمامًا، دون أي ضبط دقيق، أو في الحالات التي يكون فيها الانحراف عن قواعد تصنيف النموذج الأصلي متطلبًا صعبًا.	١	١,٤								

٦	تأثيرات الضوضاء	٣	٤,٢								
٧	نماذج معالجة اللغة الطبيعية معظمها مُصمم للغات مُحددة، ووثائق معاصرة غير مُحسّنة للتعامل مع المواد التاريخية، والتي قد تحتوي، على سبيل المثال، على اختلافات لغوية.	١	١,٤								
٨	معالجة ملايين صفحات الصحف الرقمية من منشورات مُختلفة حول العالم، والتي تُقدّر بمئات التيرابايت من البيانات.	١	١,٤							١	١,٤
٩	صعوبة الوصول إلى المواد المصدرية؛ تجعل التحليلات الكمية لمحتوى الأخبار نادرة في تاريخ الصحافة.	١	١,٤								
١٠	المناهج الشاملة والتركيبية، تُعاني من قيود شديدة؛ بسبب الغياب شبه التام للدراسات المنهجية حول الدوريات الفردية.									١	١,٤
١١	النموذج الرقمي المهيمن للبنى التحتية الاجتماعية والتقنية، المبني على النصوص الإنجليزية، واللاتينية؛ يُسهم في خلق انقسام استعماري جديد، بين وفرة المواد الثقافية الرقمية في الشمال العالمي واختفاء أي شيء يتجاوزها تقريبًا.									١	١,٤
١٢	يُعدّ التعرف الآلي على الكيانات المسماة (NER) راسخًا للغات الغربية، ولكن لا يوجد حاليًا برنامج متاح بسهولة، يدعم اللغة العربية.			١	١,٤						
١٣	مُدخلات غير مُتجانسة وتاريخية	١	١,٤								

	ومُشوّشة.									
١٤	عدم توافر موارد لغوية مناسبة.	١	١,٤							
١٥	استعادة المعلومات من مجموعات التصوير الصحفي؛ تثير تساؤلات أخلاقية حول الملكية الفكرية، والحق في حماية الهوية، عند إتاحة السجلات عبر الإنترنت.	١	١,٤							
١٦	أن المحتوى المرئي للدوريات التاريخية، لا يزال يمثل تحديًا للاستكشاف المنهجي.							١	١,٤	
١٧	جمع المحتوى القائم على الصور من البيانات الضخمة عملية معقدة تمامًا، في ظل تنسيق بيانات الصور غير المنظم	١	١,٤							
١٨	كتابة مجموعة البيانات المتاحة بلغة أو مجال ذي موارد محدودة.	١	١,٤							
١٩	لا تتوفر إلا على بيانات وصفية على مستوى الصفحة؛ مما يُصعّب استخراج المعلومات من سياقات محددة، ويعتمد البحث الواعي للسياق -بشكل كبير- على تحسين البيانات الوصفية؛ لذلك، عند استخدام البيانات الوصفية على مستوى الصفحة؛ يصبح تحديد المواقع الجغرافية للمعالم الأقل شهرة، أكثر صعوبة.							١	١,٤	

يتضح من الجدول أن أكثر المشكلات التي تواجه رقمنة وأرشفة الصحف التاريخية وفقًا للدراسات موضع التحليل، تمثلت في؛ تشويش وانخفاض جودة النصوص فيما يتعلق بالتعرف الضوئي على الحروف، والتي تكررت ١٥ مرة بنسبة ٢١,١%، ومن ثم يمثل عدم دقة التعرف الضوئي على الحروف الإشكالية الأكثر تكرارًا والأكثر حاجة إلى المعالجة فيما يتعلق برقمنة وأرشفة الصحف التاريخية، وتمثلت المشكلة التالية التي تواجه رقمنة وأرشفة الصحف التاريخية من حيث التكرار في أن رقمنة الصحف التاريخية تُواجه تحديات؛ نظرًا لتعقيد تصميماتها،

وكثرة محتواها، والتي تكررت ١٠ مرات بنسبة ١٤,١%، بينما جاءت تأثيرات الضوضاء باعتبارها أحد المشكلات التي تواجه رقمنة وأرشفة الصحف التاريخية ٤ مرات بنسبة ٥,٦%، بينما تكررت باقي المشكلات التي تم رصدها في الدراسات موضع التحليل في الدراسة مرة واحدة فقط.

كما يتضح من الجدول ندرة الدراسات التي تناولت إشكالية رقمنة وأرشفة الصحف التاريخية العربية باستخدام تطبيقات الذكاء الاصطناعي، في عينة الدراسات موضع التحليل في الدراسة .

الخاتمة ومناقشة نتائج الدراسة

رصدت الدراسة وحللت تقنيات الذكاء الاصطناعي المستخدمة في رقمنة وتحليل الصحف التاريخية والعناصر الصحفية المقروءة والمصورة، من خلال التحليل الكيفي باستخدام أسلوب تحليل المستوى الثانيMeta-Analysis ، للدراسات التي تناولت موضوع الدراسة، في الفترة من ٢٠٢٠، حتى ٢٠٢٥، وجاءت نتائج الدراسة على النحو التالي:

تمثلت تقنيات الذكاء الاصطناعي المستخدمة في رقمنة الصحف التاريخية وتحليلها، في الدراسات موضع الدراسة في التقنيات التالية:

أساليب الاستدلال، والرؤية الحاسوبية، وتعرف الأنماط، والتعلم الآلي، نماذج اللغة الكبيرة LLMs، ونموذج وصفي دلالي، وتقنيات التعلم العميق المتقدمة، ونموذجBERT ، ومناهج الرؤية الحاسوبية -التعلم الآلي، وأساليب التصنيف الخاضعة للإشراف وغير الخاضعة للإشراف، والكيانات المسماة(NER) ، ونماذج كشف الأجسام((Faster RCNN ، وبنيتا الشبكات العصبية العميقة Faster R-CNN المستندة إلى بنية الشبكة العصبية التلافيفية (CNN)، وTransformers، ونموذج لغوي واسع النطاق؛ لتصحيح أخطاء التعرف الضوئي على الحروف، والشبكات العصبية العميقةDNNs ، والتعليم الانتقالي، لتجزئة المقالات من صور الصحف، وفصل المقالات فيNewsEye، والتعلم العميق، وتطبيق ChroniclingAmericaQA، وأركانوم: واجهة برمجة تطبيقات لتجزئة الصحف، ونموذج YOLOv8، والتعرف الضوئي على الحروف، باستخدام نماذج الكشط والتعلم العميق، ونهج فصل المقالات SRRAS، كذلك نموذج التخطي(sgSTRAS)، ونموذج حقيبة الكلمات المستمرة(cbowSTRAS)، ونموذج(FastText (ftSTRAS، ونموذج SpaCy المُدرّب مسبقًا(preSTRAS)، ومورد تعليمي مفتوح OER حول البحث التاريخي الرقمي في الصحف التاريخية، ونظام قائم على القواعد، يتم تقييمها ومقارنتها مع نموذجين من نماذج التعلم الآلي، وتحليل أخطاء التعرف الضوئي على الحروف، واستخدام تقنيات التعلم الآلي على صفحات Chronicling America، ونهج متعدد الوسائط لتحليل تخطيطات الصحف التاريخية، ونموذجًا جديدًا لتحليل تخطيطات الصحف التاريخية يُسمى.Panoptic-DLA

وتمثلت الصحف التاريخية التي تم تطبيق التقنية عليها لرقمنتها أو تحليلها في الدراسات موضع التحليل في الدراسة، وفقا لنتائج الدراسة في الصحف التالية :

مجلتان برتغاليتان تاريخيتان، هما "دياريو دي نوتيسياس" و"جورنال دي نوتيسياس"، مجموعات الصحف التاريخية الرقمية، مثل مجموعات المكتبة الملكية البلجيكية(KBR) ، والملاحق الأدبية من مجلة "لو بيبول"، والصحف البنجابية، صحف بلغات مختلفة (مثل العربية، الإنجليزية، الفرنسية، الألمانية)، تنظيم أرشيف الصحف الرقمي في مختبر KBLab في المكتبة الوطنية السويدية، وأرشيفات الصحف الرقمية الكولومبية، و صفحات صحف، معظمها من غرناطة الجديدة - وهي دولة، كانت تضم كولومبيا، وبنما، وفنزويلا والإكوادور، ٣٠٣٨ صحيفة رقمية من ٥٨ مجموعة رقمية في جميع أنحاء المكسيك، والأرجنتين، وكولومبيا، وبيرو، وتشيلي، وبنما، وفنزويلا، وأوروغواي، وبوليفيا، وكوبا، والإكوادور، وبعض البلدان، مثل بوليفيا، وكوبا، وفنزويلا، و١٩٧ عنوان صحيفة و٢٣٥٢٢ صفحة من الصور الممسوحة ضوئيًا، معظمها من مدينة مكسيكو، و تشمل أيضًا منشورات من مدن أخرى في أمريكا اللاتينية، مثل بوينس آيرس، وليما، وبوغوتا، وسانتياغو دي تشيلي، ومجموعة الصحف التاريخية Chronicling America، ومجموعة صحف المكتبة الوطنية السويسرية، صحيفةBündner Nachrichten، مجموعة صحف ضمن) Annolyzer منصة عرض NewsEyeالتي تم دمجها في مختبرات(ÖNB ، وصحيفة-Illustrierte Kronen Zeitung، وأركانوم؛ واجهة برمجة تطبيقات لتجزئة الصحف؛ مثال "صحيفة نيوز- فيلت-بلات، وتصنيف الإعلانات، ونعي الوفيات في صحيفة"Slovenski narod"، وصحيفة "هونغ كونغ نيوز"، والصحف الفرنسية والفنلندية، والصحف التاريخية الأوروبية، وصحف ودوريات ومجلات فرنسية، نُشرت في منطقة فرانش كونتيه، و١٦ مليون صفحة صحيفة تاريخية في قاعدة بيانات Chronicling America والصحف السويسرية واللوكسمبورغية.

تمثلت تقنيات الذكاء الاصطناعي، المستخدمة في رقمنة وتحليل النصوص الصحفية التاريخية، في الدراسات موضع التحليل في الدراسة، وفقًا لنتائج الدراسة في التقنيات التالية :

تقنياتLDA – NMF ، والنماذج العصبية -Top2Vec ل BERTopic، ونموذج مجموعة بيانات من المقالات، باستخدام تسميات متعددة، ونماذج اللغة الكبيرة(LLM) ؛ للحدّ من تشويش التعرف الضوئي على الحروف، ونموذج لغة مُدرّب مسبقًا؛ لتحويل الصور إلى نص Pixtral 12B، ونماذج اللغة الكبيرة (LLMs) - نموذجي BERT وGPT-4o ، وخط الأنابيب تحليل تخطيط المستندات.

والتعرف الضوئي على الحروف(OCR)، وتقنيات توليد الذكاء الاصطناعي لاستخراج البيانات، تكييف برامج التعلم الآلي التوليدية، وتبني نهجًا قائمًا على التوجيه؛ من خلال ضبط التعليمات في برنامج Llama 2 ومقارنته ببرنامج BART مُعدّل بدقة علىBLN600، وقاعدة بياناتHistNERo، ومجموعة بيانات رومانية؛ لتعرف الكيانات المسماة (NER) في الصحف التاريخية، ونماذج تعرف الكيانات المسماة، ونموذج BiLSTM-CRF والنماذج القائمة على المُحوّل، والتحليل المعجمي عبر السياقات، والتحليل الدلالي الديناميكي، وتحليل التكافؤ

خط أنابيب معالجة معلومات التخطيط المفقودة، وأداة بحث دلالي مُبتكرة News Deja Vu تستفيد من نماذج اللغة الكبيرة المُحوّلة، ونهج المُشفّر الثنائي؛ لتحديد المقالات الإخبارية

التاريخية، ونهج لإنتاج مجموعات نصوص عالية الجودة، وتقديم واجهة قائمة على الأدوار، تُهيكل المقالات الإخبارية حول الشخصيات التاريخية، ومختبر أبحاث المكتبات الرقمية.Aida

وخط أنابيب مبتكر وعميق؛ لاستخراج نصوص المقالات الكاملة من صور الصحف، يطبق علىChronicling America، وواجهة امبريسو -Impresso نمذجة المواضيع، وأساليب التعلم العميق- نماذج الكشف والتصنيف القائمة على BERT نموذجين (للغتين الفرنسية والألمانية، ونهج ربط البيانات - EL بنية ربط كيانات متعددة اللغات لمقالات تاريخية، والتعلم الآلي المدمج- تحليل المواضيع والديناميكيات الدلالية؛ باستخدام مُشغّلات تحليل مُتوافقة ومتعددة، و -(ChallAm)التعلم الآلي، والتعلم الآلي المُشرف، ونماذج تضمين الكلمات؛ بالاعتماد على مجموعة كبيرة من مواد الصحف السويدية التاريخية، أُنتجت هذه البيانات، في إطار مهمة "سبراكبانكن تكست"، واجهة إمبريسو Impresso باستخدام نمذجة المواضيع (MALLET)، مُصنّف بايز.، وتطوير موردNewsEye، والشبكات العصبية الهجينة العميقة CNN–LSTM، ومجموعة بيانات؛ لتحديد القواعد التي تُعيّن تسميات منطقية، لكل من كتل وأسطر النص من وثائقXML ALTO ، والنماذج القائمة على الشبكات العصبية، بنية تتضمن طبقات BiLSTM مدمجة مع طبقةCRF ، ومعالجة الكيانات المُسمّاة (NE) للنصوص التاريخية، وتضمين الكلمات العصبية (اللغويات الحاسوبية) والنهج التاريخي للخطابDHA))، نماذجTDA word2vec ، وخط أنابيب متكامل، للتعرف الضوئي على الحروف (OCR) باسم "أوريغامي"، ومشروع "إمبريسو - رصد وسائل الإعلام للماضي"، مع مجموعات النصوص والمعايير والتعليقات الدلالية ونماذج اللغة، ونماذج المواضيع (LDA) و(DTM

وتمثلت الصحف التاريخية، التي تم تطبيق أن تقنيات الذكاء الاصطناعي على نصوصها التاريخية؛ لرقمنتها، أو تحليلها في الدراسات موضع التحليل في الدراسة، _ والتي تعكس تنوعًا كبيرًا، من حيث العدد والكم والكيف واللغات التي تصدر بها الصحف الواردة في الدراسات موضع التحليل، وفقًا لنتائج الدراسة في:

106صحيفة نيوزيلندية، ومقالات تاريخية من صحف جنوب الولايات المتحدة، و ٨٤ ألف صفحة من الصحف والدوريات الإنجليزية، وصحف أمريكا اللاتينية، ومشروع PortADa وصول السفن إلى موانئ برشلونة، ومرسيليا، وهافانا، وبوينس آيرس، وصحيفة "Le Sémaphore de Marseille"التاريخية، مقالات الصحف البريطانية، عينات من إحدى المناطق التاريخية الأربع التالية في رومانيا، وهي؛ بيسارابيا، ومولدوفا، وترانسيلفانيا، ووالاشيا، واستخراج عناوين الكتب من أرشيفات الصحف التاريخية الممسوحة ضوئيًا؛ باستخدام مجموعة بيانات من صحيفةLeeuwarder Courant ، صحف بروكويست التاريخية؛ لدراسة البناء الاجتماعي لجنوح الأحداث خلال هذه الفترة، الصحف التاريخية الصادرة باللغة الألمانية، ومجموعة بيانات Chronicling Germany وتحتوي ٥٨١ صفحة صحفية تاريخية مُعلقة، ومجموعة ضخمة من الصحف التاريخية السلوفينية، ونصوص المقالات الكاملة من صور الصحف الأمريكية، من ٢٠ مليون عملية مسح ضوئي في مجموعة Chronicling America،التابعة لمكتبة الكونغرس والمتاحة للملكية العامة، والصحف

السويسرية، واللوكسمبورجية، يتكون مجموعها من ٧٦ صحيفة، نُشرت باللغات الفرنسية، والألمانية، واللوكسمبورغية ، والمقالات الصحفية السويسرية، واللوكسمبورغية (مجموعة المطبوعات)؛ للكشف عن دور وكالات الأنباء في الماضي، مجموعتين من الوثائق التاريخية، تغطيان خمس لغات أوروبية (الإنجليزية، والفنلندية، والفرنسية، والألمانية، والسويدية)، وتطبيق أدوات تحليل نصوص متطورة، على مجموعات صحف تاريخية ضخمة، والصحف التاريخية التي جُمعت على بوابةChronicling America، والصحف السويدية التاريخية، ومجموعة كبيرة من مقالات الصحافة السويسرية (١٨٣,٢٧٠ مقالة)، ومواد صحفية تاريخية مُتعاقبة، نُشرت باللغات الفرنسية، والألمانية، والفنلندية والسويدية، والمكتب الرئيسي لمكتبات جامعة إسطنبول، والتوثيق؛ للوصول إلى قاعدة بيانات ضخمة من الصحف التاريخية الممسوحة ضوئيًا، والصحف التاريخية السويدية ، وصحف باللغة الفرنسية، وأربع دوريات عربية، من بغداد، و بيروت، والقاهرة، ودمشق، والصحف التاريخية، باللغات الفرنسية، والألمانية، والإنجليزية، والفضاء الدلالي للخطاب العام للهجرة في المملكة المتحدة، وتستخدم أرشيف تايمز الرقمي(TDA)، ومشروع رقمنة صحيفة "برلينر بورسن تسايتونغ (BBZ) "الألمانية وتحليلها الكمي، والإسهام في موارد اللغة الفرنسية، والألمانية، واللكسمبورجية، ورصد ديناميكيات الخطاب وفهمها، الصحف والدوريات المنشورة في فنلندا.

تشير نتائج التحليل في الدراسة؛ إلى أن تقنيات الذكاء الاصطناعي المُستخدمة في رقمنة تحليل العناصر الصحفية المصورة التاريخية، في الدراسات موضع التحليل، تمثلت فيما يلي:

التعلم الآلي، ومجموعة أدوات ذكاء اصطناعي؛ لنسخ ملاحظات الأرشيفيين، واستعادة أسماء المصورين وتواريخ الصور، وخط أنابيب؛ للكشف عن الوجوه وتمويهها، ونموذج مبتكر قائم على المحول، وتطبيق نموذجYOLOv8 ؛ لضبط نموذج الكشف عن الكائنات؛ لتحويل تعليقات الرسوم التوضيحية؛ إلى نص قابل للقراءةTesseract OCR ، واستخدام برنامج Open-لاسترجاع النصوص والصور متعدد الوسائط، تتيح مجموعة البيانات الناتجة وتطبيقها، تحليلًا مرنًا للتمثيلات Flask آليًا، ورقمنة المجلات باللغة الإسبانية، وآلية حسابية جديدة؛ للكشف عن :FPC تطوير نموذج YOLOv5_FPCP ؛ من خلال الضبط الدقيق لنموذج التعلم العميق(YOLOv5) You Only Look Once Version 5، واستخراج كيانات مسماة لبيانات وصفية بتنسيقXML، وتواجه مجموعة البيانات، المتاحة عبر الإنترنت وعلى Zenodoللتنزيل وإعادة الاستخدام، مشكلات في الرؤية الحاسوبية للصور التراثية، وميزة Docstrumالهندسية، تمثيل مرئي، يعتمد على فسيفساء Voronoi وDCNN ، وتقليص حجم الصورة التكيفي، الذي يجمع بين نقاط القوة في الاستراتيجيات المستقلة عن المحتوى، والاستراتيجيات الواعية بالمحتوى، وتطوير نموذج تعرف الكيانات المسماة(NER) ، والمشاهدة عن بعد؛ تقنيات حسابية لدراسة الوسائط المرئية واسعة النطاق، والرؤية الحاسوبية، وتطوير مسار؛ لتحديد المواقع الجغرافية، وتصوير الصور التاريخية، ونهجًا هجينًا يستخدم كلًا من البيانات الوصفية النصية، وخصائص الصور، والتعلم الانتقالي؛ لتحسين كشف المشاهد، ونموذج Places-365 ResNet-50 المُدرّب مسبقًا، والشبكات العصبية التلافيفية(CNNs) ؛

لتصنيف المصادر البصرية التاريخية الرقمية وتحليلها، ومناهج للكشف عن السمات الخاصة بالوسيلة (فصل الصور عن الرسوم التوضيحية)، والاستعلام عن الصور؛ بناءً على الجوانب البصرية المجردة (تجميع الإعلانات المتشابهة بصريًا)، وتدريب شبكة عصبية؛ بناءً على فئات بصرية، طورها خبراء في هذا المجال، وأساليب التعلم العميق، وبنية شبكة عصبية ملتوية بالكامل(FCN) ، ونموذج يعتمد بنية كتلة بوابة متعددة المقاييس (MSGB) جديدة.

تمثلت الصحف التاريخية التي تم تطبيق تقنيات الذكاء الاصطناعي على المواد المصورة بها، والمواد المصورة التي وردت في الدراسات موضع التحليل، في :

صحيفة بوسطن غلوب، ومجلة أخبار لندن المصورة، وصور النساء في ثلاث مجلات رائدة، من أواخر القرن التاسع عشر في إسبانيا" فيدا نويفا "(مدريد)، و"لا فيدا ليتريباريا" (مدريد)، و"لا فيدا جالانتي" (برشلونة)، وأرشيف مكتبة أخبار تشوسون إلبو - كوريا، إنتاج الفيلم الوثائقي "ذاكرة الطباعة". هذا الإنتاج السمعي البصري ، مخصص لتاريخ الصحافة الرومانية، النزاعات المسلحة من تسعينيات القرن التاسع عشر إلى عام ١٩١٨، مع التركيز على السياقات خارج أوروبا، والصحيفة الفرنسية المصورةSur le Vif ، التي وثقت الحرب العالمية الأولى، وصور من الصحف تسمى (Diario de Madrid DM) من الصحافة الإسبانية ، ومجموعة من ١٧٨٩٧ قصة مصورة من سلسلة Peanuts لتشارلز شولتز، أشهر قصة صحفية تاريخية مصورة، ومجموعات مكتبة بروكسل الملكية(KBR)، ومجموعة من الصور التاريخية، التصوير الصحفي، مكتب فوتو برسو دي بور، وأرشيف الصحف الهولندية الرقمية، مجموعة الصحف الرقميةDelpher ، و٤٢٦,٧٧٧ إعلانًا نُشر في صحيفتين هولنديتين مؤثرتين Algemeen Handelsblad و .NRC Handelsblad

تشير نتائج الدراسة، إلى تعدد وتنوع القضايا الصحفية والمواد الصحفية المصورة، التي تم معالجتها في رقمنة الصحف التاريخية، وتحليلها بتقنيات الذكاء الاصطناعي، وتكررت كل قضية منها مرة واحدة فقط بنسبة ١,٤%، باستثناء قضايا الهجرة التى تكررت مرتين بنسبة ٢,٨%، وقضية دور وكالات الأنباء في الماضي والتي تكررت مرتين بنسبة ٢,٨% في عينة الدراسات موضع التحليل في الدراسة، وقد تمثلت هذه القضايا والعناصر المصورة في :

- دراسة حالة على زلزال ميسينا الكارثي في جنوب إيطاليا.
- دراسات حالة الكوارث الطبيعية/البيئية.
- الهجرة (العودة)
- الفضاء الدلالي للخطاب العام للهجرة في المملكة المتحدة.
- استخراج عناوين الكتب من أرشيفات الصحف التاريخية الممسوحة ضوئيًا؛ باستخدام مجموعة بيانات من صحيفة.Leeuwarder Couran
- دراسة البناء الاجتماعي لجنوح الأحداث الفترة من ١٧٩٠ إلى ٢٠٠٦.
- الكشف عن دور وكالات الأنباء في الماضي.

- استعادة أسماء المصورين، وتواريخ الصور، وتمويه الوجوه؛ لحماية حقوق الملكية الفكرية والخصوصية.
- إنتاج الفيلم الوثائقي "ذاكرة الطباعة"، هذا الإنتاج السمعي البصري، مخصص لتاريخ الصحافة الرومانية.
- الأطر الدلالية المختلفة، المرتبطة بالمرأة، وأدوارها المجتمعية، وتمثيلاتها في المجتمع الإسباني، في القرن التاسع عشر، الكشف عن مدى تشابك الدلالات الطبية، والسياسية، والاجتماعية في الخطاب المحيط بهوية المرأة وأدوارها في ذلك الوقت.
- تتبع التطور التاريخي لخطاب وسائل الإعلام؛ بوصفه وسيلة للتحقيق في بناء المعنى الجماعي، استنادًا إلى نظرية الدلالات التوزيعية، ونظرية الخطاب النقدي.
- النزاعات المسلحة من تسعينيات القرن التاسع عشر إلى عام ١٩١٨، مع التركيز على السياقات خارج أوروبا.
- مجموعة من ١٧٨٩٧ قصة مصورة من سلسلة Peanuts ، لتشارلز شولتز، إحدى أشهر القصص الصحفية المصورة في العالم.
- تحليل حاسوبي للمحتوى المرئي؛ لاستكشاف صحافة لغة اللادينو.
- مجموعة من الصور التاريخية، والتصوير الصحفي، مكتب فوتو برسو دي بور.

وأشارت نتائج الدراسة، إلى تعدد المشكلات التي تواجه رقمنة وأرشفة وتحليل الصحف التاريخية وعناصرها المقرورة والمصورة، وفقًا للدراسات موضع التحليل، وجاءت من هذه المشكلات، وفقًا للأعلى في نسب التكرار في الدراسات موضع التحليل في الداسة، علي النحو التالي:

جاء تشويش وانخفاض جودة نصوص التعرف الضوئي على الحروف، بوصفه أحد تحديات رقمنة وتحليل الصحف التاريخية بتقنيات الذكاء الاصطناعي، والذي تكرر ١٥ مرة في الدراسات موضع التحليل، بنسبة ٢١,١%، بينما جاء تشكيل رقمنة الصحف التاريخية تحديًا؛ نظرًا لتعقيد تصميماتها وكثرة محتواها ١٠ مرات بنسبة ١٤,١ .%

بينما جاء تحيز خوارزميات الذكاء الاصطناعي في تحليل ورقمنة الصحف التاريخية ٤ مرات، بنسبة ٥,٦%، وأشارت إحدى الدراسات موضع التحليل في هذا السياق، إلى إن الأدوار المتزايدة للتعلم الآلي، والذكاء الاصطناعي، في بناء مجموعات بيانات التراث الثقافي، والعلوم الإنسانية، تتطلب فحصًا نقديًا، للتحيزات العديدة التي أدخلتها الآلات، والخوارزميات، والبشر، الذين يبنونها وينشرونها، من تصنيف الصور، إلى التعرف الضوئي على الحروف، وتتراكم آثار القرارات التي تتخذها الآلات ظاهريًا؛ من خلال خط أنابيب الرقمنة، وتتضاعف في كل خطوة؛ مما يتوسط تفاعلات الباحثين، مع القطع الأثرية المقدمة رقميًا؛ من خلال عملية البحث والاكتشاف، ونتيجة لذلك؛ بدأ العلماء داخل مجتمع العلوم الإنسانية الرقمية؛ في الدعوة إلى وضع مجموعات بيانات التراث الثقافي، في سياقها الصحيح، داخل الأنظمة الاجتماعية، والتقنية التي يتم إنشاؤها، واستخدامها فيها، أحد هذه الأساليب لهذا السياق؛ هو علم آثار البيانات، وهو شكل من أشكال التنقيب الإنساني لمجموعة بيانات، يعرّفها بول فايف، بأنها "استعادة وإعادة بناء

كائنات الوسائط داخل بيئاتها المتغيرة، في دراسات البيانات الحرجة، أثبت هذا التنقيب في مجموعة البيانات - بما في ذلك بنائها ومعالجتها عبر التعلم الآلي - أنه نهج واسع النطاق، وعلى الرغم من ذلك، لم يُعتمد علم آثار البيانات بوصفه ممارسة قياسية بعدُ، بين ممارسي التراث الثقافي، الذين ينتجون مثل هذه المجموعات من البيانات باستخدام التعلم الآلي.

وجاءت تأثيرات الضوضاء في الصحف التاريخية، بوصفها أحد تحديات رقمنة وتحليل الصحف التاريخية الممسوحة ضوئيًا، بتقنيات الذكاء الاصطناعي ٣ مرات، بنسبة ٤,٢%، وقد أشارت دراسة Marna Giardinetti et all (2024)[72] في تناولها لهذه الإشكالية، إلى أن ثمة مشكلات في الرؤية الحاسوبية للصور التراثية، مثل التدهور الناتج عن البهتان، وتغير اللون، والخدوش، والضوضاء؛ مما يُضعف الخوارزميات المعتمدة على السمات البصرية، ويُؤدي نقص تمثيل الثقافات الفوتوغرافية المبكرة في مجموعات البيانات؛ إلى تحيز في تطبيق الحلول القياسية على المواد الأرشيفية.

بينما تكرر كم الصحف التاريخية بوصفه تحديًا في رقمنة وتحليل الصحف التاريخية بتقنيات الذكاء الاصطناعي؛ حيث إنه يتضمن ملايين صفحات الصحف الرقمية، من منشورات مُختلفة، حول العالم، والتي تُقدّر بمئات التيرابايت من البيانات مرتين، بنسبة ٢,٨ %، إحداهما يتعلق بتحدٍ تقني، في تحليل ورقمنة الصحف التاريخية بتقنيات الذكاء الاصطناعي، والأخرى تتعلق بمشكلة البحث العلمي في الصحف العربية التاريخية ورقمنتها، وتحليلها، وفقًا لدراسة Till Grallert (2021)[73]، كما تناولت الدراسة أيضًا، أن النموذج الرقمي المهيمن للبنى التحتية الاجتماعية والتقنية، المبني على النصوص الإنجليزية واللاتينية، يُسهم في خلق انقسام استعماري جديد، بين وفرة المواد الثقافية الرقمية في الشمال العالمي، واختفاء أي شيء يتجاوزها تقريبًا، وهذا يُجبر الباحثين الذين يعملون على نصوص بلغات غير غربية، أو مكتوبة بنصوص غير لاتينية، على بذل جهود كبيرة؛ لبناء مجموعات نصوص، ويُحدّ- بشدة- من نطاق تدقيق الدراسة الحاسوبي .

قدم عدد من الدراسات موضع التحليل في الدراسة، تقنيات ذكاء اصطناعي، تتعلق بتحسين البحث العلمي في الصحف التاريخية، في مجال الدراسات الإنسانية والاجتماعية، تمثلت في؛ مقترح يتيح استخراج المقالات تلقائيًا، بما في ذلك الكشف عن نوع محدد منها، مثل مجموعة من المقالات، أو ملحق أدبي، يُعد هذا مفيدًا بشكل خاص لباحثي العلوم الإنسانية؛ إذ يُحسّن إمكانية البحث في هذه المجموعات، ويُمكّن من بناء مجموعات مقالات حول مواضيع محددة.

كما قدمت إحدى الدراسات، مجموعة البيانات الخاصة بالدراسة، والرمز الأساسي وهما متاحان مجانًا على الإنترنت، وأشارت إلى أن هذا العمل يشكل نقطة انطلاق لأبحاث مستقبلية في مجال التاريخ الرقمي، ومعالجة الصحف التاريخية باللغة الألمانية، علاوة على ذلك، يوفر فرصة لدراسة مهمة منخفضة الموارد في مجال الرؤية الحاسوبية.

كانت أبرز معالجات إحدى الدراسات، موضع الدراسة لإشكاليات البحث العلمي في الصحف التاريخية بالذكاء الاصطناعي، تتمثل في الدعوة إلى تكامل التخصصات في دراسات رقمنة

وتحليل الصحف التاريخية بالذكاء الاصطناعي؛ حيث إن إحدى الدراسات، تشير إلى أن التعاون بين التخصصات المختلفة، يمكن أن يكون ذا فاعلية عند الفهم الجيد لسير العمل، والتقاليد الخاصة بكل تخصص من التخصصات المعنية، ولكن يجب إيجاد مناهج متكاملة؛ لاستغلال الإمكانات الكاملة للمصادر الرقمية بصورة ناجحة، علاوة على ذلك، قدمت الدراسة نظرة ثاقبة على الأدوات والأساليب الرقمية والتأويلات في التطبيق، موضحةً أن البحث المتكامل متعدد التخصصات، يحتاج إلى بناء شيء ما بين التخصصات، مع احترام وفهم خبرات وتوقعات كل منها.

وتشير نتائج التحليل في الدراسة، في تحليلها للدراسات التي تتناول تطبيقات الذكاء الاصطناعي، في رقمنة الصحف التاريخية وتحليلها، وعناصرها المقروءة والمصورة، إلى أن أغلب الدراسات موضع التحليل، قد اعتمدت على دراسة الحالة، ويُعدُّ ذلك، ناتجًا - في أحد الجوانب- عن عدم توظيف تقنيات الذكاء الاصطناعي في رقمنة وتحليل الصحف، والكم الكبير من الصحف التاريخية، وفي جانب آخر؛ نتيجة التحديات والمشكلات التي تواجهها هذه الرقمنة.

قد أشارت دراسة Marcel Broersma and Frank Harbers (2021)[74] إحدى الدراسات، موضع التحليل في الدراسة، إلى أن طبيعة التحليل اليدوي للمحتوى، التي تتطلب جهدًا مكثفًا، وصعوبة الوصول إلى المواد المصدرية؛ تجعل التحليلات الكمية لمحتوى الأخبار نادرة في تاريخ الصحافة، وعلى الرغم من ذلك، فإن رقمنة أرشيفات الصحف، تتيح الآن أساليب رقمية مبتكرة؛ لإجراء أبحاث طولية منهجية، تتجاوز نطاق دراسات الحالة العرضية، نجادل بأن التعلم الآلي المُشرف، يُقدم مناهج واعدة؛ لتحليل المواد المصدرية الوفيرة، والتحليلات الميدانية للبيانات الضخمة، ورسم خريطة للتحول الهيكلي للخطاب الصحفي طوليًا؛ من خلال التحليل التلقائي لأعراف الشكل والأسلوب، التي تعكس المعايير والممارسات المهنية الأساسية، يمكن دراسة بنية التغطية الإخبارية عن كثب، وعلى الرغم من ذلك، فإن التصنيف التلقائي لفئات الترميز الكامنة والخاصة بفترة زمنية معينة، أمر معقد للغاية، ويزيد هيكل أرشيفات الصحف الرقمية (مثل التجزئة، والتعرف الضوئي على الحروف)، من تعقيد هذا الأمر، في حين أن خوارزميات التعلم الآلي، غالبًا ما تكون بمثابة صندوق أسود، توضح هذه الورقة البحثية، كيف أن جعل عمليات التصنيف شفافة؛ يُمكّن باحثي الصحافة، من استخدام هذه الأساليب الحسابية، بطريقة موثوقة وصحيحة، توضح الدراسة موضع التحليل ذلك؛ بالتركيز على المشكلات التي واجهتها، عند تصنيف أنواع الأخبار تلقائيًا، وهي فئة ترميزية مُنيرة، وإن كانت مُعقدة للغاية، ويُمكن لمثل هذا النهج؛ أن يُعزز مراجعة تاريخ الصحافة، لا سيما التحول من التقارير القائمة على الرأي، إلى التقارير المُركزة على الحقائق، والذي غالبًا ما يُفترض أنه لم يُدرس جيدًا.

توصيات الدراسة

ثمة مجموعة من التوصيات، التي يمكن تقديمها في ختام الدراسة، تتمثل في:

- وضع دراسات الإعلام والصحافة خطة بحثية، تقتضي تكامل التخصصات في مجال دراسة تاريخ الصحافة، في ظل رقمنة الصحف التاريخية وأرشيفاتها، التي تم جمعها في وقت سابق، في صورة رقية، أو على وسائط ميكروفيلم أوميكروفيش، وتم أو يتم تحويلها إلى وسائط رقمية حاليًا؛ حتى يتم دراسة هذه الأرشيفات الرقمية؛ في ظل وعي تام بجوانب القصور فيها.
- تحفيز التعاون والتكامل بين التخصصات وإجراء الأبحاث المشتركة ذات التخصصات البينية، بين باحثي الإعلام والصحافة من جانب، وباحثي علوم الحوسبة، وعلم البيانات والذكاء الاصطناعي؛ لتطوير آلات البحث بالذكاء الاصطناعي في الصحف التاريخية الورقية، أو المحفوظة في وسائط ميكروفيلم، أو ميكروفيش، والتي تم نقلها إلى وسائط وأرشيفات رقمية، وذلك بناءً على رؤية، تجعل باحثي الإعلام والصحافة، على معرفة دقيقة بتأثيرات أدوات ومناهج البحث بالذكاء الاصطناعي، على مجال البحث في الصحف التاريخية، وتاريخ الصحافة، وما يتضمنه من تحديات.
- إجراء مزيد من الدراسات، حول المجالات البحثية، المتعلقة بتطبيقات آليات ومنهجيات الذكاء الاصطناعي، في رقمنة الصحف التاريخية، في إطار ما يطرحه هذا المجال البحثي، من مشكلات وإشكاليات بحثية .
- حث المؤسسات الصحفية المصرية والعربية؛ على إعطاء المزيد من الاهتمام للأرشفة الصحفية بالذكاء الاصطناعي؛ حيث اعتمدت العديد من الدراسات السابقة، على دراسات الحالة العرضية؛ نظرًا لأنها تقوم بعمل تجارب، لإنتاج آليات مناهج بالذكاء الاصطناعي، لإتاحة أرشيفات الصحف التاريخية للدراسة، ويعد ذلك مؤشرًا؛ لضرورة استكمال أرشيفات الصحف التاريخية؛ حتى يتسنى للباحثين؛ إجراء دراسات مطولة في مجال تاريخ الصحافة وخطابها التاريخي، وموادها المقروءة والمصورة .

المصادر والمراجع

1. Bruno Frutuoso Costa et all (2025): Looking Back to 1850 in 2025: Historascan to Digitize Historical Journals, **The Impact of Digitalization on Communication Dynamics.**
   Available at: https://www.igi-global.com/chapter/looking-back-to-1850-in-2025/366101
2. Johanna Mauermann and Sarah Oberbichler (2025): LLM Biases: Expected and Unexpected Model Design Effects in Historical Newspaper Article Extraction on the Messina Earthquake, *DH Lab*. Abgerufen.
   Available at: https://doi.org/10.58079/137qr
3. Sun, Shaodan (2025): "Unearthing historical insights: semantic organization and application of historical newspapers from a fine-grained knowledge element perspective", Aslib Journal of Information Management, Vol. 77 No. 2.
   Available at: https://doi.org/10.1108/AJIM-05-2023-0180
4. Dilawar Ali et all (2024): "Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections", Journal of Documentation, Vol. 80 No. 5.
   Available at: https://doi.org/10.1108/JD-01-2022-0029
5. Atul Kumar and Gurpreet Singh Lehal (2024): Detection of Punjabi Newspaper Articles Using a Deep Learning Approach,.
6. **Electrical and Electronic Engineering Conference paper, Electrical Engineering, vol 1115**
   Available at: https://doi.org/10.1007/978-981-99-8661-3_30
7. Abdullah Almutairi (2024): Newspaper elements detection and newspaper pages categorization using CNNs and transformers, **International Journal on Document Analysis and Recognition, 07 October.**
   Avilable at: **https://doi.org/10.1007/s10032-024-00503-9**
8. Justyna Sikora and Chris Haffenden (2024): AI, Data Curation and the Data Readiness of Heritage Collections: Exploring the Swedish Newspaper Archive at KBLab, **Huminfra Conference, Gothenburg, 10–11 January**.
   Available at: https://uu.diva-portal.org/smash/get/diva2:1825045/FULLTEXT01.pdf
9. Laura Manrique-Gómez et all (2024): Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction, ACL, Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities.
   Available at: https://doi.org/10.18653/v1/2024.nlp4dh-1.13
10. Nancy Girdhar et all (2024): Leveraging Transfer Learning for Article Segmentation in Historical Newspapers, **HAL Id: hal-04676213, 23 Aug.**
    Available at: https://hal.science/hal-04676213v1
11. Bhawna Piryani et all (2024): ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages, 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.
    Available at: https://doi.org/10.1145/3626772.3657891
12. Sarah Oberbichler (2024): Large-Scale Research with Historical Newspapers: A Turning Point through Generative AI Presentation, **Conference: ÖNB Labs Symposium Newspapers as Datasets, December.**
    Available at: http://dx.doi.org/10.13140/RG.2.2.28035.31527/1
13. Filip Dobranić and Matevž Pesek, (2024): Estimating the Number of Annotations Required to Detect Content Types in Historical Newspapers, **Linking Theory and Practice of Digital Libraries, Conference paper Computer Science, Vol. 15178.**
    Available at: https://doi.org/10.1007/978-3-031-72440-4
    https://link.springer.com/chapter/10.1007/978-3-031-72440-4_11
14. Vincent Wai-Yip LUM and Michael Kin-Fu YIP (2023): Data Extraction, Visualization, and Storytelling: A Case Study in Headline Analysis on "The Hongkong News" with Deep Learning, **Society For Image science and Technology, Volume. 20 Article 6, June.**

Available at: https://doi.org/10.2352/issn.2168-3204.2023.20.1.6

15. Nancy Girdhar et all (2023): STRAS: A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers, Leveraging Generative Intelligence in Digital Libraries: **Towards Human-Machine Collaboration, Conference paper Computer Science, vol 14457. Springer, Singapore.**
Available at: **https://doi.org/10.1007/978-981-99-8085-7_8**
16. Cyrille Suire and Nicolas Sidère (2023): An OER on digital historical research on European historical newspapers with the NewsEye platform, **Education for Information Volume 39, Issue 2.**
Available at: https://doi.org/10.3233/EFI-230026
17. Nicolas Gutehrlé and Iana Atanassova (2022): Processing the Structure of Documents: Logical Layout Analysis of Historical Newspapers in French, **Journal of Data Mining and Digital Humanities**
Available at: arXiv:2202.08125v2 [cs.CL] 25 May
18. Sarah Oberbichler et all (2022): Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians, **Assoc Inf Sci Technol, February.**
Available at: https://asistdl.onlinelibrary.wiley.com/toc/23301643/2022/73/2
19. **Benjamin Lee** (2021): Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset, Digital Humanities Quarterly, Volume 15 Number
Available at: https://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html
20. Raphaël Barman et all (2021): **Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers, Journal of Data Mining and Digital Humanities, 19 January.**
Available at: https://doi.org/10.46298/jdmdh.6107
21. Min Lu et all (2021): Panoptic-DLA: Document Layout Analysis of Historical Newspapers Based on Proposal-Free Panoptic Segmentation Model, **14th International Conference, KSEM, Tokyo, Japan, August 14–16, Proceedings, Part II.**
Available at: https://doi.org/10.1007/978-3-030-82147-0_15
22. Keerthana Murugaraj et all (2025): Mining the Past: A Comparative Study of Classical and Neural Topic Models on Historical Newspaper Archives**, 5th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, Albuquerque United States 3-4 May.**
23. Karin Stahel, Irenie How et all (2025): A Bit of This, a Bit of That: Building a Genre and Topic Annotated Dataset of Historical Newspaper Articles with Soft Labels and Confidence Scores, **5th International Conference on Natural Language Processing for Digital Humanities, May 3-4,**
Available at: https://aclanthology.org/2025.nlp4dh-1.33/
24. Michael McRae (2025): Southern Newswire Corpus: A Large-Scale Dataset of Mid-Century Wire Articles Beyond the Front Page.
Available at: https://doi.org/10.48550/arXiv.2502.11866
25. Jonathan Bourne (2025): Reading the unreadable: Creating a dataset of 19th century English newspapers using image-to-text language models, 18 Feb.
Available at: https://arxiv.org/abs/2502.14901
26. Kevin Cohen et all (2025): Historical Ink: Exploring Large Language Models for Irony Detection in 19th-Century Spanish**, 5th International Conference on Natural Language Processing for Digital Humanities, May 3-4.**
Available at: https://aclanthology.org/2025.nlp4dh-1.48/
27. Nikos Kontonasios et all (2025): AI-driven Extraction of Structured Data from a Large Newspaper Archive**, ERCIM NEWS 141, April.**
28. Alan Thomas et all (2024): Leveraging LLMs for Post-OCR Correction of Historical Newspapers, **The Third Workshop on Language Technologies for Historical and Ancient Languages** (LT4HALA) @ LREC-COLING, Torino, Italia. ELRA and ICCL.

29. ndrei-Marius Avram et all (2024): HistNERo: Historical Named Entity Recognition for the Romanian Language, Computer Science, vol 14806.
Available at: **https://doi.org/10.1007/978-3-031-70543-4_8**
Niels Bijl (2024): EXTRACTING BOOK TITLES FROM HISTORICAL NEWSPAPER ARCHIVES: A NAMED ENTITY RECOGNITION APPROACH, Master Thesis, Utrecht University, 7 July.
30. Yu Zhang et all (2024): Understanding the social construction of juvenile delinquency: insights from semantic analysis of big-data historical newspaper collections,, Archiving Conference Journal of Computational Social Science Vol.
Available at: **https://doi.org/10.1007/s42001-024-00254-x**
31. Christian Schultze et all (2024): Chronicling Germany: An Annotated Historical Newspaper Dataset, June 7.
Available at: https://arxiv.org/html/2401.16845v2
32. Brevin Franklin et all (2024) : News Déjà Vu: Connecting Past and Present with Semantic Search, Sixth Workshop on Natural Language, Processing and Computational Social Science (NLP+CSS), June 21.
Available at: https://doi.org/10.18653/v1/2024.nlpcss-1.8
33. Filip Dobranić, et all (2024): A Lightweight Approach to a Giga Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection, **International Conference on Computational Linguistics, Language Resources and Evaluation, Torino, Italia, May**.
Available at: https://aclanthology.org/2024.lrec-main.61/
34. Hermann Kroll et all (2023): Aspect-Driven Structuring of Historical Dutch Newspaper Archives, Theory and Practice of Digital Libraries, Computer Science, Vol 14241. Available at: **https://doi.org/10.1007/978-3-031-43849-3_4**
35. Leen-Kiat Soh et all (2023): Applying Image Analysis and Machine Learning to Historical Newspaper Collections, AMERICAN HISTORICAL REVIEW, ARTIFICIAL INTELLIGENCE FORUM.
Available at: https://academic.oup.com/ahr/article/128/3/1382/7282221
36. Melissa Dell et all (2023): American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers. 24 August.
Available at: https://doi.org/10.48550/arXiv.2308.12477
37. Marten Düring et all (2023): Impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers, Front. Big Data Vol. 6:1249469. 03 November
Available at: https://doi.org/10.3389/fdata.2023.1249469
38. LEA MARXEN (2023): Where Did the News come from? Detection of News Agency Releases in Historical Newspapers**, Master Thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE DIGITAL HUMANITIES LABORATORY, 18 August.**
Available at: https://infoscience.epfl.ch/entities/publication/0ca6e53d-d37e-4cdf-a360-2dd9f34a8271
39. Elvys Linhares Pontes et all (2022): MELHISSA: A multilingual Entity Linking Architecture for Historical Press Articles, **International Journal on Digital Libraries,** 23.
Available at: https://doi.org/10.1007/s00799-021-00319-6
40. Sandeep Puthanveetil Satheesan (2022): Toward a big data analysis system for historical newspaper collections research, PASC 22: Proceedings of the Platform for Advanced Scientific Computing Conference.
Available at: https://doi.org/10.1145/3539781.3539795
41. Filip Grali´ nski et all (2021): Challenging America: Digitized Newspapers as a Source of Machine Learning Challenges**, 35th Conference on Neural Information Processing Systems (NeurIPS).**
Availble at: https://openreview.net/pdf?id=gx7TEqAogg8
42. Marcel Broersma and Frank Harbers (2021) : Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency. In H. Bødker (Ed.), **Journalism History and Digital Archives Taylor & Francis Group.**

Available at: https://doi.org/10.4324/9781003098843-4

43. **Simon Hengchen and Nina Tahmasebi (2021): A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data,** Journal of Open Humanities Data, 7: 2.
Available at: **https://doi.org/10.5334/johd.22**
44. Estelle Bunout (2021): Grasping the Anti-Modern Discourse on Europe in the Swiss Digitised Press, or can Text Mining Generate a Research Corpus from an Article Collection? **Journal of Open Humanities Data, 7: 21**.
Avilable at: https://doi.org/10.5334/johd.37
45. Ahmed Hamdi et all (2021): A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers, *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11 July.
Available at: **https://doi.org/10.1145/3404835.3463255**
46. Houssem Menhour et all (2021): Searchable Turkish OCRed historical newspaper collection 1928–1942, **Journal of Information Science, 49(2).**
Available at: https://doi.org/10.1177/01655515211000642
47. Molly Brandt Skelbye (2021): OCR Processing of Swedish Historical Newspapers Using Deep Hybrid CNN–LSTM Networks, **Proceedings of Recent Advances in Natural Language** Processing, 1-3 September.
Available at: https://doi.org/10.26615/978-954-452-072-4_023
48. Nicolas Gutehrlé and Iana Atanassova (2021): Logical Layout Analysis Applied to Historical Newspapers, Conference papers, Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), 19-12
Available at: https://hal.science/hal-03468972
49. Till Grallert (2021): Catch Me If You Can!, Approaching the Arabic Press of the Late Ottoman Eastern Mediterranean through Digital History, **Geschichte und Gesellschaft, Vol. 47, Issue 1. June.**
Available at: https://doi.org/10.13109/gege.2021.47.1.58
50. Ghaith Dekhili and Fatiha Sadat (2020): Hybrid Statistical and Attentive Deep Neural Approach for Named Entity Recognition in Historical Newspapers, **4.0 International (CC BY 4.0), CLEF, 22-25 September, Thessaloniki, Greece.**
Available at: https://impresso-project.ch/
51. Maud Ehrmann et all (2020): Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers**, Experimental IR Meets Multilinguality, Multimodality, and Interaction Conference paper**, **Computer Science, Vol. 12260. Springer, Cham.**
Available at: https://doi.org/10.1007/978-3-030-58219-7_21
52. Lorella Viola and Jaap Verheul (2020): One Hundred Years of Migration Discourse in The Times: A Discourse-Historical Word Vector Space Approach to the Construction of Meaning, **Front. Artif. Intell. Vol. 3, Article 64. 10 September.**
Available at: https://doi.org/10.3389/frai.2020.00064
53. Bernhard Liebla and Manuel Burghardt (2020): From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline**, Workshop on Computational Humanities Research, November 18–20,, Amsterdam, The Netherlands.**
Available at: https://ceur-ws.org/Vol-2723/long20.pdf
54. Ehrmann, Maud et all (2020): Language Resources for Historical Newspapers: the Impresso Collection. In: Proceedings of the **12th Language Resources and Evaluation Conference, Marseille, European Language Resources Association (ELRA)**.
Available at: https://doi.org/10.5167/uzh-191270
55. Jani Marjanen et all (2020): Topic modelling discourse dynamics in historical newspapers,
Available at: https://doi.org/10.48550/arXiv.2011.10428

56. Giulia Taurino et all (2025): Copyright, Privacy, and Public Access in News Archives: a proof of concept on the Boston Globe photograph morgue, AI & Society 20 March.
Available at: **https://doi.org/10.1007/s00146-025-02208-x**
57. William Mocaër et all (2025): Hierarchical Structure Extraction from Newspaper Images Using a Transformer-Based Model
58. Conference "Newspapers, Magazines & AI Models: Training and (Re-)Use in the Digital Humanities, 7–8 May.
Available at: https://www.oeaw.ac.at/acdh/newsevents/events/conference-newspapers-magazines-ai-models-training-and-re-use-in-the-digital-humanities#c274272
59. Thomas Smits et all (2025): A Fully-Searchable Multimodal Dataset of the Illustrated London News, 1842–1890, **Journal of Open Humanities Data, 11: 10**.
Available at: DOI : https: https://doi.org/10.5334/johd.284
60. Adriana Rodríguez-Alfonso (2025): AI-Driven Analysis of Female Representations in Fin-de-Siècle Spanish Magazines, Conference "Newspapers, Magazines & AI Models: Training and (Re-)Use in the Digital Humanities, 7–8 May.
Available at: https://www.oeaw.ac.at/acdh/newsevents/events/conference-newspapers-magazines-ai-models-training-and-re-use-in-the-digital-humanities#c274272
61. Seojoon Lee et all (2024): Automatic Detection of Four-Panel Cartoon in Large-Scale Korean Digitized Newspapers using Deep Learning, Journal of Open Humanities Data, 10: 36.
Available at: https://doi.org/10.5334/johd.205
62. Precup A. (2024): "Instrumente IA în producția documentarului**, Symbolon, Vol. 25 No. 2 (47).**
Available at: https://doi.org/10.46522/S.2024.02.5.
63. MARINA GIARDINETTI et all (2024): The EyCon Dataset: A Visual Corpus of Early Conflict Photography**, Journal of Open Humanities Data, 10: 40.**
Available at: https://doi.org/10.5334/johd.213
64. Pack, Chulwoo (2023). "Enhancing Document Layout Analysis on Historical Newspapers: Visual Representation, Pseudo-Ground-Truth, and Downscaling" *ETD collection for University of Nebraska-Lincoln. AAI30485052*.
Available at: https://digitalcommons.unl.edu/dissertations/AAI30485052
65. Eva Sánchez-Salido et all (2023): Seeking Information in Spanish Historical Newspapers: The Case of Diario de Madrid (18th and 19th Centuries), **Digital Humanities Quarterly**, Volume 17.
Available t: http://digitalhumanities.org:8081/dhq/vol/17/4/000735/000735.html
66. Arnold, Taylor, et all (2023): "Understanding Peanuts and Schulzian Symmetry: Panel Detection, Caption Detection, and Gag Panels in 17,897 Comic Strips Through Distant Viewing." **Journal of Cultural Analytics, Vol. 8, No. 3, Sept**.
Available at: https://doi.org/10.22148/001c.87560.
67. Dilawar Ali et all (2022): Context-Aware Querying, Geolocalization, and Rephotography of Historical Newspaper Images, Applied Sciences, 2022, 12, 11063.
Available at: https://doi.org/10.3390/app122111063
68. Benjamin Charles Germain Lee (2022): The Digital Humanities and the Ladino Press: Using Machine Learning to Extract and Analyze Visual Content in Historic Ladino Newspapers, **Jewish Studies in the Digital Age, Studies in Digital History and Hermeneutics, Volume 5.**
Available at: https://doi.org/10.1515/9783110744828-010
69. Melvin Wevers et all (2022): What to do with 2.000.000 Historical Press Photos? The Challenges and Opportunities of Applying a Scene Detection Algorithm to a Digitised Press Photo Collection, **TMG Journal for Media History, Volume 25 No (1).**

Available at: https://dx.doi.org/10.18146/tmg.815

70. Melvin Wevers (2021): Scene Detection in De Boer Historical Photo Collection, the **13th International Conference on Agents and Artificial Intelligence (ICAART - Volume 1**
Available at: https://hdl.handle.net/11245.1/7a0b6c4a-ba46-4367-ba4a-f2258c9ac70a
71. Melvin Wevers and Thomas Smits (2020): The visual digital turn: Using neural networks to study historical images, **Digital Scholarship in the Humanities, Volume 35, Issue 1, April.**
Available at: https://doi.org/10.1093/llc/fqy085
72. Yann Soullard et all (2020): Multi-scale Gated Fully Convolutional DenseNets for semantic labeling of historical newspaper images, Pattern Recognition Letters 131,
Available at: https://doi.org/10.1016/j.patrec.2020.01.026
73. Marna Giardinetti et all (2024), Op., cit.
74. Till Grallert (2021), op., cit.
75. Marcel Broersma and Frank Harbers (2021), Op., cit.