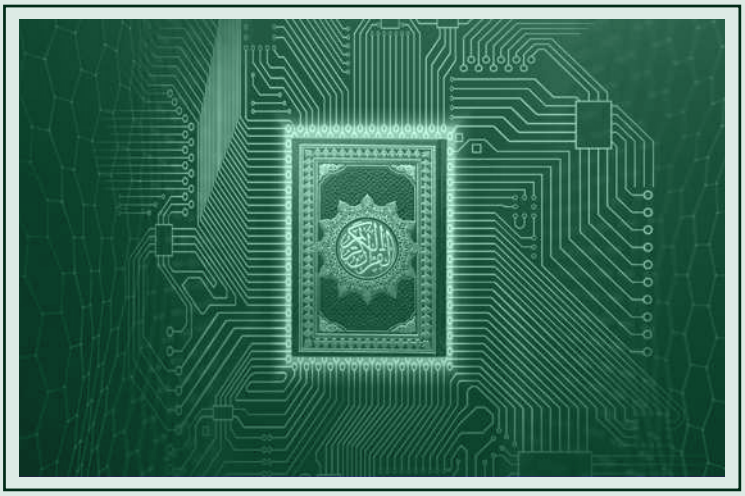
مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language

الإحصاءُ وتقنياتُ الذَّكاءِ الاصطِناعيّ للبياناتِ اللُّغويَّة

مُدَوَّنَةُ القُرآنِ الكَرِيمِ أُنمُوذجًا



سلطان بن ناصر المجيول

٢٨
الدراسات

مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language

الإحصاءُ وتقنياتُ الذَّكاءِ الاصطِناعيّ للبياناتِ اللُّغويَّة

مُدَوَّنَةُ القُرآنِ الكَرِيمِ أُنمُوذجًا

سلطان بن ناصر المجيول

٢٨
الدراسات

مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language

الإحصاءُ وتقنياتُ الذَّكاءِ الاصطِناعيّ
للبياناتِ اللُّغويَّة - مُدَوَّنَةُ القُرآنِ الكَرِيمِ أُنمُوذجًا

الطبعة الأولى
١٤٤٥هـ ٢٠٢٤م

البريد الإلكتروني: nashr@ksaa.gov.sa

ح / مجمع الملك سلمان العالمي للغة العربية، ١٤٤٥هـ
المجيول، أ.د. سلطان بن ناصر
الإحصاء وتقنيات الذكاء الاصطناعي للبيانات اللغوية - مدونة القرآن الكريم أنموذجًا. / أ.د. سلطان بن ناصر المجيول. - الرياض، ١٤٤٥هـ

٢٩٤ ص؛ ١٧ × ٢٤ سم

رقم الإيداع: ١٥٩٩٩/١٤٤٥
ردمك: ٩٧٨-٦٠٣-٨٤١٣-٩٢-٠

بسم الله الرحمن الرحيم

قائمة المحتويات

قائمة الجداول

قائمة الأشكال

مقدمة المجمع

ينشط مجمع الملك سلمان العالمي للُّغة العربيّة في مسارات عملٍ متنوعة، ويتولى مهامًا متعددةً تتّصل بنشر اللُّغة العربيّة، ودعمها، وتعزيز مكانتها، والمحافظة على سلامتها نطقًا وكتابةً، والنّظر في فصاحتها، وأصولها، وأساليبها، وأقيستها، ومفرداتها، وقواعدها، وتيسير تعلُّمها بداخل المملكة العربيّة السعوديّة وخارجها؛ لتواكب المتغيرات في جميع المجالات، ويتمثّل طموح المجمع في أن يصبح مجمعًا متميزًا يخدم اللُّغة العربيّة، وينطلق من قلب العالم الإسلامي والعربي، ومن مهد العروبة الأول، وأن يصبح رائدًا ومرجعيّةً عالميّةً في مجال اللُّغة العربيّة وتطبيقاتها المتنوعة.

وضمن توجيهات سمو وزير الثّقافة، رئيس مجلس الأمناء الأمير بدر بن عبد الله بن فرحان آل سعود -حفظه الله- في دعم أعمال المجمع، وبرامجه: العلميّة، والثّقافيّة، والبحثيّة، أطلق المجمع مشروعَ (المسار البحثي العالمي المتخصّص)؛ لتلبية الحاجات العلميّة، ومواجهة المشكلات اللُّغويّة، وسدّ الفجوات المتعلّقة بالبحث والنّشر العلمي، وفتح الآفاق العلميّة والمعرفيّة المتنوعة، واستكمال مسارات النّشر اللُّغوية المتخصّصة.

ويهدف المشروع إلى تعزيز دور المجمع، وإيصال رسالته؛ بتغطية مساحاتٍ متنوعة من التخصّصات، والفنون المتعلقة باللُّغة العربيّة، وإثراء المحتوى العلمي ذي العلاقة بمجالات اهتمام المجمع، ودعم الإنتاج العلمي المتميّز وتشجيعه، وفتح المجال أمام الباحثين والمختصّين، وتوثيق صلتهم بالمجمع؛ وذلك بإشراكهم في أعمال هذا المشروع.

ويضمُّ المشروع مجالاتٍ بحثيّة متنوعة، ويغطي الموضوعات التي تعزّز موقع العربيّة ضمن اللُّغات الحضاريّة العالميّة، ومن أبرزها: (دراسات التّراث اللُّغوي العربي وتحقيقه، والدّراسات حول المعجم، وقضايا المصطلح، وقضايا الهويّة اللُّغويّة، ومكانة العربيّة وتعزيزها، واللّسانيّات التّطبيقيّة، والتّخطيط اللُّغوي، والسّياسة اللُّغويّة، واللّسانيّات الحاسوبيّة، والتّرجمة، والتّعريب، وتعليم اللُّغة العربيّة للنّاطقين بها وبغيرها، والدّراسات البينيّة).

وقد بدأ المشروع باستقبال الدّراسات النّوعيّة الجادّة، وتواصل مع: (المختصّين، والباحثين، والمؤسّسات العلميّة داخل المملكة العربيّة السعوديّة وخارجها)، ودعاهم إلى المشاركة في المشروع، واتخذ الإجراءات المتصلة بتحكيم الأعمال والنظر في جديتها وأصالتها ومدى إضافتها للمكتبة العربية واستنادها إلى المعايير المتعارف عليها في البحث والمنهج والتوثيق قبل طباعتها ونشرها.

ويهدف هذا الكتاب (الإحصاءُ وتقنياتُ الذَّكاءِ الاصطِناعيّ للبياناتِ اللُّغويَّة، مُدَوَّنَةُ القُرآنِ الكَرِيمِ أُنمُوذجًا) إلى إثراء المكتبة العربيَّة بالكتب العلميَّة في مجال اللسانيَّات الحاسوبيَّة، ويسلّط الضّوء على العديد من القضايا المتعلّقة بالإحصاء اللُّغوي للبيانات اللُّغوية، وبتقنيات الذكاء الاصطناعي المرتبطة بتنقيب البيانات اللُّغوية.

ويجتهد المجمع في انتقاء الكتب التي يكون في نشرها إضافة معرفية نوعية، ويأمل أن يكون هذا الكتاب مفتاحًا لمشروعات علمية وعملية ويحقق إثراء معرفيًا لافتًا.

ويشكر المجمع مؤلف الكتاب سعادة الأستاذ الدكتور سلطان بن ناصر المجيول؛ لما تفضل به من عمل علمي جاد، ويدعو الباحثين إلى التواصل مع مشروعات المجمع، ومنها: مسار البحوث والنشر العلمي؛ للمشاركة فيه والمساهمة في إثرائه.

الأمين العام للمجمع

أ.د. عبد الله بن صالح الوشمي

المقدمة

قبل أن أُشيرَ إلى الهدف من هذا الكتابِ وإلى محتوياتِهِ المتعلقة بالإحصاء اللغوي للبيانات اللغوية في الباب الأول، وبتقنيات الذكاء الاصطناعي المرتبطة بتنقيب البيانات اللغوية في الباب الثاني؛ أودّ أن أُلفتَ القارئ إلى نقطةٍ مهمَّةٍ، أَجدُها غامضةً عند كثيرين في العالم العَرَبِيّ، تتعلَّقُ بطبيعةِ تخصُّصِ علمِ الإحصاء من جهةٍ أُولى، وتخصُّص معالجة اللغةِ الطَّبيعِيّةِ وعِلْم البَيَانَاتِ من جهةٍ ثانيةٍ، وهي طبيعةٌ تجري في الجامعاتِ الرَّائدة، والمراكز العِلْمِيّة والبحثِيَّة المميَّزة عالميًّا في هذين المجالين. ما يجري في هذين الاختصاصَين هو أنَّ علمَ الإحصاء مجالٌ تطبيقيٌّ، إنْ استخدمَه اللُّغَوِيُّ أو الطَّبيبُ أو الاقتصاديُّ، إلى آخِرِهِ؛ فإنَّه يُجريه بالتَّطبيق في مجالِه، وأنَّ معالجة اللغة الطَّبِيعِيّة وعِلْم البَيَانَاتِ مجالٌ تطبيقي، إنْ دخل فيه غيرُ المتخصِّص في علوم الحاسب؛ فإنه يُجريه بالتَّطبيق في مجاله، سواءً كان لُغَوِيًّا، أم اقتصاديًّا، أم خبير تسويقٍ، إلى آخِرِهِ. ومن مَهَمَّة المتخصِّصين في علم الإحصاءِ، أو المتخصِّصين في علوم الحاسبِ (المجال الفرعي: عِلْم البَيَانَاتِ أو معالجة اللغة الطَّبِيعِيّة)؛ أنْ يُطوِّروا، ويبتكروا، ويُوجدوا المفاهيمَ البنائيّةَ المتَقَدّمة والمستجدَّة في هذين المجالَين، ولغيرهم ممَّن هم ليسوا متخصِّصين في علوم الحاسبِ تعلُّمُ تطبيقاتِ هذين المجالين، وإجرائِها فيما يخدمُ اختصاصاتِهم ومشاريعَهم.

يستهدفُ هذا الكتابُ المتخصِّصين والباحثين في العلوم اللُّغَوِيّة والإنْسَانِيّة والاجْتِمَاعِيّة، ما دامت المادَّةُ محلَّ التَّحْلِيل لديهم هي اللغَةُ ونصوصُها المكتوبةُ والمنطوقة. ظهرت دوافعُ تأليف هذا الكتاب من عدم وجودِ مصدرٍ عربيٍّ في مجال الإحصاءِ اللُّغَوِيّ أوَّلًا، ثم مجال تنقيب النُّصُوص والبَيَانَات العَرَبِيَّة ومُعالَجتها ثانيًا،

وكِلا المجالَـين مكمِّـلٌ للآخـر، فالممارسـاتُ الدقيقـة في الإحصـاء اللُّغَـوِيّ بـدءًا مـن إعـداد البَيَانَـات؛ تعتمـد اعتمـادًا كبـيرًا علـى الممارسـاتِ الدَّقيقـة في تنقيـب النُّصـوص العَرَبِيَّـة ومعالجتِهـا معالجـةً أساسـيَّةً دقيقـةً. ومـن دوافـع تأليـفِ هـذا الكتـابِ -أيضًـا- الرَّغبـةُ في تسـهيل موضوعـاتِ هذيـن المجالـين، وفي فتـحِ أبـواب الممارسـات العِلْمِيّـةِ الرَّقَمِيّـة الـتي باتـت اليـومَ محـورَ المشـاريع البحثيَّـة في العُلُـومِ اللُّغَوِيّـةِ والإنسـانيّةِ والاجتماعيَّـةِ في الجامعـاتِ والهيئـات والمؤسَّسـات العالَمِيَّـة.

وينقسـمُ هـذا الكتـابُ إلى بابـين أساسـيَّين؛ البـاب الأَوَّل: الإحصـاء اللُّغَـوِيّ، والبـاب الثَّـانِي: تنقيـب النُّصُـوص والبَيَانَـات العَرَبِيَّـة. وفي كلِّ بـابٍ فصـولٌ ومباحـثُ تتنـاولُ خطـواتِ التَّحليـلِ بـأدواتٍ قائمـةٍ بذاتهـا، لا تتطلَّـب مـن القـارئِ معرفـةً بلغـةِ البرمجـةِ (آر) للإحصـاء اللُّغَـوِيّ (البـاب الأوَّل)، أو بلغـة البرمجـة (البايثـون) لتنقيـب البَيَانَـات العَرَبِيَّـة ومعالجـة اللغـة العَرَبِيَّـة الطَّبِيعِيّـة (البـاب الثَّـاني).

والمصـادر العَرَبِيَّـة تخلـو مـن الفائـدةِ للقـارئ العَـرَبِيّ، فضـلًا عـن المتخصِّـص في العلـوم الإِنْسَـانِيّةِ والاجْتِمَاعِيّـة؛ بتطبيقـات هذيـن المجالـين علـى نصـوصٍ عربيـةٍ. وعلـى الرّغـم مـن المحـاولاتِ في ترجمـة بعـض الكتـب أو المقـالات عـن الممارسـات البحثيَّـة والأسـاليب التَّحْلِيليَّـة التَّطْبِيقِيّـة في هذيـن المجالـين مـن الإنجليزيَّـة إلى العَرَبِيَّـة؛ لا يـزال المحتـوى يكتنفُـه الغمـوضُ؛ بسـبب الجوانـبِ النظرِيَّـةِ المعَقَّـدةِ، والشُّـروحات الرياضيَّـة للمعـادلات أو الخوارزميَّـاتِ، كحـالِ كتـاب (علـم الإحصـاءِ في لُغَوِيّـات المدوّنـة الحاسـوبِيَّة: دليـلٌ عملـيٌّ، المؤلـف: Vaclav Brezina)، الـذي ترجمتُـه ونشـرتُه في عـام ٢٠٢١م. ولـن يتنـاولَ هـذا الكتـابُ شـروحاتٍ نظرِيَّـةً، بـل سـيتناولُ خطـوات الأسـاليب التَّطْبِيقِيّـة الإِحْصَائِيّـةِ بـأدوات (لانكسـتر) الإِحْصَائِيَّـة في فصـولِ البـاب الأوَّل، وخطـوات تنقيـب النُّصُـوص ومعالجـة البَيَانَـات للغـةِ العَرَبِيَّـة بـأدوات (أورانج) في فصـول البَـاب الثَّـانِي، وذلـكَ باسـتخدام مُدَوّنَـة القـرآن الكريـمِ. إذن؛ هـذا الكتـابُ هـو مدخـلٌ علـميٌّ وتطبيقـيٌّ أسـاسٌ، إيمانًـا مـنِّي بـأنَّ البـدءَ بالكتـبِ التَّطبيقيَّـة يفتـحُ مغاليـقَ الكتـبِ النظريَّـةِ والعِلْمِيّـة أسـرع مِـن لـو كان الحـالُ

على عكس ذلكَ. ولقد ضمَّنتُ أكوادَ لغة البرمجة (آر) للإحصاءات اللُّغَوِيّة في ملحق الكتاب، تحسّبًا لتوقف أدوات (لانكستر) الإِحْصَائِيَّة (الشبكية) عَن العمل، أما أكوادُ تنقيب النُّصُوص العَرَبِيَّة ومعالجة اللغة العَرَبِيَّة الطَّبِيعِيّة المدمجة في أدوات (أورانج)؛ فيُمكنُ الوصولُ إليها عبر https://github.com/biolab/orange3، أو في موقعِ الأداة نفسِها https://orangedatamining.com في خياراتِ الوثائقِ (Docs). وفي ظهور هذه الأدوات القائمة بذاتها مزيّةٌ، كونها تخدم الباحثين في العلومِ ممَّن ليس لهم معرفةٌ بَرْمَجيَّةٌ.

وقد اخترتُ القرآنَ الكريمَ؛ ليكون مادَّةَ التَّطبيقِ لهذين المجالين لسببين؛ الأَوَّل: كونهُ مُدَوّنَةً معلومةً لكلِّ قارئٍ عربيٍّ، يعرفُ عددَ سورها، ويَعرفُ آياتِهَا وكلماتِهَا، وكونه نَصًّا مُبِيْنًا وتنزيلًا حكيمًا من عند الله -عزَّ وجلَّ- تألفُه ألسنتُنا منذ نعومةِ أظافرنا بالتعبُّد والتِّلاوةِ. والثَّاني: لأنَّ التَّركيزَ على أساليبِ الإحصاء اللُّغَوِيّ وأدواتِ تنقيبِ النُّصُوص والبَيَانَات هو الغايةُ؛ حتَّى تسهل معرفتُها، ويَتَغيَّا القارئُ نيلَ تطبيقَاتِها على أيِّ مُدَوّنَة لُغَوِيَّةٍ يريد جمعَها وتحليلها مستقبَلًا وفقَ اهتماماتِه، وما ذلكَ إلا لأجل تطويرِ البحث اللُّغَوِيّ العَرَبِيِّ باستخدام البَيَانَات اللُّغَوِيّة الرَّقَمِيّة، التي صارت محورَ الحياة المستَقْبِلَة اليومَ، وستكونُ كذلك في العقودِ القادمة.

عَرضتُ في البَاب الأَوَّلِ (الإحصاءُ اللُّغوِيّ) تمهيدًا لتعريف أهَمِّيَّة المُدَوّنَات اللُّغَوِيّة والبَيَانَات اللُّغَوِيّةِ؛ بوصفِها علمًا، وبوصفها منهجًا وتطبيقاتٍ وأساليب صارت تُمارسُ دورًا محوريًّا في فروعِ اللُّغَوِيّات واللُّغَوِيّات التَّطْبِيقِيّة والعلوم الاجْتِمَاعِيّة والإِنْسَانِيّة، ثم لتعريف مصادر المُدَوّنَة ومعايير الجمعِ والأدواتِ. تلا التَّمهيدَ فصولُ تطبيقاتِ علم الإحصاءِ على مُدَوّنَة القرآن الكريمِ وفقَ ما يَأتِي:

الفصل الأَوَّل: البَيَانَات اللُّغَوِيّة (أنواع الإحصاءات وأنواع تمثيل البَيَانَات).

الفصل الثَّانِي: تحليل الخطابِ (مُعاملات وشبكات تحليل التَّصاحب وحساب موافقةِ المقدّر البينيّ).

الفصل الثَّالِث: الارتباط والتَّجْمِيع ومقارنة المجموعات.

الفصل الرَّابع: الانحدار الخطّيّ والانْحِدَار اللُّوجسْتِيّ

الفصل الخامس: تحليل البَيَانَات اللُّغَوِيّة التَّاريخِيّة.

وعرضتُ في البَاب الثَّانِي (تنقيب النُّصُوصِ والبَيَانَات العَرَبِيَّة) تمهيدًا للمعالجاتِ الأساسيَّة للغة العَرَبِيَّة الطَّبِيعِيّة، وقياس أدائِها بمصفوفة الارتباك، وتطبيقاتِ معالجة اللغة العَرَبِيَّة الطَّبِيعِيّة المتعدِّدة كاستخراج النُّصُوص، وتلخيص النُّصُوص، والتَّعرُّف إلى الكياناتِ، وتصنيف النُّصُوص، وتحليل المشاعر، وكشف النِّيَّة والأخبار الزَّائفة، وروبوتات المحادثة، والدَّعم الافتراضيّ، والتَّصحيح الآليّ، والتَّشكيل الآليّ، والتَّعرّف إلى الكلام، والتَّرجمة الآليَّة. تلا ذلك فصولُ الباب التَّطبيقيَّة على مُدَوّنَة القرآن الكريم، وذلك وفقَ التَّسلسل الآتي:

الفصل الأَوَّل: تنقيب النُّصُوص (إعداد البَيَانَات اللُّغَوِيّة، ثم الإحصاءات، ثم تكرار الكلمة-تكرار المستند العكسيّ).

الفصل الثَّانِي: حقيبة الكَلِمَات، وتضمين الكَلِمَات.

الفصل الثَّالِث: استخراج الموضوعات، واستخراج الكَلِمَات المفتاحيَّة.

الفصل الرَّابع: الأنماط النَّحويَّة.

الفصل الخامس: تنقيب البَيَانَات للنُّصوص، ونماذج التَّصنيف، والانْحِدَار الخَطِّيّ.

ونظرًا إلى عدم إمكانيَّة تمثيل الكَلِمَات العَرَبِيَّة في البَيَانَات باستخدام لغة (آر)؛ فقد نُقحرت بالحروف الأجنبيَّة أو تُرجمت، على عكس لغة (البايثون). ولذلك ستجد أن اللغة العَرَبِيَّة منقحرة في بيانات (الآر)، وغير منقحرة في بيانات (البايثون).

وللاطِّلاع على البَيَانَات التي أَعْدَدتُها للتحليل في الباب الأول (الإحصاء اللُّغَوِيّ) وفي الباب الثاني: (تنقيب النُّصُوصِ والبَيَانَات العَرَبِيَّة)، ومن أجل استخدامها للتَّدريبِ العَمَليِّ وللتَّنَاتُجِ (تطابق النتائج)؛ يُمكن العودةُ إلى الرابطَين الآتِيَين:

بيانات الباب الأَوَّل: www.kaggle.com/datasets/sultanalmujaiwel/statistics-in-quran

بيانات الباب الثَّانِي: www.kaggle.com/datasets/sultanalmujaiwel/quran-mining

وأخيرًا؛ أرجو أن أكونَ قد وُفِّقت في التَّسهيل، فإن وُفِّقت، فذاك خيرٌ من الله.

المؤلف

الباب الأَوَّل

الإحصاء اللُّغَوِيّ[١]

(١) يُمكـنُ للقـارئ تطبيـقُ كُلِّ الإحصـاءاتِ اللُّغويـةِ الـواردةِ في فصـول هـذا البـاب، باسـتخدام أدوات لانكسـترا الإحصَائِيَّـةِ وباسـتخدام البيانـات اللغويـة الـتي أعدّهـا المؤلـف، مـن أجـل تكـرار التَّحليـلِ والتَّدريـبِ والتأكّـدِ مـن تَطابقِ النَّتيجة. ولتسـهيلِ العـودةِ إلى بَيَانـاتِ كلِّ مُعالجةٍ آليـةٍ في فصـولِ هـذا البـاب، فقـد جَمعتُهـا في ملـفٍ واحـدٍ (إكسـل Excel)، وسَـمّيتُ كُلَّ ورقـة sheet داخـل هـذا الملـف برقـم الشـكل الـوارد في فصـول هـذا البـاب. موقـع البيانـات اللغويـة:
.www.kaggle.com/datasets/sultanalmujaiwel/statistics-in-quran

التَّمهيد

تعـدَّدت مصطلحـات (corpus) في العَرَبِيَّـة، فظهـرت الذَّخـيرة والمكـنز والمدوّنـة، ثـم اسـتقرّ المصطلـح الأخـير في مشـاريع الـمُدَوّنَة اللُّغَوِيّـة الـتي ظهـرت في السَّـنوات الماضيـة، وظهـرت أبحـاثٌ في مصادرهـا وبنائهـا وأسـاليب توظيفِهـا في البحـثِ اللُّغَـوِيّ والتَّعْلِيـم. وشـاع مفهـوم الـمُدَوّنَة اللُّغَوِيّـة بمجـرد ظهـور أدواتٍ سـهلة الاسـتخدام، تقـوم علـى إظهـارِ تكـرارات الكَلِمَـات والكشـافات السـياقيَّة، وبمجـرد إنشـاء مُدَوّنَـات شـبكية تحتـوي علـى النُّصُـوص الخاصـة بهـا، فظُـنّ بذلـك كلِّـه أن هـذا المجـال يرتبـط بـأدواتٍ محـددةٍ فقـط؛ كأداة غـواص[١] في العالـم العَـرَبِيّ، أو أدوات (وورد سـميث[٢] WordSmith Tools) في العالـم الغـربِيّ، أو بالـمُدَوّنَات اللُّغَوِيّـة الشـبكيَّة الموجـودة والقائمـة بذاتهـا؛ مثـل الـمُدَوّنَة اللُّغَوِيّـة العَرَبِيَّـة لمدينـة الملـك عبدالعزيـز للعلـوم والتّقنيَّـة[٣]، والـمُدَوّنَة اللُّغَوِيّـة العَرَبِيَّـة الدوليَّـة[٤]، ومدوّنـة عـربِيّ[٥] (arabiCorpus). كمـا شـاع -أيضًـا- اعتقـادٌ آخـر، وهـو أن لُغَوِيّـات الـمُدَوّنَات مقتصـرةٌ علـى معالجـة الـمُدَوّنَة اللُّغَوِيّـة مـن أجـل بنـاء المعاجـمِ، وهـو اعتقـادٌ متوقّـعٌ، نتـج عـن طبيعةِ مـا تقـوم بـه الـمُدَوّنَات وأدواتهـا مـن اسـتخراجٍ للكلمـاتِ والعبـاراتِ والتّكْـرَاراتِ والكشـافات السـياقيَّة، إلا أنَّ واقـع أبحـاث الـمُدَوّنَات اللُّغَوِيّـة ولغويَّـات الـمُدَوّنَات في المشـاريع العالميَّـة قـد تطـوّر كثـيرًا بحكـم التَّطـوُّر الحاصـل في مشـاريع معالجـة اللغـة الطَّبِيعِيّـة والإنسـانيَّات

(١) غـواص: أداة لمعالجـة الـمُدَوّنَات اللُّغَوِيّـة العَرَبِيَّـة (النسـخة ٤,٦). تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢، مـن: https://sourceforge.net/ghawwasv4.

(2) Scott, M. (2020). WordSmith Tools version 8 [Lexical Analysis Software]. Retrieved October 24, 2022, from https://lexically.net/wordsmith

(٣) مدينـة الملـك عبدالعزيـز للعلـوم والتقنيـة. الـمُدَوّنَة اللُّغَوِيّـة العَرَبِيَّـة. تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢، مـن: https://corpus.kacst.edu.sa.

(٤) مكتبـة الإسـكندرية. الـمُدَوّنَة اللُّغَوِيّـة العَرَبِيَّـة الدوليـة. تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢، مـن: https://www.bibalex.org/ica/ar/login.aspx.

(5) Parkinson, D.)2001(. arabiCorpus. Retrieved October 24, 2022, from https://arabicorpus.byu.edu.

الرَّقَمِيّة في السَّنوات الأخيرة، فصارت الـمُدَوّنَات اللُّغَوِيّة، والبَيَانَات اللُّغَوِيّة، وتطبيقات معالجة اللغة الطَّبِيعِيّة؛ أدواتٍ بحثيةً حاسوبيَّةً، يُستفاد منها في كلّ العلوم اللُّغَوِيّة والإِنْسَانِيّة والاجْتِمَاعِيّة.

لا غرابةَ من اقتصار مجال الـمُدَوّنَة اللُّغَوِيّة في هذه الأَيَّام على أدواتٍ معينةٍ دون غيرها، أو على مُدَوّنَةٍ لُغَوِيَّةٍ شبكيَّةٍ دون مُدَوّنَةٍ أُخرى، أو على المقاييس الأوليَّة لتكرارات الكَلِمَات ومقاييسِ التَّصاحب؛ لأنَّ معظم الأبحاث العَرَبِيَّة إلى يومنا هذا مقيدةٌ بهذه الحدودِ، وهو قيدٌ يعود سببه بالدرجة الأولى إلى عدم وجود مصادر عربيَّة تُفيد المهتمِّين بتطبيقاتِ لُغَوِيّات الـمُدَوّنَات ومعالجة اللغة الطَّبِيعِيّة.

سأُعرِّف الـمُدَوّنَة اللُّغَوِيّة، وسأذكرُ مصادرَها المتعدِّدة، وطرائقَ بنائها المختلفة باختلاف أنواع الـمُدَوّنَات، وسأُعرّج إلى أفضل أدوات معالجة الـمُدَوّنَات اللُّغَوِيّة العَرَبِيَّة، وسأكتفي في التطبيقات التحليليَّة في هذا الباب بأداة واحدة[1]؛ وهيَ أدوات لانكستر الإِحْصَائِيَّة على الشبكة (Lancaster Stats Tools Online)، إلى جانب أداة غَوَّاص في قليلٍ من المواضع.

وفضّلتُ هذِهِ الأدواتِ دون غيرها؛ لأَنَّنا لو نظَرنا إلى جميع أدواتِ معالجة الـمُدَوّنَات اللُّغَوِيّة؛ لألفيناها أقلَّ من هذه الأدوات من ناحية تنوُّع خياراتِ المعالجة، كأن تكون أقل من حيث مُعاملات الإحصاء للتَّصاحب اللَّفْظِيّ، أو من حيث التَّحليل العام للمدونة اللُّغَوِيّة لتكرار الكَلِمَات والتّكرارات النِّسْبِيّة لها، ونسبة الكَلِمَات النَّوْعِيّة إلى الكَلِمَات الفِعْلِيّة (type-token ratio)، أو من حيث عدم إتاحة خاصِّيَّة حفظ نتائج كلِّ وظائف وخيارات المعالجة في الأداة. وستُستخدم في عرض جُلِّ أساليب التَّحليل الإِحْصَائِيّ في فصول

(١) حول أدوات معالجة الـمُدَوّنَة اللُّغَوِيّة العَرَبِيَّة الأخرى، انظر: المجيول، سلطان. البحث اللُّغَوِيّ في الـمُدَوّنَات العَرَبِيَّة الحَاسُوبِيَّة بين الممكن والمحتمل والمأمول. في: الـمُدَوّنَات اللُّغَوِيّة العَرَبِيَّة بناؤها وطرق الإفادة منها. مركز الملك عبدالله الدوليّ لخدمة اللغة العَرَبِيَّة (مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللُّغَوِيّة حاليًّا)، الرياض، الطبعة الأولى، ٢٠١٥، الصفحات ٢٤٣-٢٥٥.

هـذا البَـاب أدواتُ لانكسـترا الإِحْصَائِيَّـة (Lancaster Stats Tools) دون بقيَّـة الأدوات؛ لأنَّ الهـدفَ -كما أسـلفت القـول في المقدمـة- هـو إفـادةُ القـارئ بأفضل إجـراءات التَّحليل وأيسـرها ممارسـةً، وأدواتُ لانكسـترا الإِحْصَائِيَّـة مـن أكـثر الأدوات شـموليّةً، ولا حاجـةَ إلى اسـتعراض اسـتخدامات كلِّ الأدوات الـتي تنقُصُهـا الإمكانـاتِ الإحصائيـة المهمّـة لمعالجـة الـمُدَوّنَات اللغويّـة.

أَوَّلًا: تعريف الـمُدَوّنَة اللُّغَوِيّة

ليـس هنـاك اختـلافٌ في تعريـف الـمُدَوّنَة اللُّغَوِيّـة[1]. هـي ببسـاطةٍ نصـوصٌ لُغَوِيَّـةٌ مكتوبـةٌ أو منطوقـةٌ مُحوسَـبةٌ، وكلمـة (محوسـبة) هـي أسـاسُ المدوّنَـة اللُّغَوِيّـة. أمـا لُغَوِيّـات الـمُدَوّنَة (الحَاسُـوبِيَّة)؛ فهـي العلـم الـذي يـدرسُ ويحلِّـلُ اللغـةَ الطَّبِيعِيّـة. وقـد مـرّت الـمُدَوّنَاتُ اللُّغَوِيّـة بثـلاث مراحـل منـذ ظهـور أوَّل مُدَوّنَـةٍ لُغَوِيَّـة في عـام ١٩٦١م، وهـي مُدَوّنَـةُ (بـراون) المتاحـة في اللُّغـة الإنجليزيَّـة، الـتي احتـوت علـى مليـون كلمَـة[2].

المرحلـة الأولى: مرحلـة التَّأسـيس الـتي قامـت علـى جمـعِ النُّصُـوص اللُّغَوِيّـة كعينـاتٍ لُغَوِيَّـة. وأقـرب مثـالٍ لهـذه المرحلـة للغـة العَرَبِيَّـة هـو مُدَوّنَـة[3] (arabiCorpus).

المرحلـة الثَّانيـة: مرحلـة تطويـر أدوات المعالجـة، إمّـا في الـمُدَوّنَات اللُّغَوِيّـة أو للمدوّنـات اللُّغَوِيّـة. والفـرق بينهمـا هـو أن الأدوات في الـمُدَوّنَات اللُّغَوِيّـة تكـون مدمجـةً داخـل الـمُدَوّنَة الشـبكيَّة، وهـي أدواتٌ لا تُتيـح إلا البحـث في عيِّنـات نصـوص تلـك الـمُدَوّنَة؛ مثـل الـمُدَوّنَة

(١) حـول هـذه التعريفـات، انظـر: صالـح، محمـود إسـماعيل. (٢٠١٥). الـمُدَوّنَات اللُّغَوِيّـة وكيفيـة الإفـادة منهـا. في: الـمُدَوّنَات اللُّغَوِيّـة العَرَبِيَّـة بناؤهـا وطـرق الإفـادة منهـا. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حاليًـا)، الريـاض، الطبعـة الأولى، ٢٠١٥، الصفحـات ١٩-٢١.

(٢) مكانـري، تـوني، وأنـدرو هـاردي. لُغَوِيّـات الـمُدَوّنَة الَحاسُـوبِيَّة: المنهـج والنظريـة والتطبيـق. ترجمـة: سـلطان المجيـول. الريـاض: دار جامعـة الملـك سـعود للنشـر، ٢٠١٦، صفحـة ١٨٢.

(3) Parkinson, D. (2001). arabiCorpus. Retrieved October 24, 2022, from https://arabicorpus.byu.ede.

اللُّغَوِيّة العَرَبِيَّة لمدينة الملك عبدالعزيز للعلوم والتقنية[1]. أما الأدوات للمدوّنات اللُّغَوِيّة؛ فهي تطبيقاتٌ قائمةٌ بذاتها، وقد بُنيت فيها أدواتُ معالجة الـمُدَوّنَة دون نصوص، وتحتاجُ إلى أن يجمعَ المحلِّلون تلك النُّصُوصَ لغرض التّحليل، وتُسمّى هذه الـمُدَوّنَات التي تُجمع من قبل المستخدمين بمدونة الاكتفاء الذاتيّ (do-it-yourself corpus)، وتُختصر غالبًا بـ(DIY corpus)، ومثل هذه الأدوات أداةُ غواص[2] (الشكل ١). ونجد في أدوات هذه المرحلة خاصيّة الاستعلام والبحث عن الكلمةِ وحدها، أو عن الكلمة ومتصاحباتها، وعن تكراراتها الحقيقِيَّة والنِّسْبِيّة، وعن إظهارِ الكشافات السياقية (أسطر تَتَضَمَّن الكلمة المراد تكشيفها والكَلِمَات السابقة واللاحقة لها)، وعن إحصاءات التَّصاحباتِ اللَّفْظِيّة.

شكل ١: واجهة أداة غواص (أداة لمعالجة الـمُدَوّنَات اللُّغَوِيّة العَرَبِيَّة)

(١) مدينة الملك عبدالعزيز للعلوم والتقنية. الـمُدَوّنَة العَرَبِيَّة (الـمُدَوّنَة اللُّغَوِيّة العَرَبِيَّة لمدينة الملك عبدالعزيز للعلوم والتقنية). تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://corpus.kacst.edu.sa.

(٢) غواص: أداة لمعالجة الـمُدَوّنَات اللُّغَوِيّة العَرَبِيَّة (النسخة ٤,٦). تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://sourceforge.net/ghawwasv4.

المرحلة الثَّالثة: مرحلة البَيَانَات اللُّغَوِيّة واللوحات التَّفاعليَّةِ، وهي مرحلةٌ تعتمدُ على ما يزيدُ على أدوات المعالجة في المرحلةِ الثانية. من أمثلة أدواتِ معالجة البَيَانَات اللُّغَوِيّة أدوات (لانكستر) الإِحْصَائِيّة[1]، التي اعتمدت على لغةِ البرمجة (R) في التَّحْلِيل اللُّغَوِيّ الإِحْصَائِيّ، ووُفّرت كأداة شبكية جاهزة لا تتطلب معرفةً أو استخدامًا لأكواد (R). أما من أمثلة اللوحاتِ التَّفَاعُلِيّة فأدوات (فويانت[2] Voyant Tools) (الشكل ٢)، فبمجرد تحميل أو لصق النَّصِّ في هذه الأداة والبدء بمعالجتِها، تُظهر معلوماتٍ عديدةً عن النصِّ الخامِّ، مثل الارتباطاتِ بين الكَلِمَات، واستخراج الموضوعاتِ، وسحابة الكَلِمَات.

شكل ٢ : واجهة فوينت وواجهة بعض النَّتائج لتحليل مُدَوّنَة القرآن الكريم

ثَانِيًا: أهمِّيَّة المُدَوّنَات اللُّغَوِيّة

للمدوّنات اللُّغَوِيّة أهمِّيَّةٌ للدِّراساتِ التَّجريبيَّة (experimental)، والعمليَّة (empirical)، كيف لا ومادَّتها اللغة التي تُعدّ الوسيطَ الأساسَ للمعارف، سأوجزُ في الآتي الاتّجاهاتِ

(١) أدوات (لانكستر) الإِحْصَائِيَّة الشبكية. تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: http://corpora.lancs.ac.uk/stats/toolbox.php. شُرحت التَّطْبِيقَات الإِحْصَائِيَّة لهذه الأدوات في كتاب ألّفه (Vaclav Brezina)، وقد تُرجم إلى العَرَبِيَّة. انظر: بريزينا، فاكلاف. علم الإحصاء في لُغَوِيّات المُدَوّنَة الحَاسُوبِيَّة. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، الطبعة الأولى، ٢٠٢١.

(2) Sinclair, S. & Rockwell, G. (2022). Voyant Tools [v. 2.5.3]. Retrieved October 24, 2022, from https://voyant-tools.org.

الحديثـةَ منـذ مطلـعِ القـرن الحـادي والعشـرين لِلـدَّورِ الأسـاس الـذي تؤدِّيـه الـمُدَوّنَاتُ اللُّغَوِيّـة في العلـوم.

أ. اللُّغَوِيّات التَّطْبِيقِيّة والتَّخْطِيط اللُّغَوِيّ والسياسة

للغويـات التَّطْبِيقِيّـة فـروعٌ تتمثَّـل في اللُّغَوِيّـات التَّعْلِيمِيّـةِ، وتعليميّـة اللغـاتِ، ومـا يتعلَّـق بهـا مـن مـواد التَّعْلِيـم وفعـلِ التَّعْلِيـم، وفي اللُّغَوِيّـات النفسـيَّة، ومـا يتعلَّـق بهـا مـن دراسـاتٍ للجوانـب العقليَّـة، ولجوانـب معالجـة الدِّمـاغ عنـد تمثيـل اللغـة وتهيئـة الـكلام، وفي اللُّغَوِيّـات الاجْتِمَاعِيّـةِ، ومـا يتعلَّـق بهـا مـن دراسـاتٍ عـن اللغـةِ في المجتمـع، وفي التَّخْطِيـط اللُّغَـوِيّ، ومـا يتعلـقُ بـه مـن دراسـاتٍ في تخطيـط الوضـع اللُّغَـوِيّ، وتخطيـطِ الـمُدَوّنَة اللُّغَوِيّة، وتخطيـط الاكتسـابِ اللُّغَـوِيّ، وتخطيـط المعجـمِ، ومـا يَتَضَمَّـن هـذه الأنـواع مـن التَّخْطِيـط مـن عُنصرَيْـن أساسـيَّين يتمثَّـلان في التَّنفيـذ والقبـول[1].

هـل تفيـدُ الـمُدَوّنَات اللُّغَوِيّـة في هـذه المجـالات؟ الجـواب: نعـم، ودليـلُ ذلـك بـيّنٌ في الحجـم الهائـل مـن الدِّرَاسَـات، في تلـك الفـروع الـتي اعتمـدت علـى تطبيقـاتِ معالجـة الـمُدَوّنَة اللُّغَوِيّـة. وحصـرُ كُلّ هـذه الدِّرَاسَـات ليـسَ مجالَنـا هُنـا، لأنَّ عـددَ الأدوات يَزيـد عـن ٢٦٠ أداة[2]، ناهيـك عـن تلـك الدِّراسـات الـتي اعتمـدت علـى لغـات البرمجـة مثـل (البايثـون)،

---

(١) تأخـذ المؤسسـات حيزالتنفيـذ (implementation) لسياسـات الدولـة في أعمـال التَّخْطِيـط. أمـا القبـول (acceptance)؛ فهـو يعتمـد علـى قبـول المجتمـع لمـا خُطّـط لـه. تُعـدّ عوامـل التنفيـذ والقبـول أساسـيةً، بـل حاسـمةً، عنـد إعـادة تقييـم أيِّ وضـع لُغَـوِيّ في التَّخْطِيـط اللُّغَـوِيّ والسياسـة. نوقشـت هـذه العوامـل قديمًـا، وعُرفـت بنظريـة (هوجـين Haugen's Theory)، ولمـن يرغـب في الاطـلاع علـى أمثلتهـا الواقعيَّـة النظـرُ إلى الدِّرَاسَـات الآتيـة:

Haugen, E. (1966). Dialect, language, nation. American Anthropologist,68, 922–935.

Haugen, E. (1972). Language planning, theory and practice. In E. Haugen (Ed.), The ecology of language: Essays (pp. 287–295). Stanford: Stanford University Press.

Haugen, E. (1987). Blessings of Babel. Bilingualism and language planning. Problems and pleasures. Berlin, New York: Mouton de Gruyter.

(٢) للاطّـلاع علـى هـذه الأدوات انظـر في الرابـط الآتي: https://corpus-analysis.com (تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢).

و(الآر)، وغيرهما. لكن يكفي أن أشيرَ إلى العدد الذي تناولَ أداةً واحدةً فقط، وهي أدوات (وورد سميث WordSmith Tools)، والتي بلغت ما يقارب ٦٠٠ دراسة؛ من مقالاتٍ علميَّةٍ، وأطروحات جامعيَّةٍ في اللُّغَوِيّات التَّعْلِيمِيّة، واللُّغَوِيّات النفسيَّة، واللُّغَوِيّات الاجْتِمَاعِيّة، والتَّخْطِيطات اللُّغَوِيّة والسياسيَّة[١]. وصار العمل البحثيُّ في هذه الفروع يعتمدُ كثيرًا على الدِّراسات التَّجريبيَّةِ والدِّرَاسَات العمليَّةِ، التي تحوَّلت من البَيَانَات التَّنَاظُريَّة (analogical data) إلى البَيَانَات الرَّقَمِيّة[٢] (digital data). لم يعد من الجدوى -على سبيل المثال- العملُ على التَّخْطِيط اللُّغَوِيّ من دون أدوات ومناهج وتطبيقات المُدَوّنَة اللُّغَوِيّة، فالعملُ فيه من دونها مثلُ العمل على التَّخْطِيط في الاستثمار والمضاربة في السوق الماليَّة من دون أدواتٍ تحليليَّةٍ عن اتجاهاتِ السُّوق، والتَّنبُّؤ بأدائِها.

ب. العلوم الإنْسَانِيّة والاجْتِمَاعِيّة

بدأت بداياتُ ارتباط المُدَوّنَة اللُّغَوِيّة في العلوم الإنْسَانِيّة والاجْتِمَاعِيّة مع ظهور الاهتمام بما يُسمَّى بالإنْسَانِيّات الرَّقَمِيّة (digital humanities)، وبالبحث الاجْتِمَاعِيّ الرَّقَمِيّ على وجه الخصوص[٣]. كانت البداية الأكثر تأثيرًا في عام ٢٠١٤م، عندما نشر باحثون ورقةً علميَّةً عنوانها: (البرهنة التجريبيَّة الواسعة النِّطاق في شبكات التَّواصل الاجْتِمَاعِيّ)[٤]، شملت تحليلَ مئاتِ آلافِ المستخدمين في شبكة التَّواصل (فيسبوك)، وحظيت باهتمامِ الباحثين، وتضمَّن هذا النَّوعُ من التَّحْلِيل أساليبَ تحليليَّةٍ غير

(١) موقع حصر الدِّرَاسَات التي اعتمدت على أدوات وورد سميث. تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://lexically.net/wordsmith/research.

(٢) الفرق بين البَيَانَات التَّنَاظُرية والبَيَانَات الرَّقَمِيّة في العلوم الاجْتِمَاعِيّة هو أنَّ الأولى لا تعتمد على أيّ أداةٍ من أدوات الذَّكَاء الاصْطِنَاعِيّ في الجمع والتَّحْلِيل، بينما تعتمد الثانية على ذلك. من أمثلة البَيَانَات التَّنَاظُرية: الاستبانات.

(3) Veltri, G. A. (2020). Digital Social Research. Medford, MA: Polity Press.

(4) Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 87888790-.

مسـبوقةٍ؛ أَوَّلًا: أُجـرِي التَّحْلِيـل علـى مئـات آلافِ المسـتخدمين. ثَانِيًـا: جـرى التَّحْلِيـل علـى بيانـاتٍ لُغَوِيَّـةٍ ضخمـة. ثالثًـا: جـرى التَّحْلِيـل دون علـمِ المسـتخدمين. وقـد يكـون هـذا النَّـوع مـن البَيَانَـاتِ إما بيانـاتٍ خفيَّـةً (unobtrusive)، لا يُعـرف أصحابهـا أو لا يُهتـم بتحليلهـم، كـون الاهتمـام منصبًّـا علـى النـصِّ دون غـيره، وإمـا بيانـاتٍ ظاهـرةً (obtrusive) تتضمَّـن بيانـات منتـجي تلـك النُّصُـوص مـن أسـماء، وأعمـارٍ، وخلفيَّـاتٍ تعليميـةٍ، وثقافيَّـةٍ، وعرقيَّـةٍ، إلى آخِـرِهِ. ولا تقتصـر معالجـةُ الـمُدَوّنَة اللُّغَوِيّـة في العلـوم الإِنْسَـانِيّة والاجْتِمَاعِيّـة علـى اللغـة في الويـب والتواصـل الاجْتِمَاعِـيّ، بـل علـى كُلِّ مُدَوّنَـةٍ رَقَمِيّـةٍ مرقمنـةٍ في كلِّ مجـالات هـذه العلـومِ؛ كالدِّيـن والأدب والثَّقافـة والتَّاريـخ والأنثروبولوجيـا. أمـا أدواتُ المعالجـةِ؛ فتكـون إمّـا باسـتخدام أدوات معالجـة الـمُدَوّنَة اللُّغَوِيّـة؛ مثـل غـواص وغـيره، أو بتطبيقـات معالجـة اللغـة الطَّبِيعِيّـة.

ج. تطبيقات معالجة اللغة الطَّبِيعِيّة

ظهـرت هـذه التَّطْبِيقَـات مـن مشـاريع معالجـة اللغـة الطَّبِيعِيّـة، الـتي تقـوم علـى تدريـب الآلـة علـى فهـم النَّـصّ اللُّغَـوِيّ مثـل فهـم الإنسـان، وتُعتـبر معالجـة اللغـة الطَّبِيعِيّـة حقـلًا مـن حقـول الـذَّكَاء الاصْطِنَاعِـيّ، وتعتمـدُ علـى ثلاثـة مناهـج لخوارزميّـات معالجـة اللغـة الطَّبِيعِيّـة، وهـي منهـج القواعـد الـذي تعتمـد عليـه الآلـة في التحليـل، ومنهـج التَّعلُّـم الآليِّ الـذي تعتمـد فيـه الآلـة علـى احتمـالاتٍ رياضيَّـةٍ تُسـاعد في تأديـة مهـام المعالجـة لبيانـاتٍ جديـدةٍ، ومنهـج التعلُّـم العميـق الـذي تَتَعلَّـمُ بـه الآلـة فهـمَ المكوّنـاتِ الدقيقـة؛ مثـل الشَّـبكات العصبِيَّـة في دمـاغ الإنسـان.

هنـاك مرحلتـان في معالجـة اللغـةِ الطَّبِيعِيّـة؛ الأُولى: مرحلـة التَّحليـل الـتي فيهـا يـدرِّبُ الحاسـوبِيُّون الآلـةَ علـى تحليـل اللغـة وفهـمِ خصائصهـا. والثَّانيـة: مرحلـة التَّوليـد الـتي فيهـا يُنتـج الحاسـوبيُّون المخرجـاتِ، ودقَّـة أداء الآلـة في المعالجـة، وهـي المرحلـةُ الـتي تُنتـج تطبيقـات معالجـة اللغـة الطَّبِيعِيّـة. وتتطـوَّرُ المرحلـةُ الثَّانيـةُ بتطَـوُّر المرحلـة الأولى في كلِّ مـدَّةٍ زمنِيَّـةٍ، معتمـدةً في الآونـة الأخـيرةِ علـى منهـج تعلّـم الآلـة ومنهـج التعلّـم العميـق. مـن

أكثر تطبيقات معالجة اللغة الطَّبِيعِيّة انتشارًا في أيامنا هذه هي أدوات تنقيبِ النُّصُوصِ والبَيَانَات اللُّغَوِيّة، وتحليل المشاعرِ، واستخراج النُّصُوص، واستخراج الموضوعاتِ، ونمذجة الموضوعاتِ، وتصنيف النُّصُوصِ، وتلخيص النُّصُوصِ، والتَّرجمة الآلِيَّة، والكتابة الآليَّة، وفلترة الرَّسائل البريديَّة غير المهمَّة، والتَّصحيح الإمْلَائِيّ، والتعرُّف إلَى الكلام، وروبوتات المحادثةِ والدعم الافتراضي، وذكاء الأعمال في التَّسويق (فهم السلوك اللُّغَوِيّ للمستهلكين). وما يهمّ الباحثين في العلوم اللُّغَوِيّة والإنْسَانِيّة والاجْتِمَاعِيّة هنا هو فهمُ آليَّاتِ المعالجة الأساسيَّة للغة العَرَبِيَّة الطَّبِيعِيّة؛ من أجل تحقيق ممارساتٍ متقدّمةٍ في معالجة المُدَوّنَة اللُّغَوِيّة، حيث لا يُستغنى عنها في تطبيقات المعالجة الآلية للغة وتنقيب النصوص والبيانات اللغوية (موضوع فصول البَاب الثَّانِي من هذا الكتاب).

ثالثًا: مصادر المُدَوّنَة اللُّغَوِيّة ومعايير الجمع[1]

يمكن جمعُ كلِّ مصدرٍ يَتَضَمَّن لغَةً، سواءً كانت نصوصُ هذه اللغة من مصادِر رَقَمِيّةٍ، أم من صفحات الشَّبكة العنكبوتيَّة (الويب)، أم من اللغة الشفهيَّةِ المنطوقةِ. أما معايير الجمع فتتحدَّد من ثلاثة أهداف رئيسةٍ:

الهدف الأوَّل: الرصد (monitor)، وبه يُجمع كُلّ ما أمكن ويُمكن وسيُمكن جمعُه رَقَمِيًّا من النُّصُوص اللُّغَوِيّة، وتُمثّل واقع النموِّ اللُّغَوِيّ في جميع الأزمنة والأوعية والمجالات والموضوعات.

الهدف الثَّانِي: المرجعيَّة، وبه تُجمع مُدَوّنَات مرجعيَّة (reference)، تُستخدم بوصفها المعيار أو المقياس لاستخدامها في بناء تطبيقات تقنيَّات اللغة، ويتحدّد معيارُها بحسب التَّطْبِيقَات التي يحتاجُها المجتمعُ اللُّغَوِيّ.

(١) لا حاجة إلى ذكر تفاصيل أنواع المُدَوّنَات اللُّغَوِيّة ومعايير جمعها، وقد أُكتُفي بالخطوط العريضة. للاطلاع على التفاصيل: مكانري، توني، وأندرو هاردي. لُغَوِيّات المُدَوّنَة الحَاسُوبِيَّة: المنهج والنظريَّة والتَّطبيق. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، ٢٠١٦، الصفحات ١٠-٣٨.

الهـدفُ الثَّالِـث: الاسـتنادُ إلى عَيِّنـةٍ أو عيِّنـاتٍ، وبـه تُجمـع مُدَوّنَـة لُغَوِيَّـة محـدودةُ النُّصُـوص، وقـد تُجمـع هـذه النُّصُـوص بعنايـة فائقـة لتحقِّـق الأغـراض مـن مضمونهـا في الأبحـاث، ولتعميمهـا علـى المجمـوع اللُّغَـوِيّ، أو قـد تُجمـع دون عنايـةٍ؛ فيُنتهـز فيهـا بالجمـع رصـدُ كُلّ مـا يُمكـن رصـدُه.

يقـوم الإحصـاءُ علـى البَيَانَـات اللُّغَوِيّـة الـتي تُهيَّـأُ مـن النُّصُـوص اللُّغَوِيّـة الخـام، وتَتَضَمَّـن هـذه البَيَانَـات اللُّغَوِيّـة معلومـاتٍ تُقـاس. يُتطلـب مـن كلِّ باحـثٍ أن يعـرف طـرقَ إعـداد البَيَانَـات اللُّغَوِيّـة، وعـلامَ تحتـوي مـن مُتَغَـيّرَات، مادامـت المـادة الـتي سـتُحلّل هـي اللغـة ونصوصهـا سـواءً المكتوبـة أَم المنطوقـة.

رابعًا: أدوات معالجة المُدَوّنَات اللُّغَوِيّة العَرَبِيَّة

تتعـدَّد الأدواتُ، وتتفـاوتُ خصائـصُ معالجتهـا للنـصِّ العَـرَبِيّ تفاوتًـا نسـبيًّا، لكنهـا تشـترك في معظـم وظائـف تلـك الخصائـصِ؛ كاسـتخراج تكـرار الكَلِمَـات النَّوْعِيّـة، وتكـرار الكَلِمَـات الفِعْلِيّـة، والتّكْـرَار النِّسْـبِيّ للكلمـات الفِعْلِيّـة، وبعـض الإحصـاءات المتعلقـة بالتّصاحـب اللَّفْظِـيّ. ومـن أهـمّ هـذه الأدواتِ نظـرًا لتعـدد خصائـص ووظائـف المعالجـة فيهـا؛ مـا يَـأتِي:

١. محـرِّك التَّخْطِيـط[١] (Sketch Engine)، هـو أداةٌ شـبكيَّةٌ ضخمـةٌ، تَتَضَمَّـن مُدَوّنَـات لُغَوِيَّـة شـبكيَّةٍ وغـير شـبكيَّةٍ للُغـاتٍ عديـدةٍ، ويهمنـا فيهـا وظائـفُ معالجـة اللغـة للمدونـة اللُّغَوِيّـة الخاصـة بالمسـتخدم.

(١) مُحـرِّك التَّخْطِيـط. تـاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢، مـن: https://www.sketchengine.eu. شـهدت هـذه الأداة مـراحـل عـديـدةً مـن التطويـر والبنـاء، ويُمكـن الاطـلاع علـى الأوراق العِلْمِيّـة في كُـلّ مـراحل التَّطـوير، بالإضافـة إلى كُـلّ خصائـص البنـاء وأدواتهـا، في الـرابـط الآتي: https://www.sketchengine.eu/bibliography-of-sketch-engine/#toggle-id-1.

٢. أدوات معالجة اللغة العَرَبِيَّة الطَّبِيعِيّة (غواص) والمشذّب العَرَبِيّ لتنظيف النُّصُوص العَرَبِيَّة[١].

٣. أدوات (لانكستر) الإِحْصَائِيَّة[٢] (Lancaster Stats Tools)، وهي أدواتٌ شبكيَّةٌ تتميزَّ بسهولة الاستخدام للُّغويّين أو المهتمين بالتَّحليل الإِحْصَائِيّ للمدوّنة اللُّغَوِيّة.

يُتيح محرّك التَّخْطِيط الاشتراك المجانيَّ مدَّة شهرٍ، وهو أداةٌ شبكيَّة تَتَضَمَّن مُدَوّنَات ويب لأكثر من سبعين لغة، منها مُدَوّنَة الويب للغة العَرَبِيَّة التي تُعرفُ باسم (arTenTen12)، وتُتيح الأداة إنشاء مُدَوّنَة لكلِّ مستخدم، عن طريق تحميل الملفَّات النَّصِّيَّة التي يرغبُ المستخدم في تحليلها ومعالجتِها، وبعد إتمام الإنشاء؛ يُمكن للمستخدم أن يستكشف هذا النصَّ بإظهار قوائم الألفاظ (الكَلِمَات وتكراراتها)، والكشافات السياقِيَّة، وإحصاءات التَّصاحب اللَّفْظِيّ المشهورة. وما يُميِّز هذه الأداة هو أن نتائج كشافاتها السياقيَّة تُظهر المتتابعات اللَّفْظِيّة[٣] مقطَّعةً، فتجد في النَّتائج مثلًا ظهور حرف العطف (الواو) أو حرف الجر (الباء) منفصلًا عن الكلمة التَّالية له. وكل هذه الخصائص موجودةٌ في أداة غواص عدا التَّقطيع الآليّ. أما أدوات (لانكستر) الإِحْصَائِيَّة فتَتَضَمَّن كُلّ خصائص وخيارات التَّحليل المدوّنيّ، بالإضافة إلى معاملاتٍ

(١) غواص: أداة لمعالجة المُدَوّنَات اللُّغَوِيّة العَرَبِيَّة (النسخة ٤,٦). تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://sourceforge.net/ghawwasv4؛ المشذّب العَرَبِيّ، تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://sourceforge.net/Almoshatheb_Alarabi.

(٢) أدوات (لانكستر) الإِحْصَائِيَّة الشبكية. تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: http://corpora.lancs.ac.uk/stats/toolbox.php. شُرحت التَّطْبِيقَات الإِحْصَائِيَّة لهذه الأدوات في كتاب ألّفهُ (Vaclav Brezina)، وقد ترجمته إلى العَرَبِيَّة. انظر: بريزينا، فاكلاف. علم الإحصاء في لُغَوِيّات المُدَوّنَة الحَاسُوبِيَّة. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، ٢٠٢١.

(٣) «المتتابعات اللَّفْظِيّة» مصطلح يُشير إلى سلسلة من الكَلِمَات المتتابعة، ويدلُّ على الهيئة التي تنال حيزًا مستقلًّا عن سابقها ولاحقها بمسافةٍ، ومصطلحه الأجنبي n-grams، ومفرده n-gram الذي يعني (المتتابع اللَّفْظِيّ)، وقد عرّبته بـ(نغرام)، وجمعه: نغراميات.

إِحْصَائِيّـةٍ إضافيَّـةٍ لقيـاس معنويَّـة التَّصاحبـات اللَّفْظِيّـة، ناهيـك عمـا قـد دُمـج فيهـا مـن أدواتٍ إِحْصَائِيّـة أخـرى؛ مثـل تمثيـل البَيَانَـات (data visualization)، والانْحِـدَار اللُّوجسْـتِيّ للمُتَغَـيّرَات اللُّغَوِيّـة التَّصنيفيَّـة، وتحليـل الارتبـاط، ومقارنـة المجموعـات، وتحليـل التَّنَاظُـر، والتَّجْمِيـع، وتحليـل التَّغـير اللُّغَـوِيّ عـبر الزَّمـن، وغيرهـا ممـا سـنعرضه بالتَّطبيـق في جميـع مباحـث فصـول هـذا البـاب.

الفصل الأَوَّل
الإحصاءات وأنواع تمثيل البَيَانَات

وضعتُ في الملحق (أ) بياناتِ أسماء سور القرآن الكريم، وترتيبَ نزولها، وترتيبَها في المصحف الشَّريف، وعددَ الآيات والكَلِمَات والحروف، وزمنَ النُّزول (مكيَّةً أم مدنيَّةً)، ووجود السَّجْدَة من عدم وجودها بين السُّور.

سأبيّن في هذا الفصل أنواع المتغَيّرَات في البَيَانَات اللُّغَوِيّة في المبحث الأول، ومفهوم التّكْرَار والتَّوْزِيع والتَّنَوُّع والكثافة والتّكْرَار المعاير في المبحث الثاني، والإحصاءات الوصفيَّة في المبحث الثالث، والإحصاءات الاستدلاليَّة[1] في المبحث الرابع، ثم أنوع الأشكال البيانيَّة وتمثيلها من البَيَانَات اللُّغَوِيّة في المبحث الخامس.

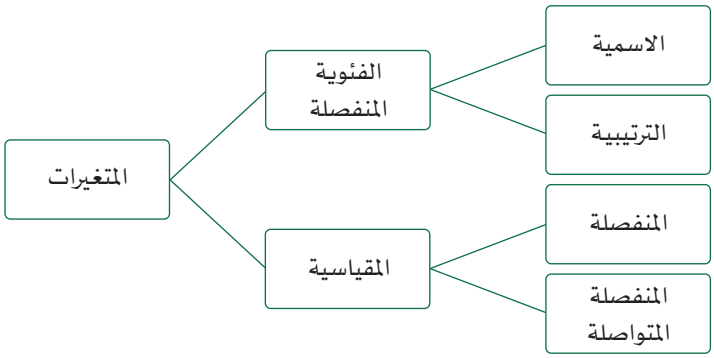
المبحث الأَوَّل: البَيَانَات اللُّغَوِيّة

يُراد بالمُدَوّنَة اللُّغَوِيّة النَّصّ اللُّغَوِيّ، أما البَيَانَات اللُّغَوِيّة فهي جدولُ خلايا يَتَضَمَّن حقائق عن تلك المُدَوّنَة اللُّغَوِيّة، وهي حقائقُ يُعبّر عنها بمصطلح مُتَغَيّرَات (variables). قد يجد الباحث بياناتٍ لُغَوِيَّةً مجدولةً وجاهزةً، تتضَمَّن حقائق عن نصٍّ لُغَوِيٍّ ما، وهذا نادرٌ في مصادر اللغة العَرَبِيَّة المفتوحة. ويعمل الباحث في الأغلب على إعداد هذه البَيَانَاتِ، وما تتضمَّنُه من مُتَغَيّرَات، وتنقسم هذه المتغَيّرَات إلى قسمين؛ الأَوَّل: مُتَغَيّرَات لُغَوِيَّة. والثَّانِي: مُتَغَيّرَات تأويلِيَّة. وتتضمَّن المُتَغَيّرَات اللُّغَوِيّة السِّمات اللُّغَوِيّة؛ مثل الكَلِمَات (الأسماء والأفعال والضمائر إلخ)، أو العبارات والجمل (الجملة الاسميَّة والجملة الفِعْلِيّة). أما المُتَغَيّرَات التأويليَّة فتعبّر عن سياقات تلك السِّمات،

(١) تنويه مهم للقارئ: عندما تقف على المواضع التي فيها ذكر قياسات حجم التأثير للإحصاءات الاستدلاليَّة في كُلّ مواضع البَاب الأَوَّل؛ فمن المهمّ أن تعرف القيم المعلنةَ في كُلّ إحصاء من تلك الإحصاءات، وهي قيمٌ تخبرك عن حجم الأثرِ، وهل هو كبيرٌ أم متوسطٌ أم صغيرٌ، وذلك هو السبيل إلى تفسير النتيجة، وهل الأثر دالٌّ أم غير دالٍّ.

كأن نقيـسَ سـمة الجنـس (ذكـر وأنـثى)، أو سـمة الأوعيـة كوعـاء الصحـف والكتـب والمجلات، أو سـمة المجالات كالمجالات المتفرّعـة في كُلّ وعـاء، مثـل مجـال المقـالات، ومجـال الأخبار داخـل وعـاء الصُّحـف، أو سـمة الموضوعـات المتفرعـة مـن كُلّ مجـالٍ، أو سـمة المواقـع الجغرافيَّـة للنـصِّ، أو سـمة أزمنـة النـصِّ.

وتكـون الـمُتَغَيّرَات اللُّغَوِيّـةُ والمُـتَغَيّرَات التأويليـة إمـا مُتَغَـيّرَاتٍ اسـميةً (nominal)، أو مُتَغَـيّرَاتٍ ترتيبيَّـةً (ordinal)، أو مُتَغَـيّرَات مقياسـيَّةً (scale). وتُسـمى الـمُتَغَيّرَات الاسـميَّة والترتيبيَّـة بالـمُتَغَيّرَات الفئويَّـة، وتكـون الـمُتَغَيّرَات المقياسـيَّة إمـا مُتَغَـيّرَات منفصلـةً أو متقطّعـةً (discrete)؛ أي: أعـداد محـدودة، أو مُتَغَـيّرَاتٍ متواصلـةً (continuous)؛ أي: أعـداد غـير محـدودةٍ[1]. في الشَّـكْل (٣) توضيـحٌ لتقسـيمات هـذه الـمُتَغَيّرَات في البَيَانَـات.



شكل ٣ : المُتَغَيّرَات في البَيَانَات (اللُّغَوِيّة)

(١) على سـبيل المثـال: لـو كان لدينـا عشـرة كتـب، فأعـداد الفصـول في كُلّ كتـاب تكـون مُتَغَـيّرَات منفصلـةً؛ كونهـا محـدودةً، وتشـير إلى قيـم فرديـةٍ محـددةٍ. أمـا أعـداد الكَلِمَـات في كُلّ فصـلٍ أو كتـابٍ؛ فهـي مُتَغَـيّرَاتٌ متواصلـةٌ؛ كونهـا غـير معروفـةٍ، وتشـير إلى قيـمٍ غـير متناهيـةٍ. وقـد تكـون الـمُتَغَيّرَات المنفصلـة أيضـا مُتَغَـيّرَاتٍ ثنائيـةً (مثـل: النكـرة والمعرفـة)، أو أعـداد في الـمُتَغَيّرَات الفئويَّـة (الاسـميَّة أو الترتيبيـة).
هـل مـا زال هنـاك غمـوضٌ في هـذه الـمُتَغَيّرَات في البَيَانَـات اللُّغَوِيّـة؟ إذا كان الجـوابُ نعـم؛ فتأكّـد بـأن وضوحهـا مرهـونٌ بالتّطبيـق والممارسـة، فكلمـا زاد اطّلاعـك علـى أنـواعٍ عديـدة مـن مجموعـة البَيَانَـات اللُّغَوِيّـة (linguistic datasets)؛ بانـت الفروقـاتُ بيانًـا جليًّـا بينهـا.

تُمثّل الـمُتَغَيّرَاتُ الاسـميَّةُ فئـاتٍ مختلفـةً تُعـبِّرعـن حـالاتٍ غـيرمتسلسـلةٍ؛ مثـل الجنـسِ (الذكـر والأنـثى)، وتُمثّـل الـمُتَغَيّرَات الترتيبيَّـة فئـاتٍ مختلفـةً تعـبِّرعـن حـالاتٍ متسلسـلةٍ؛ مثـل مراحـل العمـرِ (المراهقـة والبلـوغ والشـيخوخة). أمـا الـمُتَغَيّرَات المقياسـيَّة فهـي قيـمٌ حسـابيَّةٌ تمثِّـل تكـرارات ملحوظـةً أو تكـرارات نسـبيَّةً أو قيمًـا إحْصَائِيّـةً. وغالبًـا مـا تتـمُّ في أدوات المعالجـة الإحْصَائِيَّـة الإشـارةُ إلى الـمُتَغَيّرَاتِ الاسـميَّة والترتيبيَّـة بأرقـامٍ فئويَّـةٍ، كأن نضـع للذكـر رقـم (١)، وللأنـثى رقـم (٢)، ونضـع للمراهقـة رقـم (١)، وللبلـوغ رقـم (٢)، وللشـيخوخة رقـم (٣). في الشَّـكْل (٤) تمثيـلٌ لهيكلـة بيانـاتٍ لُغَوِيَّـة افتراضيـة لهـذه الـمُتَغَيّرَات، تَتَضَمَّـن روايـاتٍ مكتوبـةً مـن ١٠ كُتّـاب، وجنـسَ كُلّ كاتـبٍ، ومرحلتَـه العمريَّـة، وعـددَ تكـرار الأسـماء والأفعـال والضَّمائـر في كُلّ روايـةٍ.

الكاتب	الجنس	المرحلة العمرية	الأسماء	الأفعال	الصفات
هند	2	3	26740	956	287
عبدالعزيز	1	3	20121	934	278
مشاعل	2	2	6900	910	143
شيخة	2	2	6899	893	189
محمد	1	2	5008	610	178
الجوهرة	2	2	432	676	156
عبدالله	1	1	1112	198	67
خالد	1	1	1007	201	89
سارة	2	1	998	198	68
نور	2	1	108	99	98

شكل ٤ : شكل بيانات الأعمدة والصفوف لمُتَغَيّرَاتٍ لُغَوِيَّةٍ افتراضيَّةٍ

وأنواع الـمُتَغَيّرَات في هذا الشَّكْل هي على النحو الآتي:

- **المُتَغَيّرَات التأويليَّة:** الجنس ومرحلة العمر.
- **الـمُتَغَيّرَات اللُّغَوِيّة:** الأسماء والأفعال والضَّمائر.

- **المتغيّر الاسمي:** الجنس.
- **المتغيّر التَّرتيبيّ:** مرحلة العمر.
- **المتغيّر المقياسي:** تكرارات الأسماء والأفعال والضمائر.

بقي مفهومُ المُتَغَيّر المُسْتَقِلّ (independent variable)، ويُسمى -أيضًا- المُتَغَيّر التأويلي (explanatory variable) والمتغير التَّابِع (dependent variable)، والفرق الفاصلُ بينهما هو أنَّ المُتَغَيّر التَّابِع يتأثَّر ويتغيَّر ويتبدَّل بالمتغيّر المُسْتَقِلّ، والعكسُ غير صحيحٍ.

لو أردنا أن نُهيّئَ بياناتٍ لُغَوِيَّةً تَتَضَمَّن هذه الأنواع من المُتَغَيّرَات من مُدَوّنَة القرآن الكريم[1]؛ بوصفها المُدَوّنَة محلَّ التَّطبيق الإِحْصَائِيِّ لباب الإحصاء اللُّغَوِيّ من هذا الكتاب؛ فكيف سيكونُ شكلُ هيكلة البَيَانَات؟ الجواب: نحتاج إلى التَّفكير الإِحْصَائِيّ عند استخدام كُلّ نوعٍ من أنواع الإحصاءات اللُّغَوِيّة. كيف نفكِّر إحْصَائِيًّا؟ لنمعنَ النَّظر دائمًا فيما يمكن أن يكون مُتَغَيّرَاتٍ لُغَوِيَّةً ومتغيّراتٍ تأويليَّةً أوَّلًا. المُتَغَيّرَات اللُّغَوِيّة

(١) تتوفر بيانات للقرآن الكريم في موقع (Kaggle)، تَتَضَمَّن رقم كُلّ سورة حسب ترتيبها في المصحف الشريف، ورقم الآية في كُلّ سورة، ونص الآية الكريمة: https://www.kaggle.com/zusmani/the-holy-quran (تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢). رأيت أن هذه البَيَانَات ينقصها بعض المعلومات التي أحتاج إليها في التَّحْلِيل، مثل: تصنيف السُّوَر من حيث مكان النزول (مكية ومدنية)، وترتيب نزول القرآن الكريم زمنيًّا، وقد تطلّب مثل هذا النقص البحث عن معلومات جديدة عن بيانات السُّوَر وإدخالها في بيانات التَّحْلِيل. قد يتبادر سؤالٌ إلَى ذهن الباحثِ عن إمكانات الحصول على البَيَانَات اللُّغَوِيّة. يندر غالبًا أن نجد بيانات لُغَوِيَّة عربيَّةً جاهزةً ومفتوحةَ المصدر. ويتطلب من الباحث في هذه الحالة أن يُعدّ هذه البَيَانَاتِ، ويُمكن جمع البَيَانَات اللُّغَوِيّة المهيكلة بطرقٍ عديدةٍ، منها ما هو بَرْمَجِيٌّ (مثل: مكتبة Beautiful Soup في البايثون)، ومنها ما هو عن طريق أداةٍ شبكيةٍ جاهزةٍ (مثل https://www.parsehub.com: تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢) تعمل على جمع النُّصُوص والمعلومات عنها، وهيكلتها من مواقع الويب. وفي حال صعبت هذه الوسائل على الباحثِ؛ فمن أسهل الطرق هو أن يعتمد على أدوات معالجة المُدَوّنَات اللُّغَوِيّة؛ مثل غواص (وورد سميث WordSmith Tools)، لتحويل النَّصِّ الخام إلى بيانات مهيكلةٍ تَتَضَمَّن الكَلِمَات أو العبارات، وتَتَضَمَّن المُتَغَيّرَات التي يريد الباحثُ قياسَها.

عديـدةٌ، ولا تقتصـر فقـط علـى أقسـام الـكلام أو علـى أقسـام الجمـل، إذ يُمكـن أن تكـون موضوعًا نحويًّا أو استعمالًا لُغَوِيًّا. إذن؛ نحتاج هنا إلى أن نفكر فيما نرغب قياسه إحْصَائِيًّا مـن الـمُتَغَيّرَات، وفي الآتي فكـرةٌ إحْصَائِيّـة مـن أجـل التَّوضيـح:

- الـمُتَغَيّرَات التأويليَّـة الاسـمية، وهـي مُتَغَـيّرَاتٌ غـير متسلسـلةٍ، أو لا تربطهـا علاقـةٌ هرميَّـةٌ، ومـن الأمثلـة الممكنـة: السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة.
- الـمُتَغَيّرَات التأويليَّـة الترتيبيَّـة، وهـي مُتَغَـيّرَاتٌ متسلسـلةٌ، وتربطهـا علاقـةٌ هرميّـةٌ، ومـن الأمثلـة الممكنـة: ترتيـب سـور القـرآن الكريـم حسـب الـنزول.
- الـمُتَغَيّرَات اللُّغَوِيّـة: اسـتخدام النـداء للنـاس (يـا أيهـا النـاس)، وللمؤمنـين (يـا أيهـا الذيـن آمنـوا).
- المتغـيّر المقيـاسيّ: التّكْـرَارات الملحوظـة، أو التّكْـرَارات النِّسْـبِيّة، أو أيّ قيمـةٍ عدديَّـة، نضعهـا للمُتَغَـيّرَات اللُّغَوِيّـة.

بعـد تحديـد الـمُتَغَيّرَات الـتي يُـراد تحليلُهـا، وإعـداد بياناتها (الشـكل ٥)، يتطلّـب بالضرورة أن نصـوغ الأسـئلة الـتي نقيـسُ بهـا تلـك الـمُتَغَيّرَات. وفي الآتي مثـالٌ لثلاثـة أسـئلةٍ محدّدةٍ:

السـؤال الأَوَّل: هـل هنـاك علاقـةٌ بـين السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة (متغـيّر تأويلـيّ اسـميّ مسـتقلّ) وبـين ترتيـب نـزول السُّـوَر (متغـيّر تأويلـيّ ترتيـبيّ تابـع)؟

السـؤال الثَّـانِي: هـل هنـاك علاقـةٌ بـين السُّـورَة المكِّيَّـة والسُّـورة المدَنِيَّـة (متغـيّر تأويلـيّ اسـميّ مسـتقلّ) وبـين اسـتخدام المنـادى (متغـيّر لُغَـوِيّ مقيـاسيّ تابـع)؟

السـؤال الثَّالِـث: هـل هنـاك علاقـةٌ بـين ترتيـب نـزول السُّـوَر (متغـيّر تأويلـيّ ترتيـبيّ مسـتقلّ) وبـين اسـتخدام المنـادى (متغـيّر لُغَـوِيّ مقيـاسيّ تابـع)؟

السورة	مكان النزول	ترتيبها فى المصحف	الترتيب حسب النزول	عدد الآيات	يا أيها الناس	يا أيها الذين آمنوا
البقرة	مدنية	2	87	286	2	11
آل عمران	مدنية	3	89	200	0	7
النساء	مدنية	4	92	176	3	9
المائدة	مدنية	5	112	120	0	16
الأنعام	مكية	6	55	156	0	0
الأعراف	مكية	7	39	206	1	0
الأنفال	مدنية	8	88	75	0	6
التوبة	مدنية	9	113	129	0	6
...	...	...	...	...	...	...
...	...	...	...	...	...	...

شكل ٥ : بيانات المتغيّرَيْن اللُّغَوِيّين (يا أيها الناس) و(يا أيها الذين آمنوا) من مُدَوّنَة القرآن الكريم

سيعتاد الباحث على طرائق تحديد الـمُتَغَيّرَات بأنواعها، وعلى طرائق قياس الـمُتَغَيّرَات الاسميَّة والترتيبيَّة والمقياسيَّة، وذلك ليس على قياس التّكْرَار وحسب، بل على أنواعٍ أُخرى من الإحصاءات التي تردُ في المباحثِ والفصول القادمةِ من هذا البابِ، وسيتجلَّى من هذا الاعتياد ازديادٌ ملحوظٌ في فهم أنواع المـتَغَيّرَاتِ، وأسلوب توزيعها في أيِّ نصٍّ لُغَوِيّ.

المبحث الثَّانِي: التّكْرَار والتَّوزيع والتَّنوُّع والكثافة

سيُشار هنا إلى حساباتٍ أوليَّةٍ تُجرى في بداية معالجة الـمُدَوّنَة اللُّغَوِيّة العَرَبِيَّة؛ وهي:

١. التّكْرَار المطلق والتّكْرَار النِّسْبِيّ.

٢. الكلمة النَّوْعِيّة والكلمة الفِعْلِيّة.

٣. التَّوْزِيع (distribution)، والتَّشَتُّت (dispersion).

٤. التَّنَوُّع المعجَمِيّ.

٥. الكثافة المعجَمِيّة.

٦. التّكْرَار المُعاير.

باستعمال أدوات (لانكستر) الإحْصَائِيَّة؛ يبلغ عدد الكَلِمَات في القرآن الكريم ٧٧,٧٩٩ كلمة، وذلك باعتبار الرَّسْم الإمْلَائِيّ، أما باعتبار الرَّسْم العُثْمَانِيّ فالعدد يبلغ ٧٧,٤٣٩ كلمة[١]. يتغيرَّ العدد في الرَّسْم الإمْلَائِيّ عند تقطيع (segmenting) الكَلِمَات في المعالجة القبلِيَّة للنصِّ، وهذا طبيعيٌّ، ويعود إلى شكلِ الكلمة الفِعْلِيّة الذي ينال حيِّزًا مستقلًّا في الخطِّ الكتابِيِّ، ويكونُ مفصولًا بمسافة (الجدول ١)، كما يتغير عدد الكَلِمَات الفِعْلِيّة -أيضًا- في النُّصُوص المشكلة، باعتبار اختلاف الحركاتِ للكلمة الواحدةِ.

جدول ١ : الرَّسْم العُثْمَانِيّ والرَّسْم الإمْلَائِيّ قبل التَّقطيع وبعد أسلوبٍ معينٍ من التَّقطيع

الآية	الرَّسْم	عدد الكَلِمَات الفِعْلِيّة
يَٰيَحۡيَىٰ خُذِ ٱلۡكِتَٰبَ بِقُوَّةٖۖ وَءَاتَيۡنَٰهُ ٱلۡحُكۡمَ صَبِيّٗا	العُثْمَانِيّ	٧
يا يحيى خذ الكتاب بقوة وآتيناه الحكم صبيًّا	الإم ٮلَائِيّ	٨
يا يحيى خذ الكتاب ب قوة و آتيناه الحكم صبيًّا	الإمْلَائِيّ (مقطّع)	١٠

(١) ذكر السيوطي في الإتقان قول الدَّاني: «أجمعوا على أنَّ عدد آيات القرآن ستة آلاف آية، ثم اختلفوا فيما زاد على ذلك، فمنهم من لم يزدْ، ومنهم من قال: ومئتا آية وأربع آياتٍ، وقيل: وأربع عشرة، وقيل: وتسع عشرة، وقيل: وخمس وعشرون، وقيل: وست وثلاثون، كما أن الاختلافَ في عدد الآيات قد نبع من اختلافه بين أهل المدينة ومكة والشام والبصرة والكوفة، ولهذه الاختلافات أسبابها الخاصَّة بها، التي قد يكون مردّها إلى اجتهاد الكتّاب في وضع الفواصل عند سماعهم للنبي محمدٍ -صلى الله عليه وسلم- في بداية مراحل كتابة القرآن. السيوطي. الإتقان في علوم القرآن، تحقيق: شعيب الأرنؤوط، مؤسسة الرسالة، دمشق، ٢٠٠٨، صفحة ١٤٧.

وأحـوال التَّقْطِيـع في النُّصُـوص العَرَبِيَّـة متباينـةٌ، فهنـاك مقطّعـات تفصـل (أل) التعريـف وغيرهـا مـن المبـاني الصرفيَّـة والنَّحْوِيّـة. وعلـى اختلافنـا في مسـائل التَّقْطِيـع مـن وجهـة نظـر نَحْوِيَّـةٍ؛ إلا أنَّـه يتطلـب مـن اللُّغَـوِيّ الإفصـاح عـن آليَّـة التَّقْطِيـع عندمـا تكـون هنـاك آليَّـة مدمجـةٌ في أداةٍ مـن أدوات المعالجـةِ؛ كـي يتبـيَّن واقـع حسـاب الكَلِمَـات في النـصِّ، وكـي تُبـنى المعامـلات الإِحْصَائِيَّـة عليهـا؛ نظـرًا لاختـلاف النَّتـائج نسـبِيًّا بـين النُّصُـوص المعالجـة بالتَّقطيـع والنُّصُـوص المعالجـةِ دونـه.

وفيما يَأتي تَعريفاتٌ لهذه الحساباتِ الأوَّليَّة.

١. التّكْـرَار المطلـق: هـو تكـرار الكلمـة الفِعْلِيّـة في النـصِّ، أو تكـرار جميـع الكَلِمَـات الفِعْلِيّـة في النـصِّ. والتّكْـرَار النِّسْـبِيّ: نسـبة ورود الكلمـة في كامـل الـمُدَوّنَة، أو نسـبة ورودهـا في كُلّ معيـارٍ محـددٍ مـن الحجـم، كأن يُقـال: نسـبة ورود الكلمـة في كُلّ ١٠٠٠ كلمـة، أو في كُلّ ١٠,٠٠٠ كلمـة، أو في كُلّ ١٠٠ ألـف كلمـة، أو في كُلّ مليـون كلمـة.

٢. الكلمـة النَّوْعِيّـة: الصِّيغـة الفريـدة، أي كُلّ كلمـةٍ فريـدةٍ في تهجِئتهـا مـن دون تكـرار. والكلمـة الفِعْلِيّـة: ورود كُلّ صيغـةٍ كمـا هـي في كامـل الـمُدَوّنَة مـع تكراراتهـا.

٣. التَّوْزِيـع والتَّشَـتُّت، وهمـا قياسـان يُسـتعملان لمعرفـة انتشـار الكَلِمَـات في الـمُدَوّنَة، وعـادةً يُفصـح عـن التَّوْزِيـع ولا يُفصـح عـن التَّشَـتُّت مـع أهمِّيَّتـه. لننظـر إلى الفـرق بينهمـا مـن خـلال تطبيـق مفهومهمـا علـى تكـرار الكلمـة المعرّفـة (الكتـاب) في القـرآن الكريـم كمـا في الجـدول (٢). نلحـظ أن توزيـع هـذه الكلمـة يغطـي مـا نسـبته ٠,٢١ مـن المائـة مـن عـدد الكَلِمَـات الفِعْلِيّـة في القـرآن الكريـم، وهـي نسـبِيَّةٌ قليلـةٌ جـدًّا، أمـا التَّشَـتُّت فهـو منتشـرٌ في كُلّ أجـزاء القـرآنِ الكريـم (٣٠ جـزءًا) عـدا الجـزء السـابع عشـر. إذن؛ نسـتطيعُ أن نقـول عـن كلمـة (الكتـاب) بـأنَّ توزيعهـا قليـلٌ وتشـتُّتها كثـيرٌ. قِـسْ علـى ذلـك قياسًـا عكسـيًّا علـى بعـض الكَلِمَـات الـتي قـد يكـون توزيعُهـا كثـيرًا وتشـتتها قليـلًا.

جدول ٢ : التَّوْزِيع والتَّشَتُّت لكلمة (الكتاب)

الكلمة الفِعْلِيّة	التّكْرَار الملحوظ	التّكْرَار النِّسْبِيّ للكلمة من ١٠٠	التّكْرَار النِّسْبِيّ للأجزاء	الأجزاء
الكتاب	163	0.210	96.667	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
كتاب	47	0.060	86.667	1, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30
كتابا	12	0.015	36.667	4, 5, 6, 7, 15, 17, 22, 23, 25, 26, 30
والكتاب	8	0.010	20.000	2, 4, 5, 20, 25, 27

٤. التَّنَوُّع المعجَمِيّ: ويقاس إما بنسبة تكرار الكلمة الفِعْلِيّة إلى تكرار الكلمة النَّوْعِيّة (type-token ratio TTR) وهي (ناتج قسمة عدد الكلمة الفِعْلِيّة على عدد الكلمة النَّوْعِيّة)، أو بمتوسط نسب تكرار الكلمة النَّوْعِيّة إلى الكلمة الفِعْلِيّة في كُلّ ١٠٠٠ كلمة[١] (STTR1000).

٥. الكثافة المعجَمِيّة: نسبة الكَلِمَات المعجَمِيّة (الكَلِمَات ذات المحتوى فقط، أي: لا بدَّ من إزالة الكَلِمَات الوظيفيَّة من قائمة الكَلِمَات) إلى الكَلِمَات الفِعْلِيّة كلها.

(١) تعتمد نسبة الكلمة الفِعْلِيّة إلى الكلمة النَّوْعِيّة (أي: النسبة التي تنتج عن قسمة عدد تكرار الكلمة الفِعْلِيّة على عدد تكرار الكلمة النَّوْعِيّة) على أسلوب آخر يُعرف بالنسبة الموحدة للكلمة الفِعْلِيّة إلى الكلمة النَّوْعِيّة standardized type-token ratio STTR التي تعمل على تجزئة النصِّ إلى أجزاء عَديدةٍ، كُلّ جزءٍ يكون فيه عدد الكلمة الفِعْلِيّة موحّدًا، وتكون فيه نسبة الكلمة الفِعْلِيّة إلى النَّوْعِيّة خاصةً بها، ثم يُستخرج متوسط نسب كُلّ الأجزاء لإظهار قيمة النسبة الموحدة. وعند استخدام صيغةِ (STTR1000)؛ يكون كُلّ جزءٍ متضمنًا ١٠٠٠ كلمةٍ، ويُمكن اختيار غير الألف، حسب حجم المُدَوّنَة، وعادةً يكون التجزيء على عشرةِ أجزاء، فإن كان حجم المُدَوّنَة مليون كلمةٍ؛ تكون صيغة النّسبة الموحدة (STTR100000)، أي: ١٠٠ ألف كلمة في كُلّ جزءٍ منفصلٍ. وتفيد هذه النسبة المعرفة الدقيقة لنسبة توزيع الكلمة الفِعْلِيّة إلى الكلمة النَّوْعِيّة في أجزاء النصِّ.

تُتيح أداة غواص خاصية (قائمة الإيقاف)، وبها تستطيع أن تُضيف قائمة الكَلِمَات الوظيفيَّة (حروف الجر، وحروف العطف، وأدوات الربط، وأسماء الاستفهام، إلخ). كُلّ ما عليك فعله هو أن تفتح ملفًّا نَصِّيًّا جديدًا (notepad)، ثم تكتب كُلّ كلمةٍ في سطر، ثم تحفظ الملف بالترميز(ANSI)، (تنبّه إلى الترميز حيث إن الترميز utf-8 لن يعمل في خاصية قائمة الإيقاف في أداة غواص)، ثم تُحمِّل الملف في خيار قائمة الإيقاف في أداة غواص.

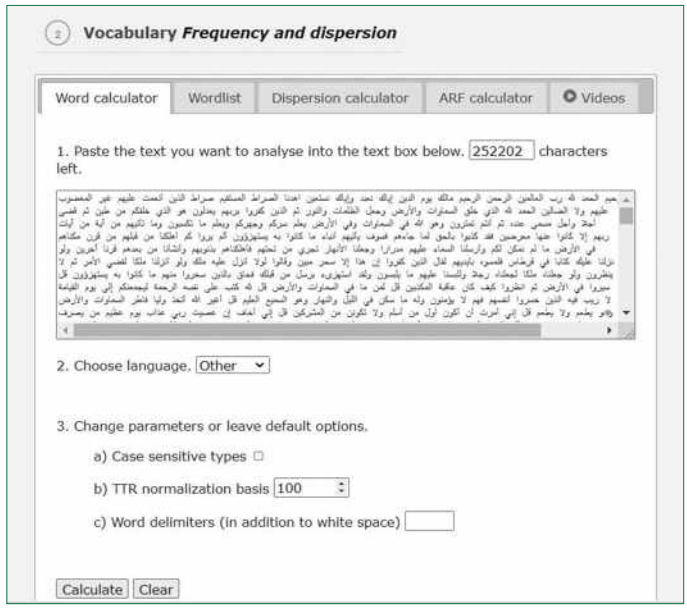
٦. التّكْرَار المُعايَر (normalized): ويُستخرجُ من بين عيناتٍ مختلفة الحجم في المُدَوّنَات اللُّغَوِيّة؛ لأجل تحقيق الدِّقة، فلو وردت كلمة (الرجل) ألف مرة في مُدَوّنَة حجمها عشرة آلاف كلمةٍ، ثم وردت عشرة آلاف مرة في مُدَوّنَة أخرى حجمها مليون كلمةٍ، فلا يمكن القول بأن ورودها في المُدَوّنَة الثانية أعلى. نحتاج هنا إلى أن نُعاير تكرار الرجل في كُلّ مائة كلمة أو ألف كلمة في كلتا المدونتين؛ لإظهار نسبة التّكْرَار المعاير. وفي المثال هنا عن كلمة (رجل)، تكون نسبتها في مُدَوّنَة الـ ١٠,٠٠٠ كلمة (١٠٪)، ونسبتها في مُدَوّنَة المليون كلمة (١٪)، وعليه يكون التّكْرَار في الأول أعلى من التّكْرَار في الثاني.

لننظر إلى الجدول (٣) لنتائج حسابات التّكْرَار من القرآن الكريم.

جدول ٣ : التّكْرَارات والتَّنوُّع المعجَمِيّ والكثافة المعجَمِيّة للرسم الإِمْلَائِيّ قبل التَّقْطِيع

المُدَوّنَة	الكَلِمَات الفِعْلِيّة	الكَلِمَات النَّوْعِيّة	التَّنَوُّع المعجَمِيّ		الكثافة المعجَمِيّة
القرآن الكريم	77799	14952	TTR	STTR1000	0.687
			0.19	0.54	
السُّوَر المكِّيَّة	47642	10665	TTR	STTR1000	0.683
			0.22	0.60	
السُّوَر المدَنِيَّة	30157	7336	TTR	STTR1000	0.693
			0.24	0.56	

يُمكن الحصول على هذه النتائج من خلال نسخ النَّصّ في أدوات (لانكستر) الإِحْصَائِيَّة، في مربع حساب الكَلِمَاتِ (word calculator)، في حزمة المفردات (التّكْرَار والتَّشَتُّت) كما في الشَّكْل (٦).



شكل ٦ : واجهة حساب الكَلِمَات

أزيلت علاماتُ التشكيل من مُدَوّنَة القرآن الكريم، حيث إن هذه العلامات تعطي في أدوات معالجة الـمُدَوّنَة عددًا مختلفًا لتكرارات الكَلِمَات النَّوْعِيّة. على سبيل المثال؛ لو جاءت كلمة (الناس) في النَّصّ ثلاث مراتٍ، فعدد تكرار الكلمة الفِعْلِيّة لها هو ثلاث مراتٍ، وعدد تكرار الكلمة النَّوْعِيّة لها مرَّة واحدة. ولكن، لو وردت هذه الكلمة ثلاث مرَّاتٍ بثلاث علامات تشكيلٍ في آخرها؛ مثل: النَاسُ، النَاسَ، النَاسِ، فتخبرنا الأداة بأن عدد تكرارها فعليًّا ثلاث مرَّات، وعدد تكرارها نوعيًّا ثلاث مرَّاتٍ -أيضًا-، وهذا غير دقيقٍ في حساب عدد تكرار الكلمة النَّوْعِيّة، وفي حساب نسبة عدد الكَلِمَات النَّوْعِيّة إلى الكَلِمَات الفِعْلِيّة (type-token ratio).

يجب أن يُدرك واقع عدد الكَلِمَات الفِعْلِيّة في النَّصّ العَرَبِيّ بعد التَّقْطِيع (segmentation)، كما يجب الإفصاحُ عن ذلك. ويُجرى التَّقْطِيع في معالجةِ اللغة الطَّبِيعِيّة لضمان الدِّقة الأعلى في نتائج حالات الكَلِمَات من ناحيةِ مبانيها النَّحْوِيّة، ويُجرى التَّقْطِيعُ قبل التَّوسيم النَّحْوِيّ (part-of-speech POS tagging)؛ لضمان دقَّة التَّوسيم. استَعْمَلْتُ نظامَ التَّقطيع والتَّوسيم النَّحويّ لستانفورد -انظر إلى قائمة الوسوم النَّحويَّة للغة العَرَبِيَّة في الملحق (ب)-؛ كونه مفتوح المصدر في مكتبة (البايثون) للغة العَرَبِيَّة، وذلك في التَّحليلِ الذي يتطَلَّبُ معالجةً نَحْوِيّةً في بعض مباحث هذا البابِ، ولمن يرغب في تقطيع النَّصّ وتوسيمِه، فإنَّ ذلك يُمكن إجراؤُه أيضًا في أداة محرك التَّخْطِيط (Sketch Engine).

إليكم في الجدول (٤) كيف أنَّ عددَ الكَلِمَات الفِعْلِيّةِ والكَلِمَات النَّوْعِيّةِ والتَّنَوُّع المعجَمِيّ والكثافة المعجَمِيّة قد تغيَّر بعد تقطيع النَّصِّ القرآني. ويجب أن نضع في الاعتبار طبيعةَ اختلاف نماذج التَّقْطِيع الآليِّ للغة العَرَبِيَّة، فنموذج (ستانفورد) يختلفُ عن نماذج آليَّةٍ أُخرى، فبعضها تقومُ على التَّقْطِيع القائم على المباني الصرفِيَّة، كقطع (أل) التَّعريف عن الكلمة، أو قطع الضَّميرِ المتَّصلِ من آخر الكلمة، وهذا سيؤدِّي لاختلافِ النَّتائج الإحْصَائِيَّة عن المُدَوَّنَة، ولا مشكلة في ذلك من وجهةِ نظري، بشرط أن يُفصحَ عن كُلّ ما اختاره المحلّل من المعالجات القبليَّةِ من تقطيعٍ، أو توسيمٍ، أو إيقافٍ للكلمات الوظيفِيَّة.

جدول ٤: التّكْرَارات والتَّنَوُّع المعجَمِيّ والكثافة المعجَمِيّة للرَّسم الإمْلَائِيّ بعد التَّقْطِيع

المُدَوَّنَة	الكَلِمَات الفِعْلِيّة	الكَلِمَات النَّوْعِيّة	التَّنَوُّع المعجَمِيّ		الكثافة المعجَمِيّة
القرآن الكريم	87457	13342	TTR	STTR1000	0.602
			0.15	0.48	
السُّوَر المكِّيَّة	53519	9724	TTR	STTR1000	0.603
			0.18	0.52	
السُّوَر المدَنِيَّة	33938	6573	TTR	STTR1000	0.599
			0.19	0.48	

المبحث الثَّالِث: الإحصاءات الوصفِيَّة

يُستخدمُ هذا النَّوعُ من الإحصاء لوصف البَيَانَات اللُّغَوِيّة وصفًا خاصًّا بها، بحيث نُعمّم نتائجها على المجتمع اللُّغَوِيّ الكُلِّيّ، ويختلفُ هذا النوع عن الإحصاءات الاستدلاليَّة (المبحث القادم) التي تقومُ على الطَّرائق الاحتماليَّة التي تُساعدُ على تعميمِ نتائج البَيَانَات اللُّغَوِيّة المحَدَّدة على المجموعةِ اللُّغَوِيّة الكاملة. ومن أكثر الطَّرائق شيوعًا لهذا النَّوعِ من الإحصاءاتِ مقاييسُ النَّزعةِ المركزيَّة (المتوسّط والوسيط)، ومقاييس التَّشَتُّت (المدى، والانحراف المعياري، ومعامل التباين[1]، وتوزيع جولاند، وانحراف النّسب).

في الشَّكل (٧) تضمينٌ لكلِّ هذه الإحصاءات الوصفيَّةِ[2] عن تكرار الكَلِمَات الفِعْلِيّة في كُلّ جزءٍ من أجزاء القرآنِ الكريمِ، وعن تكرار كلمة (الناس) في كُلّ جزءٍ، مع ملاحظةِ أن النَّصّ القرآني المُعالج مكتوبٌ بالرَّسْم الإمْلَائِيّ، ومن دون تقطيعٍ للمباني النَّحْوِيّة.

كيف هُيِّئَت البَيَانَات عن أجزاء القرآنِ كما في الشَّكْل (٧)؟ أَوَّلًا: وُضع كُلّ جزءٍ في ملفٍّ خاصٍّ به، ليكون عددُ الملفَّاتِ النَّصِّيَّة ٣٠ ملفًّا. ثَانِيًا: باستخدام أداةِ غواص، اُستُخرِجَ مجموعُ تكرارات الكَلِمَات الفِعْلِيّة في كُلّ جزءٍ، ثم استُخرجَت التّكْرَارات المطلقة

(١) ترجمتُ هذا المصطلح بـ (معامل التنوع coefficient of variation) في: بريزينا، فاكلاف. علم الإحصاء في لُغَوِيّات المُدَوّنَة الحَاسُوبِيَّة. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، ٢٠٢١، الصفحات ٦٤-٦٦. لهذا المصطلح مقابلاتٌ عربيَّةٌ كثيرةٌ، ومنها: معاملُ التَّغيُّر، ومعامل التَّفاوُت، ومعامل الانحرافِ، ومعامل الفَرْق. أَعْدِلُ هنا عن المقابل (معامل التنوع)، ليكون المقابل (معامل التَّباين)، والسَّبب في ذلك يعود إلى انتشار ترجمةِ مصطلحِ (ANOVA) بتحليل التَّباين، وهو مصطلحٌ مختصرٌ من (Analysis of Variation). إذن؛ أميل إلى المقابل (التباين)؛ لانتشارهِ في الكتب الإحْصَائِيَّة، فعنصر الانتشار وكثرة الاستعمال هو ما أميلُ إليه في اختيار المقابل العَرَبِيّ مقابلَ المصطلح الأجنبيّ، وإن كان المقابل العَرَبِيّ غير دقيقٍ.

(٢) يمكن إجراء هذه الإحصاءات باستخدام أداة صندوق (لانكستر LancsBox).

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Retrieved October 24, 2022, from http://corpora.lancs.ac.uk/lancsbox.

(التّكْـرَارات الملحوظـة) لكلمـة (النـاس) مـن كُلّ جـزءٍ. ثالثًـا: اسـتُخلِصَت النِّسَـب المتوقَّعـة (النّظريَّـة) بقسـمة كُلّ تكـرارٍ فعلـيٍّ للكلمـات الفِعْلِيّـة علـى مجموعِهـا، علـى سـبيل المثـال: في الجـزء الأوَّل في الصـف (١) في الشَّـكْل (٧)، عندمـا قُسِّـم التّكْـرَار ٢٥٣٢ علـى المجمـوع ٧٧٧٩٩؛ ظهـر النَّـاتج ٠,٣١٦، وهكـذا دواليـك علـى بقيَّـة النّسـب المتوقّعـة في بقيَّـة الأجـزاء. رابعًـا: اسـتُخلِصت النسـب الملحوظـة بقسـمة كُلّ تكـرارٍ ملحـوظٍ لكلمـة النَّـاس في كُلّ جـزءٍ علـى مجمـوع التّكْـرَارات الملحوظـةِ للكلمـة، علـى سـبيل المثـال؛ في الجـزء الأوَّل في الصَّـفِّ (١)، في الشَّـكل (٧)، عندمـا قُسِّـم التّكْـرَار (٨) علـى المجمـوعِ (١٨٤)؛ ظهـرَ النَّـاتجُ (٠,٠٤٣)، وهكـذا دواليـك علـى بقيَّـة النِّسـب الملحوظـةِ في بقيَّـةِ الأجـزاء. أمـا الفروقـاتُ النِّسـبِيَّة فهـي قيمـةُ الفـارق بـين النِّسـبة المتوقَّعـة والنِّسـبة الملحوظـةِ في كُلّ صَـفٍّ.

الجزء	تكرار الكلمات الفعلية	(الناس) التكرار المطلق	(الناس) التكرار النسبي	النسبة المتوقعة	النسبة الملحوظة	الفروقات بين النسب
1	2532	8	0.316	0.033	0.043	0.011
2	2586	16	0.619	0.033	0.087	0.054
3	2636	7	0.266	0.034	0.038	0.004
4	2517	7	0.278	0.032	0.038	0.006
5	2589	11	0.425	0.033	0.060	0.027
6	2529	8	0.316	0.033	0.043	0.011
7	2788	2	0.072	0.036	0.011	-0.025
8	2583	3	0.116	0.033	0.016	-0.017
9	2500	5	0.200	0.032	0.027	-0.005
10	2398	4	0.167	0.031	0.022	-0.009
11	2665	13	0.488	0.034	0.071	0.036
12	2731	11	0.366	0.035	0.060	0.025
13	2624	9	0.343	0.034	0.049	0.015
14	2503	2	0.080	0.032	0.011	-0.021
15	2688	4	0.149	0.035	0.022	-0.013
16	2777	2	0.072	0.036	0.011	-0.025
17	2453	14	0.571	0.032	0.076	0.045
18	2646	0	0.000	0.034	0.000	-0.034
19	2624	3	0.114	0.034	0.016	-0.017
20	2562	6	0.234	0.033	0.033	0.000
21	2592	12	0.463	0.033	0.065	0.032
22	2655	9	0.339	0.034	0.049	0.015
23	2639	1	0.038	0.034	0.005	-0.028
24	2531	5	0.198	0.033	0.027	-0.005
25	2673	4	0.150	0.034	0.022	-0.013
26	2622	3	0.114	0.034	0.016	-0.017
27	2531	3	0.119	0.033	0.016	-0.016
28	2643	2	0.076	0.034	0.011	-0.023
29	2667	0	0.000	0.034	0.000	-0.034
30	2315	10	0.432	0.030	0.054	0.025
المجموع	77799	184		1	1	0.032

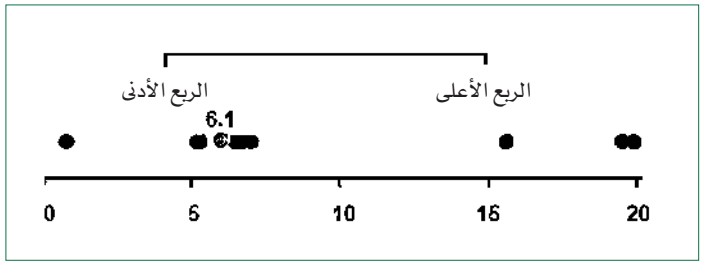
المتوسط		6.133	ناتج مجموع التكرار المطلق للكلمة ÷ عدد الأجزاء
الوسيط		4.500	الرقم المتوسط بين التكرار المطلق لكل كلمة
الانحراف المعياري		4.369	الجذر التربيعي لمجموع مربعات المسافات من المتوسط ÷ مجموع أجزاء المدونة (أجزاء القرآن) -١
المدى ١		الربع الثاني	الرقم ٦ يقع في الربع الثاني بين الصفر والعشرين
المدى ٢		0.933	عدد الأجزاء التي وردت فيها الكلمة ÷ عدد أجزاء المدونة
معامل التباين		0.130	الانحراف المعياري ÷ المتوسط = معامل التباين. (ثم) معامل التباين ÷ عدد أجزاء المدونة -١
توزيع جولاند		0.976	معامل التنوع ÷ الجذر التربيعي لـ (عدد أجزاء المدونة -١)
انحراف النسب		0.016	مجموع الفروقات بين النسب ÷ ٢

شكل ٧ : الإحصاءات الوصفيَّة ومقاييس التَّشَتُّت لكلمة (النَّاس) في القرآن الكريم

يُستخرجُ المتوسّط بجمعِ كُلّ تكرارٍ مطلقٍ مع الآخر، ثم بقسمةِ النَّاتج على عددِ الحالات (٣٠ حالة). ونضطرُّ إلى استخدام الوسيط بدلًا من المتوسّط عندما تكونُ هناك قيمٌ متطرِّفةٌ ستؤثِّرُ في المتوسِّط، ولا تعطي متوسطًا دقيقًا. في البَيَانَات في الشَّكْل (٧)؛ لا نجد قيمًا متطرّفةً؛ وعليه فلا حاجةَ بِنا إلى النَّظر إلى الوسيطِ. لقد استُخرج الوسيطُ لأجل توضيح كيفيَّةِ استخلاصه. وهوَ لا يتطلَّب أيَّ حسابٍ، إنما يتمُّ البحثُ عن العددِ الواقع في الوسط لكلِّ القيمِ. لدينا ثلاثون قيمةً، وآخر قيمةٍ في النِّصف الأوَّل هي (٤)، وأوَّل قيمةٍ في النِّصف الثَّانِي هي (٥)، وما بينهما بطبيعة الحالِ هي قيمة الوسيط، وهي (٤,٥). لو افترضنا بأن بعد القيمة ١٤ جاء تكرار الكلمةِ في الجزءِ (٣٠)، بواقعِ (٩٠٥) مرات، فبمثل هذه الحالات نستعملُ الوسيط بدلًا من المتوسّط الذي ستكونُ قيمتُه (٣٦)، وهي قيمةٌ مضَلِّلةٌ. أما الانحراف المعياري فيُحسب عن طريقِ (الجذر التربيعي لمجموع مربَّعات مسافات قيم التّكْرَار من المتوسّط) مقسوم على (مجموع الأجزاء - ١).

0، 0، 1، 2، 2، 2، 2، 2، 2، 3، 3، 3، 3، 4، 4، 4	⬇	5، 5، 6، 7، 7، 8، 8، 9، 9، 10، 11، 11، 12، 13، 14، 16
	4.5	

ولقياس المدى الذي يعدُّ أوَّلَ مقاييس التَّشَتُّت طريقتان؛ الأولى تُعرف بـ: المدى١ (الشكل ٨) الذي يُبيّن موقع المتوسط بين الربع الأدنى والأعلى للمدى، وهو يفيدُ في معرفة طبيعة الانحراف من المتوسّط، وموقع المتوسّط على وجه التَّحديدِ، أما الثَّانية فتُعرف بـ: المدى٢ الذي يقيس تشتُّت الكلمة بين أجزاء الـمُدَوّنَة، وتُعرف قيمتُه عن طريق المعادلة الآتية: (عدد الأجزاءِ التي وردت فيها الكلمة ÷ إجمالي عدد أجزاء الـمُدَوّنَة × ١٠٠)، لتكون قيمتُه (٠,٩٣)، وهي قيمةٌ عاليةٌ، ذات دلالةٍ على جودة تشتُّت الكلمة في الـمُدَوّنَة.



شكل ٨ : المدى١

كيف تُعرف القيم المتطرفة؟ الجواب: باستخدام منطق الربعيّات (quantiles)، وفق المعادلة الآتية: (Q1 – Q3 × 1.5). في بيانات الشَّكْل (٧) لا توجد قيم متطرفة. لننظر مثلا إلى تكراراتِ كلمة (النَّاس)، وهي قيمٌ تبدأ من (١) إلى (١٦). الربع الأول (Q1) يمتدُّ من (١) إلى (٤)، ويمتدُّ الربع الثاني (Q2) من (٥) إلى (٨)، ويمتدُّ الرُّبعُ الثَّالثُ (Q3) من (٩) إلى (١٢)، ويمتدُّ الرُّبعُ الرابعُ من (١٣) إلى (١٦). لنجري معادلة منطق القيمة المتطرّفة على النَّحو الآتي: (١ – ٩ × ١,٥) يساوي (١٢). يُضاف هذا النَّاتج إلى أوَّل قيمةٍ للرُّبع الثّالث (أي: ١٢ + ٩)، لتكونَ النتيجة (٢١)، وعليه تكونُ القيمةُ (٢١) فأعلى هي القيم المتطرفة في هذا المثال، وتُعرف هذه الطريقة بمعدَّل الربعيَّات البينيَّة (Inter Quantile Rate -IQR)، وتُجرى آليًّا على البَيَانَات.

يُحدِّد لنا معاملُ التَّباينِ حجمَ تنوُّعِ الكلمة في الـمُدَوّنَة اللُّغَوِيّة، وبينه وبين التَّشَتُّت علاقةٌ طردِيَّةٌ، فكلما زادَت قيمةُ معامل التَّباين زادت قيمةُ التَّشَتُّتِ، ويعتمد حسابُ التَّنَوُّع على ناتج قسمةِ (الانحراف المعياري) على (المتوسّط)، ثم يُقسمُ النَّاتجُ على (الجذر التَّربيعيّ لعدد أجزاء الـمُدَوّنَة – ١) للحصول على قيمة معاملِ التَّباين.

- 0.712 = 6.133 ÷ 4.369
- 0.13 = 5.385 ÷ 0.712 = $\sqrt{1 - 30}$ ÷ 0.712
- لتحويل القيمة إلى نسبيَّة مئويَّة: (0.13 × 100) = 13%

كلما كانت قيمة معامل التَّباين أقلّ من ٥٠٪؛ كان التَّنَوُّع أقلّ والتَّشَتُّت أفضل، وكلما ازدادَت دلَّ ذلك على وجود تباينٍ في التّكْرَارات أو بين التّكْرَارات، ويُعيدنا هذا إلى الكلامِ عن الفرق بين التَّوْزِيعِ (distribution)، وبين التَّشَتُّتِ (dispersion)، ومن المهم أن يُفصح عنهما، فازديادُ قيمة معامل التباين يعني أن التَّوْزِيع غير عادلٍ، وأن التَّباين عالٍ بين التّكْرَارات، وأن التَّشَتُّت ليس جيدًا. أما توزيع (جولاند)[1]؛ فهو على عكسِ معامل التَّباين، إذْ كلما زادت قيمةُ توزيع (جولاند)؛ دلّ ذلكَ على أن التَّوْزِيع عادلٌ، وأن التَّباينَ أقل، وأنَّ التَّشَتُّت أفضل. تكون قيمة توزيعِ (جولاند) بين (٠) و(١)، حيث يشير (٠) إلى أن التَّوْزِيع غير عادلٍ أبدًا، ويشير (١) إلى أن التَّوْزِيع عادلٌ تمامًا، ويتمُّ استخراجُ قيمة توزيع (جولاند) بالإجراء الحسابِيِّ الآتي:

- 1 - (معامل التباين) ÷ (الجذر التربيعي لعدد أجزاء المُدَوّنَة – ١)
- 1 - (0.13) ÷ ($\sqrt{30 - 1}$)
- 1 - (0.13) ÷ (5.385)
- 1 - 0.024 = 0.976

وفي آخر مقاييس التَّشَتُّتِ؛ نجد ما يُسمى بالانحراف النّسبيِّ، ويُستخرج بقسمةِ (٢) على (مجموع الفروقات النِّسْبِيّة)، ونتيجته ٠,٠١٦، ويُفسَّر على أن التَّوْزِيع بين الأجزاء عادلٌ جدًّا، إذْ كلما اقتربت القيمة إلى (٠)؛ دلّ ذلك على تساوي التّكْرَارات وتكافؤِها بين الأجزاء، وكلما اقتربت القيمة إلى (١) دلّ ذلك على أنَّ التَّوْزِيع غير متساوٍ ومتكافئٍ بين الأجزاء.

(1) Juilland, A. G., Brodin, D. R. & Davidovitch, C. (1970). Frequency dictionary of French words. The Hague: Mouton.

المبحث الرَّابع: الإحصاءات الاستدلالِيَّة

تُستخدمُ هذه الإحصاءاتُ لوصفِ البَيَانَات اللُّغَوِيّة العشوائِيَّة، وللحصولِ على معلوماتٍ وأحكامٍ عامَّةٍ عن المجتمع اللُّغَوِيّ كُلِّه من خلال عيِّنةٍ من المجتمع اللغوي. وتُجرى هذه الإحصاءاتُ لغرض اختبار فرضيةٍ ما، وفي الآتي الاختبارات التي ستُناقش في هذا المبحث.

- كاي تربيع (chi-squared)، وكاي تربيع بتصحيح ييتس (chi-squared Yates's correction)، والاحتمال اللوغارثمي log likelihood، واختبار فيشر المنضبط Fisher exact test.
- معامل (كريمر ڤ Cramer V).
- نسبة الاحتمال.

نحتاج بدايةً أن نفترض فرضِيَّةً بحثِيَّةً في مُدَوّنَة القرآن الكريم؛ لنفترض أنَّ الفعل (يخشون) والفعل (يخافون) لا يختلفُ ورودهما بين السُّوَر المكِّيَّة والسُّوَر المدَنِيَّة. فيما يلي صياغة الفرضيَّة الصِّفْرِيّة، ثم صياغة الفرضِيَّة البديلة.

الفرضِيَّة الصِّفْرِيّة: ليس للسياق أثرٌ في استخدامات الفعل (يخشون) والفعل (يخافون).

الفرضِيَّة البديلة: للسياق أثرٌ في استخدامات الفعل (يخشون) والفعل (يخافون).

وبعد صياغة الفرضيّتين؛ تُهَيَّأُ البَيَانَاتُ إما على صيغة جدولةٍ متقاطعةٍ cross-tabulation على نحو (أ) في الشَّكْل (٩)، أَوْ على شكل مجموعة بيانات (data set)، على نحو (ب) في الشَّكْل (٩).

لاحظْ أن الكَلِمَات قد أُدخلت بحروفٍ لاتينِيَّةٍ بدلًا من الحروف العَرَبِيَّة، وسبب ذلك أن أدوات (لانكستر) الإِحْصَائِيَّة لا تقبل الحروف العَرَبِيَّةَ، ويُمكنُ معالجة ذلك باستعمال أيّ طريقةٍ من طرائق الكتابة الصَّوتيَّة الدولِيَّة للحروف العَرَبِيَّة، وبعد الحصول على النَّتائِج؛ يُمكن إعادةُ تمثيل النَّتائج بالحروف العَرَبِيَّة.

بالنّسبة للتّكرارات الملحوظة؛ وردَ الفعلُ (يخشون) أربع مراتٍ في السُّوَر المكِّيَّة، ومرتين في السُّوَر المدَنِيَّة. وورد الفعل (يخافون) أربع مراتٍ في السُّوَر المكِّيَّة، وأربع مراتٍ في السُّوَر المدَنِيَّة. نظرتُ إلى صيغ الفِعْلِيّن من دون اشتقاقاتها، ومن دون اللواصق التَّصريفيَّة. ستعمل أدوات (لانكستر) الإِحْصَائِيَّة على تحويل التّكْرَارات الملحوظة إلى التّكْرَارات المتوقَّعة (النَّظريَّة)، لمعرفة طريقة هذا التحويل تطبيقيًّا؛ يشير الجدولُ (٥) إلى التّكْرَارات الملحوظة، ويشيرُ الجدول (٦) إلى طريقة استخراج التّكْرَارات المتوقّعة نَظَرِيًّا.

(أ)

Context	yakhŝawna	yakhāfūna
Makkiyya	2	4
Madaniyya	4	4

(ب)

ID	Variety	Outcome
1	Makkiyya	yakhŝawna
2	Makkiyya	yakhŝawna
3	Makkiyya	yakhŝawna
4	Makkiyya	yakhŝawna
5	Makkiyya	yakhŝawna
6	Makkiyya	yakhŝawna
7	Madaniyya	yakhāfūna
8	Madaniyya	yakhāfūna
9	Madaniyya	yakhāfūna
10	Madaniyya	yakhāfūna
11	Madaniyya	yakhāfūna
12	Madaniyya	yakhāfūna
13	Madaniyya	yakhāfūna
14	Madaniyya	yakhāfūna

شكل ٩ : جدولة متقاطعة للفعلين (يخشون) و(يخافون)

جدول ٥ : التّكْـرَارات الملحوظـة للفعلـين (يخشـون) و(يخافـون) في السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة

الأفعال / نوع السياق	يخشون	يخافون	المجموع
مَكِّيَّة	2	4	6
مدنيَّة	4	4	8
المجموع	6	8	14

جدول ٦ : التّكْـرَارات المتوقعـة للفعلـين (يخشـون) و(يخافـون) في السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة

الأفعال / نوع السياق	يخشون	يخافون	المجموع
مكيَّة	$2.57 = \frac{6 \times 6}{14}$	$3.43 = \frac{8 \times 6}{14}$	6
مدنِيَّة	$3.43 = \frac{8 \times 6}{14}$	$4.57 = \frac{8 \times 8}{14}$	8
المجموع	6	8	14

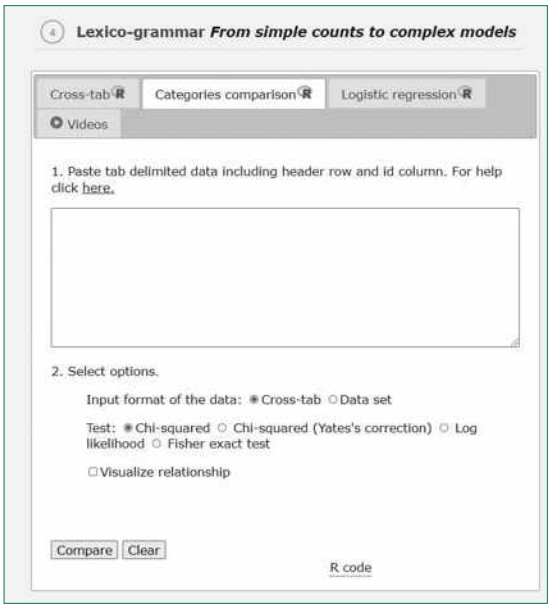
لـن أشـيرهنـا إلى طريقـة اسـتخراج كُلّ قيمـةٍ لـكل الاختبـارات الفرضِيَّـة[١]، إذْ يكفـي أن يعـرف القـارئ المفهـوم النَّظـريَّ للتّكـرارات المتوقَّعـة، كمـا أن أدوات (لانكسـتر) الإِحْصَائِيَّـة سـتُخرج القيـم الاحتماليَّـة (p values) لـكل اختبـارات الفَرضِيَّـة، والـذي يهـمُّ عندئـذٍ هـو القيمـة الاحتماليَّـة، فَمـاذا تعـني القيمـة الاحتماليَّـة؟ الجـواب ببسـاطةٍ هـو علـى النَّحـو الآتي:

- إذا جـاءت القيمـةُ الاحتماليَّـةُ أقـلَّ مـن ٠,٠٥ (أي: أقـل مـن ٥٪)؛ نرفـض الفرضيَّـة الصِّفْرِيّـة، وتكـون الفَرضِيَّـة البديلـة هـي الصحيحـة.

(١) مـن أراد الاطـلاع علـى الحسـابات النظريـة لاختبـارات الفَرضِيَّـة؛ فليعـد إلى المرجـع الآتي: بريزينـا، فـاكلاف. علـم الإحصـاء في لُغَوِيّـات الـمُدَوّنَة الحَاسُـوبِيَّة. ترجمـة: سـلطان المجيـول. دار جامعـة الملـك سـعود للنشـر، الريـاض، ٢٠٢١، الصفحـات ١٤١-١٥٣.

- إذا جاءت القيمة الاحتماليَّةُ أعلى من ٠,٠٥ (أي: أعلى من ٥٪)؛ نقبل الفرضيَّة الصِّفْرِيّة، وتكون عندئذٍ صحيحةً.

لاستخراج القيمة الاحتماليةِ (p value)، تُجري أدوات (لانكستر) اختبار الفَرضِيَّة واستخراج القيمة الاحتمالية بأربعة اختباراتٍ؛ كاي تربيع (chi-squared)، وكاي تربيع بتصحيح ييتس (chi-squared (Yates's correction)، والاحتمال اللوغارثميّ (log likelihood)، واختبار (فيشر) المنضبط (Fisher exact test) (الشكل ١٠). في الخانة الفارغة من هذا الشَّكلِ؛ نقوم بنسخ خلايا بيانات الجدولة المتقاطعة cross-tab (أ من الشكل ٩)، أو بنسخ خلايا بيانات المجموعة data set (ب من الشكل ٩)، ثم نُلصقها فيها. سنختار من صيغة إدخال البَيَانَات (input format of the data) الخيارَ الأَوَّل: cross-tab، بعد ذلكَ نختار نوعَ الاختبار من الاختبارات الأربعة المدمجة في الأداةِ، وَتظهرُ نتائجُ الاختبارات بعد الضَّغطِ على خيار المقارنةِ (Compare)، في الآتي عرضٌ لنتائج كُلّ اختبارٍ من الاختبارات الأربعة.



شكل ١٠: واجهة مقارنة المجموعات في صندوق (النحو المُعْجَمِيّ)

ستظهر نتائج كاي تربيع على النحو الآتي:

Some expected frequencies are too small. Try Fisher exact test or Log likelihood test instead.

Pearson's Chi-squared test: 0.39 (1), p = 0.532884, Cramer's V = 0.167, 95% CI [0, 0.687], SMALL EFFECT

Probability ratios (PR):

PR of verb_1 in Makkiyya vs. Madaniyya = 1.333 (95% CI: 0.545, 3.262)

PR of verb_1 in Madaniyya vs. Makkiyya = 0.75 (95% CI: 0.307, 1.835)

PR of verb_2 in Makkiyya vs. Madaniyya = 0.667 (95% CI: 0.177, 2.513)

PR of verb_2 in Madaniyya vs. Makkiyya = 1.5 (95% CI: 0.398, 5.654)

الترجمة:

بعض التّكْرَارات المتوقعة صغيرةٌ جدًّا. حاول استخدام اختبار فيشر المنضبط أو اختبار الاحتمال اللوغارثمي.

اختبار كاي تربيع بيرسون: 0.39 درجة الحرية (١)، قيمة الاحتمال p = ٠,٥٣٢، معامل كريمر ڤ = ٠,١٦٧، ٩٥٪ مجال الثقة [0، 0.687]، الأثر صغيرٌ.

نسب الاحتمال:

نسبة احتمال الفعل (يخشون) في السُّوَر المكِّيَّة مقابل الفعل (يخشون) في السُّوَر المدَنِيَّة = 1.333 (٩٥٪ مجال الثقة: 0545، 3.262).

نسبة احتمال الفعل (يخشون) في السُّوَر المدَنِيَّة مقابل الفعل (يخشون) في السُّوَر المكِّيَّة = ٠,٧٥ (٩٥٪ مجال الثقة: 0.307، 1.835).

نسبة احتمال الفعلِ (يخافون) في السُّوَر المكِّيَّة مقابل الفعل (يخافون) في السُّوَر المدَنِيَّة = ٠,٦٦٧ (٩٥٪ مجال الثقة: 0.177، 2.513).

نسبة احتمال الفعل (يخافون) في السُّوَر المكِّيَّة مقابل الفعل (يخافون) في السُّوَر المدَنِيَّة = ١,٥ (٩٥٪ مجال الثقة: 0.398، 5.654).

شكل ١١: نتائج كاي تربيع للفعلين (يخشون) و(يخافون)

أمـا بخصـوص بقيـة الاختبـارات الثَّلاثـة، فسـتظهر نتيجـةُ كاي تربيـع (تصحيـح ييتـس) = ٠,٠١، ودرجـة الحريـة (١)، وقيمـة الاحتمـال p = 0.938، وسـتظهر نتيجـة الاحتمـال اللوغارثـمي = ٠,٣٩، ودرجـة الحريـة (١)، وقيمـة الاحتمـال p = 0.531.، وسـتظهر قيمـة الاحتمـال لاختبـار (فيشـر) المنضبـط p = 0.627. أمـا نسـب الاحتمـال مـن هـذه الاختبـاراتِ الثَّلاثـة الأخـيرة؛ فهـي نسـب الاحتمـال الـواردة في كاي تربيـع نفسُـها.

سنشـرح معامـل (كريمـر ڤ)، ونسـبة الاحتمـال؛ بعـد الوقـوف علـى قيـمِ الاحتمـالِ، ودرجـة الحرِّيَّـةِ، ومجـال الثِّقـة. دعونـا نعُـدْ إلى الفرضِيَّـة الصِّفْرِيّـة والفرضيَّـة البديلـة. قلنـا إنْ كانـت قيمـةُ الاحتمـال أقـلَّ مـن ٠,٠٥؛ نرفـض الفرضيَّـة الصِّفْرِيّـة، ونعدُّهـا غـير صحيحـةٍ. وإذا نظرنـا إلى قيـم الاحتمـال في كُلّ هـذه الاختبـاراتِ الأربعـةِ؛ سـنجدها أعلـى مـن ٠,٠٥. فالفرضيَّـة الصِّفريَّـة صحيحـةٌ، ونسـتنتج مـن ذلـكَ أنَّـه لا يوجـد دليـلٌ كافٍ لرفـض الفرضِيّـة الصِّفْرِيّـة.

هـي لعبـةُ احتمـالاتٍ، وتتطلَّـبُ تـوخِّي الحـذر عنـد التَّفسـير، فهـي لا تُخبرنـا قطعًـا بصحَّـة الفرضيَّـة الصِّفْرِيّـةِ، وخطـأ الفَرضِيَّـة البديلـة، فلـو أجرينـا بحثًـا عـن الألفـاظ العاطفيَّـة بـين الرجـال والنِّسـاء، وجمعنـا مُدَوّنَـةً لروايـاتٍ كتبهـا رجـالٌ، وروايـاتٍ أُخـرى كتبهـا نسـاء، ووضعنـا الفرضيَّتـين الآتيتـين؛ الفَرضِيَّـة الصِّفْرِيّـة: لا فـرق بـين الرِّجال والنِّسـاء في اسـتخدام الكَلِمَـات العاطفيَّـة، والفَرضِيَّـة البديلـة: هنـاك فـرقٌ بـين الرِّجـال والنِّسـاء في اسـتخدام الكَلِمَـات العاطفيَّـةِ؛ فـإنَّ قِيَـمَ الاحتمـال قـد تكـونُ أقـل مـن ٠,٠٥ بـين روايـاتٍ محـدَّدةٍ، وقـد تكـون أعلـى مـن ٠,٠٥ في حـال لـو زدنـا حجـم الرِّوايـات، ولا بـدَّ أن تكـون الافتراضـاتُ سـليمةً، وأن تكـون البَيَانَـات اللُّغَويّـة (العيّنـة) متسـقةً مـع بيانـات المجموعـة اللُّغَويّـة كلِّهـا، والأخـذ في الحسـبان بِعوامـل متعلّقـةٍ بالخلفيـات الثقافيـة والعُمُريَّـة لكتّـاب هـذه الروايـاتِ؛ كـي يكـونَ الاختبـار الإِحْصَـائِيّ دقيقًـا في نتائجـه.

نـأتي إلى معامـلِ (كريمـر ڤ)، وهـي عمليَّـةٌ إحْصَائِيّـةٌ، تحسـب لنـا حجـمَ التَّأثـير لـكاي تربيـع، وتكـون مهمَّـةً في حـال صـارت قيـم إحْصَائِيّـات تربيـع كاي في تزايـدٍ؛ بسـبب تزايد حجم العَيِّنـة، وبمـا أن كاي تربيـع يقـدِّمُ لنـا دليـلَ قبـولِ أو عـدم قبـولِ الفَرضِيَّـة الصِّفْرِيّـة؛ فحجـم التَّأثـير باسـتخدام هـذا المُـعامل يقـدِّم لنـا قابليَّـةَ المقارنـة بـين أحجـام العَيِّنـة. متى يكـون حجم التَّأثـير كبـيرًا؟ الجـواب: تظهـر قيمـة هـذا المُـعامل بـين (٠) و(١). وضـع (كوهـن)[1] التَّفسـير المـوصَى بـه لتلـكَ القيـم، وهـي في الجـدول (٧).

جدول ٧ : توصية (كوهن) لتفسير حجم التَّأثير بمعامل (كريمر ڤ)

درجة الحرية	حجم التَّأثير بمعامل كريمر ڤ		
	التأثير صغير	التأثير متوسط	التأثير كبير
١	0.10	0.30	0.50
٢	0.07	0.21	0.35
٣	0.06	0.17	0.29

لـو عدنـا إلى نتيجـة المعامـل مـن أدوات (لانكسـتر) الإحْصَائِيَّـة؛ سـنجد أنَّ القيمـة هـي (٠,١٦٧) لدرجـة الحرِّيَّـة (١)، وبنـاءً علـى تفسـير كوهـن؛ فالتَّأثـير صغـيرٌ أو ضعيـفٌ. مـاذا يعـني ذلـك؟ يعـني أنَّـه لا يوجـد تأثـيرٌ في اختـلاف اسـتعمال الفعـل (يخشـون) والفعـل (يخافـون) بـين السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة.

سـؤالٌ أخـيرٌ عَـن درجـة الحرِّيَّـة؛ مـاذا تعـني؟ الجـوابُ ببسـاطةٍ: هـي تكنيـك إحْصَـائِيٌّ يقـوم علـى الاحتمـالات الـتي لدينـا عنـد حسـاب الحـالات، لدينـا حالتـان في المثـال الـذي طبقنـاه في العمـود (الفِعْلَـين) وحالتـان في الصـفّ (المكِّيَّـة والمدَنِيَّـة) في البَيَانَـات. درجـة الحرِّيَّـة هـي (١)، وتكـون درجـة الحرِّيَّـة دائمًـا هـي (عـدد الحـالات - ١).

(1) Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

لماذا ناقص واحد؟ الجواب: لأنَّ الاحتمال الثاني غير محسوبٍ، فالقياس ليس حرًّا في اختياره؛ لأنه هو ما قد تبقّى. تشبه فكرة الحرية فكرة أن تكون أمام حريةِ انتقاء سيَّارتين، أنت حرٌّ في اختيار السَّيارة الأولى، أما في مرحلة الاختيار الثانية فليست هناكَ حرِّيَّةٌ؛ لأنه لم يتبقَّ لك إلا هذه السَّيارة.

نأتي إلى نسبة الاحتمالِ؛ في كُلّ الأبحاث التَّجريبيَّة، سواءً الطبيَّة أَم غيرها؛ يهتمُّ الباحثون بنوعٍ واحدٍ من التَّأثير لنسبة الاحتمال، وهي نتيجةُ عيِّنة الدِّراسة مقابل المجموعة المرجعيَّة كلِّها. في الإحصاء اللُّغَوِيِّ؛ يُمكن أن نحسبَ أكثر من نسبة احتمالٍ واحدةٍ، حسب محاورِ الحالاتِ في البَيَانَات. ولتوضيح طريقة إظهار نسب الاحتمالِ؛ فَهي نتيجة قسمةِ التّكْرَار الملحوظِ لكلِّ فعلٍ على مجموع تكرار الأفعالِ في السِّياقِ الواحدِ. لننظر إلى الجدول (٩) ونقارنه بجدول (٨) للتّكرارات الملحوظةِ، ثم ننظرُ إلى النَّتائج التي أظهرتها أدواتُ (لانكستر) الإِحْصَائِيَّة لنسبِ الاحتمالِ من أجلِ فهم طريقةِ استخراجِها.

جدول ٨ : التّكْرَارات الملحوظة

الأفعال / نوع السياق	يخشون	يخافون	المجموع
مَكِّيَّة	2	4	6
مدنِيَّة	4	4	8
المجموع	6	8	14

جدول ٩ : التّكْرَارات المتوقعة (نَظَرِيًّا)

الأفعال / نوع السِّياق	يخشون	يخافون	المجموع
المؤلف	0.333 = 6 ÷ 2	0.666 = 6 ÷ 4	6
مدنيَّة	0.5 = 8 ÷ 4	0.5 = 8 ÷ 4	8
المجموع	8	6	14

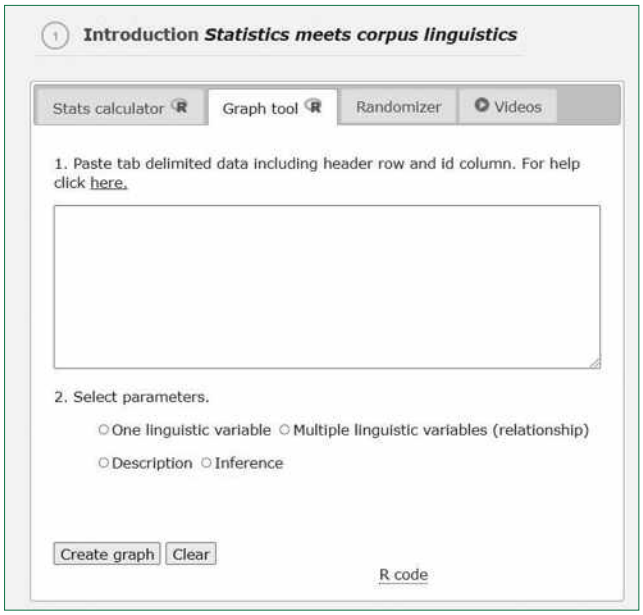
نتائج أدوات (لانكستر) الإِحْصَائِيَّة

- نسبة احتمال الفعل (يخشون) في السُّوَر المكِّيَّة مقابل الفعل (يخشون) في السُّوَر المدَنِيَّة $(\frac{0,333}{0,5})$ = (0.667).
- نسبة احتمال الفعل (يخشون) في السُّوَر المدَنِيَّة مقابل الفعل (يخشون) في السُّوَر المكِّيَّة $(\frac{0,5}{0,333})$ = (1.5).
- نسبة احتمال الفعل (يخافون) في السُّوَر المكِّيَّة مقابل الفعل (يخافون) في السُّوَر المدَنِيَّة $(\frac{0,666}{0,5})$ = (1.333).
- نسبة احتمال الفعل (يخافون) في السُّوَر المدَنِيَّة مقابل الفعل (يخافون) في السُّوَر المكِّيَّة $(\frac{0,5}{0,666})$ = (0.75).
- تُفسَّر نسب الاحتمال هذه على النَّحو الآتي:
- نسبة الاحتمال (١): لا يوجد فرقٌ بين السِّياقين.
- نسبة الاحتمال الأصغر من (١): ظهور الفعل (يخشون) في السُّوَر المكِّيَّة أقلّ احتمالًا من ظهوره في السُّوَر المدَنِيَّة، وظهور الفعل (يخافون) في السُّوَر المدَنِيَّة أقلّ احتمالًا من ظهوره في السُّوَر المكِّيَّة.
- نسبة الاحتمال الأكبر من (١): ظهور الفعل (يخشون) في السُّوَر المدَنِيَّة أكبر احتمالًا من ظهوره في السُّوَر المكِّيَّة، وظهور الفعل (يخافون) في السُّوَر المكِّيَّة أكبر احتمالًا من ظهوره في السُّوَر المدَنِيَّة.

المبحث الخامس: أنواع الأشكال البيانِيَّة

يحتوي هذا المبحثُ على أساليب تمثيلِ البَيَانَاتِ (visualization)، وفي كُلِّ نوعٍ من أنواع تمثيل البَيَانَات اللُّغَوِيّة أمثلةٌ للمُتَغَيِّرَات اللُّغَوِيّة، ونوع العيِّنة من القرآن الكريم، وعدد النُّصُوص، بالإضافة إلى هيكل البَيَانَات لكلِّ نوعٍ من أنواع الأشكال البيانِيَّة. وكُلّ ما يُحتاج

القيام به بعد تهيئة هيكل البَيَانَات في ملفٍّ (إكسل) هو نسخه ثم لصقه في أدوات (لانكستر) الإحْصَائِيَّة، ثم اختيار المقاييس (parameters): متغَيّر لُغَوِيّ واحِد (one linguistic variable)، أم مُتَغَيّرَات لُغَوِيَّة متعدِّدة (العلاقة) (Multiple linguistic variables (relationship، ووصف (description) أم استدلال (inference) (الشكل ١٢).



شكل ١٢ : واجهة أداة تمثيل البَيَانَات اللُّغَوِيّة

أَوَّلًا: المخطّط الدَّرجي histogram

يُجرَى هذا الشَّكلُ البيانيُّ للإحصاء الوصفيِّ، ولتمثيل هذا المخطّط؛ تحتوي البَيَانَات اللُّغَوِيّة على متغَيّر لُغَوِيّ واحد، ومدونةٍ واحدةٍ، ونصوص متعدّدة ضمن هذه المُدَوّنَة.

مثال:

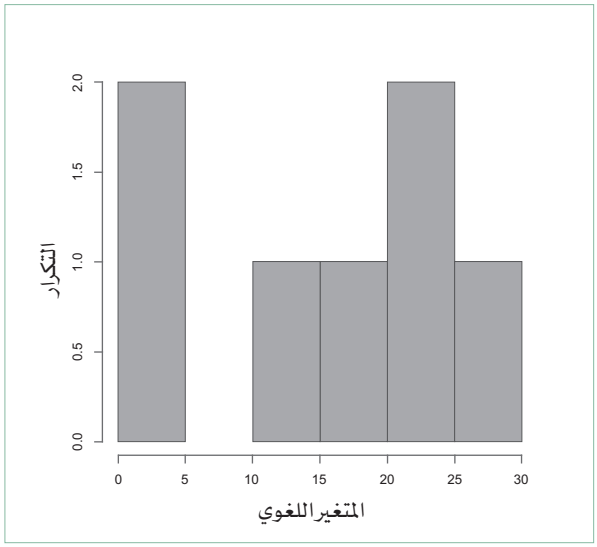
- المتغيّر اللُّغَوِيّ: عبارة (الذين آمنوا: تتابعان لفظِيَّان 2n-grams).
- الوصف description.

- الـمُدَوّنَة الواحـدة: السُّـور السَّـبع الطِّـوال (أي: بـين سـورة البقـرة، وآل عمـران، والنِّسـاء، والمائـدة، والأنعـام، والأعـراف، والأنفـال والتَّوبـة معًـا).
- النُّصُوص المتعدِّدة سبعة، باعتبار أن كُلَّ سورةٍ قد جُعلت نَصًّا مستقلًّا.

عندئـذ يُهيّـأُ هيـكل البَيَانَـات وفـق الشَّـكل (١٣)، ويخـرج تمثيـل المَخَطّـط الـدَّرجي كمـا في الشَّكْل (١٤).

Texts	Frequency
text1	26
text2	12
text3	17
text4	22
text5	1
text6	2
text7	21

شكل ١٣: تكرار (الذين آمنوا) في السُّوَر الطوال



شكل ١٤: المَخَطّط الدّرجي لتكرار (الذين آمنوا) في السُّوَر الطوال

ثَانِيًا: مخطَّط الصّندوق (أو رسم المربع) boxplot

يُجرى هذا الشَّكلُ البيانيُّ للإحصاء الوصفيِّ، ولتمثيل هذا المخطَّطِ؛ تحتوي البَيَانَات اللُّغَويّة على متغيِّرٍ لُغَوِيٍّ واحد، ومدوّناتٍ متعدِّدةٍ (أي: مدونتين أو أكثر)، ونصوصٍ متعدِّدةٍ في كلِّ مُدَوّنَة.

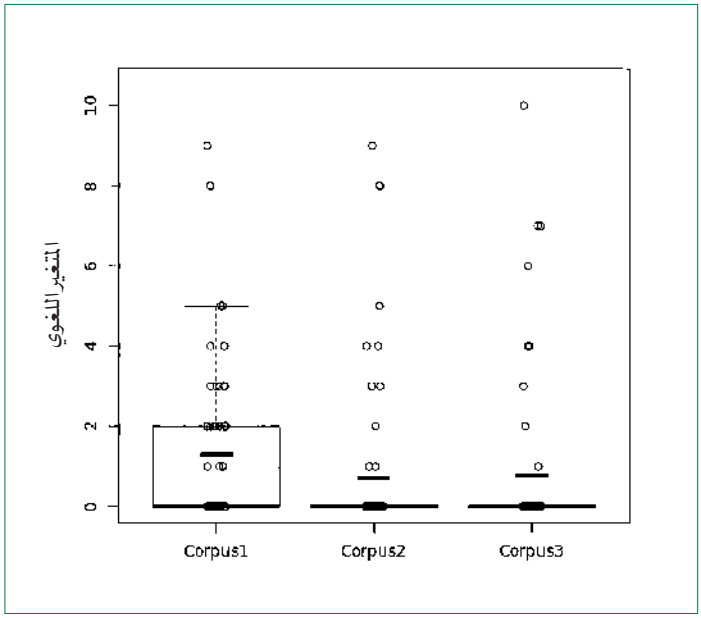
مثال:

- المتغيِّر اللُّغَوِيّ: كلمة (ما).
- الوصف description.
- المُدَوّنَات: ثلاث مُدَوّنَات منفصلة. المُدَوّنَة الأولى (corpus1): جزء عمَّ، والمُدَوّنَة الثَّانية (corpus2) جزء تبارك، والمُدَوّنَة الثَّالثة (corpus3) جزء قد سمع.
- عدد النُّصُوص: ٥٧ نَصًّا (على اعتبار عدد السُّوَر في هذه الأجزاء الثَّلاثة).

ويُهيَّأ شكلُ البَيَانَات كما في الشَّكل (١٥)، إذ يتضمَّنُ العمود الأول (النَّصّ Text)، وكل نَصّ محفوظٍ برقم ترتيب السُّورة في المصحف الشَّريف في الأجزاء القرآنيَّة الثَّلاثة الأخيرة، ويتضمَّن نَصّ السُّورة، ويخرج تمثيل المخَطّط الصّندوقي كما في الشَّكل (١٦).

Text	Corpus1	Corpus2	Corpus3
text58	0	0	7
…	…	…	…
text67	0	5	0
…	…	…	…
text78	2	0	0

شكل ١٥ : سبعةٌ وخمسون نَصًّا (٥٧ سورة) في الأجزاء الثَّلاثة الأخيرة



شكل ١٦ : مخطّط الصّندوق

ثالثًا: أعمدة الخطأ. مجال الثقة ٩٥٪

يُجرى هـذا الشَّـكل البيـانيُّ للإحصـاء الاسـتدلاليِّ، ولتمثيـل هـذا المَخَطّـط؛ تحتـوي البَيَانَـاتُ اللُّغَوِيّـة علـى متغـيِّرٍلُغَـوِيّ واحـد، ومدوّنـةٍ تتضمَّـن نصوصًـا متعـدِّدةً، أو مُدَوّنَـاتٍ متعـدِّدةٍ (أي: مدونتـين أو أكـثر) تتضمَّـن نصوصًـا متعـددةً في كلِّ مُدَوّنَـةٍ.

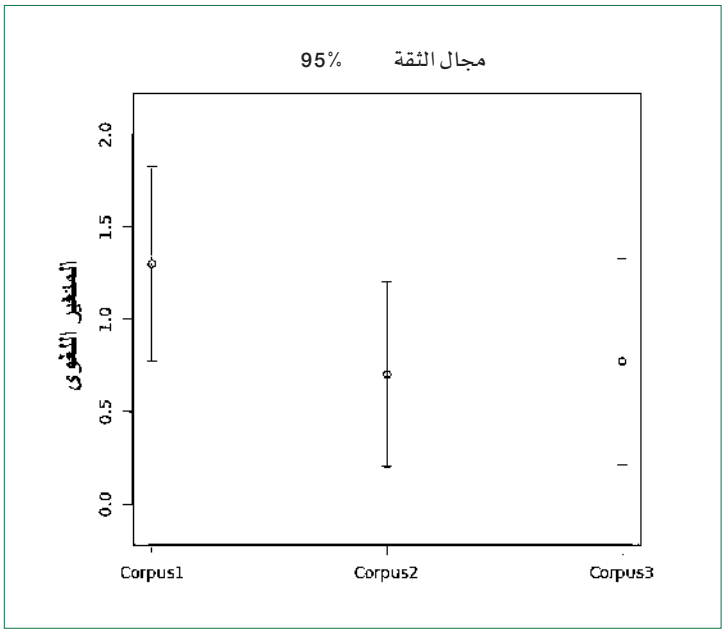
مثال:

- المتغير اللُّغَوِيّ: كلمة (ما).
- الاستدلال inference.
- الـمُدَوّنَات: ثـلاث مُدَوّنَات منفصلة. الـمُدَوّنَة الأولى (corpus1): جزء عمَّ، والـمُدَوّنَة الثَّانيـة (corpus2) جـزء تبـارك، والـمُدَوّنَة الثَّالثـة (corpus3) جزء قد سـمع.
- عدد النُّصُوص: ٥٧ نَصًّا (على اعتبار عدد السُّوَر في هذه الأجزاء الثَّلاثة).

ويُهيّـأ شـكلُ البَيَانَـات كمـا في الشَّـكل (١٧)، إذ يتضمَّـن العمـود الأوَّل (النَّـصّ Text)، وكل نَـصّ محفـوظ برقـم ترتيـب السُّـورة في المصحـف الشَّـريف في الأجـزاء القرآنيَّـة الثَّلاثـة الأخـيرة، ويتضمَّـن نَـصّ السُّـورة، وبعـد المعالجـة يخـرجُ تمثيـل المَخَطّـط الصنـدوقي كمـا في الشَّـكْل (١٨).

Text	Corpus1	Corpus2	Corpus3
text58	0	0	7
…	…	…	…
text67	0	5	0
…	…	…	…
text78	2	0	0

شكل ١٧ : سبعةٌ وخمسون نَصًّا (٥٧ سورة) في الأجزاء الثلاثة الأخيرة



شكل ١٨ : أعمدة الخطأ

رابعًا: مخطَّط التَّشَتُّت scatter plot

يُجرى هذا الشَّكْلُ البياني للإحصاء الوصفيِّ، ولتمثيل هذا المخَطّط؛ تحتوي البَيَانَات اللُّغَوِيّة على مُتَغَيّرَات لُغَوِيَّةٍ عديدةٍ، ومدوّنةٍ لُغَوِيَّةٍ واحدةٍ تَتَضَمَّن نصوصًا عديدةً.

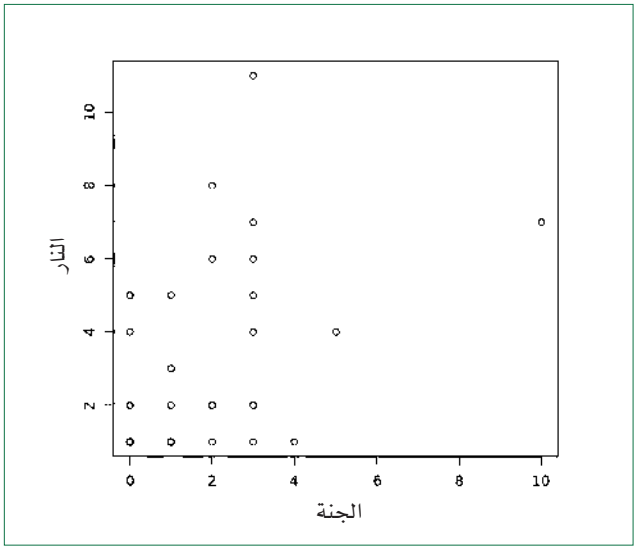
مثال:

- المُتَغَيّرَات اللُّغَوِيّة: كلمة (الجنة) heaven وكلمة (النار) hell.
- الوصف description.
- المُدَوّنَة: القرآن الكريم كاملًا.
- النُّصُوص: ٣٠ نصًّا، ويَتَضَمَّن كُلّ نَصّ جزءًا من أجزاء القرآن الكريم.

ويُهيّأ شكل البَيَانَات كما في الشَّكْل (١٩)، إذ يَتَضَمَّن العمودُ الأول (النُّصُوص Texts)، ويحمل اسم كُلّ ملفٍّ نصِّيٍّ رقمَ الجزء في المصحفِ الشَّريف، ويحتوي على نَصِّ سورِه، وبعد الضَّغط على مفتاح (أنشئْ الشَّكْلَ Create graph) يخرج تمثيلُ مخطط التَّشَتُّت كما في الشَّكْل (٢٠).

Texts	heaven	hell
1	3	5
2	2	6
3	0	5
...	...	...
30	3	2

شكل ١٩: الجنة والنار بين أجزاء القرآن



شكل ٢٠ : مخطط التَّشَتُّت

خامسًا: مخطَّط خطِّيّ line chart بخطِّ انحداريٍّ regression line

يُجرَى هذا الشَّكْلُ البيانيُّ للإحصاء الاستدلاليِّ، ولتمثيل هذا المخَطّط؛ تحتوي البَيَانَات اللُّغَوِيّة على مُتَغَيّرَات لُغَوِيَّةٍ عديدةٍ، ومدوّنةٍ لُغَوِيَّةٍ واحدةٍ تَتَضَمَّن نصوصًا عَديدةً.

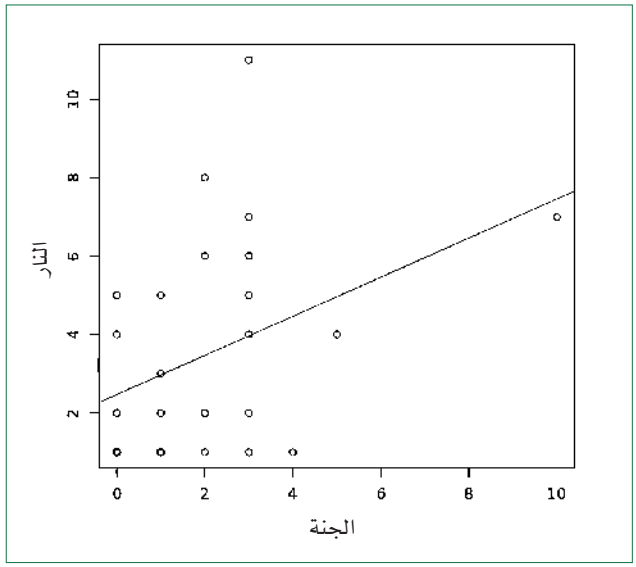
مثال:

- المُتَغَيّرَات اللُّغَوِيّة: كلمة (الجنة) heaven وكلمة (النار) hell.
- الاستدلال inference.
- المُدَوّنَة: القرآن الكريم كاملًا.
- النُّصُوص: ٣٠ نصًّا، ويَتَضَمَّن كُلّ نَصّ جزءًا من أجزاء القرآن الكريم.

ويُهيّـأ شـكل البَيَانَـات كمـا في الشَّـكْل (٢١)، إذ يَتَضَمَّـن العمـود الأول (النُّصُـوص Texts)، ويحمـل اسـم كُلّ ملـفٍّ نـصِّيٍّ رقـم الجـزء في المصحـف الشَّـريف، ويحتـوي علـى نَـصِّ سـوره، وبعـد الضَّغـط علـى إنشـاء الشَّـكْل يخـرج تمثيـلُ المَخَطّـط الخطِّـيِّ الانحـداريِّ كمـا في الشَّـكْل (٢٢).

Texts	heaven	hell
1	3	5
2	2	6
3	0	5
...	...	...
30	3	2

شكل ٢١ : الجنة والنار بين أجزاء القرآن



شكل ٢٢ : المَخَطّط الخطّيّ الانحداريِّ

سادسًا: مخطَّط خطِّيّ line chart

يُجرى هذا الشَّكْلُ البياني للإحصاء الوصفيِّ، ولتمثيل هذا المخَطّطِ؛ تحتوي البَيَانَاتُ على مُتَغَيّرَاتٍ لُغَوِيَّةٍ تُقاس بين نصوص داخل مُدَوّنَةٍ واحدةٍ، أو بين مُدَوّنَاتٍ متعدِّدةٍ.

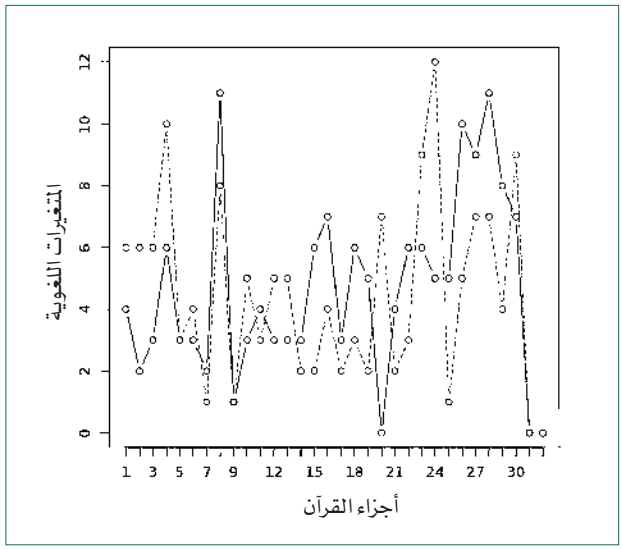
مثال:

- المُتَغَيّرَات اللُّغَوِيّة: صيغ (جنة) heaven و(نار) hell في القرآن الكريم.
- الوصف description.
- المُدَوّنَة: القرآن الكريم كاملًا.
- النُّصُوص: ٣٠ نصًّا، ويَتَضَمَّن كُلّ نَصٍّ جزءًا من أجزاء القرآن الكريم.

ويُهيّأ شكلُ البَيَانَات كما في الشَّكْل (٢٣)، ويَتَضَمَّن العمود الأول (النُّصُوص Texts)، ويشير اسم كُلّ ملفٍّ نصّيٍّ إلى رقمِ الجزء في المصحف الشَّريف، ويحتوي على نَصِّ سوره، وبعد الضغط على إنشاء الشَّكْل يخرجُ تمثيل مخطَّطٍ خطيٍّ كما في الشَّكْل (٢٤).

Corpus	heaven	hell
1	4	6
2	2	6
3	3	6
4	6	10
5	3	3
...	...	...
30	7	9

شكل ٢٣: اشتقاقات الجنة والنار بين أجزاء القرآن



شكل ٢٤ : المخَطّط الخطي

سابعًا: الخريطة الجغرافِيَّة geomapping

يُجرى هذا الشَّكْلُ البياني للإحصاء الوصفيِّ. ولتمثيل عدد تكرار Frequency الكَلِمَات في المواقع الجغرافِيَّةِ؛ تحتوي البَيَانَاتُ على مُتَغَيّرَاتٍ لُغَوِيَّةٍ بين الأماكن الجغرافيَّة Place، مع تحديد خطِّ الطُّول Longitude وخطِّ العرض Latitude لكلِّ موقعٍ جغرافِيٍّ.

مثال:

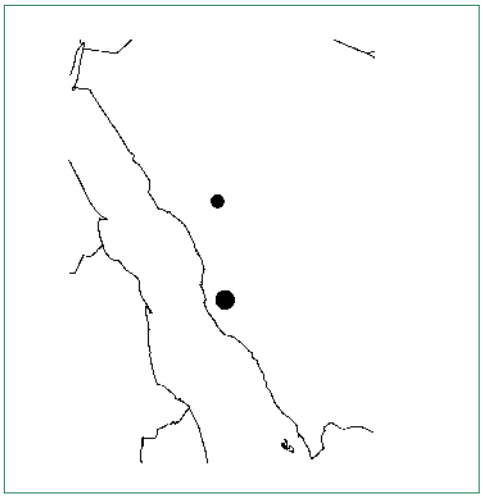
- المُتَغَيّرَات اللُّغَوِيّة (عدد الكَلِمَات بين السُّوَر المكِّيَّة والسُّوَر المدَنِيَّة).
- المُدَوّنَة: مُدَوّنَة السُّوَر المكِّيَّة، ومدونة السُّوَر المدَنِيَّة.
- النُّصُوص: ٨٦ نَصًّا (أي ٨٦ سورة مَكِّيَّة) و٢٨ نَصًّا (أي ٢٨ سورة مدنيَّة).

ويُهيّأ شكلُ البَيَانَات كما في الشَّكْل (٢٥)، إذ يَتَضَمَّن العمودُ الأول السُّوَرَ المكِّيَّة والسُّوَر المدَنِيَّة، ثم المواقع الجغرافِيَّة لمكَّة المكرَّمة والمدينة المنوَّرة، ثم إجمالي تكرارات

الكَلِمَـات الفِعْلِيّـة للمكِّيَّـة ثـم للمدنيَّـة، وبعـد الضَّغـط علـى إنشـاء الشَّـكْل يخـرجُ تمثيـل الخريطـة الجغرافيَّـة كمـا في الشَّـكْل (٢٦).

Place	Longitude	Latitude	Frequency
Makkiyya	39.826168	21.42251	47642
Madaniyya	39.5692	24.5247	30157

شكل ٢٥ : بيانات المواقع الجغرافِيَّة [١]



شكل ٢٦ : تمثيل بيانات المواقع الجغرافِيَّة

(١) القـول الـذي يذهـب إلى أن السُّـوَر المكِّيَّـة قـد نزلـت في مكَّـة، وأن السُّـوَر المدَنِيَّـة قـد نزلـت في المدينـة، قـولٌ ضعيـفٌ؛ لأن هنـاك آيـاتٍ مكِّيَّـةً قـد نزلـت في غيـر مكان مكـة، وآيـاتٍ مدنيَّـةً قـد نزلـت في غيـر مكان المدينـة، والقـول الراجـح في تصنيـف السُّـورَة المكِّيَّـة والسـورة المدَنِيَّـة هـو أن التَّصنيـف إنمـا هـو باعتبـار الزَّمـان، وليـس باعتبـار المكان، فالسُّـوَر المكِّيَّـة قـد نزلـت في الزَّمـن الـذي سـبق زمـنَ الهجـرة النبويَّـة إلى المدينـة، والسُّـوَر المدَنِيَّـة قـد نزلـت في الزمـن الـذي تـلا زمـن الهجـرة النبويَّـة، وهـو القـول الرَّاجـح، وذكـره ابـنُ النَّقيـب في مقدمـة (تفسـيره)، وأخرجـه عثمـان بـن سـعد الـرازي بسـنده إلى يحيى بـن سـلام. السـيوطي. الإتقـان في علـوم القـرآن، تحقيـق: شـعيب الأرنـؤوط، مؤسسـة الرسـالة، دمشـق، ٢٠٠٨، الصفحـات ٣٢-٣٢. إذن، فالغـرض مـن التَّصنيـف المكانيِّ في تمثيـل الخريطـة الجغرافيَّـة هـو التَّعريـف بأسـلوب تمثيـل بيانـات الـمُتَغَيِّرَات اللُّغَوِيّـة القائمـة علـى خطـوط الطُّـول وخطـوط العـرض ليـس إلا.

ثامنًا: مخطَّط الأعمدة المرصوفة stacked barchart

يُجرى هذا الشَّكْل البيانيُّ لتمثيل مُتَغَيّرَات لُغَوِيَّةٍ ضمن سجلاتٍ لُغَوِيَّةٍ Registers، أي: أكثر من متغيِّرٍ لُغَوِيّ واحدٍ يُقاس بين أكثر من سجلٍّ لُغَوِيٍّ (أدب، وسياسة، واقتصاد، إلخ).

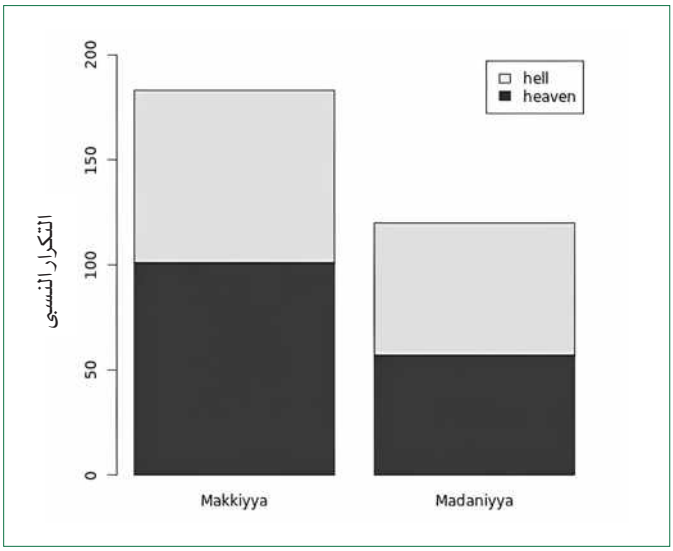
مثال:

- المُتَغَيّرَات اللُّغَوِيّة: صيغ (جنة) heaven و(نار) hell في القرآن الكريم.
  - صيغ (جنة) في السُّوَر المكِّيَّة: الجنة، جنات، جنة، جنتين، جنتان، جنتك، جنتي، جنته.
  - صيغ (جنة) في السُّوَر المدَنِيَّة: جنات، الجنة، جنة، جنتان.
  - صيغ (نار) في السُّوَر المكِّيَّة: النار، نارا، نار.
  - صيغ (نار) في السُّوَر المدَنِيَّة: النار، نارا، نار.
- المُدَوّنَة: مُدَوّنَة السُّوَر المكِّيَّة، ومدونة السُّوَر المدَنِيَّة.
- النُّصُوص: ٨٦ نَصًّا (أي ٨٦ سورة مكية) و٢٨ نَصًّا (أي ٢٨ سورة مدنية).
- المُتَغَيّرَات المقياسِيَّة: التّكْرَارات المطلقة.

ويُهيّأ شكل البَيَانَات كما في الشَّكْل (٢٧)، وبعد الضَّغط على إنشاء الشَّكْل يظهرُ تمثيل مخطَّط الأعمدة المرصوفة ك ما في الشَّكْل (٢٨).

Register	heaven	hell
Makkiyya	101	82
Madaniyya	57	63

شكل ٢٧: صيغ الجنة والنار بين السُّوَر المكِّيَّة والسُّوَر المدَنِيَّة



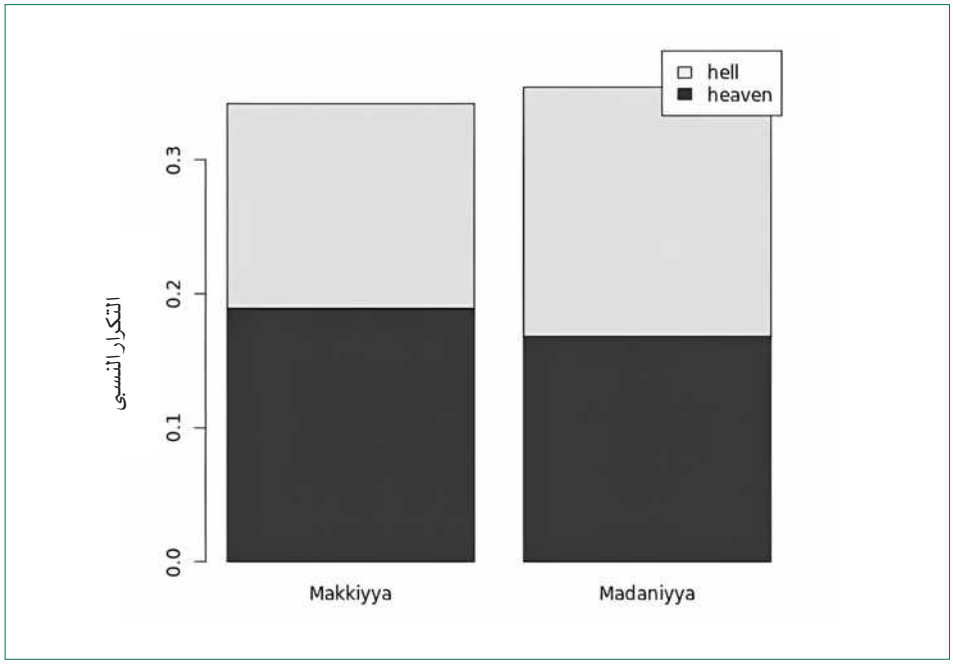
شكل ٢٨ : مخطَّط الأعمدة المرصوفة

- الـمُتَغَيّرَات اللُّغَوِيّة : صيغ (جنة) heaven و(نار) hell في القرآن الكريم.
  - صيغ (جنة) في السُّوَر المكِّيَّة : الجنة، جنات، جنة، جنتين، جنتان، جنتك، جنتي، جنته.
  - صيغ (جنة) في السُّوَر المدَنِيَّة : جنات، الجنة، جنة، جنتان.
  - صيغ (نار) في السُّوَر المكِّيَّة : النار، نارا، نار.
  - صيغ (نار) في السُّوَر المدَنِيَّة : النار، نارا، نار.
- الـمُدَوّنَة : مُدَوّنَة السُّوَر المكِّيَّة، ومدوّنة السُّوَر المدَنِيَّة.
- النُّصُوص : ٨٦ نَصًّا (أي ٨٦ سورة مَكِّيَّة) و٢٨ نَصًّا (أي ٢٨ سورة مدنيَّة).
- الـمُتَغَيّرَات المقياسِيَّة : التّكْرَارات النِّسبِيَّة.

ويُهيّأ شكل البَيَانَات كما في الشَّكْل (٢٩)، وبعد الضَّغط على إنشاء الشَّكْل يظهر تمثيل مخطَّط الأعمدة المرصوفة كما في الشَّكْل (٣٠).

Register	heaven	hell
Makkiyya	0.189	0.153
Madaniyya	0.168	0.186

شكل ٢٩ : صيغ الجنة والنار بين السُّوَر المكِّيَّة والسُّوَر المدَنِيَّة



شكل ٣٠ : مخطَّط الأعمدة المرصوفة

تاسعًا: خطوط المؤشّر sparklines

يُجرى هذا الشَّكْل البياني لتمثيل سلسلة اتجاه المتغيّر اللُّغَوِيّ، وقد تكون هذه السِّلسلة تواريخ زمنِيَّةً متتابعةً، أو حالات متتابعةً، أو أحداثًا متتابعةً، ولتمثيل هذا المخَطّط؛ تحتوي البَيَانَاتُ على متغيّرٍ لُغَوِيٍّ واحد، أو مُتَغَيّرَاتٍ لُغَوِيَّةٍ عَديدَةٍ.

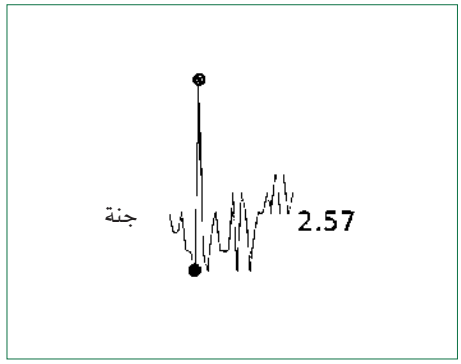
مثال:

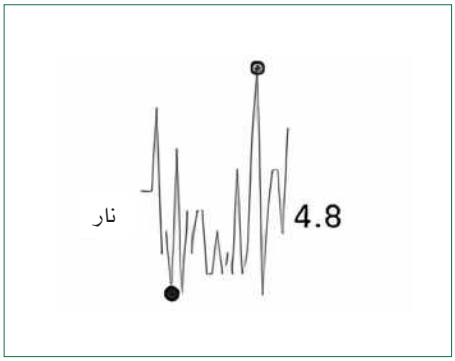
- المُتَغَيّرَات اللُّغَوِيّة: صيغ (جنة) heaven و(نار) hell في القرآن الكريم.
  - صيغ (جنة) في السُّوَر المكِّيَّة: الجنة، جنات، جنة، جنتين، جنتان، جنتك، جنتي، جنته.
  - صيغ (جنة) في السُّوَر المدَنِيَّة: جنات، الجنة، جنة، جنتان.
  - صيغ (نار) في السُّوَر المكِّيَّة: النار، نارا، نار.
  - صيغ (نار) في السُّوَر المدَنِيَّة: النار، نارا، نار.
- المُدَوّنَة: القرآن الكريم كاملًا.
- النُّصُوص: ٣٠ نَصًّا يُمثل الثلاثين جزءًا في القرآن الكريم.
- تمثيل السِّلسلة الزمنيَّة: من ٠٠٠١ إلى ٠٠٢٣ (أي: السَّنواتُ الثَّلاث والعشرون التي نزلت فيها آياتُ القرآن الكريم).

ويُهيّأ شكل البَيَانَات كما في الشَّكْل (٣١)، وبعد الضَّغط على إنشاء الشَّكْل يظهرُ تمثيل خطوط المؤشر كما في الشَّكْل (٣٢).

Candle	heaven	hell
0001_0002	3	1
0002_0003	2	3
0003_0004	0	0
0004_0005	0	0
0005_0006	2	0
...	...	...
0023_0024	12	10

شكل ٣١ : صيغ الجنة والنار بين السنوات الزمنية (١-٢٣)





شكل ٣٢ : خطوط المؤشر

عاشرًا: مخطَّط الشموع candlestick plot

يُجرى هذا الشَّكْل البياني لتمثيل سلسلة اتجاه المُتَغَيّر اللُّغَوِيّ، وقد تكون هذه السلسلة تواريخ زمنيَّةً متتابعةً، أو حالات متتابعةً، أو أحداثًا متتابعةً. ولتمثيل هذا المخَطّطِ؛ تحتوي البَيَانَاتُ على متغيِّرٍ لُغَوِيّ واحدٍ، أو مُتَغَيّرَاتٍ لُغَوِيَّةٍ عَديدةٍ.

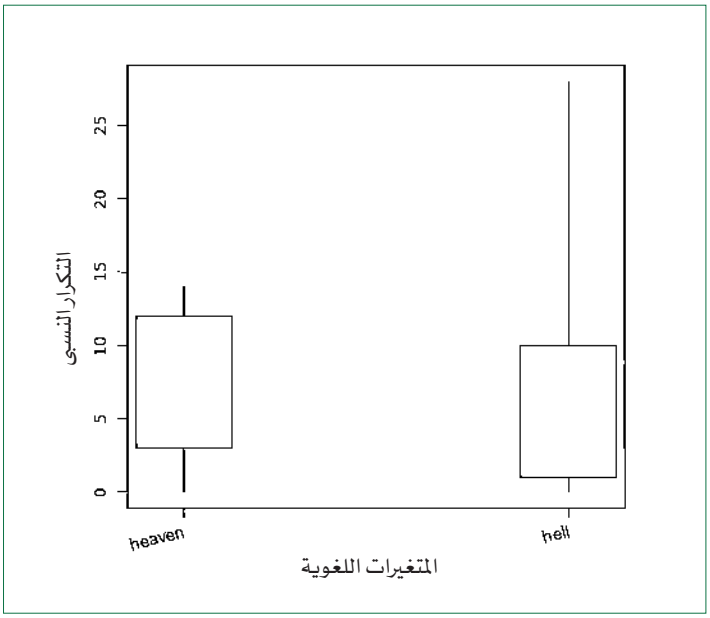
مثال:

- المُتَغَيّرَات اللُّغَوِيّة: صيغ (جنة) heaven و(نار) hell في القرآن الكريم.
  - صيغ (جنة) في السُّوَر المكِّيَّة: الجنة، جنات، جنة، جنتين، جنتان، جنتك، جنتي، جنته.
  - صيغ (جنة) في السُّوَر المدَنِيَّة: جنات، الجنة، جنة، جنتان.
  - صيغ (نار) في السُّوَر المكِّيَّة: النار، نارا، نار.
  - صيغ (نار) في السُّوَر المدَنِيَّة: النار، نارا، نار.
- المُدَوّنَة: القرآن الكريم كاملًا.
- النُّصُوص: ٣٠ نَصًّا يُمثل الثلاثين جزءًا في القرآن الكريم.
- تمثيل السِّلسلة الزمنيَّة: من ٠٠٠١ إلى ٠٠٢٣ (أي: السَّنواتُ الثَّلاث والعشرون التي نزلت فيها آيات القرآن الكريم).

ويُهيّأ شكل البَيَانَات كما في الشَّكْل (٣٣)، وبعد الضَّغط على إنشاء الشَّكْل يظهر تمثيل مخطَّط الشُّموع كما في الشَّكْل (٣٤)، الذي يشير إلى متوسط التكرار لكلمتي (جنة) و(نار) في جسم الشمعة، وإلى حجم بقية أعلى القيم وأقلِّها في الخط المتجهة من أعلى الجسم وأسفله.

Candle	heaven	hell
0001_0002	3	1
0002_0003	2	3
0003_0004	0	0
0004_0005	0	0
0005_0006	2	0
...	...	...
0023_0024	12	10

شكل ٣٣ : صيغ الجنة والنار بين السنوات الزمنية (١-٢٣)



شكل ٣٤ : مخطّط الشُّموع

الفصل الثَّانِي
تحليل الخِطاب

قبل البدءِ في تطبيقاتِ مقاييسِ حساب التَّصاحبِ اللَّفْظِيّ، أنبّه إلى أهَمِّيَّة معالجةِ النَّصّ المراد تحليله، وما سيُذكر هنا حول طبيعة هذا النَّصِّ ما هو إلا من بعض التَّجاربِ التي طَبَّقتها، والتي تبيّن منها أن النَّصّ المراد تحليله يحتاجُ إلى معالجةٍ محدَّدةٍ بحدود الغرض من تحليل الخطاب. ولأغراض تحليل الخطاب أو النَّصّ -من وجهة نظري- أربع بنياتٍ خطابيَّة، وتَأخذ كُلُّ بنيةٍ معالجةً شكلِيَّةً محددةً.

البنية الشَّكليَّة الأولى: تحليل النَّصّ الخطابيّ المكتوب كما هو، الذي لا يتطلَّب معالجةً قبليَّةً بالتَّقْطِيع، وبفصل المباني النَّحْوِيّة المتَّصلة في الكَلِمَات، مثل فصل حرف العطف (و) عن الكلمة التي بعدها.

البنية الشَّكليَّة الثَّانية: تحليل النَّصّ الخطابيِّ الذي يحتاج إلى معالجةٍ آليَّةٍ قبلِيّةٍ قبل تحليله، إما بالتَّقْطِيعِ، أو بإيقاف كلماتٍ لا تفيد في التَّحليل، مثل الكَلِمَات الوظيفيَّة أو غيرها. يفيد التَّقْطِيع مثلًا تحليل الإحالات النَّصِّيَّةِ؛ وذلك نظرًا إلى وجود ضمائر مُتَّصلةٍ يتطلَّب فصلها عن الكلمةِ؛ وسبب الفصل يتأتّى من طبيعة وظائف أدوات التَّحليل في كشفها عن الأنماط والأمثلة بخاصِّيَّة النغراميات -التتابعات اللَّفْظِيّة- (n-grams). أما استعمال كلماتِ الإيقافِ؛ فيفيد بتحليل الكَلِمَات المعجَمِيّة ذات المحتوى وذات الدّلالة المعجَمِيّة، والوقوف على كشفِ استعمالات الأنماط شبكِيًّا في البنية النَّصِّيَّة[1].

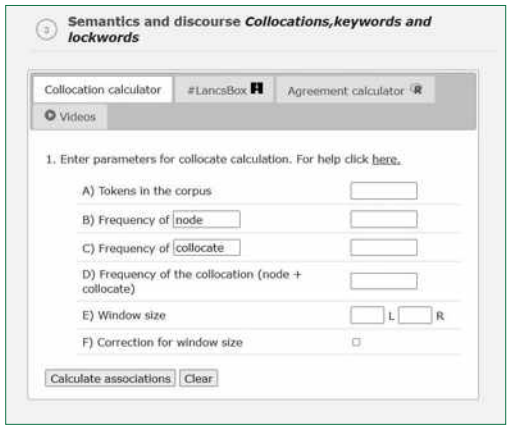
(١) لكلماتِ الإيقاف قوائمُ مختلفةٌ باختلاف ازديادِ عَدَدِها أو نقصانه، غير أن مفهومها الدقيق يشير إلى الكلمات الوظيفيَّة function words التي لا تدل على معنى بذاتها، بل تدل على وظيفةٍ محددةٍ سياقيًّا عند تلازمها مع الكلماتِ ذات المحتوى (الكلماتُ المعجميّة). يُمكن الاستفادةُ من هذه الدراسةِ التي قارَنَتْ الكلماتِ الوظيفيَّةَ بين اللغة العربية القديمة واللغة العربية المعاصرة، وفي مدونة حاسوبية بلغ عدد الكلمات فيها ١٠ ملايين كلمة:

Almujaiwel, S. (2019). Grammatical construction of function words between old and modern written Arabic: A corpus-based analysis. Corpus Linguistics and Linguistic Theory, 15(2). 267296-.

البنيـة الشـكليَّة الثالثـة: تحليـل الكيانـات والموضوعـات، ويكـون منطلـق الغـرض فيـه الوقـوف علـى أسـماء الكيانـات (الأشـخاص، والمؤسَّسـات، والمنَظَّمـات، إلـخ)، أو علـى الموضوعـات والتَّعقيبـات عليهـا (الموضوع-التَّعقيـب topic-comment)، وهـذه سـأرجئها إلى الفصـل الثَّالـث مـن البَـاب الثَّـانِي؛ نظـرًا لإمكانيـة تحليلهـا آلِيًّـا باسـتخدام أداة (أورانج).

البنيـة الشَّـكليَّة الرَّابعـة: الخطـاب الشَّـفهيّ والحـوارات، وهـي بنيـةٌ تتطلَّـبُ معالجـةً ترميزيَّـةً، وتخلـو العَرَبيَّـة مـن مُدَوّنَـةٍ حاسـوبيَّةٍ تَتَضَمَّـن نَصًّـا شـفهيًّا مرمـوزًا برمـوز النَّـبر والتَّنغيـمِ، أو نَصًّـا حوارِيًّـا مرمـوزًا برمـوز تحليـل الحـوار، وهـذه الرمـوز تفيـدُ فيمـا لـو كانـت هنـاك مُدَوّنَـة شـفويَّة؛ وفائدتهـا تتجلـى في إمكانيَّـة تحليـل عينـةٍ مفيـدةٍ للدِّراسـة، تتَّسـم بمعاييـر اختيـار العيِّنَـات والتَّمثيـل العـادل وعـدم الانحيـاز.

مـا الـذي سـيُطبّق مـن إحصـاءاتٍ؟ سـيُطبّق في هـذا الفصـل إحصـاءات تتعلَّـق بعلـم الدّلالـة وتحليـل الخطـابِ، وهـي: مقاييـس حسـاب التَّصاحـب اللَّفْظِـيّ، واسـتخراج الكَلِمَـات المفتاحِيَّـة (keywords) -المبحـث الأول-، وحسـاب مواءمـة المقـدّر البيـنِيّ -المبحـث الثـاني-، ويُمكـن إجـراء تلـك المقاييس عـن طريـق اختيـار حاسـبة التَّصاحـبِ (Collocation calculator)، وذلـك الحسـاب عـن طريـق اختيـار حاسـبةِ التَّوافـقِ (Agreement calculator) في أدوات (لانكسـتر) الإِحْصَائِيَّـة كمـا في الشَّـكْل (٣٥).



شكل ٣٥ : واجهة حساب التَّصاحب اللَّفْظِيّ في (علم الدّلالة والخطاب)

المبحث الأَوَّل: التَّصاحب اللَّفْظِيّ

التَّصاحـبُ (collocation) يعـني الكلمـة ومتصاحبهـا، والكلمـة الرئيسـة هـي منطلـقُ التَّحليـل، والكلمـة المتصاحبـةُ (collocate) هـي مـا جـاورت الكلمـة علـى نافـذة التَّصَاحُـب مـن يمـين الكلمـة الرَّئيسـة ومـن يسـارها، ويمتـدُّ مـن (يم١-يـس١) إلى (يم٥-يـس٥). يحتـوي هـذا المبحـث علـى تطبيقـاتٍ تتعَلَّـق بالتَّصَاحُـب اللَّفْظِـيّ. وينقسـم المبحـث إلى قسـمين.

الأَوَّل: مُعاملات مقاييس التَّصَاحُب اللَّفْظِيّ.

الثَّانِي: شبكات التَّصَاحُب اللَّفْظِيّ.

أَوَّلًا: مُعاملات مقاييس التَّصَاحُب اللَّفْظِيّ

تعمـل مقاييـس حسـاب التَّصَاحُـبِ (collocation calculator parameters) علـى كشـف شـبكات التَّصَاحُـب اللَّفْظِـيّ في النَّـصِّ، واعْتَمَـدتُ علـى بنيـة نَـصٍّ خاليـةٍ مـن التَّشـكيل، ومعالجـةٍ بالتَّقْطِيـع (مقطّـع سـتانفورد)؛ لتحليـل التَّصَاحُـب اللَّفْظِـيّ. سـأُوضّحُ هنـا فائـدة كُلّ مُعامـلٍ مـن معامـلات مقاييـس التَّصَاحُـب اللَّفْظِـيّ، ويبلـغ عددُهـا ١٢ مُعامـلًا: المتبـادل (mutual -MU)، والمعلومـات المتبادلـة (mutual information MI-)، والمعلومـات المتبادلـة٢ (MI2)، والمعلومـات المتبادلـة٣ (MI3)، والأرجحيَّـة اللوغارثميـة (log likelihood LL-)، وقيـاس-ز (z-score)، وقيـاس-ت (t-score)، والـدايـس (Dice)، واللـوغ دايـس (log dice). ونسـبة اللوغارثـم (log ratio)، والحساسـية الصغـرى (minimum sensitivity)، ودلتـا ب (delta P).

مـا فائـدة هـذه المقاييـس؟ كُلّ مقيـاسٍ يُخـبر عـن جانـبٍ مـن جوانـب الاعتبـار لعلاقـة التَّصَاحُـب بـين الكلمتـين، وسـيُطبّق ذلـك علـى التَّصَاحُـب اللَّفْظِـيّ (غفـور رحيـم)، والتَّصَاحُـب اللَّفْظِـيّ (الباقيـات الصَّالحـات) في مُدَوّنَـة القـرآن الكريـم.

- حجم الـمُدَوّنَة (٨٧٤٥٦ كلمة فعليَّة)، وهي مُدَوّنَة معالجة بالتَّقْطِيع كما أخبرتُ بذلك سابقًا.
- وردت كلمة (غفور) ٥٨ مرَّة، وكلمة (رحيم) ٦١ مرَّة، والتَّصَاحُب اللَّفْظِيّ (غفور رحيم) ٤٩ مرَّة.
- وردت كلمة (الباقيات) مرَّتين، وكلمة (الصَّالحات) ٦١ مَرَّة، والتَّصَاحُب اللَّفْظِيّ (الباقيات الصَّالحات) مرَّتين.

– **أسئلةٌ مهِمَّةٌ:**

كيف تعمل مقاييس التَّصَاحُبِ؟

كيف تُحسب آليًّا في أدوات (لانكستر) الإِحْصَائِيَّة؟

كيف نُفسّر كُلّ قياسٍ على حدة؟

تعمل مقاييس التَّصَاحُب وفق خمس خطواتٍ:

الأولى: يُحدّد حجم النَّافذة للتَّصاحب، فمثلًا: (R1-L1) تعني كلمةً على اليمين وكلمةً على اليسار (انظر: شبكات التَّصَاحُب في هذا المبحث)، وأختصرها بالعَرَبِيَّة بـ (يم١-يس١).

الثَّانية: تُحسب آليًّا التّكْرَارات الملحوظة (أي المطلقة كما هي) للكلمة ومتصاحبها.

الثَّالثة: تُوزّع في جدولةٍ تقاطعِيَّةٍ تلك التّكْرَارات الملحوظة، ويكون في الصُّفوف عدد مَرَّات وجود الكلمةِ، وعدد مرَّات غيابها عن المتصاحبة، ويكون في الأعمدةِ عدد مرَّات وجود المتصاحبة وعدد مرَّات غيابها.

الرَّابعة: تُحسب آليًّا التّكْرَارات المتوقَّعة النَّظرية في جدولةٍ تقاطعِيَّةٍ.

الخامسة: تَستخرج الأداةُ مقاييسَ التَّصَاحُب اللَّفْظِيّ، وقد جمعتُ كُلّ نتائج (غفور رحيم) و(الباقيات الصّالحات) في جدولٍ واحد (الجدول ١٠).

سـأقفُ علـى نتـائج مقاييـس التَّصَاحُـب اللَّفْظِـيّ للمثالـين، دون الوقـوفِ علـى خوارزمِيَّـات كُلِّ مُعامـل منهـا، وسأشـيرُ إلى مُنطلقِ تفسـيركُلّ قيمـةٍ مـن تلـك القيـم، وطريقـة الإفصـاح بشـكلٍ عـامٍّ عـن قيمـة التَّصَاحُـب لـكلِّ معامـلٍ، وكيفيَّـة الاعتمـاد عليـهِ. هنـاك سـتّ حـالاتٍ عنـد اسـتخدام هـذه المُـعاملات لقيـاس التَّصَاحُـب اللَّفْظِـيّ.

جدول ١٠ : مقاييس التَّصَاحُب اللَّفْظِيّ

غفور رحيم				الباقيات الصالحات			
التَّكْرَارات الملحوظة				التَّكْرَارات الملحوظة			
	وجود المتصاحب	غياب المتصاحب	المجموع		وجود المتصاحب	غياب المتصاحب	المجموع
وجود الكلمة	49	9	58	وجود الكلمة	2	0	2
غياب الكلمة	12	87,386	87,398	غياب الكلمة	59	87,395	87,454
المجموع	61	87,395	87,456	المجموع	61	87,395	87,456
التَّكْرَارات المتوقعة				التَّكْرَارات المتوقعة			
	وجود المتصاحب	غياب المتصاحب	المجموع		وجود المتصاحب	غياب المتصاحب	المجموع
وجود الكلمة	0.040	57.960	58	وجود الكلمة	0.001	1.999	2
غياب الكلمة	60.960	87,337.040	87,398	غياب الكلمة	60.999	87,393.001	87,454
المجموع	61	87,395	87,456	المجموع	61	87,395	87,456
مقاييس التَّصَاحُب اللَّفْظِيّ				مقاييس التَّصَاحُب اللَّفْظِيّ			
MU	MI	MI2	MI3	MU	MI	MI2	MI3
1,211.233	10.242	15.857	21.472	1,433.705	10.486	11.486	12.486
T-score	Z-score	Log Dice	Dice	T-score	Z-score	Log Dice	Dice
6.994	243.4158	13.720	0.824	1.413	53.511	10.023	0.063
LL	Log ratio	MS	Delta P	LL	Log ratio	MS	Delta P
656.733	12.587	0.803	[0.803 ;0.844]	25.141	10.534	0.033	[0.0.32 ;0.999]
p < 0.0001				p < 0.0001			

حالات الاستخدام:

١. لاسـتنتاج نـدرة وحصريّـة التَّصَاحُـب: مقاييـس المتبـادل (mutual) في عائلـة مُعامـلات المعلومـات: المتبـادل (MU– mutual)، والمعلومـات المتبادلـة (MI– mutual information)، والمعلومات المتبادلة ٢ (MI2)، والمعلومات المتبادلة ٣ (MI3). ولا يُقصد بالنَّدرة والحصريَّة هنا معنى الضّعف، بل الأهَمِّيَّة، فحصرِيَّة الذَّهب

على ندرتِه تجعل له قيمةً أعلى من غيره من المعادن[1]. يظهر ناتج المتبادل بقسمة التّكْرَار الملحوظ لكلمة (غفور؛ 49 مرة) على التّكْرَار المتوقَّع لها (0.040)[2]. يتغيَّر ناتج هذه القسمة للمعلومات المتبادلة باستخراج (log2) لناتج المتبادل، ويزيد هذا النَّاتج للمعلومات المتبادلة ٢ باستخراج أس اثنين (س٢) للتّكرار الملحوظ (٤٩) ليكون (2401)، ويزيد للمعلومات المتبادلة ٣ باستخراج أس ثلاثة (س٣) للتّكرار الملحوظ (٤٩) ليكونَ (117649). انظر إلى ناتج كُلّ مُعاملٍ كما هو في الآتي:

- المتبادلة: (49 ÷ 0.040) = (1211.233)
- المعلومات المتبادلة: (49 ÷ 0.040) = (10.242 = (log2 (1211.233)
- المعلومات المتبادلة ٢: (2401 ÷ 0.040) = (15.857 = (log2 (59350.4169)
- المعلومات المتبادلة ٣: (117649 ÷ 0.040) = (21.471) = log2 (2908170.4329)

لماذا هذا التَّضاعفُ الأسيّ في المعلومات المتبادلة ٢ والمعلومات المتبادلة ٣؟ ولماذا لا يُكتفى بالمتبادلة أو بالمعلومات المتبادلة؟ المتبادلة لا يمكن فهمها دون مُعايرةٍ وضبطٍ، ويصعب تحديد قيمة قوَّتها بين القيم الأخرى، أما المعلومات المتبادلة فهي منضبطةٌ بين قيمٍ محدودةٍ. أما سبب التَّضاعف في المعلومات المتبادلة ٢ والمعلومات المتبادلة ٣؛ فعائدٌ إلى طبيعة اللغة في الـمُدَوّنَة اللُّغَوِيّة التي يجمعها المحَلِّل للتَّحليل والاختبار، يُشار

(١) لنعد إلى تجربةٍ أجريتُها في عام ٢٠١٥ للنظر في التَّصَاحُب اللَّفْظِيّ (عاصفة شديدة) و(عاصفة قاشرة)، إذ ظهرت قيمة المعلومات المتبادلة لـ (عاصفة قاشرة) أعلى من التَّصَاحُب الأوَّل، ويُمكن تفسير الحصريَّة هنا على أن التَّصَاحُب الثَّاني أندر، فـ (شديد) ترد مع كلماتٍ عديدةٍ من غير (عاصفة) على عكسِ (قاشرة)، ويُستفاد من مثل هذه التَّطْبيقَات في بناء معاجِم التَّصَاحُب اللَّفْظِيّ، أو التَّلازم اللَّفْظِيّ كما يسميه البعض، وذلك بجعل التَّصَاحُب النادر مقدَّمًا في التَّمثيل عن التَّصَاحُبات الشَّائعة. انظر: المجيول، سلطان. (٢٠١٥). البحث اللُّغَوِيّ في الـمُدَوّنَات العَرَبِيَّة الحَاسُوبِيَّة بين الممكن والمحتمل والمأمول، صفحة ٢٥٣.

(٢) قيمة التّكْرَار المتوقَّع بالصورة الأدق هي: 0.040454644151868385356079301009 8.

في أدبيَّـات لُغَوِيّـاتِ الـمُدَوّنَة إلى أن قيمـة المعلومـات المتبادلـة الدَّالـة تكـون مـن ٣ وأعلـى. حسـنًا؛ ولكـن مـاذا لـو مثّلـت الـمُدَوّنَة عينـةً لا تُحقـق هـذه القيمـة فأعلـى؟ هنـا نحتـاج إلى أن نسـتعين بالمعلومـاتِ المتبادلـة٢، وإن كانـت القيمـةُ أقـلَّ فيهـا -وهـذا نـادرٌ- نسـتعين بالمعلومـات المتبادلـة٣. قـد لا تتحَقَّـقُ قيمـة ٣ فأعلـى للمعلومـات المتبادلـةِ في عَيِّنـةٍ لُغَوِيَّـة، وهـذا طبيعِـيٌّ، وحينئـذ يُحتكـمُ إلى القيمـة الأعلـى عندمـا لا نسـتعمل المفاهيـمَ الأُسِّـيَّة في المعلومـاتِ المتبادلـة٢ والمعلومـات المتبادلـة٣. لنأخـذ علـى سـبيلِ المثـال (السَّـماوات والأرض) بـين السُّـوَر المكِّيَّـة والسُّـوَر المدَنِيَّـة. يُخبرنـا قيـاس المعلومـات المتبادلـة بـالآتي:

التتابع	التّكْرَار في المكِّيَّة	التّكْرَار في المدَنِيَّة	المعلومات المتبادلة في المكِّيَّة	المعلومات المتبادلة في المدَنِيَّة
السماوات والأرض	92	41	0.1758	-0.3305

هنـا يُفصـح عـن أن دلالـة المعلومـات المتبادلـة لهـذا التَّصَاحُـبِ في السُّـوَر المكِّيَّـة أعلـى مـن دلالتـه في السُّـوَر المدَنِيَّـة.

٢. لاسـتنتاج حصرِيَّـة التَّصَاحُـب: الدايـس (dice)، واللـوغ دايـس (log dice)، أمـا قيمـة الدايـس فتكـون بـين (0) و(1)، وكلمـا تناقصـت القيمـةُ عـن (1) ازدادت معنويَّتـهُ الحصريَّـة. علـى سـبيل المثـال: قيمـة الدايـس للتَّصاحـب (الباقيـات الصالحـات) أقـلُّ مـن قيمتـه في التَّصَاحُـب (غفـور رحيـم)، كلاهمـا اكتسـبا الحصرِيَّـة، غـير أنَّ الأوَّل أكـثر حصريَّـةً. لا بـدَّ مـن التَّنبيـه إلى أنَّ تلـك القيمتـين دالَّتـان لـكلِّ تصاحـبٍ علـى حصريَّتهمـا في مُدَوّنَـة القـرآنِ الكريـم، ويتِـمُّ أخـذُ اعتبـارات الأقـلّ عندمـا ننظـرُ في التَّصَاحُـب بـين مدَوَّنتـين مختلفتـين، أمـا عنـد اسـتخدام معامـل اللـوغ دايـس؛ فالمفهـوم النَّظـريُّ هـو أنَّ القيمـة الـتي تظهـر تشـيرُ إلى أن التَّصَاحُـبَ قـد ورد كـذا مَـرَّة (حسـب قيمـة المـرَّات الظَّاهـرة) في كُلّ ١٤ ألـف كلمـة، ويتـم تفسـير اللـوغ دايـس مثـل تفسـير الدايـس مـن جهـة الحصريَّـة.

٣. لاستنتاج أثر وكثرة التَّصَاحُب: قياس-ت (t-score)، وقياس-ز (z-score). بالنسبة لمُعامل الأوَّل والثَّاني، فتكون قيمة ٢ فأكثر وقيمة ٣ فأكثر على التوالي بداية اعتبار الأثرِ (وليس للحصريَّة)، وينطبق حال الأمثلةِ التي تكون فيها القيمةُ كما قلنا في المعلوماتِ المتبادلة، فيُفصح عن الأعلى وإن كان أقل من القيم المقترحةِ، وعادةً تكون قيمةُ هذين المعاملين قليلةً جِدًّا لتأثُّرها بأحجام المُدَوّنَة الكبيرة، فهما مقياسان إشكاليّان عند الاستخدام.

٤. لاستنتاج معنويَّة التَّصَاحُب: الأرجحيَّة اللوغارثمية (LL)، ونسبة اللوغارثم (log ratio)، أما الأرجحيَّة اللوغارثمية فكلُّ قيمةٍ أعلى من الأخرى تدلُّ على معنويَّة التَّصَاحُب، وتُفسّر الأرجحيَّة بقيمة الاحتمال، وهل هي دالَّة أم لا، فإن كانت أقَلّ من ٠,٠٥ فهي دالَّةٌ، وإن كانت أعلى فهي غير دالَّةٍ، وتُظهر أدوات (لانكستر) الإِحْصَائِيَّة قيمةَ الاحتمال مع الأرجحِيَّة اللوغارثمية، أما نسبة اللوغارثم فهي نتيجةُ (log2) لقيمة الأرجحيَّة اللوغارثمية.

٥. لاستنتاج حجم زيادة نمُوِّ تكرار التَّصَاحُب وانخفاضه: الحساسية الصُّغرى (minimum sensitivity)، تُعطي قيمة هذا المُعامل إشارةً إلى أنَّهَا تزيد مع زيادة تكرارِ التَّصَاحُب، وتنخفض مع زيادة تكرار الكلمة وحدها، أو زيادة تكرار المتَصاحب وحده، ونلحظُ أنَّ قيمة الحساسية الصُّغرى لـ (الباقيات الصالحات) أقلُّ؛ وذلك نظرًا إلى كثرة تكرار كلمةِ (الصالحات) وحدَها، دون وجود سمةِ الكثرة للمتصاحب (الباقيات) وحده، فقِلَّةُ ورودِ كلمة (الباقيات) وحدها في مُدَوّنَة القرآن الكريم جعلت من القيمة أقلَّ مقارنةً بقيمتها لـ (غفور رحيم).

٦. لاستنتاج استعمال ثنائِيَّة الاتِّجاه: دلتا ب (delta P). يشيرُ هذا الاستنتاج إلى إمكانيَّة استعمال التَّصَاحُب استعمالًا ثنائيَّ الاتِّجاه كـ (غفور رحيم) و(رحيم غفور)، ويعطينا هذا المُعامل قيمتين اثنتين، ونلحظ أنَّ تقاربَ أو تشابهَ القيمتين إشارةٌ إلى أن التَّصَاحُب (غفور رحيم) والتَّصَاحُب العكسي (رحيم غفور) لهما

معـنى اسـتعمالِيٌّ، أي أنَّ اسـتعمالَهما مقبـولٌ، علـى عكـس (الباقيـات الصَّالحـات)، فقيمتُهمـا متباعـدةٌ، وليـس هنـاك معـنى اسـتعمالِيٌّ للتَّصاحـبِ العكـسِيِّ (الصَّالحات الباقيـات)، لمـاذا دلَّ التَّصَاحُـب العكـسِيُّ بالقيمتـين القريبتـين للتَّصاحـب (غفـور رحيـم)؟ سـبب ذلـك يعـود إلى وجـود التَّصَاحُـب العكـسِيِّ بواقـع مـرَّة واحـدَة في القـرآن الكريـم، وذلـك في قـولِ الله سـبحانه وتعـالى: ﴿ يَعۡلَمُ مَا يَلِجُ فِي ٱلۡأَرۡضِ وَمَا يَخۡرُجُ مِنۡهَا وَمَا يَنزِلُ مِنَ ٱلسَّمَآءِ وَمَا يَعۡرُجُ فِيهَاۚ وَهُوَ ٱلرَّحِيمُ ٱلۡغَفُورُ ٢ ﴾ [سـبأ: ٢].

نـأتي إلى كيفيَّـةِ اسـتخراجِ الكَلِمَـاتِ المفتاحِيَّـة؛ قبـل ذلـكَ أشـيرُ إلى أنَّ الكلمـةَ المفتاحِيَّـة هـي كلمـةٌ ممَـيَّزةٌ في النَّـصِّ، وتكتسـبُ هـذا التَّمـيُّزمـن كونهـا فريـدةً أو رئيسـةً بوجودِهـا في الـمُدَوّنَة اللُّغَوِيّـة، ولا أنصـح باسـتخدام الأرجحِيَّـة اللوغارثميـة؛ لأنهـا وحدهـا لا تكفـي في تفسـير الفـرادة أو نسـبةِ الفـرادة؛ فهـي لاسـتنتاج المعنويَّـة دون اسـتنتاج النّـدرة أو الحصرِيَّـة.

سـيُطبّق معرفـة اسـتخراج الكَلِمَـات المفتاحِيَّـة[1] في مُدَوّنَـة القـرآن الكريـم بإجـراء مقارنـةٍ بـين مُدَوّنَـة السُّـوَر المكِّيَّـة ومدوَّنـة السُّـوَر المدَنِيَّـة باسـتخدامِ أداةِ غـواص. سـنضع مُدَوّنَـة

(١) قـد يتسـاءلُ باحـث عـن كيفيـة اسـتخراج ثـلاث قوائـم مـن الـمُدَوّنَات: قائمـة الشـيوع، وقائمـة الألفـاظ الأكاديميـة، وقائمـة المصطلحـات العِلْمِيّـة؛ يُسـتعان باسـتخراج هـذه القوائـم مـن الـمُدَوّنَات التـي تَتَضَمَّـن نصوصًـا متعَـدِّدةَ الأوعيـة، ومـن ضمـن أوعيتهـا نصـوصٌ أكاديمِيّـةٌ، ومـن المعاييـر الـتي تضبـط اسـتخراجَ هـذه القوائـم هـو مـا أجرتـه الباحثـةُ حـوراء الجاسـم في رسـالتها للماجسـتير بإشـرافي. التجربـة خـير برهـان، ولغويـات الـمُدَوّنَة الحاسـوبيَّة علـمٌ تجريـبيٌّ، وقـد سـعينا إلى اختبـار المقاييـس لضبـط فواصـل تلـك القوائـم، وباسـتخدام أدوات وورد سـميث (WordSmith Tools)، وتـم الوصـول إلى مزيـة كُلّ مـن نسـبة اللـوج (log ratio)، ومعيـار معلومـات بايـزي (Bayesian Information Criterion BIC)، ومقاييـس التَّشـتُّت dispersion. ولمـن أراد الاسـتفادة مـن نتـائج هـذا التَّحليـل التَّجريـبي العـودة إلى المرجـع الآتي. الجاسـم، حـوراء. الألفـاظ الأكاديميـة في الـمُدَوّنَات العَرَبِيَّـة العِلْمِيّـة: دراسـة حاسـوبية تحليليـة. (رسـالة ماجسـتير) إشـراف: د. سـلطان المجيـول. قسـم اللغـة العَرَبِيَّـة (جامعـة الملـك سـعود)، ٢٠٢٢. الجاسـم، حـوراء، والمجيـول، سـلطان. المدونـة الأكاديميـة العربيـة (أواك) Academic Written Arabic Corpus AWAC (قائمـة الجاسـم والمجيـول للألفـاظ الأكاديميـة وقوائـم الجاسـم والمجيـول للمصطلحـات العربيـة العلميـة، ٢٠٢٢)، الرابـط: https://www.kaggle.com/hawraaljasim/datasets؛ والرابـط: https://www.arabic-ai.info/awac (تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢).

السُّـوَر المدَنِيَّـة (الـمُدَوّنَة الرئيسـة)، ومدونـة السُّـوَر المكِّيَّـة (الـمُدَوّنَة المرجعية)، ثم سـنفصح عـن القيـم الـتي تُخبرنـا بتميُّز الكلمـة أو التَّصَاحُـب اللَّفْظِـيّ في مُدَوّنَـةٍ دون أخـرى. تُمكّـن أداة غـواص حفـظ النَّتـائج في ملـف أكسـل، وسـأذكر هنـا أربعـة أمثلـة مـن الكَلِمَـات، كُلّ مثـالٍ يُعـبّر عـن مفهـومٍ مهـمٍّ، فبعـد إدراك هـذه المفاهيـم الأربعـة؛ سـيُصبح اتخـاذُ قـرار أيّ الكَلِمَـات تُعـدُّ مفتاحِيَّـةً، وأيّهـا لا تُعـدُّ؛ سـهلًا، وهـذه الكَلِمَـات هيَ: لفـظ الجلالـة (الله)، و(كلّا)، و(طاغون)، و(مولاكـم). أمـا القياسـات الـتي أوصي بهـا فهـي: الدايـس، واللـوغ دايـس، ومعامـل الغرابـةِ (weirdness coefficient) -الجـدول ١١-. سـنوضح هـذه المفاهيـم مـن خـلال هـذه الكَلِمَـات الأربَعـة، وقبـل ذلـك سـنوضِّح قيـم معامـل الغرابـة ودلالاتهـا، وتكـون قيـم معامـل الغرابـة إمَّـا (0)، أو (1)، أو قيمـة أقـلّ مـن (١)، أو قيمـة أعلـى مـن (1)، وإمـا قيمـة (Infinity).

- القيمة (1): تدل على أن الكلمة قد وردت ورودًا متساويًا في المدَوّنتين الرئيسَة والمرجعيَّة.
- قيمـة تكـون أقـل مـن (1): تـدلّ علـى أن الكلمـة قـد وردت في الـمُدَوّنَة الرئيسـة أكـثر مـن الـمُدَوّنَة المرجعِيَّـة.
- قيمـة تكـون أعلـى مـن (1): تـدلُّ علـى أنَّ الكلمـة قـد وردت في الـمُدَوّنَة المرجعِيَّـة أكـثر من الـمُدَوّنَة الرَّئيسـة.
- القيمـة (0): تـدلُّ علـى أنَّ الكلمـة لـم تـرد في الـمُدَوّنَة أبـدًا، سـواءً كانـت في الـمُدَوّنَة الرئيسـة أم في الـمُدَوّنَة المرجعِيَّـة، وإن كانـت في إحـدى المدونتـين؛ فقيمـة الـمُدَوّنَة الأخـرى سـتكون (Infinity) وسـتكون دالَّـة علـى أن الكلمـة قـد وردت فيهـا فقـط.

- **لفـظ الجلالـة (الله)**: تكَـرَّرت في السُّـوَر المدَنِيَّـة أكـثر مـن السُّـوَر المكيَّـة، وتكـون القيمـة الأقـل لـكلٍّ مـن الدايـس واللـوغ دايـس للكلمـة الـتي جـاء ورودهـا أقـلّ في إحـدى المدوّنتـين، ويُخبرنـا معامـل الغرابـة بـأنَّ ورود لفـظ الجلالـة في الـمُدَوّنَة الأولى أعلـى مـن الـمُدَوّنَة الثانيـة.
- **لفـظ (كلا)**: علـى عكـس لفـظ الجلالـة (الله)، وورودهـا في الـمُدَوّنَة الثانيـة (السُّـوَر المكِّيَّـة) أكـثر دلالـة.

- **لفظ (طاغون):** يخبرنا معامل الغرابة بالقيمة (Infinity)، وهي تدلُّ على أن الكلمة خاصَّةٌ بالـمُدَوّنَة الثَّانية (السُّوَر المكِّيَّة).
- **لفظ (مولاكم):** يُخبرنا معامل الغرابة بالقيمة (Infinity)، وهي تدلُّ على أنَّ الكلمة خاصَّةٌ بالـمُدَوّنَة الأولى (السُّوَر المدَنِيَّة).

كيف يُمكن فرزُ الكَلِمَات المَمَيّزة والكَلِمَات المفتاحِيَّة؟ الجواب: في ملف أكسل، أَعِدْ ترتيبَ الصفوف بناءً على عمود (معامل الغرابة) للمدوّنة الأولى، واجعل التَّرتيب من الأقلّ إلى الأعلى رقمًا، ويُمكن تطبيقُ ذلك على بقِيَّةِ أعمدة المُعاملاتِ؛ لتسهيلِ قراءة القيم وتفسيرها. لاحظْ أنَّ إجراءَ استخراج الكَلِمَات المفتاحِيَّة ينطبقُ على استخراج التَّصَاحُبات اللَّفْظِيّة.

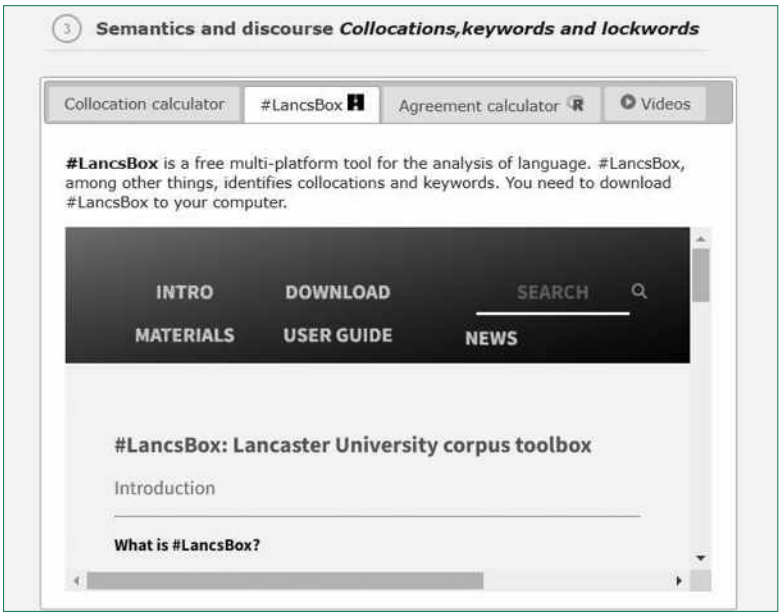
جدول ١١ : المقاييس الموصَى بها لاستخراج الغريب

الكلمة	الله	كلا	طاغون	مولاكم
تكرار (مدنية)	1547	4	0	5
تكرار (مكية)	820	45	2	0
التَّكْرَار الكلي	2367	49	2	5
دايس (مدنية)	0.0852	0.0002	0	0.0003
دايس (مكية)	0.0297	0.0017	0.0001	0
لوغ دايس (مدنية)	10.4474	1.9473	0	2.2711
لوغ دايس (مكية)	8.9093	4.7828	0.2922	0
الغرابة (مدنية)	2.9751	0.1402	0	Infinity
الغرابة (مكية)	0.3361	7.134	Infinity	0

ثَانِيًا: شبكات التَّصَاحُب اللَّفْظِيّ

تتمَيَّز هذه التّقنية بمعرفةٍ أدقّ لشبكاتِ التَّصَاحُب اللَّفْظِيّ، التي فيها يقومُ المحَلِّلُ بضبط مقاييس التَّصَاحُب بخطوةٍ واحدةٍ، وَتَعتمد هذه التّقنيةُ على تحديد

شـبكة التَّصَاحُـب للكلمـةِ بسلسـلةٍ ثابتـةٍ مـن المقاييـس، ويُعـبّر عنهـا في لُغَوِيّـات الـمُدَوّنَة بمجموعـة مقاييـس التَّصَاحُـبِ (collocation parameters notation). يُجـرى ذلـك مـن خـلال أداة صنـدوق (لانكسـتر) (النسـخة ٦,٠)[١]، ويُمكـن تحميـل هـذه الأداة مـن صنـدوق علـم الدلالـة والخطـاب Semantics and discourse كمـا هـو في الشَّـكْل (٣٦).



شكل ٣٦ أداة صندوق لانكستر

(١) هي أداةٌ قائمةٌ بذاتها، وليست شبكيَّةً، وهي مصاحبةٌ لأدوات (لانكستر) الإحْصَائِيَّة، ومن مميزاتها خاصيَّة (GraphColl شبكات التَّصَاحُب). ويشار عند الإشارة إلى أداة صندوق (لانكستر LancsBox) إلى المراجع الآتية:

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Retrieved October 24, 2022, from http://corpora.lancs.ac.uk/lancsbox.

Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. Retrieved October 24, 2022, from http://corpora.lancs.ac.uk/lancsbox.

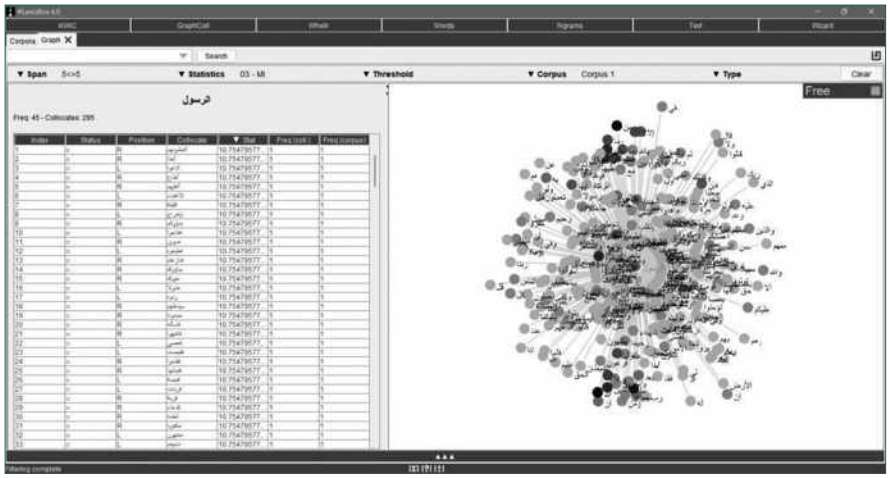
Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139173-.

يعمـل المحَلِّـلُ اللُّغَـوِيّ علـى تحديـد مجموعـةِ مقاييـس (بارامـترات) التَّصَاحُـب بالأسـلوب الخطِّـيِّ الآتي:

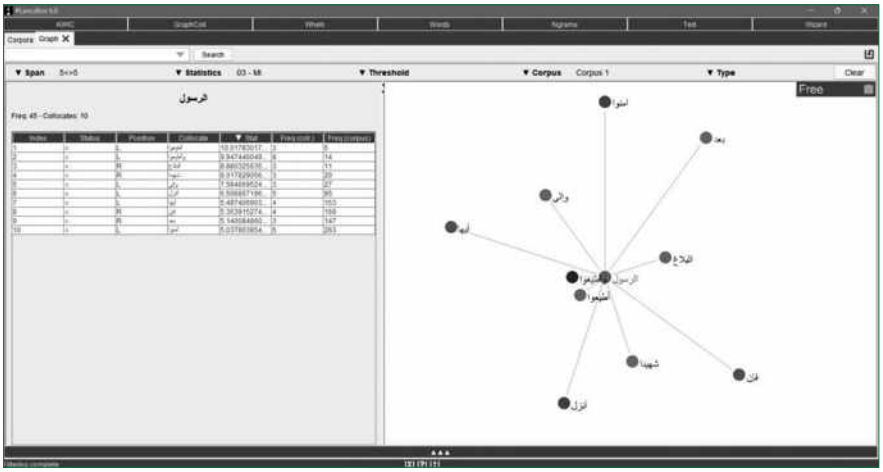
الكلمـة: الرَّسـول (المعلومـات المتبادلـة: قيمـة النُّقطـة الفاصلـة cut-off point: ٥، التّكْـرَار: ٣ فأفكـر، يم٥-يـس٥، تصفيـة: لا يوجـد، أي: لـم تُسـتخدم كلمـاتُ إيقـاف)

لننظـر إلى نتيجـةِ هـذا التَّرميـز في (ب) مـن الشَّـكْل (٣٧)؛ مـا الـذي يُمكـن ملاحظتُـه؟ الملحـوظ هـو أنَّ بهـذا التَّرميـز قـد تـمَّ التَّصـدّي لمتصاحبـات مهمَّـةٍ جـاءت مـع الكلمـةِ (الرسـول) علـى حجـم النافـذة المحَـدَّدة، كمـا أنَّ قيمـةَ المعلومـات المتبادلـة قـد حُـدِّدت مـن (٥) فأعلـى؛ لاسـتبعادِ التَّصَاحُـب ذي القيمـة (٥) فأقـل. ومـن المتصاحبـات المهمَّـة؛ المتصاحـب (وأطيعـوا)، و(شـهيدا)، و(البـلاغ)، و(أنـزل)، و(آمنـوا). مـاذا لـو لـم نضـع تصفيـاتٍ لمقاييـس التَّصَاحُبيـة؟ النَّتيجـة سـتكون كمـا في (أ) مـن الشَّـكْل (٣٧)، إذ ظهـرت كُلّ التَّصَاحُبـات، ومنهـا الكثيـر مـن التَّصَاحُبـات الـتي لا تـدلُّ علـى علاقـةٍ معنويَّـةٍ.

(أ)



(ب)



شكل ٣٧ تمثيل التَّصَاحُب من دون مقاييس مجموعة التَّصَاحُب (أ)
وباستخدام مقاييس مجموعة التَّصَاحُب (ب)

المبحث الثَّانِي: موافقة المقدِّر البينيّ

يُجرَى حسابُ موافقة المقدِّر البينيِّ (inter–rater agreement) لغرض تقدير التَّوافق والاختلاف في فهم المقدّرين، ويُطبّق كثيرًا في فهم الخطاب، ويهدفُ إلى تقريب المقصد قدر الممكنِ؛ كي يُبنى عليه التَّحليل. مثال ذلك: لو أن لدينا خطابًا يحتوي على استخدام كلمةٍ في سياقاتٍ عديدةٍ، ويصعبُ على الفرد أن يُحدِّد المقصد من هذه الكلمة في تلك السِّياقاتِ، كأَن يُطرح سؤالٌ عما إن كانت الكلمة دالَّةً على معنى إيجابِيٍّ، أم دالَّةً على معنى سلبيٍّ، يتطلب هنا من محلّلِ الخطاب أن يبحث عن مُقدّريْن اثنين، وكلما كان عدد المقدّرين أكثر من اثنين كان ذلكَ أفضل، ويطلب المحلّل من كُلّ مقدّرٍ أن يحَدِّد الاختياريَّة الثُّنائيَّة (نعم/لا)، كأنْ يكون التَّحديدُ عما إن كان معنى الكلمة إيجابيًّا أو سلبيًّا، مطابقًا أو غير مطابقٍ، صالحًا أو غير صالح، إلى آخِرِهِ، ثم بعد الحصول على التَّقديراتِ، يُجمعُ عددُ الموائم وعدد غير الموائِم، ولنفترض أن الكلمة قد جاءت في عشرةِ سياقات، وطُلب من المقدّرين تحديد ما هو إيجابِيّ وما هو سلبِيّ، ستكون عندئذ هيكلةُ البَيَانَات كما في الجدول (١٢)، وكلَّما زاد عددُ المقدّرين زادت موثوقِيَّةُ نتائِجِ المواءَمة.

جدول ١٢ : حساب موافقة المقدِّر البينيّ

رقم الكشاف السِّياقِيّ	المقدّر الأول	المقدّر الثاني	مطابق / غير مطابق
	إيجابي (١) سلبي (٢)	إيجابي (١) سلبي (٢)	
1	0	1	غير مطابق
2	1	0	غير مطابق
3	1	0	غير مطابق
4	0	0	مطابق
5	1	0	غير مطابق
6	1	1	مطابق
7	1	0	غير مطابق
8	1	1	مطابق
9	0	0	مطابق
10	1	0	غير مطابق
إيجابي	7	3	-
سلبي	3	7	-
الإجمالي	10	10	-

لتطبيق مثال من مُدَوّنَة القرآن الكريم على التَّقدير البينِيِّ؛ سأختار كلمةً، ثم سأعتبر المقدّرين هنا من أهل التَّفسير (كتب التَّفسير)، وشرحهم لمعنى الكلمةِ. في الآتي آيتان كريمَتان تَتَضَمَّنان كلمة (الغداة).

- قال الله تعالى: ﴿وَٱصۡبِرۡ نَفۡسَكَ مَعَ ٱلَّذِينَ يَدۡعُونَ رَبَّهُم بِٱلۡغَدَوٰةِ وَٱلۡعَشِيِّ يُرِيدُونَ وَجۡهَهُۥۖ وَلَا تَعۡدُ عَيۡنَاكَ عَنۡهُمۡ تُرِيدُ زِينَةَ ٱلۡحَيَوٰةِ ٱلدُّنۡيَاۖ وَلَا تُطِعۡ مَنۡ أَغۡفَلۡنَا قَلۡبَهُۥ عَن ذِكۡرِنَا وَٱتَّبَعَ هَوَىٰهُ وَكَانَ أَمۡرُهُۥ فُرُطٗا ٢٨﴾ [سور الكهف: ٢٨].

- قال الله تعالى: ﴿وَلَا تَطۡرُدِ ٱلَّذِينَ يَدۡعُونَ رَبَّهُم بِٱلۡغَدَوٰةِ وَٱلۡعَشِيِّ يُرِيدُونَ وَجۡهَهُۥۖ مَا عَلَيۡكَ مِنۡ حِسَابِهِم مِّن شَيۡءٖ وَمَا مِنۡ حِسَابِكَ عَلَيۡهِم مِّن شَيۡءٖ فَتَطۡرُدَهُمۡ فَتَكُونَ مِنَ ٱلظَّٰلِمِينَ ٥٢﴾ [سورة الأنعام: ٥٢].

في الجدول (١٣) معنى كلمة (الغداة) في بعض التَّفاسير، مع الإشارةِ إلى المطابقة من عدمها في شرح المقصود بهذه الكلمة.

جدول ١٣ : حساب موافقة تفسير كلمة (الغداة) في بعض التَّفاسير

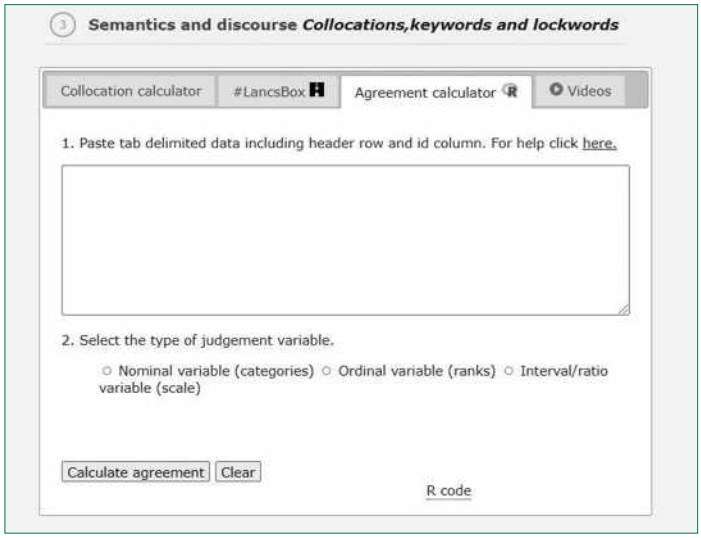
المعنى السياقِيّ	ابن عباس	ابن كثير	القرطبي	المطابقة
صلاة الصبح	1	0	0	غير مطابق
بين الفجر وطلوع الشمس	1	1	1	مطابق
أول النهار	1	1	1	مطابق

تطابقت التَّفاسيرُ في شرح كلمة (الغداة) في المعنيين السياقِيَّين: (الوقت المحدد بين الفجر وطلوع الشمس)، و(أوَّل النهار)، ولم يتطابق تفسيرُ ابن كثير وتفسيرُ القرطبِيِّ في المعنى السياقِيِّ (صلاة الصبح) مع تفسير ابن عباس. وسيكون شكلُ البَيَانَات التي تَتَضَمَّن السياقاتِ والمقَدّرين حسب الشَّكْل (٣٨)؛ نظرًا إلى ضرورة التَّرميز. الحالةُ (Case) تشيرُ إلى المعنى السياقِيِّ، و(A) يشير إلى تفسير ابن عباس، و(B) يشير إلى تفسيرِ ابن كثير، و(C) يشيرُ إلى تفسير القرطبي. وتُنسخ هذه البَيَانَاتُ، ثم تُلصقُ في موضع حاسبة التوافقِ (Agreement calculator) كما في الشَّكْل (٣٩). يُختار نوع المتغَيّرِ التَّقييميّ، والمتغيّر التَّقييمي في هذا المثال فئويٌّ ثنائِيٌّ (٠) أو (١)، وحالته كحالة (نعم/لا).

Case	A	B	C
1	1	0	0
2	1	1	1
3	1	1	1

شكل ٣٨ : ترميزات التَّفاسير الثَّلاث

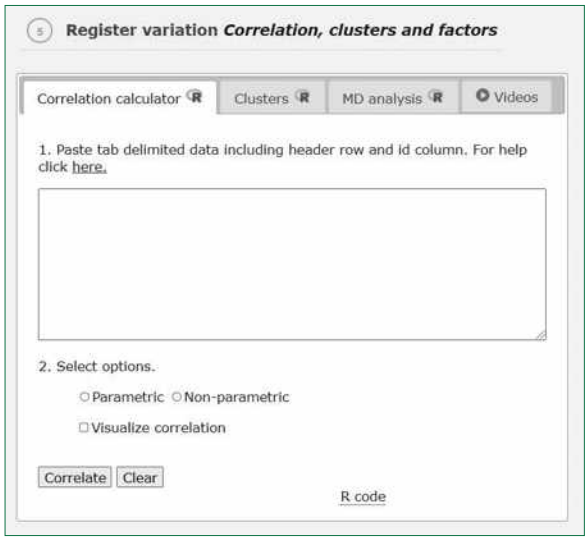
ظهـرت نتيجـة الموافقـة بنسـبة (77.78)، وهـي نتيجـة قسـمة (حـالات الموافقـة) علـى (إجمـالي عـدد الحـالات). النتيجـة في الأسـاس تظهـر بنسـبة 0.7769، ثـم تُحـوّل إلى نسـبةٍ مئوِيَّـةٍ. ونسـبة الموافقـة هنـا ليسـت عاليـةً، وأيضًـا ليسـت منخفضـةً، ويُنصـح في نتـائج موافقـة المقـدّر البيـنيِّ أن تكـون النِّسـبة ٨٠٪ فأعلـى.



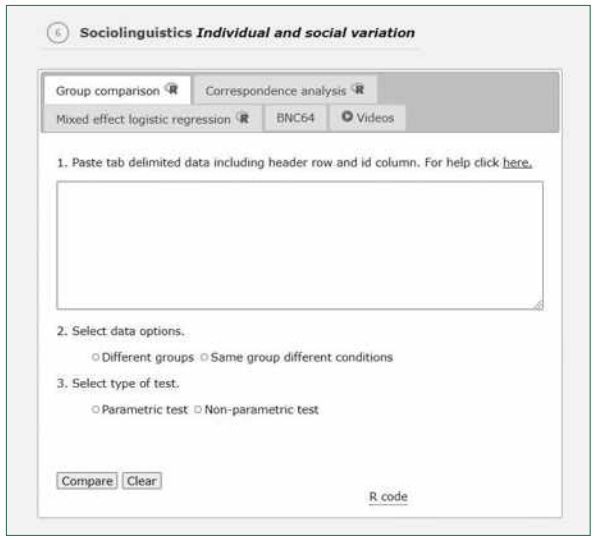
شكل ٣٩ : واجهة حسابِ الموافقة في (علم الدّلالة والخطاب)

الفصل الثَّالِث
الارتباط والتَّجميع ومقارنة المجموعات

يحتوي هذا الفصلُ على تطبيقاتٍ إحْصَائِيّةٍ؛ لكشف (الارتباط) بين الـمُتَغَيّرَاتِ، وتبيان العلاقة بينها، ولتمثيل (التَّجميع) الذي يبحث في العلاقةِ من حيث تصنيف الـمُتَغَيّرَات وتحديد المسافات بينها، ثم لعرض إحصاءات المقارنةِ بين المجموعات، ويُقصد بالمجموعاتِ هنا الـمُتَغَيّرَات الاسميَّة أو التَّرتيبيَّة؛ بوصفها مجموعاتٍ، كأن تُقارَن مجموعتان؛ الأولى: للذُّكور، والثَّانية: للإناث، في استخدامهما لضمائر المتكلِّم. يُجرى حساب الارتباطِ (Correlation calculator) والتَّجميع (Clusters) في مربع تنوع السجل اللُّغَوِيّ (Register variation) في أدوات (لانكستر) الإحْصَائِيَّة كما في الشَّكْل (٤٠)، أما إحصاء مقارنة المجموعاتِ (Group comparison)؛ فيُجرى في مربع اللُّغَوِيّات الاجْتِمَاعِيّةِ (Sociolinguistics) كما في الشَّكْل (٤١). ولكلِّ إحصاء من هذه الإحصاءات شكلٌ من البَيَانَات يُهيَّأُ في جداول محدَّدة الشكل.



شكل ٤٠ : واجهة حساب الارتباط في (تنَوُّع السِّجل اللُّغَوِيّ)



شكل ٤١ : واجهة مقارنة المجموعات في (اللُّغَوِيّات الاجْتِمَاعِيّة)

المبحث الأَوَّل: الارتباط

يكشف أسلوبُ تحليل الارتباط العلاقةَ بين الـمُتَغَيِّرَات، وَيختلفُ تحليل الارتباط عن تحليل الانْحِدَار الخَطِّيّ، فالأوَّل يكشف ما إن كان هناك ارتباطٌ أم لا بين الـمُتَغَيِّرَات اللُّغَوِيّة، أما الثَّاني فيجعلنا نفهم ونُفسّر هذا الارتباط، كما يجعلنا نتنبأ بحالاتٍ مشابهةٍ له، وسيَتَّضح الفرقُ بينهما في سياق مناقشة تحليل الانْحِدَار الخَطِّيّ في الفصل الرابع.

يُكتشف الارتباطُ من عدمه بين متغَيِّرين أو أكثر بالصِّيغ الآتية:

١. هناك ارتباطٌ إيجابيٌّ، ويعني ذلك أنَّ تزايد المتغيّر الأول مرتبطٌ بِتزايد المتغيّر الثَّاني في الـمُدَوّنَة اللُّغَوِيّة.

٢. هناك ارتباطٌ سلبِيٌّ، ويعني ذَلكَ أنَّ تزايد المتغَيّر الأوَّل مرتبطٌ بِتناقص المتغيّر الثَّاني في الـمُدَوّنَة اللُّغَوِيّة.

٣. عدم وجود ارتباطٍ أو علاقةٍ، سواءً كان إيجابيًّا أم سلبيًّا.

لـو جمعنـا مُدَوّنَـةً تَتَضَمَّـن روايـاتٍ كتبهـا روائيُّـون، ومدوّنـة ثانيـةً تَتَضَمَّـن روايـاتٍ كتبَتْهـا روائيـاتٌ، وأردنـا مثَـلًا أن نبحـث عـن إجابـةٍ لأسـئلة الارتبـاط مـن عدمـهِ لمتغَـيِّرٍ معَـيّنٍ، وليكـن هـذا المتغـيّر مثـلًا اسـتخدام الضَّمائـر، فـإن النَّتيجـة سـتحقّق فرضِيَّـةً واحـدةً مـن ثـلاث فرضيـاتٍ:

الأولى: كلمـا تزايـد اسـتخدام الضَّمائـر مـن الروائيـين تزايـدَ اسـتخدامُها مـن الروائِيَّـات (علاقـة الارتبـاط إيجابِيَّـة).

الثَّانيـة: كلمـا تزايـد اسـتخدامُ الضَّمائـر مـن الرِّوائيـين تناقـص اسـتخدامها مـن الروائيَّـات (علاقـة الارتبـاط سـلبيَّة).

الثَّالثة: لا يوجد اتجاه تزايد أو تناقص (لا توجد علاقة ارتباطٍ).

لتطبيـق تحليـلِ الارتبـاط في مُدَوّنَـة القـرآن الكريـمِ؛ تبـادر إلى ذهـني موضـوعُ الارتبـاط مـن عـدم الارتبـاط بـين اسـتخدامات الأسـماء والصِّفـات والأفعـالِ والأدوات والضَّمائـر بـين سـور القـرآن الكريـم، اُختـيرت السُّـوَر الطويلـة الآتيـة: البقـرة، وآل عمـران، والنِّسـاء، والمائـدة، والأعـراف.

سـأقف علـى النَّوْعِـيّن المعروفـين مـن تحليـل الارتبـاطِ؛ ويُعـرفُ الأول بارتبـاط بيرسـون (Pearson)، ويُعـرف الثـاني بارتبـاط سـبيرمان (Spearman). تُفسَّـر نتـائجُ الارتبـاط إما بقيـمٍ لا تكـون إلا بـين (1- و1). وقـد جـرى في الإفصـاح عـن وجـود الارتبـاط مـن عـدم الارتبـاط علـى أنَّ القيمَـة الـتي تكـون (0.8) فأعلـى دالَّـةٌ بقـوَّةٍ علـى الارتبـاط، وهـذا ليـس بالضَّـرورة ممكنًـا في البَيَانَـات اللُّغَوِيّـة، إذ تكـون القيمـة (0.2) دالّـةً علـى الارتبـاط، وعلـى أيِّ حـالٍ؛ فالقيـم بـين (1-) و(1)، ولنـا أن نتصـوّر القيـمَ بينهـا علـى المسـطرة المرسـومةِ أدنـاه، الـتي تنتصـف عليهـا القيمـةُ (0)، وفيمـا قبـل الصِّفـرِ تكـون القيـم سـالبةً، وتُعـبّر عـن ارتبـاطٍ سـلبِيٍّ، وفيمـا بعـد الصِّفـرِ تكـون القيـم موجبـةً، وتُعـبّر عـن ارتبـاطٍ إيجابِيٍّ. ضـع في ذهنـك أنَّ القيمـة بعـد الصِّفـر

كلما زادت واقتربت إلى (1) دلّ ذلك على زيادة الارتباط إيجابيًّا، ويشيرُ الارتباطُ الإيجابيُّ إلى علاقةٍ مطردةٍ بين متغيِّرين أو أكثر، والارتباط السَّلبيُّ يُشير إلى علاقةٍ غير مطّردةٍ.

1 0 -1

يقترح كوهن[1] أحجام التأثير الآتية:

- (0) لا تأثير–بدون قيمة
- (±0.1): تأثير صغير
- (±0.3): تأثير متوسط
- (±0.5): تأثير كبير

أَوَّلًا: ارتباط بيرسون

يعمل ارتباط بيرسون مع مُتَغَيّرَات المقياسِ (scale variables)، ولتهيئة البَيَانَات من أجل معالجتها في أدوات (لانكستر) الإِحْصَائِيَّة؛ تُدخَل البَيَانَاتُ كما في الشَّكْل (٤٢). علمًا بأَنَّ أرقامَ الملفات تعكسُ أرقامَ ترتيب السُّوَر في المصحف الشَّريفِ، وقيم التّكْرَار تعكس التّكْرَارات النِّسْبِيّةَ في كُلّ سورةٍ.

Sura	Nouns	Adjectives	Verbs	Particles	Pronouns
2	52.209	3.887	19.406	23.19277	1.305
3	53.024	3.755	18.800	23.23589	1.134
4	53.116	4.134	19.731	22.36142	0.626
5	54.145	4.848	18.735	21.96721	0.304
7	52.193	3.253	19.983	22.91606	1.656

شكل ٤٢ : بيانات التّكْرَارات النِّسْبِيّة لأقسام الكلام الرئيسة بين خمس سورٍ طويلة

(1) Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum, pp. 7980-.

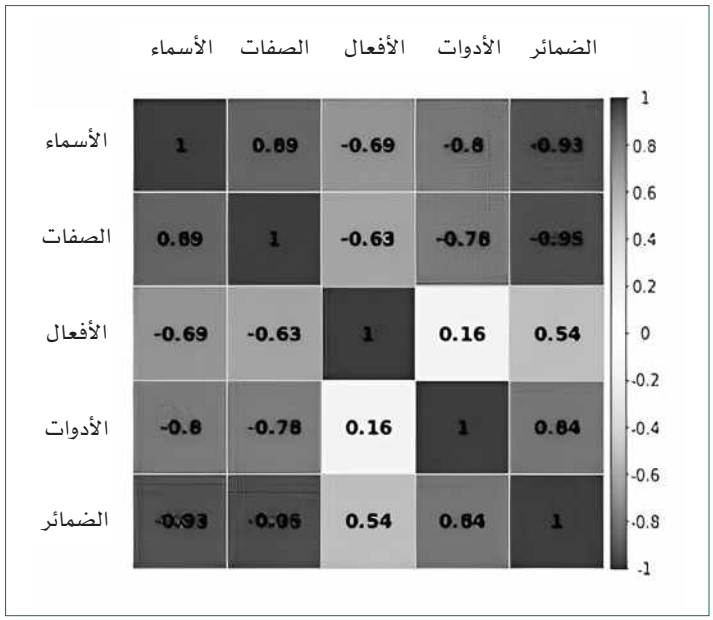
بعـد لصـق جـدول البَيَانَـات كمـا في الشَّـكْل (٤٠) في أدوات (لانكسـتر) الإِحْصَائِيَّـة في المربـع الخـاصِّ بحسـاب ارتبـاط بيرسـون؛ سـتَظهرُ النَّتـائج في جدولـين؛ في الجـدول (١٤) أَخْرَجَـت الأداةُ مُعاملَ الارتبـاط بين الـمُتَغَيّرَات (أقسـام الكلام)، وفي الجـدول (١٥) أخرجت القيـم الاحتماليَّـة للمُتَغَـيّرَات (الأقسـام النَّحْوِيّـة). أما الشَّـكْل (٤٣) فيُمثّـل خريطـةً حَرَارِيّـة لمعامـلاتِ الارتبـاط بـين الـمُتَغَيّرَات.

جدول ١٤ : معامل ارتباط بيرسون بين أقسام الكلام

الـمُتَغَيّرَات	الأسماء	الصفات	الأفعال	الأدوات	الضمائر
الأسماء	1.00	0.89	-0.69	-0.80	-0.93
الصفات	0.89	1.00	-0.63	-0.78	-0.95
الأفعال	-0.69	-0.63	1.00	0.16	0.54
الأدوات	-0.80	-0.78	0.16	1.00	0.84
الضمائر	-0.93	-0.95	0.54	0.84	1.00

جدول ١٥ : القيم الاحتماليَّة

الـمُتَغَيّرَات	الأسماء	الصفات	الأفعال	الأدوات	الضمائر
الأسماء	-	0.0448	0.1944	0.1052	0.0240
الصفات	0.0448	-	0.2534	0.1176	0.0137
الأفعال	0.1944	0.2534	-	0.7986	0.3487
الأدوات	0.1052	0.1176	0.7986	-	0.0767
الضمائر	0.0240	0.0137	0.3487	0.0767	-



شكل ٤٣ : الخريطة الحَرَارِيّة لارتباط أقسام الكلام

كيف تُفسّر قيم الارتباطات؟

يُنظر إلى القيم التي تكون فوق (٠) كونها قيمًا تدلُّ على وجود ارتباطٍ إيجابِيٍّ، ثم إلى القيمِ التي تكون أقلَّ من (٠) كونَها تدلُّ على وجود ارتباطٍ سلبِيٍّ، ويُمثل خريطة الحرارة الارتباط الإيجابِيّ بدرجات اللون الأزرق، والارتباط السلبيّ بدرجات اللون الأحمر.

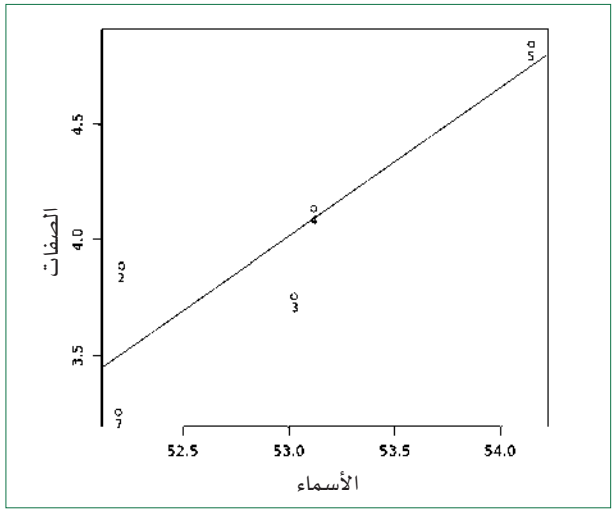
ما الـمُتَغَيّرَات التي ترتبط إيجابِيًّا والتي ترتبط سلبِيًّا؟ لا ننظر إلى القيمة (١) بين كُلّ مُتَغَيّريقابل نفسه، فلا ننظر إلى ارتباط الأسماء مع الأسماء، فذلك لا معنى لهُ، بل ننظرُ إلى الارتباطات بين الأقسام النَّحْوِيّة المختلفة، سيُرتّب الارتباط الإيجابيُّ، ثم الارتباط السَّلبيُّ بين الأقسام النَّحْوِيّة من الأقوى إلى الأضعفِ لكل الحالاتِ، وفق العرض في الجدول (١٦).

جدول ١٦ : الارتباط الإيجابِيّ والارتباط السَّلبي بين الأقسام النَّحْوِيّة

الارتباط الإيجابِيّ من الأقوى إلى الأضعف	الارتباط السلبي من الأقوى إلى الأضعف
ارتباط الأسماء مع الصفات والعكس: 0.89	ارتباط الصفات مع الضمائر والعكس: 0.95-
ارتباط الأدوات مع الضمائر والعكس: 0.84	ارتباط الأسماء مع الضمائر والعكس: 0.93-
ارتباط الأفعال مع الضمائر والعكس: 0.54	ارتباط الأسماء مع الأدوات والعكس: 0.8-
ارتباط الأفعال مع الأدوات والعكس: 0.16	ارتباط الصفات مع الأدوات والعكس: 0.78-
	ارتباط الأسماء مع الأفعال والعكس: 0.69-
	ارتباط الصفات مع الأفعال والعكس: 0.63-

الارتباط الإيجابِيُّ يعني أنَّ تكرار الـمُتَغَيّر النّسبي يزيد مع تكرار المتغيّر النّسبي الآخر، والارتباط السلبي يعني أنَّ تكرار الـمُتَغَيّر النِّسْبِيّ عندما يزيد ينقص التّكْرَار النِّسْبِيّ للمتغير الآخر. هذه طبيعة الأقسام النَّحْوِيّة في النُّصُوص اللُّغَوِيّة، فالأسماء أكثر ارتباطا بالصفات، والأفعال أكثر ارتباطا بالضمائر والأدوات، وذلك في أحوال التّكْرَارات الملحوظة أو التّكْرَارات النِّسْبِيّة. في الجدول (١٦) الذي يحتوي على القيم الاحتمالِيَّة؛ دائِمًا ستلحظ بأنَّ القيمَ التي تكون أقل من ٠,٠٥، والتي تكون دالَّةً إحْصَائِيًّا؛ تظهر مع قيم الارتباط الدَّالَّة على الارتباط الإيجابيِّ بقوة (٠,٨ فأعلى)، أو الدالَّة على الارتباط السلبيِّ بقوَّة (-٠,٨ فأعلى).

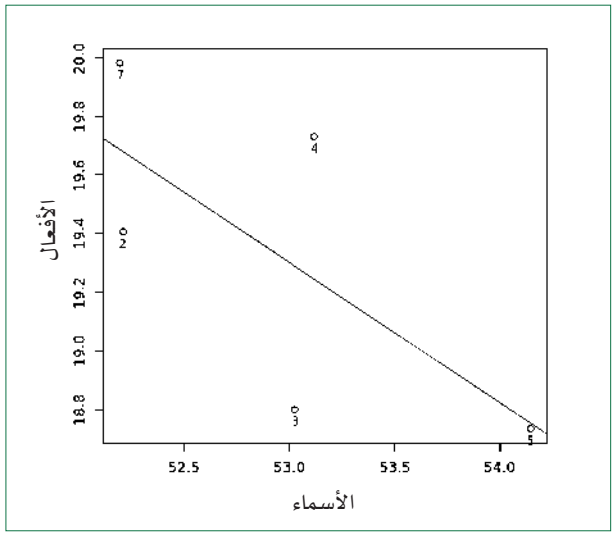
لنُقلّص حسابَ الارتباطِ ليكونَ بين الأسماءِ والصِّفات للسُّور المذكورة، وبين الأسماء والأفعال للسور المذكورةِ. في الشَّكْل (٤٤) تفسيرٌ لنتيجة ارتباط الأسماء والصفات، وفي الشَّكْل (٤٥) تفسيرٌ لنتيجة ارتباط الأسماء والأفعال.



شكل ٤٤ : ارتباط الأسماء مع الصفات

التفسير: الارتباط r = 0.887 قيمة الاحتمال: 0.0449 (مجال الثقة ٩٥٪: من 0.022 إلى 0.993

الارتباط الخطي: إيجابي



شكل ٤٥ : ارتباط الأسماء مع الأفعال

التَّفسير: الارتباط 0.693 = r قيمة الاحتمال: 0.1944 (مجال الثقة ٩٥٪: من 0.978- إلى 0.487

الارتباط الخطي: سلبي

تنبّه إلى الآتي:

- عندما يكون الخطُّ غير منحدرٍ إلى الأعلى أو إلى الأسفل، ويكون مستقيمًا في خطٍّ عرضِيّ؛ فذلك دلالةٌ على عدم وجود ارتباطٍ إيجابِيٍّ أو سلبِيٍّ.
- يُفيدك حسابُ الارتباط لمعرفة الارتباطاتِ بين الـمُتَغَيّرَات القياسيّة وبين الـمُتَغَيّرَات التأويليَّة، وقِسْ ذلك على كثيرٍ من الحالات التي يُمكن تطبيقُ حساب الارتباط عليها.

ثَانِيًا: ارتباطُ سبيرمان

يعمل ارتباط سبيرمان مع الـمُتَغَيّرَات الترتيبِيَّة (ordinal variables)، والتفسير هنا يكون بكشف الارتباط بين السُّوَر (ترتيب السورة في المصحف الشَّريف) بدلا من الأقسام النَّحْوِيّة. إذن، فهيئة البَيَانَات ستكون كما في الشَّكْل (٤٦).

POS	2	3	4	5	7
Nouns	52.209	53.024	19.406	54.145	52.193
Adjectives	3.887	3.887	4.134	4.848	3.253
Verbs	19.406	19.406	19.731	18.735	19.983
Particles	23.1928	23.1928	22.3614	21.9672	22.9161
Pronouns	1.305	1.305	0.626	0.304	1.656

شكل ٤٦ : بيانات التّكْرَارات النِّسْبِيّة لخمس سور طويلة بين أقسام الكلام الرئيسة

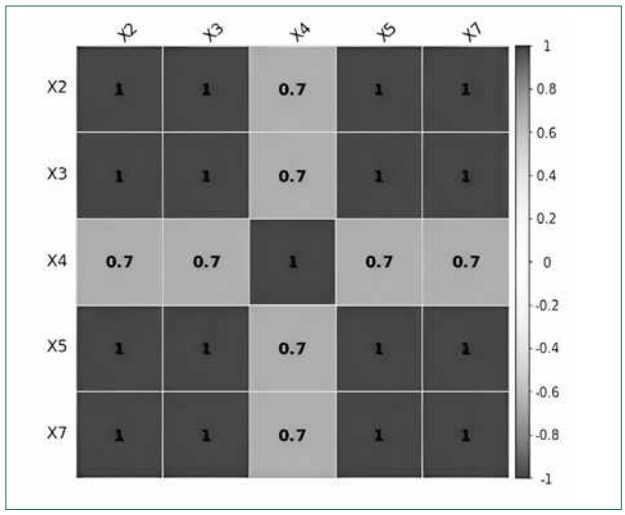
تظهرُ النَّتائجُ في جدولين؛ وفي كل جدول عرضٌ للمتغيرات x٢، وx٣ إلخ، التي تشير إلى رقم السورة حسب ترتيبها في المصحف الشَّريف. يَتَضَمَّنُ الجدولُ الأوَّل قيمَ الارتباط كما في الجدول (١٧)، والثاني يَتَضَمَّن قيم الاحتمال كما في الجدول (١٨)، وفي الشَّكْل (٤٧) خريطةٌ حَرَارِيّة عن قيمِ الارتباط.

جدول ١٧ : معامل ارتباط سبيرمان للمُتَغَيّرَات الترتيبية

المُتَغَيّرَات	X2	X3	X4	X5	X7
X2	1.0	1.0	0.7	1.0	1.0
X3	1.0	1.0	0.7	1.0	1.0
X4	0.7	0.7	1.0	0.7	0.7
X5	1.0	1.0	0.7	1.0	1.0
X7	1.0	1.0	0.7	1.0	1.0

جدول ١٨ : القيم الاحتمالية

المُتَغَيّرَات	X2	X3	X4	X5	X7
X2	-	0.0000	0.1881	0.0000	0.0000
X3	0.0000	-	0.1881	0.0000	0.0000
X4	0.1881	0.1881	-	0.1881	0.1881
X5	0.0000	0.0000	0.1881	-	0.0000
X7	0.0000	0.0000	0.1881	0.0000	-



شكل ٤٧ : الخريطة الحَرَارِيّة لارتباط أقسام الكلام

كيف نفسِّرُ هذه القيم؟ ظهرت كُلّ القيم (١) بين السُّور، وهي دلالةُ ارتباطٍ كاملٍ، عدا سورة المائدة التي ظهر ارتباطُها بقيمة (٠,٧)، وهي قيمةٌ تظلُّ ذات تأثيرٍ كبير. أما قيم الاحتمال فيها فَدالَّةٌ بقوّةٍ، عدا سورة المائدةِ مع بقِيَّة السُّور، يُفسّر ارتباط سبيرمان مثل التفسير المعروض في ارتباطِ بيرسون.

المبحث الثَّانِي: التَّجميع

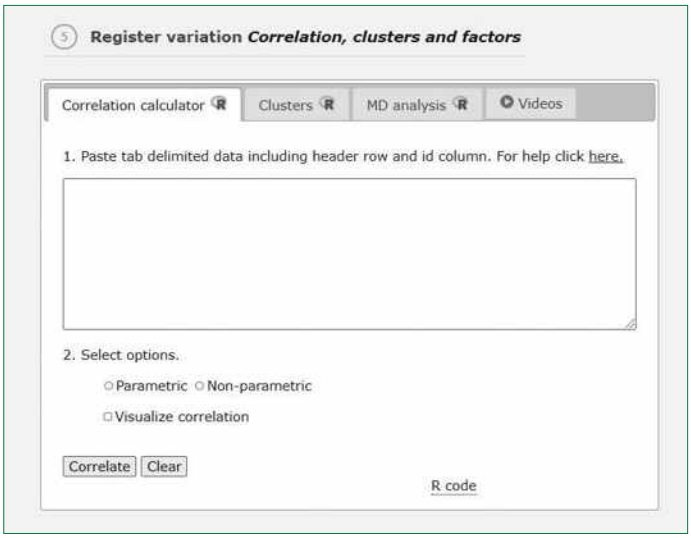
يُقصد بالتَّجميعِ (clustering) إيجاد المسافاتِ، والوصول إلى إنشاء التَّجميعات للكلمات ذات العلاقة. لقد نظرنا في المبحث الأوَّل (الارتباط) إلى العلاقة بين الخصائص، وكانت تلك الخصائص أقسام الكلامِ. وفي التَّجميع، سيكون البحثُ عن العلاقة بين الموضوعاتِ أو المسافات، وكيف تنتمي الكَلِمَاتُ إلى بعضها البعض على أساس ما من حيثُ قُرْب الانتماءِ أو بُعْدِه.

كيف يُمكنُ تصنيفُ الكَلِمَات في الـمُدَوّنَة اللُّغَوِيّة بناءً على ما نعرفُه عن عدد مرَّات تكراراتها؟ نعمل على تهيئة البَيَانَات في جداول أكسل بجعل العمود الأول محتويًا على الـمُتَغَيّرَات التأويليَّة التي نريد قياس المسافات بينها، بناءً على الـمُتَغَيّرَات المقياسيَّة لها. اخترتُ للمُتَغَيّرَات القياسِيَّة في العمود الثاني (التّكْرَار Frequency) وفي العمود الثَّالث (التّكْرَارات النِّسْبِيّة في كل ١٠٠ كلمة Relative Frequency x 100)، وفي الشَّكْل (٤٨) هيئةُ البَيَانَات التي أُدخلت في خيار التَّجْمِيع (Clusters) في مرَبَّع تنَوُّعِ السِّجلِّ اللُّغَوِيِّ (Register variation) في أدوات (لانكستر) الإِحْصَائِيَّة (الشَّكْل ٤٩).

ID	Frequency	Relative Frequency (x100)
Moses	136	0.175
Ibraham	69	0.089
Noah	43	0.055
Joseph	27	0.035
Adam	25	0.032
Jesus	25	0.026
Haroon	20	0.026
David	16	0.021
Soliman	17	0.022
Jacob	16	0.018
Isaac	15	0.017
Ishmael	12	0.014
Saleh	7	0.008
Zakariya	7	0.008
Yahya	7	0.008
Hud	4	0.005
Yunus	4	0.005
Mohammad	4	0.005
Job	4	0.005
Shuaib	4	0.005
Edris	2	0.002
Ilyas	2	0.002
Dhu_Alkifl	2	0.002
Elisha	2	0.002

شكل ٤٨ : بيانات التّكْرَار المطلق والتّكْرَار النِّسْبِيّ لأسماء الأنبياء في القرآن الكريم[١]

(١) مرة أخرى، لا يمكن تمثيل الكلمات العربية في أعمدة وصفوف جداول البيانات عندما نستخدم لغة آر R أو أداة لانكستر الشبكية القائمة بذاتها والمصممة من أكواد لغة آر، وعليه فقد نُقلت أسماء الأنبياء إلى الحروف اللاتينية. ينطبق هذا أيضا على شكل (٥٠). في الآتي ترجمة رؤوس الجدول: التكرار (Frequency)، والتكرار النسبي في كل ١٠٠ كلمة ((Relative Frequency (x100))، وتكرار قيمة زد (Frequency z score)، والتكرار النسبي لقيمة زد (Relative Frequency z score).



شكل ٤٩ : واجهة حساب الارتباط في (تنوع السجل اللُّغَوِيّ)

تُقاسُ المسافاتُ بناءً على نقاط البَيَانَات، فيتمُّ تجميع النقاط ذات القيم الصغرى أَوَّلًا، فالكُبرى ثانيًا. هناك أربعة مناهج للتَّجميع؛ الأَوَّلُ الرَّبط المنفرد (single linkage)، الذي يعتمدُ في منطلقِه من التَّجميع على أقرب نقطةٍ. الثَّانِي: الرَّبط التامّ (complete linkage)، الذي يعتمد في منطلقِه من التَّجميع على أبعد نقطة. الثَّالِث: الرَّبط المتوسّط (average linkage)، الذي يعتمدُ في منطلقه من التَّجميعِ على النُّقاط المتوسطة. الرابع: طريقة وارد (Ward)، التي تعتمد على مراعاة مسافات كُلّ النقاطِ، وتَحسب مجموعَ المسافات المربعة. يُنصح في المُدَوّنَات اللُّغَوِيّة بإجراء طريقةِ وارد[1].

لتوضيح أسلوبِ الإحصاء في نمذجة المسافاتِ؛ يُشار هنا إلى أهَمّ أسلوبين؛ وَهما المسافة الإقليدية (Euclidian Distance)، ومسافة منهاتن (Manhattan Distance). لنأخذ

(1) Gries, S. Th., Newman, J., Shaoul, C. & Dilts, P. (2009). N-grams and the clustering of genres. Presented at workshop on Corpus, Colligation, Register Variation at the 31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft, March. Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. Corpus Linguistics and Linguistic Theory, 2(1), 23–60.

على سبيل المثال كلمة (موسى) وكلمة (هارون). يعمل التَّجميعُ على البدء بالخطوة الأولى التي بها يُحدّد موقعُ الكلمة بينَ محور (س) x ومحور (ص) y. تُمثَّلُ قيمةُ كُلّ كلمة في كُلّ محور وفق ما يُراد قياسه من المسافات. لنفترض أنَّ القيمة التي يُراد قياسها للكلمتين هي التّكْرَار لمحور (س)، والتّكْرَار النِّسْبِيّ لمحور (ص): (موسى): [١٣٦:٠,١٧٥]؛ (هارون): [٢٠:٠,٠٢٦]. تُحول هذه القيم إلى قيمة ز (z-score) كما في الشَّكْل (٥٠). تُستخرج هذه القيمة بـ (القيمة) - المتوسط ÷ (الانحراف المعياريّ للعينة)؛ أي: (قيمة التّكْرَار للكلمة - قيمة متوسط التّكْرَارات لكلِّ الكَلِمَات) ÷ (الانحراف المعياريّ لمتوسّطات التّكْرَارات لكلِّ الكَلِمَات محلّ التَّحليل)، لتكون على النَّحوِ الآتي:

(موسى Moses): [4.067:3.978]

(هارون Haroon): [0.036:0.014]

تُستخرج المسافة الإقليدِيَّة بين كُلّ كلمةٍ وكلمةٍ أخرى بالحصول على ناتج الجذر التَّربيعي لـ (مربع مجموع قيمة س للكلمتين) + (مربع مجموع قيمة ص للكلمتين)، وتُستخرج مسافة منهاتن بالحصول على النَّاتج لـ (طرح قيمتي س لكلِّ كلمةٍ) + (طرح قيمتي ص لكلِّ كلمةٍ).

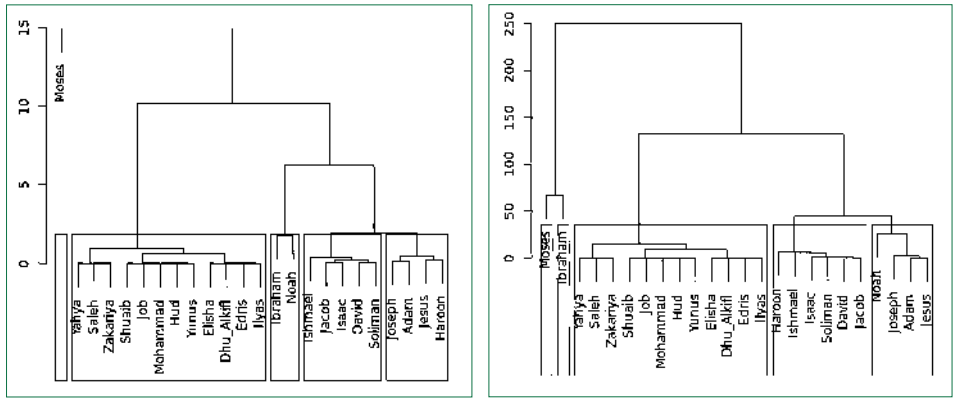
المسافة الإقليدية: $\sqrt{(4.067 - 0.036)^2 + (3.978 - 0.014)^2}$

مسافة منهاتن: $|4.067 - 0.036| + |3.978 - 0.014|$

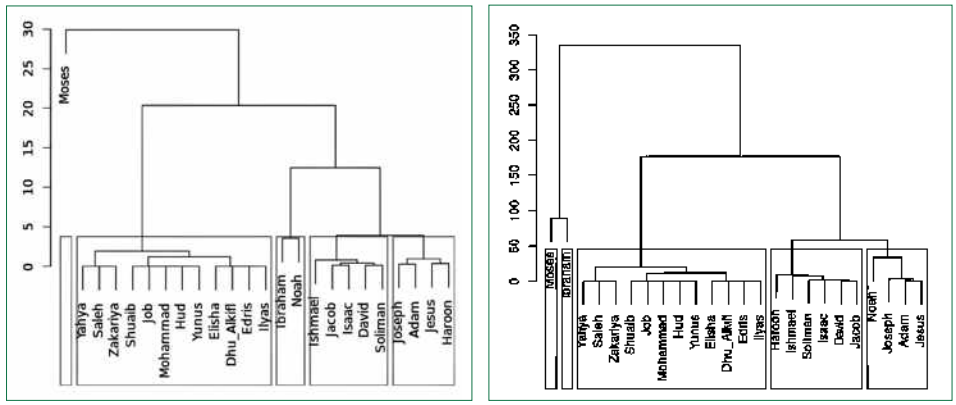
يُجرى هذا آلِيًّا عند نمذجة المسافاتِ بين المُتَغَيّرَات، فيعمل النّموذج على استخراجِ نتيجة المسافة الإقليدِيَّة أو مسافة منهاتن، ثم يقيس المسافة بين كُلّ كلمةٍ وكلمةٍ، ثم يجمع الكلماتِ ذوات النِّقاطِ الأقرب من بعضها البعض لتكون ضمن تجميعٍ تفريعيٍّ واحد كما في الشَّكْل (٥١) للمسافة الإقليديَّة، والشَّكْل (٥٢) لمسافة منهاتن، وفي كلا الشكلين؛ مُثِّلَ الرَّسمُ التَّخطيطيُّ في الجهة اليمنى بدون التحويل إلى قيم ز؛ ومُثِّلَ الرَّسمُ التَّخطيطيُّ في الجهة اليسرى باختيار التحويل إلى قيم ز، وينتج عن اختيار التحويل تجميعًا أكثر دقة في الحالات التي تكون قيم التكرارات فيها متطرفة، أو التي تكون للمقارنة بين عينات مختلفة تتباين فيها القيم التكرارية.

ID	Frequency	Relative Frequency (x100)	Frequency _z_score	Relative Frequency_z_score
Moses	136	0.175	3.978	4.067
Ibraham	69	0.089	1.689	1.739
Noah	43	0.055	0.800	0.835
Joseph	27	0.035	0.253	0.279
Adam	25	0.032	0.185	0.210
Jesus	25	0.026	0.185	0.052
Haroon	20	0.026	0.014	0.036
David	16	0.021	-0.122	-0.103
Soliman	17	0.022	-0.088	-0.069
Jacob	16	0.018	-0.122	-0.165
Isaac	15	0.017	-0.157	-0.196
Ishmael	12	0.014	-0.259	-0.288
Saleh	7	0.008	-0.430	-0.443
Zakariya	7	0.008	-0.430	-0.443
Yahya	7	0.008	-0.430	-0.443
Hud	4	0.005	-0.533	-0.536
Yunus	4	0.005	-0.533	-0.536
Mohammad	4	0.005	-0.533	-0.536
Job	4	0.005	-0.533	-0.536
Shuaib	4	0.005	-0.533	-0.536
Edris	2	0.002	-0.601	-0.597
Ilyas	2	0.002	-0.601	-0.597
Dhu_Alkifl	2	0.002	-0.601	-0.597
Elisha	2	0.002	-0.601	-0.597

شكل ٥٠ : حساب التّكْرَارات المطلقة والنِّسْبِيّة المتعدّدة لأسماء الأنبياء في القرآن الكريم



شكل ٥١ : نتائج التَّجميع بالمسافة الإقليدية



شكل ٥٢ : نتائج التَّجميع بمسافة منهاتن

سؤالٌ سيتبادر حتمًا إلَى الأذهان؛ وهوَ ما فائدةُ التَّجْمِيع؟ يعتمد التَّجْمِيعُ على وضع صورةٍ عامَّةٍ للمُتَغَيّرَات اللُّغَوِيّة وفق قيمِها، إذ قد تكونُ هذه الـمُتَغَيّرَات بالمئات، وتَتَضَمَّن قيمًا مختلفةً، والتَّجْمِيع يعمل آلِيًّا على وضع كُلّ مجموعةٍ من الـمُتَغَيّرَات في تجميعٍ واحدٍ ليكون مُعينًا في تحليلاتٍ لاحقةٍ ترومُ فهمَ العلاقاتِ بين المتغيرات اللغوية أو تفسيرها، فلو نظرنا إلى الشَّكْل (٥١)، والشكل (٥٢)؛ سنجد خمسَ مجموعاتٍ في كُلِّ رسم تخطيطي من الرسوم الأربعة، تتضمن كل مجموعة منها تجميعًا لأسماء الأنبياءِ المتقاربة، ويعتمد تحديدُ عدد هذه المجموعات على المحَلّل البشريِّ، إذ يُمكن أن يُحدِّدَ المجموعاتِ لتكون مجموعتين،

أو ثلاث، وقد اخترتُ خمس مجموعات بتفكيرٍ إِحْصَائِيٍّ أوليٍّ؛ كون عدد المُتَغَيّرَات (أسماء الأنبياء) خمسة وعشرين اسمًا، ولأجل توضيح المجموعاتِ؛ كان الهدف أن يكون عددُ التَّجْمِيعات خمسة، التي ستُفيد بطبيعة قرب وبعد نقاط أسماء الأنبياء.

المبحث الثَّالِث: مقارنة المجموعات اللُّغَوِيّة

سأُناقش هنا تطبيقَ الإحصاءات المتعَلّقَة بمقارنة المجموعاتِ اللُّغَوِيّة من المُدَوّنَات اللُّغَوِيّةِ، وتنقسم هذه الإحصاءاتُ إلى قسمَين رئيسَين.

القسم الأَوَّل: الاختبارات البارامترية وغير البارامترية.

تنقسم الاختباراتُ البارامتريَّة (المعتمدة على مقاييس أوَّليَّةٍ من المتوسّطات والانحرافات المعيارِيَّةِ) إلى نوعين من الاختبارات؛ الأَوَّل: اختبار-ت (ويلش Welch) للمقارنة بين مجموعتين مختلفتين في استخدامهما متغَيّرًا لُغَوِيًّا واحدًا (التكرار في كُلّ مجموعة). والثَّانِي: تحليل التباينِ (ANOVA) للمقارنة بين مجموعتين مختلفتين (أو أكثر) في استخدام متغيِّرٍ لُغَوِيٍّ (التكرار والرتب في كُلّ مجموعة). وتنقسم الاختبارات غير البارامترية إلى نوعين من الاختباراتِ؛ الأَوَّل: اختبار مان-ويتني U، المعروف -أيضًا- باسم اختبار مجموع رتب ويلكوكسون (Wilcoxson rank-sum test) للمقارنة بين مجموعتين مختلفتين في استخدام متغير لُغَوِيّ (التكرار في كُلّ مجموعة) أو اختبار مجموع رتب ويلكوكسون للمقارنة داخل مجموعة واحدة فيها حالات مختلفة. والثَّانِي: اختبار كروسكال-والاس للمقارنةِ بين مجموعتين مختلفتين (أو أكثر) في استخدام متغَيِّر لُغَوِيٍّ (التكرار والرتب في كُلِّ مجموعة). يعني مصطلح غير بارامتري، أو (غير مقياسي non-parametric)، أَنَّنا لسنا بحاجةٍ إلى أيِّ معرفةٍ عن البارامترات كالمتوسَّط أو الانحرافِ المعيارِيِّ للمتغيّر المعنيِّ بالتَّحليل في المجموع اللُّغَوِيّ[1]. لا تحتوي هذه الاختباراتُ على

(1) Dodge, Y. (2008). The concise encyclopedia of statistics. New York: Springer.

افتراضاتٍ عن التَّوْزِيع الطَّبِيعِيّ الأساسِيِّ للمتغَيّر اللُّغَوِيّ الاجْتِمَاعِيّ في المجموع اللُّغَوِيِّ، وعن تجانس التَّباين، ويمكنُ استخدامُها في حالات بياناتِ المقياس الشَّديدة الانحراف، أو في حالات البَيَانَات التَّرتيبِيَّة، وبدلًا من التّكْرَارات النِّسْبِيّة يعملُ هذان الاختباران بصيغة (الرتب)، ويمكن بسهولةٍ تحويل التّكْرَارات النِّسْبِيّة (متغيّر المقياس) إلى رُتبٍ، على الرّغم من أنَّ هذا يأتي على حسابِ فقدان بعض المعلوماتِ عن الاختلافاتِ الدَّقيقة بين الحالاتِ للمتحدِّثين.

تَتَضَمَّن المجموعتان اللتان ستُقارنان بهذهِ الاختباراتِ البارامترية وغير البارامترية من مُدَوّنَة القرآنِ الكريمِ ما يَأتي:

المجموعة الأولى: السُّوَر المكِّيَّة.

المجموعة الثَّانية: السُّوَر المدَنِيَّة.

وتحتوي كُلُّ مجموعةٍ على التّكْرَاراتِ النِّسْبِيّة لكلمة (الناس)، وبواقع ١١٦ مرَّة في السُّوَر المكِّيَّة و١٢٢ مرَّة في السُّوَر المدَنِيَّة[١]. اُعتُمِد على أداةِ غواص لاستخراجِ التّكْرَارات النِّسْبِيّة لكلمةِ (الناس) بين السُّوَر المكِّيَّة والسُّوَر المدَنِيَّة، وتكوّنت النُّصُوصُ الخامّ من مجلّدين؛ الأَوَّل: ويحتوي على ١٣ ملفًّا نَصِّيًّا، وفي كُلّ ملفٍّ نَصِّيٍّ الجزء الذي يُمثّل تقريبًا سنةَ نزول الآياتِ على مرّ ١٣ سنة (السُّوَر المكِّيَّة). الثَّانِي: ويحتوي على ١٠ نصوصٍ، وفي كُلّ ملَفٍّ نَصِّيٍّ الجزء الذي يُمثِّل تقريبًا سنة نزول الآيات على مرّ ١٠ سنواتٍ (السُّوَر المدَنِيَّة). في (أ) من الشَّكْل (٥٣) التّكْرَارات النِّسْبِيّة لكلمة (الناس) على فترة ٢٣ سنة، وهي هيئة البَيَانَات بين المجموعتين التي تُستخدم من أجل إظهار اختبار-ت البارامتري، واختبار مان-ويتني غير البارامتري، أما في (ب) من الشَّكْل (٥٣)؛ فهي التّكْرَارات النِّسْبِيّة لتلك الفترة، بالإضافة إلى رتبة الحالات من أجل إظهار تحليل

(١) وردت صيغة (الناس) في السُّوَر المكِّيَّة ٩٢ مرة، وفي السُّوَر المدَنِيَّة ٩٢ مرة. اُستعمل أداة ستانفورد للتَّقطيع الآلِيِّ من أجل دِقَّة نتائج البحثِ. ظهرت التَّتابع (ل الناس) ٢٣ مرة في السُّوَر المكِّيَّة و٢٨ مرة في السُّوَر المدَنِيَّة، ثم التتابع (ب الناس) مرة واحدة في السُّوَر المكِّيَّة ومرتان في السُّوَر المدَنِيَّة، ليكون مجموع تكرار الكلمة ١١٦ مرة في السُّوَر المكِّيَّة و١٢٢ مرة في السُّوَر المدَنِيَّة.

التبايـنِ (ANOVA) البارامـتري، واختبـار كروسـكال-والاس غـير البارامـتري، وتُوضـع الرُّتـب بشـكلٍ تصاعـدِيٍّ بالتَّطابق مع تصاعـد التّكْـرَاراتِ، فمثـلًا؛ أعلى تكـرارٍ نسـبِيٍّ هـو ٠,٨٩٦٠٦، ورتبتـه (١)، وأقَـلّ تكـرارٍ نسـبِيٍّ هـو ٠,١٤٦٩٤، ورتبتـه (١٦)، أمـا القيـم الصِّفْرِيّـة فهـي تُشـير إلى عـدم ظهـورِ الكلمـة بـين الآيـاتِ الكريمـةِ في رقـم السَّـنة في عمـودِ (ID).

(أ) بيانات اختبار-ت واختبار مان-ويتني

ID	MK	MD
1	0.00000	0.00000
2	0.00000	0.00000
3	0.00000	0.00000
4	0.00000	0.00000
5	0.89606	0.00000
6	0.40161	0.00000
7	0.14694	0.00000
8	0.00000	0.00000
9	0.23271	0.00000
10	0.22211	0.00000
11	0.33351	0.00000
12	0.29775	0.00000
13	0.15051	0.00000
14	0.00000	0.15473
15	0.00000	0.00000
16	0.00000	0.00000
17	0.00000	0.64725
18	0.00000	0.52619
19	0.00000	0.43380
20	0.00000	0.21834
21	0.00000	0.45934
22	0.00000	0.23962
23	0.00000	0.21952

(ب) بيانات تحليل التباين واختبار كروسكال-والاس

ID	MK	Rank_MK	MD	Rank_MD
1	0.00000	0	0.00000	0
2	0.00000	0	0.00000	0
3	0.00000	0	0.00000	0
4	0.00000	0	0.00000	0
5	0.89606	1	0.00000	0
6	0.40161	6	0.00000	0
7	0.14694	16	0.00000	0
8	0.00000	0	0.00000	0
9	0.23271	10	0.00000	0
10	0.22211	11	0.00000	0
11	0.33351	7	0.00000	0
12	0.29775	8	0.00000	0
13	0.15051	15	0.00000	0
14	0.00000	0	0.15473	14
15	0.00000	0	0.00000	0
16	0.00000	0	0.00000	0
17	0.00000	0	0.64725	2
18	0.00000	0	0.52619	3
19	0.00000	0	0.43380	5
20	0.00000	0	0.21834	13
21	0.00000	0	0.45934	4
22	0.00000	0	0.23962	9
23	0.00000	0	0.21952	12

شكل ٥٣ : بيانات اختبارات المجموعة البارامترية وغير البارامترية

يحتوي اختبار-ت على افتراضاتٍ ينبغي مراجعتها قبل إجراء الاختبار، والافتراض الأهمّ هو استقلال الملاحظاتِ (independence of observations)، ويشيرُ هذا الاستقلالُ في السِّياق اللُّغَوِيّ الاجْتِمَاعِيّ إلى كُلّ ملاحظةٍ تنتمي إلى متحدّثٍ واحدٍ أو مجموعةٍ لُغَوِيَّةٍ واحدةٍ، شرط أن تكون كُلّ ملاحظةٍ من متحدثٍ معَيَّنٍ أو مجموعةٍ معينةٍ غير متأثرةٍ بمتحدثٍ آخر، أو بمجموعةٍ لُغَوِيَّة معينةٍ.

نتيجة اختبار-ت

- ت t (43.9) = -0.15، القيمة الاحتمالية = 0.878 (غير دال إحْصَائِيّا)
- كوهن د (حجم الأثر) = 0.05- (مجال الثقة ٩٥٪ [0.64-، 0.55] الأثر متوسط)
- الارتباط (حجم الأثر) = 0.02- (مجال الثقة ٩٥٪ [0.32-، 0.28] الأثر متوسط)

تشبه افتراضاتُ تحليلِ التباين (ANOVA) اختبارَ-ت، وهذهِ الافتراضات هي استقلال الملاحظات، والحالة الطَّبِيعِيّة، والتَّفاوت المتماثل، والأهمُّ من بَينها هو استقلال الملاحظات.

نتيجة تحليل التَّباين (ANOVA) اتجاه أحادي One-way

- التباين الاتجاه الأحادي: فF (٣، ٨٨) = ٥,٠٥، القيمة الاحتمالية = ٠,٠٠٣ (دالٌ إحْصَائِيّا)
- قياس حجم الأثر (أوميغا ω) = ٠,٣٤١ (حجم الأثر متوسِّط)
- الاختبار البعديُّ (post-hoc tests الذي يهدف إلى تقليل احتمالات الدلالة الإحْصَائِيَّة عند استعمال فرضياتٍ اختبارِيَّةٍ عَديدةٍ): اختبار ت المعدّل (القيم الاحتمالِيَّة).

المجموعات	المدّنيَّة	المكّيَّة	رتب المدَنِيَّة
المكّيَّة	1.000	-	-
رتب المدَنِيَّة	0.092	0.090	-
رتب المكّيَّة	0.023	0.022	1.000

نتيجة اختبار مان-ويتني

- اختبار مان-وتني = 258.5، والقيمة الاحتمالية = 0.887 (غير دال إحْصَائِيًّا)
- حجم الأثر (الارتباط r) - 0.209- (الأثر صغير)

نتيجة اختبار كروسكال-والاس

- اختبار كروسكال-والاس = 1.39، والقيمة الاحتمالية = 0.707 (غير دال إحْصَائِيًّا)

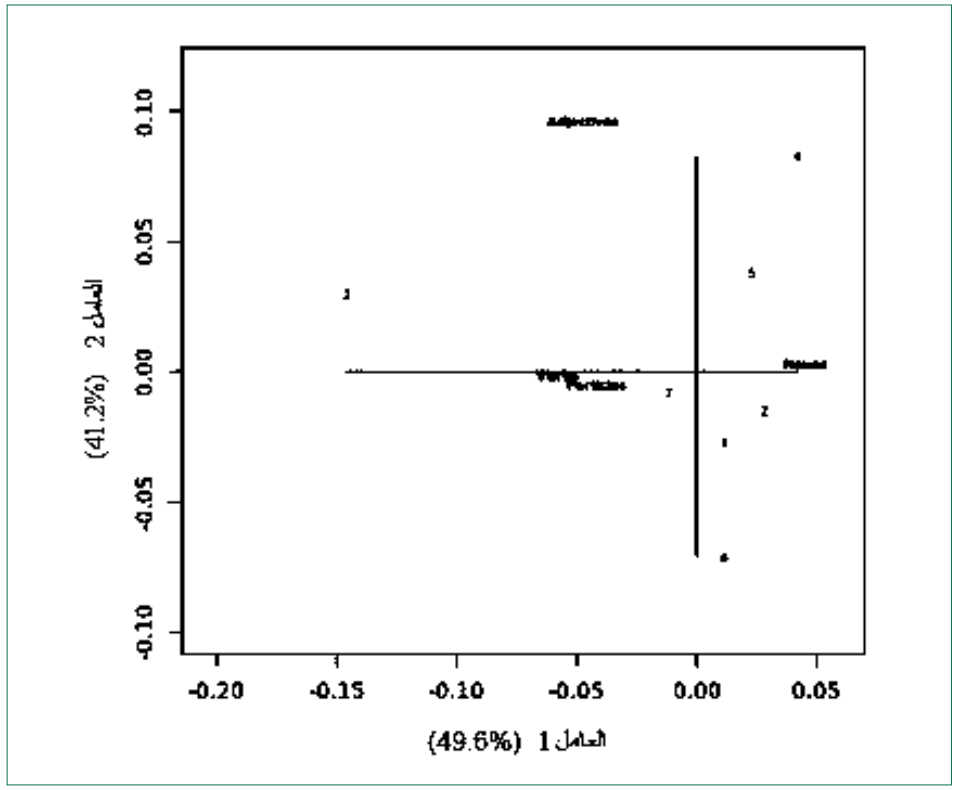
القسم الثَّانِي: تحليل التَّناظر

هو لتحليل المُتَغَيّرَات اللُّغَوِيّة التي تتقارب أو تتباعدُ، ويهدف إلى الكشفِ عن التفرّد الأسلوبِيِّ اللُّغَوِيّ من المصدر اللُّغَوِيّ لتلك المُتَغَيّرَات اللُّغَوِيّة، سواءً كان هذا المصدرُ جنس المتحَدّث، أو طبقات المتحدثين الاجْتِمَاعِيّة، أو أوعية نَصِّيَّة مختلفَة (المُتَغَيّرَات التأويليَّة) إلخ، وبالنسبة للمُتَغَيّرَات التي ستُحلّل فهي التّكْرَارات المطلقة لأقسام الكلام الأربعة (الأسماء، والصفات، والأفعال، والأدوات)، بين السُّوَر السبع الطوال: (البقرة، وآل عمران، والنساء، والمائدة، والأنعام، والأعراف، والأنفال-التوبة).

الهدفُ هنا هو الكشفُ عن تفَرُّدِ الأسلوب اللُّغَوِيّ بين أقسامِ الكلام الأربعة، وستُطبّق هنا المقارنة بين السُّوَرِ السبع الطّوال، وملاحظة تجمّع هذه السُّوَر كَسَبعِ عَيِّناتٍ، وحجم تفَرُّدِ كُلِّ عَيِّنةٍ في استخدام الأسماءِ والصِّفاتِ والأفعال والأدوات (الشكل ٥٤).

File	Nouns	Adjectives	Verbs	Particles
1	3640	271	1353	1617
2	2104	149	746	922
3	1207	132	630	714
4	2312	207	800	938
5	1696	132	630	714
6	1797	112	688	789
7	4099	372	1636	1793

شكل ٥٤ : بيانات أقسام الكلام الرَّئيسة بين السُّوَر الطَّويلة



شكل ٥٥ : تحليل التَّنَاظُر

نلحـظ في الشَّـكْل (٥٥) أنَّ العامـل الأول هـو عامـل السُّـور السـبع الطـوال، ونسـبة التَّنَاظُـر بينهـا ٤٩,٦، وَالعامـل الثـاني هـو عامـل أقسـام الـكلام، ونسـبةُ التَّنَاظُـر بينهـا ٤١,٢، ويُقصـد بالتناظـر هنـا مسـتويات التماثـل مـن عدمـه. كيـف نفسّـر هـذا التَّمثيـل البيـانيّ؟ نقـول الآتي:

- تميزت السُّوَر ٤ و٥ بالأسماء أكثر من بقية السور السبع.
- كان تميز السُّورَة ٤ ثم ٥ بالأسماء أعلى من السُّورَة ٢ ثم ١ ثم ٦.
- تميزت السُّورَة ٣ بالصفات أكثر من بقية السُّوَرِ.
- تميزت السُّورَة ٧ بالأفعال والأدوات.

يُسـتنتج مـن هـذا التفسـير بـأن الأسـماء بـين السُّـوَرِ السـبع الطـوال قـد ظهـر بُـرُوز اسـتخدامِها في سـورة المائـدة والأنعـام أكـثر مـن سـورة البقـرة وآل عمـران والأعـراف، كمـا يُسـتنتج بـأن الصفـات في سـورة النسـاء أكـثر بـروزًا مـن بقيـة السـور، أمـا السـورة السـابعة (الأنفال-التوبـة) فأكـثر أقسـام الـكلام فيهـا بـروزًا هـو الأفعـال والأدوات.

الفصل الرَّابع
الانحدار

بما أنَّ الارتباط يُخبرنا بالعلاقة غير السببية بين الـمُتَغَيِّرَات، وبنسبة هذا الارتباط فيما إن كان إيجابِيًّا أم سلبِيًّا أم منعدمًا؛ فإنَّ الانحدار يُخبرنا عن العلاقةِ السَّببيَّة بين الـمُتَغَيِّرَات، وعن شكل هذا الخطّ الانحداريّ، بين الـمُتَغَيِّر المُسْتَقِلّ (المُتنبِّئ) والمتغير التَّابِع (النَّاتِج)، الذي يفيدُ إما بوجود علاقةٍ خَطِّيَّةٍ، أو بعدم وجود علاقةٍ خَطِّيَّةٍ. وتُفيدُ هذه العلاقة الخَطِّيَّة بأنَّ هناك علاقةً سببِيَّةً، أي أن الـمُتَغَيِّر المُسْتَقِلَّ (المُتنبِّئ) يؤَثِّر في الـمُتَغَيِّر التَّابع (النَّاتِج) وبأنَّه من الممكن التَّنبُّؤ بهذا التَّأثير، أو تُفيد بعدم وجود علاقة خَطِّيَّةٌ سببِيّةٌ. ويُبنى نموذجُ الانحدار لأغراض التَّفسير والتَّنبُّؤ، وهو من أهم النَّماذج في التَّنَبُّؤ، وينقسمُ الانحدار إلى نوعين؛ الأَوَّل: الانْحِدَار الخَطِّيّ (المبحث الأول)، والثَّانِي: الانْحِدَار اللُّوجسْتِيّ (المبحث الثاني)، وسأتناول هذين النَّوْعِيّن من الانحدار، وكيفِيَّة الاستفادة منهُما في تحليل البَيَانَات اللُّغَوِيّة، وذلك بالتَّطبيق على مُدَوّنَة القرآن الكريم.

المبحث الأَوَّل: الانْحِدَار الخَطِّيّ

لم يضع مصمّم أدوات (لانكستر) الإِحْصَائِيَّة أداةً جاهزةً تُنسخُ فيها البَيَانَات لاستخراج نتائج تحليل الانْحِدَار الخَطِّيّ، ولذلك سأشيرُ هنا إلى مفهوم الانْحِدَار الخَطِّيّ، وطبيعة شكل البَيَانَات اللُّغَوِيّة، وما تَتَضَمَّنه هذه البيانات من مُتَغَيِّرَاتٍ، مع أكواد التَّحليل من الآر للتَّطبيق (الملحق ج: الفصل أكواد الفصل الرَّابع).

يُستخدم تحليلُ الانْحِدَار الخَطِّيّ في مهمَّتين؛ الأولى: أن يتعَلَّم كيف يتنَبَّأ بالقيمة مقابل قيم حقيقِيَّةٍ، وهذا يُطبق في تعَلُّم الآلة. وَالثَّانية: التنبُّؤ بقيمة مُتَغَيِّرٍ بناء على قيمة مُتَغَيِّرٍ آخر، وهذا يُسمى باختيار السِّمة (feature selection)، وهو ما يُطبّق في اللُّغَوِيّاتِ التجريبِيَّة والنَّفسِيَّة كثيرًا، ويكون الـمُتَغَيِّر الذي يُراد التنَبُّؤ به هو الـمُتَغَيِّر التَّابع، ويكون الـمُتَغَيِّر الذي يُعتمدُ عليه في التنَبُّؤ هو الـمُتَغَيِّر (أو المُتغيّرات) المُسْتَقِلّ (المُتنبِّئ)، وتكون العلاقةُ خَطِّيَّةً عند تأَثُّرِ الـمُتَغَيِّر التَّابِع (النَّاتِج) بالمتغَيِّرِ المُسْتَقِلّ.

الانْحِدَار الخَطِّيّ إما انحدارٌ خَطِّيٌّ بسيطٌ (متغير مستقلٌّ واحدٌ ومتغيرٌ تابعٌ واحد)، أو انحدارٌ خَطِّيٌّ متعَدِّدٌ (مُتَغَيِّرَات مستقِلَّة ومتغيّر تابع واحد)، ودائمًا يكون الـمُتَغَيِّر التَّابِع واحدًا، ويُفيد الانْحِدَار الخَطِّيّ دِقَّةَ الاستدلال فيما إن كانت هناك علاقةٌ سببِيَّةٌ واضحةٌ بين الـمُتَغَيِّر المستقِلِّ والمتغَيِّر التَّابع.

يجب أن تكون الـمُتَغَيِّرَات المُسْتَقِلّة مستمرةً (continuous)، وفي حال كانت هناك مُتَغَيِّرَاتٌ مستقلةٌ فئوِيَّةٌ فلا يتعامل معها الانْحِدَار الخَطِّيّ، ويُحتاج في هذه الحالِ إلى تحويلِها إلى مستمِرَّةٍ عن طريقِ التَّرميزِ (encoding)، الذي يُحول البَيَانَات الفئوِيَّةَ إلى بياناتٍ رَقَمِيّةٍ بواسطة ما يعرف بـ (label encoding)، أو (one hot encoding)، أو (dummy variable trap)، ولن أقف على الـمُتَغَيِّرَات الفئوِيَّة، وكيفِيَّة تحويلها إلى مُتَغَيِّرَاتٍ رَقَمِيّة[1]؛ والسَّبب يعودُ إلى عدم فائدتها في اللُّغَوِيّات التَّجريبِيَّة والـمُدَوّنَات اللُّغَوِيّة، إذ يكونُ البديل الأمثل لها هو باستعمالِ الانْحِدَار اللُّوجسْتِيّ (المبحث الثاني من هذا الفصل).

هُيِّئَت بياناتُ الانْحِدَار الخَطِّيّ من مُدَوّنَة القرآن الكريم كما في الشَّكْل (٥٦) على النَّحو الآتي:

- الزَّمن chronology (ترتيب السُّورَة حسب نزولها زمنِيًّا): مُتَغَيِّرٌ مستقلٌّ مستمِرٌّ.
- الوقت time (نزول السُّوَر المكِّيَّة ونزول السُّوَر المدَنِيَّة): مُتَغَيِّر مستقلٌّ فئويٌّ.
- الآيات verses (عدد آيات كُلّ سورة): مُتَغَيِّر مستقلٌّ مستمرٌّ.
- الكَلِمَات words (عدد الكَلِمَات في كُلّ سورةٍ): مُتَغَيِّر مستقلٌّ مستمِرٌّ.
- السَّجْدَة prostration (السُّوَر التي تَتَضَمَّن سجدةً): مُتَغَيِّر مستقلٌّ فئويٌّ.

---

(١) لمن يطلب الفائدة في التَّعامل مع الـمُتَغَيِّرَات الفئوِيَّة في الانْحِدَار الخَطِّيّ:
Kumar, Sudhir. 2020. Linear Regression Tutorial (Medical Cost Personal Datasets). Retrieved October 24, 2022, from https://www.kaggle.com/sudhirnl7/linear-regression-tutorial

- الطول length (سور قصيرة، وسور متوسطة، وسور طويلة): مُتَغَيّر مستقلٌ فئويٌّ.
- الحروف characters (عدد الحروف في كُلّ سورة): مُتَغَيّر تابعٌ مستمرٌّ.

أسلفنا القولَ عن عدم تعامل الانْحِدَار الخَطّيّ مع الـمُتَغَيّرَات المُسْتَقِلّة الفئوِيَّة، إذْ لها حالةٌ خاصَّةٌ غير مهمَّةٍ في تحليلِ البَيَانَات اللُّغَوِيّة، ولذلك تُستبعد من البَيَانَاتِ، ويُطرح السُّؤال المتعلّق بإمكان التنَبُّؤ من عدم الإمكان على النَّحو الآتي: هل يُمكن التنبُّؤ بوجود علاقة سببٍ تُفيد بتأثُّر الـمُتَغَيّر التَّابِع (النَّاتِج) من الـمُتَغَيّرَات المُسْتَقِلّة (المتنبِّئة)؟ لننظر إلى الـمُتَغَيّرَات المُسْتَقِلّة (المتنبِّئة): الزَّمن (chronology)، وعدد الآياتِ (verses)، وعدد الكماتِ (words)، ثم إلى الـمُتَغَيّرَات التَّابعة (النَّاتِج): عدد الحروفِ (characters). ويتطَلّب كشف الارتباط بين هذه المـتَغَيّرَات قبل بناءِ نموذج الانْحِدَار الخَطّيّ. والسُّؤال المصوغُ هنا: (هل تأثُّر الـمُتَغَيّر التَّابِع يُمكن تفسيره والتنَبُّؤ به بقيمِ الـمُتَغَيّرَات المُسْتَقِلّة؟).

نتيجة الارتباط: ترتيب الارتباط الإيجابي والسلبي بدءًا من الأقوى حتى الأضعف:

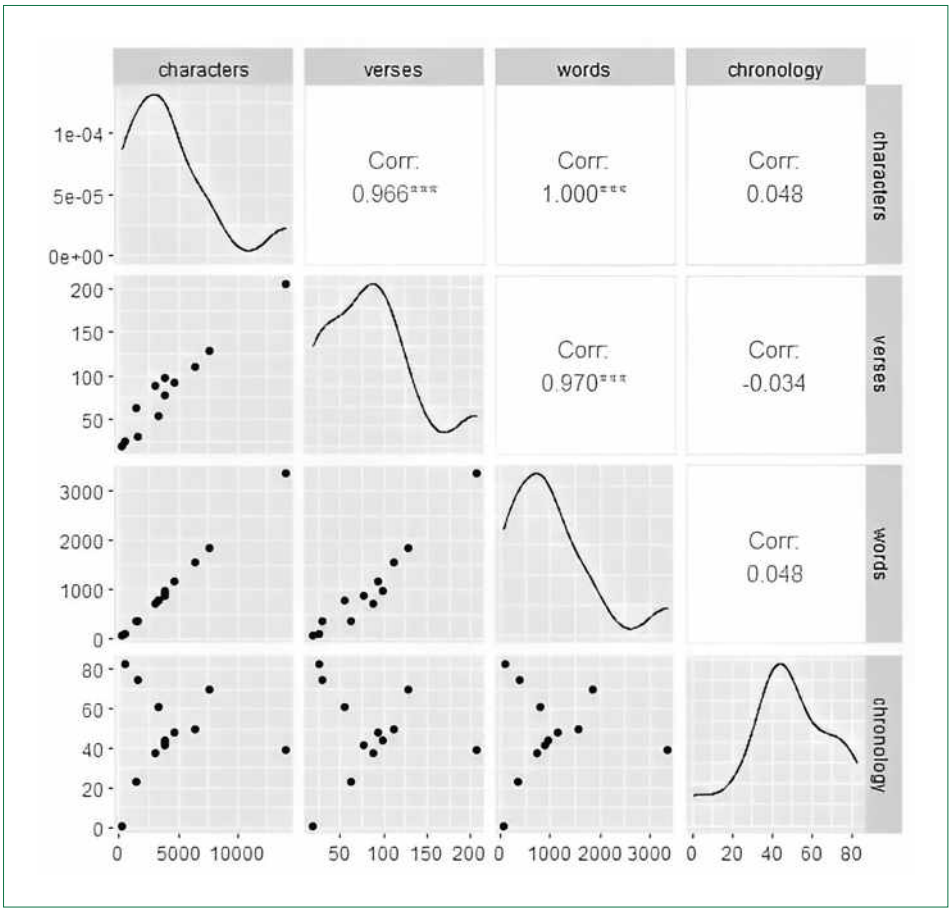
- ارتباط عدد الكَلِمَات مع عدد الحروف: 1.0
- ارتباط عدد الكَلِمَات مع عدد الآيات: 0.970
- ارتباط عدد الآيات مع عدد الحروف: 0.966
- ارتباط الزَّمن مع عدد الحروف: 0.048
- ارتباط الزَّمن مع عدد الكَلِمَات: 0.048
- ارتباط الزَّمن مع عدد الآيات: 0.34-

نتيجة النّموذج الخَطّيِّ: بناءً على الشَّكْل (٥٧) الذي يتضمنُ تمثيلًا للارتباط بين الـمُتَغَيّرَات؛ لا يُمكن التنَبُّؤ بتغير عدد الحروف (المتغير التَّابِع والنَّاتِج)، وفقًا للمُتَغَيّرَات

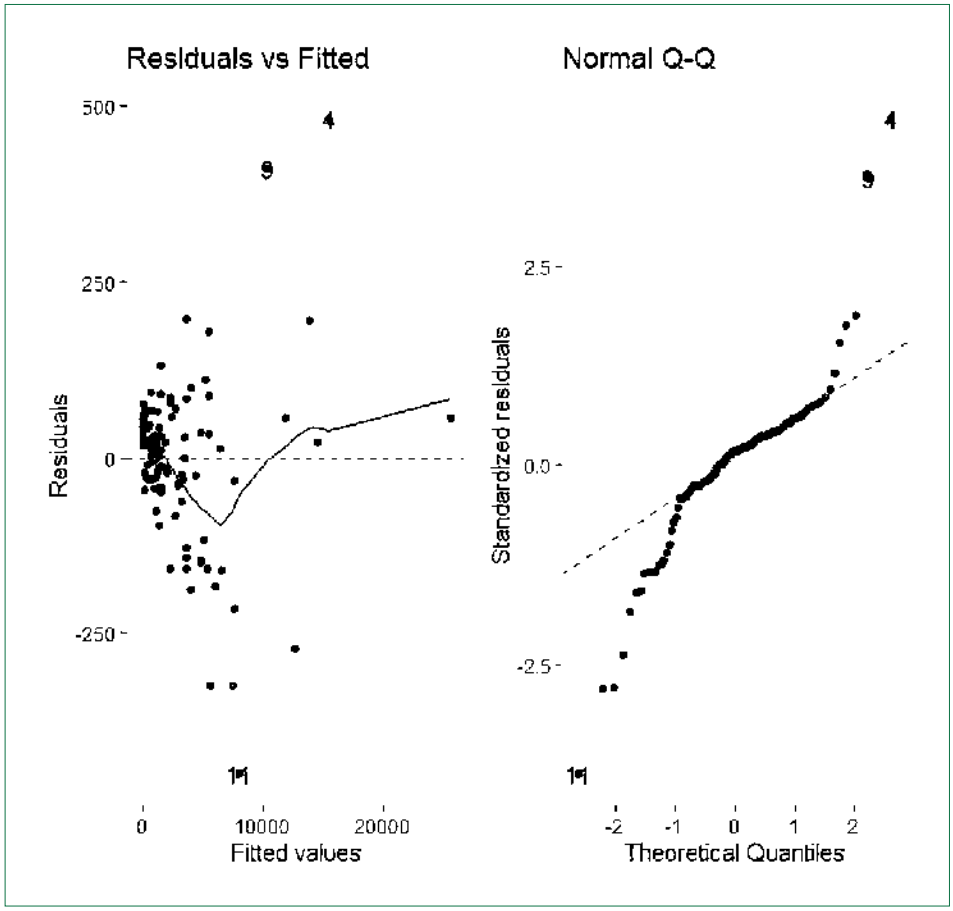
المُسْـتَقِلّة (المتنبِّئـة)، ولكـن كيـف يُعـرف فيمـا إن كانـت هنـاك علاقـةٌ خَطِّيّـةٌ أم لا؟ بالنظر في المتبقيَّـات -أو المختلفـات- (residuals) مـن النَّمـوذج الموفّـقِ -أو المتوائـم- (fitted model) وفي الاختلافـاتِ بـين قيـم الـمُتَغَيّر النَّـاتج الملحـوظ في كُلّ مجموعـةٍ من قيـم الـمُتَغَيّرَات المُسْـتَقِلّة (الشَّـكْل ٥٨)؛ يُشـيرُ التَّمثيـلُ البيـانيُّ في الجهـة اليسـرى إلى الأنمـاطِ غـير الخَطِّيَّـةِ في القيـم المتبقِّيـةِ (residuals)، والعلاقـة هنـا غـيرُ خَطِّيّـة، وعندمـا تكـون العلاقـةُ خَطِّيَّـةً؛ فـإنَّ الخـطَّ باللـون الأزرق سـيكون منحـدرًا أو مائـلًا (slope) مـن الأسـفل إلى الأعلـى، ومتراكبًـا مـع قيـم الـ (residuals). إذن؛ عندمـا يكـونُ الخَـطُّ باللـون الأزرق متراكِبًـا مـع نقـاط القيـمِ المتبقِّيـة بشـكلٍ خطـيٍّ؛ فذلـك يـدلُّ علـى العلاقـة الخَطِّيَّـةِ؛ أي علـى تأثـير الـمُتَغَيّرَات المُسْـتَقِلَّة (المتنبّئـة) علـى الـمُتَغَيّر التَّابـع (النَّـاتج)، وعندمـا يظهـر هـذا التأثـير فذلـك يُمكِّـن أوَّلًا مـن تفسـير التَّأثيرِ؛ كونـه واقعًـا، ومـن التَّنَبـؤ ثانيًـا بـأنَّ هذا الـمُتَغَيّر المُسْـتَقِلّ، أو تلـك الـمُتَغَيّرَات المُسْـتَقِلّة، لها تأثيرٌ سـببيٌّ في تغَـيُّر قيـم الـمُتَغَيّر النَّـاتج التَّابـع. ومثـل هـذا الأسـلوبِ الإِحْصَـائِيّ يُطبّـقُ بالدرجة الأولى علـى البَيَانَـات اللُّغَوِيّـة المرتبطـة باللُّغَوِيّـات النفسِـيَّةِ، وباكتسـاب اللغـة، كأن تُقيّـم وجـودَ العلاقـة الخَطِّيَّـة مـن عـدم وجودِهـا بـين مُتَغَـيّرَاتٍ مسـتقلةٍ مـن قبيـل (عـدد سـاعات التعلّـم والأعمـار)، وبـين الـمُتَغَيّر التَّابـع مـن قبيـل (الزَّمـن المسـتغرق في قـراءة واسـتيعاب فقـرةٍ في قطعـةٍ نَصِّيَّـةٍ للقـراءة)، في التَّمثيـل البيـانيِّ الأيمـن مـا يُسّـمى بـ (Q-Q plot)، وهـو اختصـارٌ لـ (quantile-quantile plot)، الـذي يُقـارن الكمِّيَّـات النظريَّـةَ والكمِّيَّـات التَّجريبِيَّـةَ الـتي نحصـل عليهـا مـن البَيَانَـاتِ، ونلحـظ في المَخَطّـط الـذي ظهـر في البَيَانَـات المسـتخدمةِ، بعـد مقارنـة توزيـع النّقـاط للبقايـا (residuals) مـع التَّوْزِيـع الطَّبِيعِـيّ (الخـط المرجعـي reference line)، وجـودَ نقـاطٍ بعيـدةٍ نسـبيًّا علـى يسـار ويمـين الخـط المرجعِـيِّ، لكنهـا لا تُشـكّل انحرافـاتٍ حـادّةً عـن التَّوْزِيـع الطَّبِيعِـيّ.

chronology	time	verses	words	prostration	length	characters
5	MK	7	10	no	short	139
87	MD	286	1000	no	long	25613
89	MD	200	1000	no	long	14605
92	MD	176	1000	no	long	15937
112	MD	120	1000	no	long	11892
55	MK	156	1000	no	long	12418
39	MK	206	1000	yes	long	14071
88	MD	75	1000	no	long	5299
113	MD	129	1000	no	long	10873
51	MK	109	1000	no	long	7425
...	...	....	...	...	...	...

شكل ٥٦ : بيانات الانْحِدَار الخَطِّيّ



شكل ٥٧ : تمثيل الارتباط والنموذج الخطي



شكل ٥٨ : مخطط الربعيّات Normal Q-Q والتمثيل الخطي للمتبقيات residuals مقابل القيم الموفقة للنموذج

المبحث الثَّانِي: الانْحِدَار اللُّوجسْتِيّ

يعمـلُ هـذا الانحـدار مثـل التَّصنيـف، وتَتَضَمَّـن بياناتُـه مُتَغَـيّرَاتٍ فئويـةً للمتغـيِّر المُسْـتَقِلّ (المُـتنبِّئ)، والمتغَـيّر التَّابـع (النَّـاتج)، وعندمـا يكـون الـمُتَغَيّر التَّابِـع فئويًّـا ثنائيًّـا؛ فعندئـذٍ يُسـمّى بالانْحِـدَار اللُّوجسْـتِيّ الثُّنـائي الاتجـاه (binomial logistic regression)، وعندمـا يكـون في الـمُتَغَيّر التّابـع أكـثر مـن فئـةٍ؛ فعندئـذ يُسـمّى بالانْحِـدَار اللُّوجسْـتِيّ متعـدِّد الاتجـاهِ

(multinomial logistic regression)، وغالبًا يُجرَى الانْحِدَار اللُّوجسْتِيّ الثنائي الاتجاه، ويُجرى على موضوعاتِ النَّحوِ المعجَمِيّ في الـمُدَوّنَات اللُّغَوِيّةِ، وهدف الانْحِدَار اللُّوجسْتِيّ مثل هدفِ الانْحِدَار الخَطّيّ، يُخبرنا عن الـمُتَغَيّر المُسْتَقِلّ الذي يؤثّر أكثر في الـمُتَغَيّر التَّابِع، ولا تكون المتغيرات في الانْحِدَار اللُّوجسْتِيّ إلا فئوية (سنرى في البَاب الثَّانِي في الفصل الخامس جانب تعلُّم الآلة للانحدار). سأُطبّق مثالَيْن للانحدار اللوجستيِّ لأجل التَّوضيح.

المثال (أ): تتكوّن فيه البَيَانَاتُ من متغيّرين مستقلّيْن (مُتنبّئات)، ومتغَيِّر تابع (ناتج) واحد.

- المتنبئ (الوقت time): (المكي a_MK) و(المدني b_MD)
- المتنبئ (الطول length): (طويلة a_long) و(متوسطة b_medium) و(قصيرة c_short)
- النَّاتج (السَّجْدَة prostration): (لا a_no) و(نعم b_yes)

المثال (ب): تتكَوَّن فيه البَيَانَاتُ من ثلاثة مُتَغَيّرَات مستقِلَّةٍ (مُتنبّئات)، ومتغيّر تابع (ناتج)

- المتنبئ (التَّنَوُّع Variety): (سورة النساء a_3) و(سورة الأعراف b_7)
- المتنبئ (التركيب Syntaxt): (جملة الفعل الماضي a_past) و(جملة الفعل المضارع b_present) و(الجملة الاسمية c_nominal)
- المتنبئ (الطول Length): (طول الكلمة الأولى من الجمل وتُرواح بين ٢ و٧)
- النَّاتج (الاسم الموصول Relative): (ما a_mā) و(الذي b_alladhī)

في الشَّكْل (٥٩)؛ سترى شكلَ البَيَانَات اللُّغَوِيّة للمثال الأوَّل كما هو في (أ)، وشكل البَيَانَات اللُّغَوِيّة للمثال الثاني كما في هو في (ب). يُلصق الشَّكْل (أ) في خيار الانْحِدَار اللُّوجسْتِيِّ (Logistic regression) من مربَّع النَّحوِ المعجَمِيِّ (Lexico-grammar) في أدوات (لانكستر) الإِحْصَائِيَّة كما في الشَّكْل (٦٠). في الخيار (١) سؤالٌ يَتَضَمَّن الآتي: اخترْ ماذا

تريـد أن تفعـل؟ (Select what you want to do). تظهـرُ قائمـةٌ منسـدلةٌ فيهـا (بنـاء نمـوذج Build a model)، و(بنـاء نمـوذج تدريـجي Build a model stepwise). يُظهـر (بنـاء النمـوذج) نتيجـة الـمُتَغَيّرَات كلها، ويُظهر بناء النموذج التَّدريـجي أي الـمُتَغَيّرَات المتنبِّئة (المُسْـتَقِلَّة) أفضـل مـن ناحيـة تأثيرهـا في الـمُتَغَيّر النَّـاتج (التابـع). عنـد بنـاء نمـوذج؛ تُلصـق البَيَانَـات في الخيـار (٢)، ثـم يُكتـبُ اسـمُ الـمُتَغَيّر النَّـاتج (التابـع) في الخيـار (٣)، ثـم يُكتـب في الخيـار (٤) أسـماء مُتَغَـيّرَات المتنـبّئ (المُتَغَـيّرَات المُسْـتَقِلَّة) مفصولـةً بفاصلـةٍ منقوطـةٍ (;) مـن دون مسـافاتٍ، ثـم اختيـار (no) في الخيـار (٥).

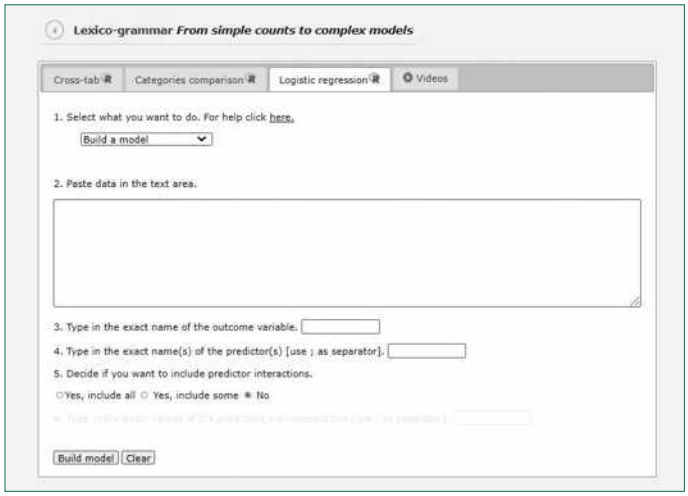
(أ)

ID	time	length	prostration
1	a_MK	c_short	a_no
2	b_MK	a_long	a_no
3	b_MK	a_long	a_no
4	b_MK	a_long	a_no
5	b_MK	a_long	a_no
6	a_MK	a_long	a_no
7	a_MK	a_long	b_yes
8	b_MK	a_long	a_no
9	b_MK	a_long	a_no
10	a_MK	a_long	a_no

(ب)

ID	Variety	Syntax	Length	Relatives
1	a_3	a_past	3	a_mā
2	a_3	a_past	3	a_mā
3	a_3	a_past	3	a_mā
4	a_3	a_past	3	a_mā
5	a_3	a_past	3	a_mā
6	a_3	a_past	3	a_mā
7	a_3	a_past	3	a_mā
8	a_3	a_past	3	a_mā
9	a_3	a_past	3	a_mā
10	a_3	a_past	3	a_mā
11	a_3	a_past	3	a_mā
12	a_3	a_past	3	a_mā
13	a_3	a_past	3	a_mā
14	a_3	a_past	3	a_mā
…	…	…	…	…
98	b_7	b_present	4	b_alladhī
99	b_7	b_present	4	b_alladhī
100	b_7	b_present	4	b_alladhī
101	b_7	c_nominal	7	b_alladh

شكل ٥٩: الانْحِدَار اللُّوجسْتِيّ



شكل ٦٠ واجهة الانْحِدَار اللُّوجسْتِيّ في (النَّحو المعجَمِيّ)

أَوَّلًا: نتيجة المثالِ (أ)

- بناء النَّموذج للانحدار اللوجستيِّ

إحصاءات النَّموذج: الأرجحِيَّة اللوغارثمية: 6.11 (القيمة الاحتمالية = 0.106: غير دالٍّ إحْصَائِيًّا)، C-index: 0.69 (غير مقبول)، ناغلكيرك ر٢: ١,٠، AIC: 86.82.

جدول ١٩ : مُعاملات النَّموذج للبيانات (أ)

	تقدير الأرْجحِيَّة اللوغارثمية	خطأ المعيار	قيمة-ز (والد)	قيمة الاحتمال	تقدير الأرْجحِيَّة	مجال الثقة ٩٥٪ (الأقل)	مجال الثقة ٩٥٪ (الأعلى)
(Intercept)	-1.093	0.540	-2.026	0.043	0.335	0.105	0.909
timeb_MK	-1.004	0.815	-1.233	0.218	0.366	0.053	1.519
lengthb_medium	-0.568	0.640	-0.887	0.375	0.567	0.163	2.103
lengthc_short	-2.207	1.147	-1.924	0.054	0.110	0.005	0.774

- بناء النَّموذجِ التَّدريجي stepwise (النَّموذج الأفضل)

الإحصـاءات الكُلِّيَّـة: الأرجحِيَّـة الاحتماليَّـة: 4.3 (القيمـة الاحتماليَّـة = 0.116 غـير دالٍّ إحْصَائِيًّـا)، C–index: 0.64 (غـير مقبـولٍ)، ناغلكـيرك ر٢: 0.07، AIC: 88.62.

جدول ٢٠ : مُعاملات النّموذج الأفضل للبيانات (أ)

	تقدير الأَرْجَحِيَّة اللوغارتمية	خطأ المعيار	قيمة-ز (والد)	قيمة الاحتمال	تقدير الأَرْجَحِيَّة	مجال الثقة ٩٥٪ (الأقل)	مجال الثقة ٩٥٪ (الأعلى)
lengthb_medium	-0.466	0.628	-0.742	0.458	0.627	0.186	2.287

- تفسير النَّتيجة

مـا النَّمـوذج الأفضـل الـذي يُبـيِّن حجـم التَّأثـير الأعلـى مـن المـتَغَيّرَات المُسْـتَقِلَّة (المتنبئـة) الفئويَّـة علـى الـمُتَغَيّر التَّابِع (النَّـاتج)؟ بالنظـر إلَى المثـالِ؛ يكـون السُّـؤال كالآتي: ما الـمُتَغَيّر المُسْـتَقِلّ، سـواء كان السُّـورَة المكِّيَّـة أو السُّـورَة المدَنِيَّـة بوصفهمـا متغيرَيـن فئويَّـين للوقـتِ (time)، أو كان قصـر أو توسُّـط أو طـول السُّـورَة كمـا هـو في الطـول (length)، الـذي سـيؤثِّر ويحـدد فيمـا إن كانـت هنـاك سـجدةٌ أم لا (yes أو no لـ prostration)؟ لننظـرْ إلى جـدولِ (١٩) لمعامـلات النّمـوذج. سـنلحظُ أَنَّ النَّمـوذج قـد أظهـر لنـا المنطلـقَ (intercept)، والوقـت (المكِّيَّـة time b_MK)، والطـول المتوسّـط (length b_medium)، والطّـول القصـير (length c_short)، ويُلحـظ أنَّ نمـوذج المنطلـقِ (intercept) أفضـل مـن نمـوذج الطُّـول القصـيرِ (length c_short)، ونعـرف ذلـك مـن تقديـر الأَرْجَحِيَّـة، فقيمـة التَّقديـر للمنطلـق أعلـى من قيمـة التَّقديـر لنموذج الطُّـول القصـير. نبحـثُ هنـا عـن أعلـى قيمـةٍ في عمـود تقديـر الأَرْجَحِيَّـة، وسـنجدها لنَمـوذج الطُّـول المتوسـطِ (length b_medium)، إذن؛ نقـول مبدئيًّـا بـأَنَّ أفضـلَ نمـوذج هـو نمـوذج الطول بفئـة المتوسـط (أي السُّـوَر متوسّـطة الطـول). ثقتنـا ليسـت كبـيرةً؛ لأنَّ مجـال الثقـة لهـا يُـراوح بـين 0.163 و2.103، مـن المهـمِّ أن نذكـر مجال الثِّقـة في نتيجة الانْحِـدَار اللُّوجسْـتِيّ، دون أن

نذكرَ قيم الاحتمال. لا نجد في قيمِ الاحتمال أيَّ دلالةٍ إحْصَائِيّةٍ للنماذج عدا المنطلق (كلها أكبر من 0.05)، ويُخبرنا جدول مُعاملات النموذج الأفضل بأنَّ الطُّولَ المتوسّطَ هو أفضلُ نموذج، وذلك بعد بنائِه باستخدام بناء النَّموذج التَّدريجِيِّ (build a model stepwise) في أدوات (لانكستر) الإحْصَائِيَّة، حتى وإن كانت قيمةُ الاحتمال غير دالَّةٍ؛ إلا أنَّ أفضل نموذج يُفسّر لنا وجودَ السَّجْدَة من عدم وجودها هو السُّوَر المتوسّطة، ويتجَلَّى هذا في معظمِ السُّوَر المتوسّطةِ التي وردت فيها السَّجْدَةُ؛ وهيَ الرعد، والنَّحل، والإسراء، ومريم، والحج (سجدتان)، والفرقان، والنَّمل، والسَّجْدَة، وص، وفصلت، والنَّجم، عدا سورة واحدة طويلة، وهي الأعراف، وسورتَين قصيرتَين؛ وهما الانشقاق والعلق.

ثَانِيًا: نتيجة المثال (ب)

- بناء النَّموذج للانحدار اللوجستي

إحصاءات النَّموذج: الأَرْجَحِيَّة اللوغارثمية: 6.66 (القيمة الاحتمالية = 0.155: غير دالٍّ إحْصَائِيًّا)، C–index: 0.72 (مقبول)، ناغلكيرك ر٢: 0.11، AIC: 91.62.

جدول ٢١ : مُعاملات النموذج للبيانات (ب)

	تقدير الأَرْجَحِيَّة اللوغارثمية	خطأ المعيار	قيمة-ز (والد)	قيمة الاحتمال	تقدير الأَرْجَحِيَّة	مجال الثقة ٩٥٪ (الأقل)	مجال الثقة ٩٥٪ (الأعلى)
(Intercept)	-3.075	1.247	-2.466	0.014	0.046	0.004	0.585
Varietyb_7	1.230	0.629	1.956	0.050	3.422	1.069	13.308
Syntaxb_present	0.028	0.967	0.029	0.977	1.029	0.143	7.047
Syntaxc_nominal	-1.193	1.133	-1.053	0.292	0.303	0.015	1.915
Length	0.199	0.339	0.586	0.558	1.220	0.594	2.393

- بناء النموذج التدريجِيّ stepwise (النموذج الأفضل)

الإحصـاءات الكليـة: الأَرْجَحِيَّـة الاحتماليَّـة: 4.04 (القيمـة الاحتماليـة = 0.0443 دالٌّ إِحْصَائِيًّـا)، C–index: 0.63 (غـير مقبـول)، ناغلكـيرك ر٢: 0.07، AIC: 88.24.

جدول ٢٢: مُعاملات النموذج الأفضل للبيانات (ب)

	تقدير الأَرْجَحِيَّة اللوغاريتمية	خطأ المعيار	قيمة-ز (والد)	قيمـة الاحتمال	تقـدير الأَرْجَحِيَّة	مجال الثقة ٩٥٪ (الأقل)	مجال الثقة ٩٥٪ الأعلى)
Varietyb_7	1.169	0.617	1.896	0.058	3.220	1.029	12.247

- تفسير النَّتيجة

مـا النمـوذج الأفضـل الـذي يُبـيّن حجـمَ التَّأثـيرِ الأعلـى مـن الـمُتَغَيّرَات المُسْـتَقِلَّة (المتنبئـة) الفئويَّـة علـى الـمُتَغَيّر التَّابِـع (النَّـاتج)؟ بالنَّظـر إِلَى المثـالِ؛ يكـونُ السـؤال كالآتي: مـا الـمُتَغَيّر المُسْـتَقِلّ الفئـوي؛ سـورة النسـاء (Variety a_3)، أو سـورة الأعـرافِ (Variety b_7)، أو الزمـن المـاضي (Syntax a_past)، أو الزَّمـن المضـارع (Syntax b_present)، أو الجملـة الاسـمِيَّة (Syntax c_nominal)؛ الـذي يؤثِّـر أكـثر علـى الـمُتَغَيّر النَّـاتج (اسـتخدام «مـا»[1] الموصوليَّـة، واسـتخدام اسـم الموصـول «الـذي»)؟ بالنظـر إلى جـدولِ معامـلات النَّمـوذج؛ فـإنَّ القيمـة الأعلـى لتقديـر الأَرْجَحِيَّـة هـي ٣,٤٢٢ لنمـوذج سـورة الأعـرافِ، وقيمـة احتمـال دالـة: ٠,٠٥٠، ومجـال الثِّقـة كبـيرٌ. يُكتَفـى هنـا بذكـر نتيجـة مُعامـلات النمـوذج الأفضـل، وأن سـورة الأعـراف كمتغـير مسـتقلٍّ ذي تأثـيرٍ أكـبر علـى اسـتخدام (مـا الموصوليـة) أكـثرُ أرجحِيَّـةً

(١) وردت (مـا) في سـورة الأعـراف ٥٨ مـرةً، منهـا ٤١ حالـة جـاءت موصوليَّـةً، مثـل هـذه الحـالات الـتي تتطلـب تصنيفًـا فئويًّـا دقيقًـا؛ يحتـاج المحلّـل إلى أن يُهيئ بياناتـه وفـق الحـالات الصَّحيحـة، إذ احتجـت إلى تتَبُّـع (مـا)، واسـتخراج مـا جـاءت موصوليَّـةً منهـا. كذلـك الحـال لَـدَى الـمُتَغَيّرَات التَّركيبيـة، وهـل مـا جـاء بعدهـا جملـةٌ فعليَّـةٌ أم جملـة اسـمِيَّةٌ، وهـل الجملـة الفِعْليّـة قـد جـاءت علـى صيغـة المـاضي أم المضـارع، وكـم عـدد حـروف الكلمـة الأولى بعـد (مـا الموصوليَّـة)، سـواء كانـت فعـلًا أم اسـمًا، وهـذا مـا تتَطَلَّبُـه بيانـاتُ الانحـدار.

لسـورة الأعـراف، ولا علاقـةَ لاسـتخدامها تحديـدًا في سـورة النسـاء، أو في كـون الجملـة الـتي تـأتي بعدهـا جملَـةً فعلِيَّـةً (ماضيـة أو مضارعـة) أو جملـةً اسـمِيَّةً.

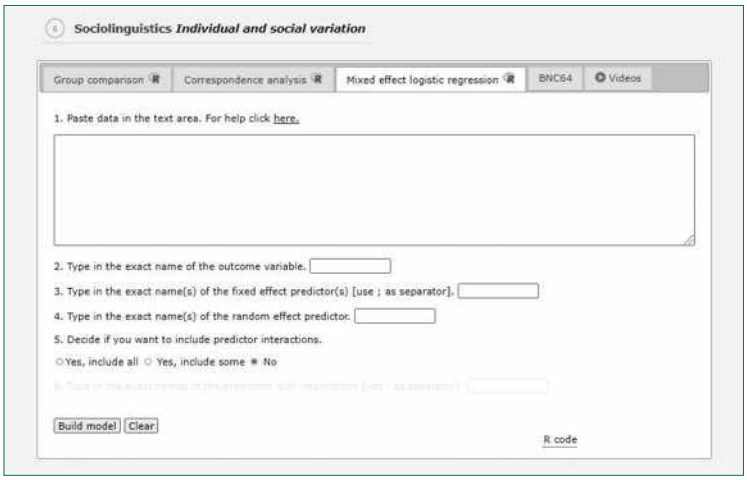
المبحث الثَّالِث: التَّأثيرات الممتزجة

هـو أسـلوبٌ مهِـمٌّ، ويعمـلُ مثـل الانْحِـدَار اللُّوجسْـتِيّ، ويُسـتخدمُ لكشـف حجـم تأثـير مُتَغَـيّرٍ مسـتقِلٍّ نختـاره عشـوائيًّا مـن بـين الـمُتَغَيّرَات المُسْـتَقِلَّة الأخرى علـى الـمُتَغَيّر التَّابِع (النَّـاتِج)، سـنعيد بيانـاتِ الانْحِـدَار اللُّوجسْـتِيّ (الشـكل ٦١)، لننظـر في نتيجـةِ نمذجـة التَّأثـيرات الممتزجـةِ (mixed-effects modeling).

- المتغيّر التَّابِع (النَّاتِج): الأسماء الموصولة (Relatives).
- الـمُتَغَيّرَات المُسْـتَقِلَّة (المتنبِّئـة) للتَّأثـيرات الممتزجـة: يعتمـد علـى الاختيـار وفـق سـؤال التَّحْلِيـل.
- المتغـير المُسْـتَقِلّ (المتنـبّئ) للتَّأثـير العشـوائِيِّ: يعتمـدُ علـى تحديـد أحـد الـمُتَغَيّرَات المُسْـتَقِلَّة (المتنبِّئـة).

ID	Variety	Syntax	Length	Relatives
1	a_3	a_past	3	a_mā
2	a_3	a_past	3	a_mā
3	a_3	a_past	3	a_mā
4	a_3	a_past	3	a_mā
5	a_3	a_past	3	a_mā
6	a_3	a_past	3	a_mā
7	a_3	a_past	3	a_mā
8	a_3	a_past	3	a_mā
9	a_3	a_past	3	a_mā
10	a_3	a_past	3	a_mā
11	a_3	a_past	3	a_mā
12	a_3	a_past	3	a_mā
13	a_3	a_past	3	a_mā
14	a_3	a_past	3	a_mā
…	…	…	…	…
98	b_7	b_present	4	b_alladhī
99	b_7	b_present	4	b_alladhī
100	b_7	b_present	4	b_alladhī
101	b_7	c_nominal	7	b_alladhī

شكل ٦١ : بيانات الانْحِدَار اللُّوجسْتِيّ



شكل ٦٢ : واجهة الانْحِدَار اللُّوجسْتِيّ للتَّأثير الممتزج في (اللُّغَوِيّات الاجْتِمَاعِيّة)

تُوضع البَيَانَات في خانة الانْحِدَار اللُّوجسْتِيّ للتأثير الممتزج Mixed effect logistic regression في مربع اللُّغَوِيّات الاجْتِمَاعِيّة Sociolinguistics (الخيار ١) في أدوات (لانكستر) الإحْصَائِيَّة كما في الشَّكْل (٦٢)، ثم يُكتب اسم الـمُتَغَيّر النَّاتج في الخانة المخصصة له في الخيار (٢)، ثم يُكتب اسم الـمُتَغَيّرَات المُسْتَقِلَّة مفصولةً بفاصلةٍ منقوطةٍ، دون مسافاتٍ، في الخانة المخَصَّصة لها في الخيار (٣)، ثم يُكتبُ اسم الـمُتَغَيّر المُسْتَقِلّ المتنَبّئ العشوائِيّ في الخانة المخصَّصة له في الخيار (٤)، ثم يُختار (Yes, include all)، أي: تضمين كُلّ التفاعلاتِ بين كُلّ الـمُتَغَيّرَات المُسْتَقِلَّة المتنبّئة (predictor interactions)؛ من أجل دقة النتيجة، سنستخلص ثلاثَ نتائج، وذلك بجعل الطولِ (Length) في المرة الأولى متنَبِّئ الأثر العشوائي، ثم بجعل التَّنَوُّع (Variety) في المرة الثَّانية متنبئ الأثر العشوائيّ، ثم بجعل التركيبِ (Syntax) في المرة الثَّالثة متنبئ الأثر العشوائي. لاحظ أننا في كُلّ مرةٍ سنجعل المتنبئات المنضبطَة ما يتبقى منها من دون متنبئ الأثر العشوائي. أنبِّهُ -أيضًا- إلى أنَّ التَّحليل قد يعتمد على متنبئ منضبطٍ واحدٍ مع متنبئٍ عشوائِيٍّ، وكل ذلك يعتمدُ على السُّؤال البحثيّ الآتي: أيُّ الـمُتَغَيّرَات المتنبّئة العشوائيَّة أكثر تفاعُلًا مع الـمُتَغَيّر المنضبط، أو مع المُتَغَيّرَات المنضبطة في حجمِ التّأثير في الـمُتَغَيّر التَّابِع (النَّاتج)؟ تحديد ذلك هو ما سيُحدِّد الـمُتَغَيّر العشوائيَّ، والمتغير أو الـمُتَغَيّرَات المنضبطة.

- نتيجة الـمُتَغَيّر المُسْتَقِلّ المتنَبِّئ العشوائِيّ -الطول Length-: يُلحظُ أَنَّ أَعلى قيمة لتقدير الأَرْجَحِيَّة اللوغارثمية من الطولِ هو لسورة الأعرافِ، وللتَّركيب الاسمِيِّ (Variety_7:Syntaxt_nominal) اللذين كانَ حجمُ أثرهما لاستعمال ما (الموصولِيَّة) أكبر.

جدول ٢٣ التَّأثيرات الممتزجة من الـمُتَغَيّر المتنبّئ العشوائي (الطول Length)

	تقدير الأرجحيات اللوغارثمية	الخطأ المعياري	قيمة-ز (والد)	قيمة الاحتمال
(Intercept)	-2.548560	1.3787328	-1.84848025	0.0645329
Varietyb_7	1.100914	0.8959125	1.22881924	0.2191396
Syntaxb_present	2.414647	2.2230061	1.08620794	0.2773870
Syntaxc_nominal	-16.745291	482.7189460	-0.03468953	0.9723273
Varietyb_7:Syntaxb_present	-2.195673	2.2992855	-0.95493725	0.3396094
Varietyb_7:Syntaxc_nominal	15.525913	482.7190753	0.03216345	0.9743417

- نتيجة الـمُتَغَيّر المُسْتَقِلّ المتنبئ العشوائي: (التَّنَوُّع Variety): يُلحظ بأن أعلى قيمة لتقدير الأَرْجَحِيَّة اللوغارثمية من التَّنَوُّع (سورة النساء أو سورة الأعراف) هو في التركيب الفِعْلِيّ (المضارع)، الذي كان حجم أثرِه لاستعمال ما (الموصولِيَّة) أكبر.

جدول ٢٤ التأثيرات الممتزجة من الـمُتَغَيّر المتَنبِّئ العشوائِيّ (التَّنَوُّع Variety)

	تقدير الأرجحيات اللوغارثمية	الخطأ المعياري	قيمة-ز (والد)	قيمة الاحتمال
(Intercept)	-6.527617	2.3180065	-2.816047727	0.0048618444
Length	1.485655	0.6818446	2.178875697	0.0293409037
Syntaxb_present	15.840394	4.5401480	3.488959759	0.0004849041
Syntaxc_nominal	-25.518920	7308.5967095	-0.003491631	0.9972140875
Length:Syntaxb_present	-3.545045	1.0418110	-3.402771926	0.0006670594
Length:Syntaxc_nominal	3.092422	1044.0854628	0.002961848	0.9976367908

- نتيجـة الـمُتَغَيّر المُسْـتَقِلّ المتنـبّئ العشـوائي -التركيـب Syntax-: يُلحـظ أَنَّ أعلـى قيمَـة لتقديـر الأَرْجَحِيَّـة اللوغارثميـة مـن التَّركيـب هـو لسـورة الأعـرافِ، الـتي كان حجـمُ أثرهـا لاسـتعمال مـا (الموصولِيَّـة) أكـبر.

جدول ٢٥ التأثيرات الممتزجة من الـمُتَغَيّر المتنبئ العشوائي (التركيب Syntax)

	تقدير الأرجحيات اللوغارثمية	الخطأ المعياري	قيمة -ز (والد)	قيمة الاحتمال
(Intercept)	-4.2245138	1.6338685	-2.5855898	0.009721255
Varietyb_7	2.6768476	1.9472614	1.3746730	0.169232857
Length	0.4578589	0.3517511	1.3016558	0.193034059
Varietyb_7:Length	-0.3732201	0.4403539	-0.8475458	0.396690975

الاسـتنتاج العـام: عنـد اسـتعمال (مـا الموصولِيَّـة) يغلـبُ مـجيءُ المرَكَّـبِ الاسـمِيّ، ويغلـب عنـد اسـتعمال (الـذي) مـجيءُ المركَّـبِ الفِعْلِـيّ.

الفصل الخامس
تحليلُ البَيَانَات اللُّغَوِيّة التَّاريخِيّة

يُطبَّق أسلوبان لتحليل التَّغيُّر اللُّغَوِيّ عبر الزَّمن؛ الأَوَّل: تحليل القمم والقعور. والثَّانِي: تحليل التقلّب الاستعماليّ. للتغير اللُّغَوِيِّ أوجُه عَديدَةٌ في الدِّرَاسَاتِ اللُّغَوِيّة التَّاريخِيّة، كأن يكون وجه التَّغير اللُّغَوِيّ تكراريًّا (أي: زيادة أو نقصان أو اختفاء ظهور الكلمة عبر الزَّمن)، أو أن يكون وجه التغير اللُّغَوِيّ في البنية اللُّغَوِيّة (أي: البنية الصَّوتية، أو البنية الصَّرفية، أو البنية التَّركيبيَّة، أو البنية الدّلالِيَّة). لا شكّ أن وجهَ التغير اللُّغَوِيّ في البنيةِ من حيث تهيئة بياناتِه أصعبُ من وجه التغَيُّر اللُّغَوِيّ في تذبذبِ الاستخدام للكلمةِ، وعلى أيِّ حالٍ؛ أعيدُ أهَمِّيَّةَ تأسيس البَيَانَات اللُّغَوِيّة من حيث الجمع والإعداد والتَّهيئة والعنايةِ؛ كونها المسألةَ التي تنالُ الجهدَ الأكبر في مجال عِلْمِ البَيَانَاتِ (اللُّغَوِيّة).

المبحث الأَوَّل: القمم والقعور

في هذا المبحث الأَوَّل؛ احتجتُ إلى تفكيرٍ إحْصَائِيٍّ عما يُمكن تحليلُه في مُدَوّنَة القرآن الكريم. ظهرَت لي فكرةُ النَّظر في استعمالِ كلمة (رحمة) واشتقاقاتها، واستعمال كلمةِ (عذاب) واشتقاقاتها[1] عبر سور القرآن الكريم، حسب نزولها زمنِيًّا. في القرآنِ الكريم ١١٤ سورة، نزلت على النبيِّ محمدٍ -صلَّى الله عليه وسَلَّمَ- على مدى ٢٣ سنَةً. عملتُ

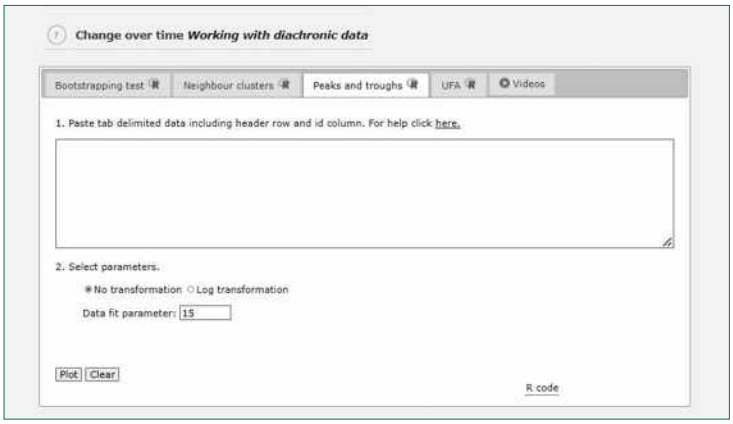
(١) اشتقاقات رحمة: (رحمة)، و(رحمته)، و(ترحمون)، و(الرحمة)، و(رحمتنا)، و(أرحم)، و(رحمت)، و(رحم)، و(ارحمنا)، و(يرحمكم)، و(رحمتك)، و(يرحم)، و(رحمناهم)، و(المرحمة)، و(سيرحمهم)، و(ارحمهما)، و(ارحم)، و(ارحمنا)، و(رحمه)، و(رحما)، و(ترحمني)، و(ترحمنا)، و(رحمتي)، و(يرحمنا)، و(رحمنا)، و(رحماء). واشتقاقات (عذاب): (عذاب)، و(يعذب)، و(يعذبكم)، و(معذبين)، و(أعذبه)، و(يعذبه)، و(معذبهم)، و(عذّب)، و(تعذبهم)، و(معذبوها)، و(سنعذبهم)، و(عذبناها)، و(يعذبهم)، و(عذَبهم)، و(عذّبنا)، و(معذبين)، و(يعذبنا)، و(تعذب)، و(نعذب)، و(المعذبين)، و(أعذبنه)، و(العذاب)، و(بعذاب)، و(لعذاب)، و(عذابه)، و(عذابها)، و(أفبعذابنا)، (بعذابكم)، و(عذابهما)، و(كعذاب).

علـى تصنيـفِ السُّـور، ووضـع كُلّ خمـس سـورٍ توالـت بالتَّرَاتـب في نزولهـا زمنيًّـا في كُلّ سـنةٍ، ليكـون عـددُ المجلـداتِ الـتي عولجـت ٢٣ مجلـدًا، واسـتعنتُ بـأداة غـواص في اسـتخراج تكـرارات الكلمتـين مـن كُلّ مجلـدٍ. يُبـيِّن الشَّـكْل (٦٣) شـكلَ البَيَانَـات المهَيَّـأة، وتُلصـق كمـا هـي في خانـة القمـم والقعـورِ (Peaks and troughs) في مربـع التغَـيُّرِ عـبر الزمـنِ (Change over time)؛ لاسـتخراجِ تمثيـل الانحـدار غـير الخَطِّـيِّ، وفي مربـع التَّجْمِيعـات المتجـاورةِ (Neighbour clusters) لاسـتخراج تمثيـل مسـافات نقـاط ورود الكلمـة عـبر الزَّمـنِ في أدوات (لانكسـتر) الإِحْصَائِيَّـةِ كمـا في الشَّـكْل (٦٤). ظهـرت النَّتـائج كمـا في الشَّـكْل (٦٥)، وذلـك بتمثيـل الانحـدار غـير الخَطِّـيِّ لكلمـة (رحمـة) علـى مـدى ٢٣ عامًـا كمـا في (أ)، وتمثيـل مسـافات نقـاطِ ورودِهـا عـبر الزَّمـنِ كمـا في (ب)، ثـم كمـا في الشَّـكْل (٦٦)، وذلـك بتمثيـلِ الانحـدار غـير الخَطِّـيِّ لكلمـة (عـذاب) علـى مـدى ٢٣ عامًـا كمـا في (أ)، وتمثيـل مسـافاتِ نقـاط ورودِهـا كمـا في (ب).

بيانات كلمة (رحمة)		
ID	time	Variable
1	1	0
2	2	0
3	3	0
4	4	0
5	5	0
6	6	0
7	7	1
8	8	17
9	9	8
10	10	17
11	11	29
12	12	7
13	13	10
14	14	10
15	15	8
16	16	1
17	17	8
18	18	13
19	19	7
20	20	1
21	21	5
22	22	1
23	23	6

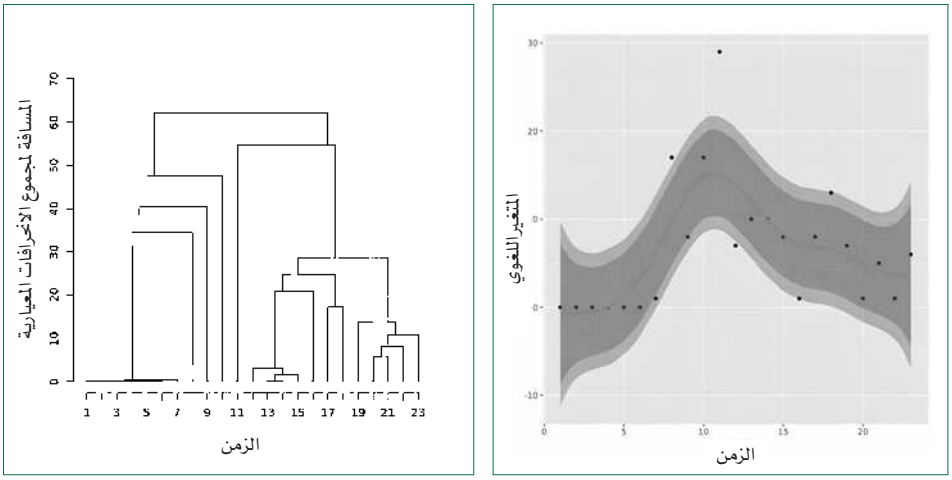
بيانات كلمة (عذاب)		
ID	time	Variable
1	1	3
2	2	3
3	3	0
4	4	0
5	5	0
6	6	2
7	7	1
8	8	22
9	9	21
10	10	20
11	11	36
12	12	34
13	13	26
14	14	25
15	15	16
16	16	13
17	17	10
18	18	52
19	19	15
20	20	6
21	21	24
22	22	2
23	23	42

شكل ٦٣ : بيانات (رحمة) و(عذاب)



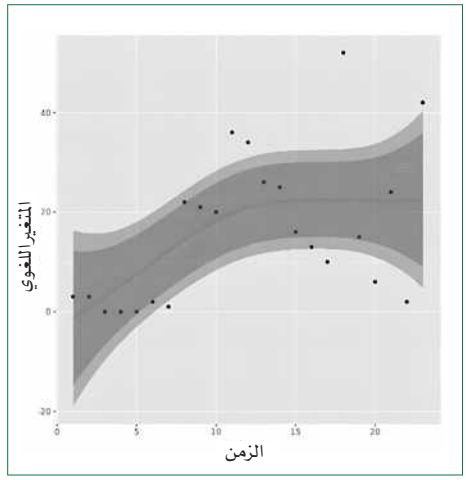
شكل ٦٤ : واجهة القمم والقعور في (التغير عبر الزمن)

(أ) (ب)

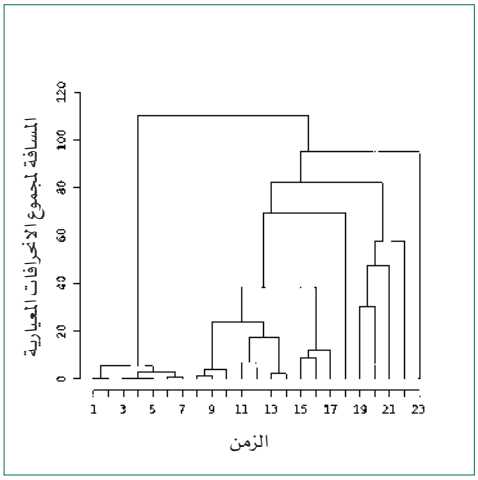


شكل ٦٥ : تمثيل نتيجة استعمال (رحمة) عبر الزمن في سور القرآن الكريم

(أ)



(ب)



شكل ٦٦ : تمثيل نتيجة استعمال (عذاب) عبر الزمن في سور القرآن الكريم

كيـف نفسّـر القمـمَ والقعـور لاسـتخدام الكلمتـين عـبر الزَّمـن في القـرآنِ الكريـم؟ في الشَّـكْل (٦٣) سـيُلحظُ بـأن الـمُتَغَيّر اللُّغَـوِيَّ قـد مُثّـلَ بالتّكْـرَار الملحـوظ مـن دون تحويـلٍ، ويُقصـد بالتَّحويـل إمكانِيَّـة تحويـل التّكْـرَار إلى قيـمٍ لوغارثميـة، أو باسـتخدام التّكْـرَارات النِّسْـبِيّةِ، وتتمَثَّـلُ فائـدةُ التَّحويـل عندمـا تكـونُ قيـم التّكْـرَارات الملحوظـة متطرفَـةً أو متباينَـةً، ولـم يُحتـج إلى مثـل هـذا التَّحويـلِ؛ نظـرًا إلَى قِلَّـةِ التّكْـرَارات الملحوظـةِ. وبالنَّظـر إلى الانحـدار غـير الخَطِّـيِّ للكلمتـين عـبر الزمـنِ؛ فـإن كُلَّ نقطـةٍ تُمثِّـلُ التّكْـرَار الأعلـى أو الأقـلّ، والـذي يُمكـن قولـه مـن خـلال هـذه النِّقـاطِ الـتي نراهـا أَنَّ كلمـتَي (رحمـة) و(عـذاب) في آيـات السُّـوَر الـتي نزلـت في السَّـنوات السـبعَةِ الأُولى متقاربَـةَ الـورودِ، ثـم ازدادت في السَّـنواتِ الـتي تلَـت، واسـتقَرَّ الـورودُ لكلمـة (عـذاب)، وبـدأ ورود كلمـة (رحمـة) في الـنزول في بقِيَّـة السَّـنوات، ونلحـظُ -أيضًـا- في الخَطَّـين البيانِيَّـين أَنَّ أَعلـى قِمَّـة لـورود كلمـة (رحمـة) قـد كانـت في السَّـنة الثَّانيـة قبـل الهجـرةِ، (الـتي تمثِّـل السَّـنة الـ١١ مـن السَّـنوات الـ٢٣)، وأنَّ أعلـى قمـة لـورود كلمـة (عـذاب) قـد كانـت في السَّـنة الثَّانيـة والثَّالثـة قبـل الهجـرةِ، (الـتي تمثـل السَّـنة الـ١١ والـ١٢ مـن السَّـنوات الـ٢٣)، ثـم بـين السَّـنة الخامسـة والسَّـادسـة مـن

الهجرة (التي تُمثِّل ما بين السنة الـ ١٨ والـ ١٩ من السَّنوات الـ ٢٣)، ويُقصد بنقطة القمَّة تلك النُّقطة التي ظهرت خارج الانحدار غير الخَطِّيِّ، ويُلحظ من هذا كلِّه أنَّ نتائج تحليل القمم والقعور في هاتين الكلمتين عبر الزَّمن في القرآن الكريم متقاربةً نوعًا ما، عدا نقاط القمم المتفاوتة نسبيًا بينهما.

المبحث الثَّانِي: التَّأرجُح الاستِعمَالِيّ

يهدفُ هذا النَّوعُ من التَّحليلِ إلى كشفِ تطَوُّرِ معاني الكَلِمَاتِ، أو تطور استعمالاتها، وهو تطوُّرٌ يطرأ من تأرجح المتصاحبات اللَّفْظِيّة مع الكَلِمَات عبر الزمن. ولتبسيط طريقة تحليل هذا النوع؛ لنفترض أن الكلمةَ هي (س)، وأن هذه الكلمةَ قد تصاحبت معها الكَلِمَات (أ)، و(ب)، و(ج)، و(د)، ونريد كشفَ التَّقَلُّب الاستعماليّ لهذه الكلمةِ في مُدَّةٍ تُراوح بين نقطتين، تُمثِّل عامين، أو أعوامًا، أو قرونًا من الزَّمنِ، ويتطَلّب هذا بناء مجموعة بياناتٍ لكشف حالات بداية الظُّهور، والثَّبات، والانقطاع، والتحوُّل للمتصاحبات اللَّفْظِيّة (أ) و(ب) و(ج) و(د) مع (س).

يُخيَّر المحلّل بين طريقتين لتحليل التَّأرجح الاستعماليّ؛ وهما على النَّحو الآتي:

الأولى: التَّحليل النَّوْعِيّ القائم على ما يُقيِّمه المحَلِّل البشري (الشكل ٦٧ والشكل ٦٨).

الثَّانِي: التَّحليل الكَمِّيّ القائم على ما يُقيِّمه المحَلِّل الآلِيّ (الشكل ٦٩ والشكل ٧٠).

بالنِّسبة للطَّريقة الأولى؛ سننظر إلى كلمة (رحمة) وكلمة (عذاب) من ناحية التَّأرجح الاستعماليّ لهما مع متصاحباتهما اللَّفْظِيّة على نافذةِ تصاحب (يم١-يس١). تَصَدَّيتُ لتلك المتصاحبات عبر الزَّمنِ، ثم صنَّفتُها إلى تلكَ التي لُوحظت بدايتُها، والتي لُوحظَ ثباتُها، والتي لُوحظ انقطاعُها، والتي لُوحظ تحَوُّلُها. يُمكن فهمُ بداية التَّصَاحُب زمنِيًّا، وثبات التَّصَاحُب زمنِيًّا، وانقطاع التَّصَاحُب زمنِيًّا، لكن ماذا عن تحوُّلِ المتصاحب؟ الجواب: هو تَحَوُّلُ المتصاحب إلى مصاحبةِ كلمةٍ أخرى زمنِيًّا، ويُستخرج بوضع متصاحبات الكلمة المعنية في قائمة شمول، ثم استخراج التَّتابعِ (2n-grams)

[نافـذة التَّصَاحُـب: يم١-يـس١]، ثـم الوقـوف علـى المتصاحبـةِ مـع الكَلِمَـات الـتي جـاءَت معهـا مـن غـير كلمـةِ (رحمـة) و(عـذاب).

٢٣ سنة (كل رقمٍ مع المتصاحب يشير إلى السَّنةِ)

الكلمة: (رحمة). نوع التحليل: تحليل التأرجح الاستعماليّ للمتصاحبات

المتصاحبات التي بدأت: «خزائن» (٨)، «ذكر»، «ربك» (٩)، «ابتغاء»، «ربي» (١٠)، «واسعة»، «ربه» (١١)، «الله» (١٢)، «للعالمين» (١٥)، «الناس» (١٧).

المتصاحبات التي ثبتت: لم ينطبق.

المتصاحبات التي انقطعت: «ذكر» (٩).

المتصاحبات التي تحَوَّلت: «خزائن» (١١) (٢١)، «ابتغاء» (١٩) (٢٠)، «للعالمين» (٩) (١١) (١٥) (١٧) (١٨).

شكل ٦٧: تمثيل نتيجة استعمال تصاحبات (رحمة) عبر الزَّمن في سور القرآن الكريم

٢٣ سنة (كلُّ رقمٍ مع المتصاحب يشيرُ إلى السَّنةِ)

الكلمة: (عذاب). نوع التحليل: تحليل التأرجح الاستعماليِّ للمتصاحبات

المتصاحبات التي بدأت: «سوط» (٢)، «الحريق»، «جهنم» (٦)، «شديد»، «بكرة»، «مستقر» (٨)، «أليم» (٩)، «ربك» (١٠)، «الله»، «الخلد»، «الآخرة»، «مقيم»، «الهون»، «غليظ»، «قريب» (١١)، «السعير»، «مهين»، «النار»، «الجحيم»، «واصب»، «مقيم» (١٢)، «عظيم»، «الحميم»، «الخزي» (١٣)، «الهون» (١٤)، «ربهم»، «يومئذ»، «السموم» (١٦).

المتصاحبات التي ثبتت [ثبات منقطع]: «شديد» (٨) (٩) (١٢) (١٣) (١٥) (١٨)، «أليم» (٩) (١١) (١٤) (١٥) (١٦) (١٧) (١٨). [ثبات متصل]: لا ينطبق.

المتصاحبات التي انقطعت: «سوط» (٢)، «بكرة» (٨)، «مستقر» (٨).

المتصاحبات التي تحَوَّلت: «قريب» (١١)، «عظيم» (١) (١٨) (٢٢) (٢٣)، «الخلد» (٩) (١٣)، «شديد» (١٠) (١٩) (٢٣).

شكل ٦٨ : تمثيل نتيجة استعمال تصاحبات (عذاب) عبر الزَّمن في سور القرآن الكريم

لنتذكَّر في (شبكات التَّصَاحُب اللَّفْظِيّ) في المبحث الأَوَّل من الفصل الثَّاني إجراء (مجموعة مقاييس التَّصَاحُبِ collocation parameters notation)، إذْ لا بدَّ من تطبيقِه على تحليل التَّأرجحِ الاستعماليِّ بين الكَلِمَات والمتصاحباتِ؛ من أجل ضبطِ المخرجات بدِقَّةٍ في حال ظهور كمٍّ هائلٍ من المتصاحبات في المُدَوّنَات اللُّغَوِيّة الكبيرة.

أما بالنِّسبة إِلَى الطَّريقة الثَّانية، التي تعتمد على ما يُقيّمه المحلّل الآلِيُّ؛ فيتطلَّبُ ذلكَ إعداد بيانات مهيكلةٍ، كما في الشَّكْل (٦٩)، تَتَضَمَّن سبعةَ أعمدةٍ، وعنوان كُلّ عمودٍ هو: التَّسلسل (No.)، ثم الكلمة (Word) المتصاحبة، ثم تكرارها لوحدها في كامل المُدَوّنَة

(Total no. in whole corpus)، ثـم التّكْرَار المتوقَّـع للتَّصاحـبِ (Expected collocate frequency)، ثـم التّكْرَار الملحُـوظ للتَّصاحـب (Observed collocate frequency)، ثـم تكـرار النَّصِّ للتصاحب (In no. of texts)، ثـم قيمـة مُعامـل المعلومـات المتبادلـةِ (Mutual information value). ولإتمـام المعالجـة في أدوات (لانكسـتر)، أو باسـتخدام الأكـواد المذكـورةِ في الملحـق (ج)، لا بـدَّ أن يكـون المجَلَّـدُ مضغوطًـا، ويَتَضَمَّـنُ ملَفَّـاتٍ عَديـدَةً، وفي كُلّ ملَـفٍّ مجموعـةُ المصاحبـات اللَّفْظِيّـة لكلمـةِ (رحمـة)، أو (عـذاب)، أو أي كلمـةٍ أُخـرى معنِيَّـة بالتَّحليـل لِمُـدَّةٍ زمنِيَّـةٍ محَـدَّدةٍ. وفي مثالنـا هنـا؛ تمتـدُّ المُـدَّةُ الزمنِيَّـة لثـلاث وعشـرين سـنةً لتلـك الكلمتَـيْن، وعليـه يَكـون عـددُ الملَفَّـات ٢٣ ملَفًّـا، ثـم تُـترك الخطـوات التاليـة بعـد تحميـل الملـف في خانـة تحليـل التَّأرجـح الاسـتعماليِّ (UFA) في مربـع التَّغـيُّر عـبر الزمـنِ (Change over time) في أدوات (لانكسـتر) الإحْصَائِيَّـة كمـا في الشَّـكْل (٧٠).

No.	Word	Total no. in whole corpus	Expected collocate frequency Observed collocate frequency	In no. of texts	Mutual information value
1	خزائن	0	1	1	6.0747
2	ذكر	0	2	2	6.0747
3	ربك	0	3	3	6.0747
4	.....	.....	.....	.....	.....
5	.....	.....	.....	.....	.....
6	.....	.....	.....	.....	.....

شكل ٦٩ : بيانات استخدام منهج تحليل التأرجح الاستعمالي

شكل ٧٠ : واجهة تحليل التأرجح الاستعمالي UFA في (التغير عبر الزمن)

لهذا الأسلوبِ الآليِّ قوَّةٌ متينةٌ في استخراج خارطة التَّصَاحُباتِ، وتأرجحِها، وتقلُّبِها الاستعماليِّ عبر الزَّمنِ، من حيث المصاحبة التي بدأ استعمالها، والتي ثبتَ استعمالها، والتي غاب استعمالها، والتي تحوّل استعمالها، وهي معالجةٌ قويَّةٌ تُستخدم في بناء المعاجم التَّاريخِيّةِ، صحيحٌ أنَّ إعداد البَيَانَات يأخذُ جهدًا؛ ولكنه لا يأخذ زمنًا طويلًا؛ إذْ كلما زاد الجهد في إعداد البيانات قلّ زمن الإعداد، والعكس صحيح، فلو افتَرضنا أنَّ الهدفَ من بناء معجمٍ تاريخِيٍّ هو الوقوفُ على التَّأرجُحاتِ والتقلُّبات الاستعماليَّة عبر ١٤ قرنًا، ولدينا مُدَوّنَةٌ عربيَّةٌ تاريخِيَّةٌ تَتَضَمَّن مليارَ كلمة، وأن كُلّ نقطةٍ تاريخِيَّةٍ نودُّ تصميم الخارطة بناءً عليها تمتدُّ لكلِّ عشر سنواتٍ، أو لكُلِّ مائة سنةٍ، فسنحتاج إلى إعدادِ ١٤٠ ملفًّا للسَّنوات العَشر، أو ١٤ ملَفًّا فقط للسَّنوات المائةِ، وذلك لا يستغرقُ زمنًا، إذْ يُمكن إعدادُها في غضون شهر من محلل بشريٍّ واحدٍ (هذه تجربَة مؤلّف الكتاب)، ولكَ أن تتصوَّر ما ستُظهره الآلةُ بهذه الأداةِ أو بأكوادها في الآر في ثوانٍ معدودةٍ من قوائم الكَلِمَاتِ ومتصاحباتها من بداية الظُّهور زمنِيًّا، وهل ثبتت زمنِيًّا، وهل غابت في نقطةٍ زمنِيَّةٍ، أو في نقاطٍ زمنِيَّةٍ معينةٍ، وهل تحوَّلَت في نقطةٍ أو في نقاطٍ زمنِيَّةٍ معَيَّنةٍ.

المبحث الثَّالِث: حوسبة المعجم

تُجـرَى حوسبـةُ المعجـم في مجالَيـن مختلفـين؛ الأَوَّل: صنـاعـةُ المعجـم الحـاسـوبِيِّ (computational lexicography)، والثَّانِي: علم الوجـودِ (ontology)، وتـقنيـات دلالة الويـب (web semantic technologies).

يهدفُ المجالُ الأَوَّلُ إلى تصميم وإخراج معاجم إلكترونِيَّـةٍ في تطبيقاتٍ ومواقع شـبكِيَّةٍ، ويعتمـدُ التَّصميـمُ والبنـاءُ علـى مُدَوّنَـةٍ لُغَوِيَّـةٍ، وهـو الأسـاس والمنطلـقُ، ويتحَـدَّدُ مجـالُ الـمُدَوّنَة اللُّغَوِيّـة وزمانُهـا بتحديـد غـرض المعجـم، فـإن كان تاريخِيًّـا؛ فالـمُدَوّنَة بطبيعـة الحـال سـتكون تاريخِيَّـةً[1]، وإن كان معاصِـرًا؛ فالـمُدَوّنَة بطبيعـة الحـال سـتكون معاصـرةً. ولمناهـج لغويـات الـمُدَوّنَات دورٌ مهـمٌّ قـد بـات أساسِـيًّا في عصرنـا اليـوم لأيِّ مشـروع يُبـنى لغـرض صناعـة المعاجـمِ وإخراجهـا، كمـا أنَّ لـدور علـم المعلومـات أهَمِّيَّـةً أساسِـيَّةً في الصِّناعـة المعجَمِيّـة الرَّقَمِيَّـةِ (e-lexicography)، فيشـير تـارب[2] إلى أنَّ دور الحاسـب الآلِيِّ، وعلـم الإحصـاءِ، وتعلـم الآلـة الموجـه وغـير الموجَّـه في الـمُدَوّنَات اللُّغَوِيّـة؛ قـد ابتكـر بيئـةً عملِيَّـةً وتطبيقِيَّـةً تُطـوِّر مـن صناعـة المعاجـم، وأنَّ المعجـم في الأسـاس مـا هـو إلَّا حقـلٌ معلوماتِيٌّ، ويحتَـاجُ في العصـر الحـالِيِّ إلى الأدوات الرَّقَمِيّـة الـتي تكـونُ فيـه المنطلـق. وكل معجـم ورقِيٍّ يُحـول إلى منصَّـةٍ رَقَمِيّـةٍ في الإنترنـت أو في التَّطْبِيقَـات الذكِيَّـة سـيُعدُّ معجمًـا مرقمنًـا digitized، ولـن يُعـدَّ معجمًـا مُشَـفْرَنًا digitalized أو مُعتمـدًا علـى أدوات التحـوّل الرَّقَـمِيّ؛ لأن المنطلـقَ لـم يكـن معتمـدًا علـى مـا يسـمِّيه تـارب بـالأدوات الرَّقَمِيّـةِ (e-tools)، وهـي أدواتٌ تُبـنى لغـرض الإفـادة المعلوماتِيَّـة المعجَمِيّـة مـن المدوَّنـاتِ اللغويـةِ الحيّـةِ والطبيعـة، مثـل الـتي طبَّقتهـا -ومازالـت تُجـري تطبيقاتهـا- مجموعـةُ (Pearson) ومجموعـةُ (Longman)

(١) مـن أمثلـة هـذا النـوع ما يُعمـل عليـه في معجـم الـدوحـة التَّـاريـخِيِّ للغـة العَـرَبِـيَّـة: https://www.dohadictionary.org (تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢).

(2) Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purporses: Towards the Individualziation of Needs Satisfaction. In P. A. Fuertes-Olivera & H. Bergenholtz (Eds.), e-Lexicography: The Internet, Digital Initiatives and Lexicography, (pp. 5479-). London, New York: Continuum.

في الأعـوامِ العشـرةِ الأخيرةِ في بنـاء المعاجـم العامَّـة والأكاديميَّـة والخاصَّـة الـتي تَعكـسُ واقـعَ الاسـتعمالِ اللُّغـويِّ المتنـوّع.

أمـا المجـال الثَّـانِي: (الويـب الـدلالي وتقنياتـه)؛ فيعتمِـد علـى أدواتِ وتقنيـات اتِّحاد شـبكة الويـب العالمِيَّـة (World Wide Web Consortium)، وتُعـرف بالاختصـار الشـائع لهـا: (W3C)[1]، وتُمكّـن تقنيـات هـذا الاتحـاد مـن إنشـاء قواعـد بيانـاتٍ لبنـاء المفـرداتِ، وتصميـم قواعـد بياناتٍ لتلـك المفـرداتِ بصيـغٍ اسـتعلاميَّةٍ وهيـاكل بيانيَّـةٍ حاسـوبيَّةٍ، وتعمـل مجموعـةُ هـذا الاتحـادِ علـى بنـاء مصـادر المفـرداتِ المعجَمِيّـة في الويـب، مـن خـلال صيـغ حاسـوبِيَّةٍ محـدَّدةٍ، تُبـنى عليهـا مصـادر الكَلِمَـات المعجَمِيّـة، مثـل بروتوكـول بنيـة المعجـم (DICT) الـذي يُبـنى بصيغـة لغـة الترمـيز الموسـعةِ (Extensible Markup Language XML)، ومثـل إطـار وصـف المصـادرِ (Resource Description Framework) الـذي يُسـتخدم لوصـفِ البَيَانَـات المعجَمِيّـةِ، ثـم يُدمـج مـع الإطـار المنهـجِيِّ النظـريِّ الـذي يُسَـمَّى إطـارَ الترمـيز المعجَـمِيّ (Lexical Markup Framework). ومـن أدوات مصـادر الويـب الـدلالي وسـحابة البيانـات مـا يُعـرف بـ: ليمـون المعجـم الوجـوديِّ (OntoLex Lemon)، وهـو نمـوذجٌ يعمـلُ علـى حَوْسَـبَةِ مفـرداتِ الويـب؛ بهـدف بنـاء معاجـم أُنطُولُوجِيَّـة، ويَتَضَمَّـنُ هيكلـةً مـن الإدخـالات المعجَمِيَّـة الـتي تَتَكَـوَّنُ مـن بِنيتَـيْن أساسـيَّتين: بنيـة مفـرداتٍ تُبـنى لهـا الشـبكاتُ الدلاليَّـة، وبنيـة مفاهيـمَ تُربـط فيهـا الكلمـاتُ المتعلقـةُ بـكلِّ مفهـومٍ علـى حِـدَة[2].

(١) مـن أمثلـة هـذا النـوع للغـة العَرَبِيَّـة مشـروع (الأنطولوجيـا العَرَبِيَّـة) مـن جامعـة بيرزيـت، وهـو مشـروع لشـبكة الكَلِمَـات العَرَبِيَّـة، وصنفـت الكَلِمَـات فيهـا حسـب المعـاني والعلاقـات الدلاليـة: https://ontology.birzeit.edu (تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢).

(2) Bosque-Gil, J., et al. (2019). The OntoLex Lemon Lexicography Module.

الباب الثاني

تنقيب النُّصُوصِ والبَيَانَات العَرَبِيَّة[١]

(١) يُمكنُ للقارئ تطبيقُ كُلِّ المعالَجَات الآلية الواردةِ في فصول هذا الباب، باستخدام أدوات أوراخ وباستخدام البيانات اللغوية التي أعدّها المؤلف، من أجل تكرار التَّحليلِ والتَّدريبِ والتأكّدِ من تَطابقِ النَّتيجة. ولتسهيلِ العودةِ إلى بَيَاناتِ كلِّ مُعالجةٍ آليةٍ في فصولٍ هذا الباب، فقد جَمعتُها في ملفٍ واحدٍ (إكسل Excel)، وسَمّيتُ كُلَّ ورقة sheet داخل هذا الملف برقم الشكل الوارد في فصول هذا الباب. موقع البيانات اللغوية: www.kaggle.com/datasets/sultanalmujaiwel/quran-mining.

التَّمهيد

مـن الضَّـروريّ التمهيـدُ لمفاهيـمَ أساسِـيَّةٍ في ممارسـاتِ تنقيـب النُّصُـوص والبَيَانَـات والمعالجـة الآليَّـةِ للُّغـة قبـل الوقـوفِ علـى أدواتِ المعالجـةِ المدمَجـة في أداة (أورانج). وكمـا أسـلفنا في مقدمـة الكتـابِ؛ فالهـدفُ مـن شـرح تطبيقـاتِ هـذه الأداة هـو إفـادةُ الباحـث اللُّغَـوِيِّ بمـا امتـازت بـه مـن أدواتٍ مفيـدةٍ في تحليـل النَّـصِّ العَـرَبِيّ ومُعالجَتِـه، وفـق تقنيـاتِ الـذَّكَاء الاصْطِنَاعِـيّ.

وهنـاك هـدفٌ آخـر لا يقـلُّ أهميـةً عـن الأوَّل، وهـو أنَّ الممارسـةَ علـى اسـتخدام تِقنيَـاتِ هـذه الأدواتِ في التَّحليـل والمعالجـة دون برمجـةٍ؛ سـتجعلُ منهـا عونًـا للباحـث اللُّغَـوِيّ في فهـم أُسُـسِ معالجـةِ اللُّغـة العَرَبِيَّـة الطَّبِيعِيّـة، وفي التآلـفِ مـع تطبيقاتهـا في المشـاريع اللُّغَويَّة-التِّقْنيَّـة.

وبمـا أن هـذا الكتـابَ مُوَجَّـهٌ إلى اللُّغَوِيّـين وإلى المهتمـين بتحليـل النُّصُـوص في العلـومِ الإِنْسَـانِيّة والاجْتِمَاعِيّـةِ؛ فبرنامـج (أورانج) بمكتباتِـه المدمجـة فيـه لتنقيـب النُّصُـوص والبَيَانَـات لا يتطَلَّـبُ معرفـةً بَرْمَجيَّـةً؛ كونـه برنامجًـا قائمًـا بذاته[1]. وسـأُبينّ مـا يَتَوفَّـر في هـذه الأداة مـن أدواتٍ تحليلِيَّـةٍ في الفصـل الأَوَّل، وهـي الأدواتُ الـتي سـتُتناول بالتطبيـق في الفصـلِ الثـاني والثالـث والرَّابـع والخامـس. أمـا مـا لا يَتَوفَّـرُ فيهـا مـن أدواتٍ لبعـضِ المعَالَجَـات؛ فتتطلَّـب اسـتخدام لغـة البرمجـة (البايثـون)، وقـد زودتُ القـارئَ بأكوادهـا في الملحق (د) والملحق (و).

(١) لأدوات تنقيـب النُّصُـوص (text mining) وتنقيـب البَيَانَـاتِ (data mining) الـتي سـتُذكر في هـذا البـاب في أدوات (أورانج)؛ ملفـاتُ أكـواد في البايثـون، فلمـن يرغـبُ في اسـتخدامها الرجـوعُ إلى هذيـن الرابطـين: أكـواد تنقيـب النُّصُـوص (Scripting). تاريخ الاسـترداد ٢٤ أكتوبر، ٢٠٢٢، مـن: https://orange3-text.readthedocs.io/en/latest.
أكواد تنقيب البَيَانَات. تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من:
https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/_modules/index.html.

نسمعُ عن ثلاثة مجالات في نطاق تقنيات اللغة[1]؛ وهي تنقيب النُّصُوصِ (text mining)، واللُّغَوِيّات الحاسوبِيَّة (computational linguistics)، ومعالجة اللغة الطَّبِيعِيّةِ (natural language processing).

يعتمدُ تنقيبُ النُّصُوص على جمعِ البَيَانَاتِ اللُّغَوِيّة، وكشطِ الويب، واستخراج المفاهيم، واستخراج المعلومات، وتصنيفِ المستندات، ومعالجة اللغةِ الطَّبِيعِيّة. ويُعدُّ تنقيبُ النُّصُوص من المراحل الأوَّلِيَّة لجمع البَيَانَات اللُّغَوِيّة، وتصنيفِها، واستخراج المعلوماتِ منها، واستبعاد ما ليس مهمًّا فيها في مهمَّةٍ لاحقةٍ، وإعدادها لمعالجاتٍ متعددةٍ في مجال اللُّغَوِيّات الحاسوبِيَّة التَّطْبِيقِيّة ومعالجة اللغة الطَّبِيعِيّة.

ويُعدُّ اللُّغَوِيّات الحَاسُوبِيَّة حقلًا مُتداخلَ الاختصاصات بين اللُّغَوِيّات وحوسبة اللغة والذَّكَاء الاصْطِنَاعِيّ والرياضيات التَّطبيقيَّة وعلم الإدراك، ويهتم بكلِّ ما يتعلَّق بالنمذجة الحاسوبِيَّة للغة، ودراسة أفضل المناويل الحاسوبيَّة للأسئلة اللُّغَوِيّة، وينقسم إلى اللُّغَوِيّات الحَاسُوبِيَّة النَّظرية، واللُّغَوِيّات الحَاسُوبِيَّة التَّطْبِيقِيّةِ، وفي الجانب النظريِّ؛ يُهتم بالنَّظريات الصارمة التي تُعمّق من فهم النَّحو والدلالةِ، أما الجانب التَّطْبِيقِيّ فهو متعلقٌ بتعلم الآلة الذي بدأ باستخدامِ المناهج الإِحْصَائِيَّة، ثم تطوّر باستخدام الشبكات العصبِيَّة التي بدأت تأخذ حيّزالتطبيق الجادِّ منذ عام ٢٠١٥م في تحسين معالجة اللغة الطَّبِيعِيّة. ويصبُّ اختصاصُ اللُّغَوِيّات الحَاسُوبِيَّة التَّطْبِيقِيّة في تحسين أداء فهم الآلة لنماذج مُعَجِّم الكَلِمَات (ردّها إلى الأصل المعجميّ)، ولنماذج وُسُومِ أقسام الكلام،

(1) Miner, G., Elder, J., Fast, A., Hill, Th., Nisbet, R., & Delen, D. (2012). Practical Text Mining and Statistical Analysis for non-Structed Text Data Applications. Elsevier, pp. 3-28.

كل مجال مرتبط بالآخر، وقد أوضحتُ شكل الارتباط توضيحًا مختصرًا، بناء على مسح سبع عشرة دراسةٍ منشورةٍ في دليل أكسفورد:

Mitkov, R. (ed.), (2022). The Oxford Handbook of Computational Linguistics. Oxford University Press, 2nd edition.

انظر أيضا: Tsujii, J. (2021). Natural Language Processing and Computational Linguistics. Computational Linguistics, 47(4).

ولنماذج المحَلّلات النَّحْوِيّة، ولنماذج شبكة الدلالات للكلمات، ولنماذج التَّرجمة الآلِيَّة، وساعدت هذه التَّحسينات على تطوّر تطبيقات معالجة اللغة الطَّبِيعِيّة.

أمَّا مجال معالجة اللغة الطَّبِيعِيّة، فهو نفسُه اللُّغويات الحَاسُوبِيَّة التَّطْبِيقِيّة. وتُستعمل مصطلحات (اللُّغَوِيّات الحَاسُوبِيَّة التَّطْبِيقِيّة)، و(معالجة اللغة الطَّبِيعِيّة)، و(تقنيات اللغة language technologies) لمفهومٍ واحدٍ، ويعود اختلافُ الاستعمال إلى جِهويَّةِ المرجِعِيَّة المؤسَّسِيَّة، فالأكاديميون يستعملون مصطلح (اللُّغَوِيّات الحَاسُوبِيَّة التَّطْبِيقِيّة)، والمبرمجون والمطورون يستعملون مصطلح (معالجة اللغة الطَّبِيعِيّة)، والحكومات والمؤسَّسات السِّياسية تستعمل مصطلح (تقنيات اللغة).

يُعتبرُ مجالُ معالجةِ اللغة الطَّبِيعِيّة مجالًا مُتنَوِّعًا بتنَوُّعِ تطبيقاتِه، وهو مجالٌ يقومُ على التَّجريبِ (praxis) المستمِرِّ في تحسين أداء تعَلُّمِ الآلة لأنظمة اللغة البشرِيَّةِ. سأقدّم لمحة عن بدايةِ معالجة اللغةِ الطَّبِيعِيّةِ، وعما وصلت إليه في الآونة الأخيرة من مخرجات المطَوِّرين والمبَرْمَجيّن في بناء نماذج تعملُ على معالجة اللغة، وتصنيفها، وفهمِها، وتوليدِها.

بدأت فكرةُ معالجة اللغةِ الطَّبِيعِيّة من عَالِمِ التَّشفيرِ آلن تيورنغ (Alan Turing)، عندما أوجد للآلة طريقةً خاصَّةً بها للتَّفكيرِ في عام ١٩٤١م، ليبرهن على فكرة ما إن كانت الآلة تفكّر مثل الإنسان. تطوّرت نماذجُ الترميز التي تُساعد الآلة على فهم أقيسة التَّفكير المنطقِيِّ عند الإنسانِ، ووصلت في مطلعِ القرن الحادي والعشرين إلى مرحلةٍ عميقةٍ من فهم الآلةِ، ومعالجتِها، وتوليدِها لما فهمت. قامت معالجةُ اللغة الطَّبِيعِيّة على ثلاثة مناهج متتابعَة زمنِيًّا؛ المنهج الأوَّل: تعلُّم القواعدِ (rule-based method). والثَّانِي: تعلّم الآلةِ (machine learning). والثَّالِث: التعلّم العميق (deep learning).

قام منهج تعَلُّمِ القواعد على محدّداتٍ قطعِيَّةٍ من القواعدِ، التي تُبنى لتعليم الآلةِ عليها، وتَتَضَمَّن جملة من التَّعبيرات الشَّرطِيَّة (إن if، إذن then)، فيقوم النظام بالتخزين

والتَّطبيق بناء عليها، وطُبِّق هذا المنهج مثلًا على المحلِّل الصرفِيِّ؛ كون شروط التَّحليلِ والتَّوليدِ محدودةً.

تغيَّر التعلّم على القواعد إلى تعَلُّم الآلة الذي قام على الاحتماليةِ (probabilistic)، فأنظمة تعلُّم الآلة ونماذجها التقليدِيَّة قد صارت تعتمدُ على قواعد إحْصَائِيّةٍ بدلًا من القواعد الحتميّة، وهناك فروقاتٌ أساسِيّةٌ بين أنواع تعلّم الآلة[1]، وهي على النَّحو الآتي:

- **تعلم الآلة التَّقليدي**

بدأ هذا النوع في خمسينيَّات القرن العشرين في علم الاحتمالات الإِحْصَائِيَّة البحتة، ويقوم التَّعلُّم هنا على البحثِ عن الأنماط في البَيَانَاتِ، وتقييم نقاطِ القرب فيها، وحساب فضاءِ المتَّجهاتِ، الذي يعرف بالشُّعاع أو المصفوفة (array or matrix). معظم الخوارزميَّات اليومَ تعملُ على هذا النّوعِ من التَّعلُّم، فاقتراح مُحرِّك البحث جوجل لكلماتٍ بحثِيَّةٍ، أو اقتراح منصة اليوتيوب لمقاطع فيديو أخرى، أو اقتراح منصة تسوق إلكتروني لأصنافٍ معَيَّنةٍ؛ ما هو إلا نتيجة لكشفِ أنماط سلوك المستخدم في تلك المواقعِ، وستتَّضح نقطةُ الأنماط والنّقاط والمتَّجهات من خلال شرح نوعين من أنواع تعلُّم الآلة فيما يتعلَّق باللغة؛ وهما: التَّعلُّم الموجَّه، والتَّعلُّم غير الموَجَّه.

في التعلم الموجَّه؛ تُزوَّد الآلةُ بالإجابات، فمثلًا؛ يعمل المحَلّل البشريُّ على تقطيع الكَلِمَات في النَّصِّ، ثم تصنيف أقسام الكلام فيها، ثم تتعلَّم الآلة من هذه الأمثلة، ثم تُختبر على بيانات لم ترَها ليُقيّم بها أداءُ فهمِهَا. ينقسمُ تعَلُّم الآلةِ الموجَّه إلى التَّصنيف والانحدارِ، وفي التَّصنيف تكون السِّماتُ (features) فئات (مثل: اسم وفعل وضمير وأداة إلخ)، أما في الانحدار فتكون السِّمات أرقامًا (مثل التّكْرَارات الحقيقيَّة أو التكرارات النسبيَّة).

---

(١) لتفاصيل أنواع التعلم وخوارزمياتها العودة إلى: لحلح، محمد. (٢٠٢٠). مدخل إلى الذَّكَاء الاصْطِنَاعِيّ وتعلم الآلة. أكاديمية حاسوب، المملكة المتحدة، الطبعة الأولى. الشروحات في هذا الكتاب مختصرة ومبسطة ومقروءة ومفهومة.

في التَّعلُّمِ غيرِ الموجَّه الذي جاء متأخِّرًا في نهاية التسعينيات من القرن العشرين؛ دعت حاجةُ التعامل مع بياناتٍ ضخمةٍ يصعبُ تصنيفُ خصائصها وسماتها بشريًا بسبب الكلفة البشرية والزمنية إلى بناء نماذج خاصةٍ لها. والمهم في هذا النوع من التَّعلُّمِ هو أن نُعطي الآلة بياناتٍ منظمةً على الأقل، ومهيكلةً بصيغةٍ منطقِيَّةٍ من الناحية الشكليَّة، لتعمل على معالجتِها -على الوجه الأفضل- بالتَّجْمِيعِ (clustering)، أو بتقليل الأبعادِ (dimensionality reduction)، أو بالتعميم (generalization)، أو بتعلم قواعد الربطِ (association rule learning). خذ على سبيل المثالِ: استخراج الموضوعات من النُّصُوص، إذْ تحتاجُ هذه النُّصُوص إلى تنظيمٍ يُعين على تحقيقِ نتيجةٍ جيدةٍ، ومن أشكال التَّنظيمِ مثلًا اختيار أو توزيع النُّصُوص حسب مجالاتِها أو حسب أزمنتِها، فلو كان النَّصُّ مثلًا يَتَضَمَّن مئات الآلافِ من النُّصُوصِ؛ فمن المهِمِّ أن نفصِّل مجالاتها المختلفة لتوضيح كُلّ نَصٍّ، أو أن نعيد صياغةَ توزيعها وفق الزَّمنِ، أو وفق الكُتّاب، أو وفق السياقِ الذي يُمكننا الاعتماد عليهِ لتحقيق غرض التنظيم. وبحسب جودة البَيَانَات النَّصِّيَّة، سنلحظ أنَّ استخراجَ الآلة للموضوعاتِ ذو مغزى، وتُمكّنها هذه الجودة من تحديد الموضوعات، فتعمل على فصل النُّصُوص أو المستنداتِ بناء على كُلّ موضوع، ثم على تحديدِ الكَلِمَات ذات العلاقة بكلِّ موضوعٍ. فكيف لها أن تقوم بذلك؟ (الفصل الثالث).

- **التَّعلم المعَزّز**

يقومُ على ترك الآلة تفَكِّرُ بنفسها، فتتدرب على تعزيز ذاتها بفهم المحيط والسلوك، ونجد هذا النوع من التعلُّم في التَّطْبِيقَات العملِيَّة المستخدمة في السَّياراتِ ذاتِيَّة القيادة، وفي روبوتات تنظيف الأرضيات، وفي الألعابِ، وفي أتمتة التَّداول المالِيِّ، وأيضًا في بناء تطبيقٍ شعرِيٍّ يقوم على السِّجال والردِّ والنقائض، ويعتمدُ هذا التَّعلم على معيار المكافأةِ (award)، أو العقوبةِ (penalty)، فكلما تحسَّن التعلُّم ارتفعت المكافأة، وكلما قلَّ التعلُّم ارتفعت العقوبة.

الشَّبكات العصبِيَّة والتعلُّم العميق

تَتَضَمَّن الشبكاتُ العصبيَّةُ مجموعةً من الخلايا العصبِيَّة الاصْطِنَاعِيّة ذات الطبقات، وتعمل الطبقة الأولى (الدخل) على فهم عناصر وسمات ما تريد تعلّمه، ثم تنتقل إلى (الطبقة الخفيَّة)، التي تحوّل تلك العناصرَ إلى أوزانٍ رَقَمِيّة، ثم تنتقلُ إلى الطَّبقة الأخيرة (الخرج)، التي تعدُّ نهايةَ ما تم تعلُّمه. يُحدّد المبرمج عدد الطبقاتِ (layers)، وعدد العصبوناتِ (neurons) من بداية التعَلُّم، أما التعلُّم العميق فهو عبارةٌ عن شبكِيَّةٍ عصبيةٍ اصطناعِيَّةٍ كُبرى، ويَتَضَمَّن بنياتٍ عديدةً من الشَّبكات العصبية[1]، ففي نماذج التعلم العميقِ؛ تُحدّد بنية شبكة التَّعلم العميق، ولكلِّ شبكةٍ وظيفةٌ تعلّمِيَّةٌ.

إذا كان هناك مئاتُ الآلاف من المدخلات في طبقة الدَّخلِ؛ فستتضاعفُ أوزانها في الطبقة الخفيَّة بأرقامٍ كبيرةٍ جدًّا؛ نتيجة عملياتِ الضَّربِ، فكلُّ طبقةِ دخل ستحتوي على مُتَغَيّرَات x1, x2, x3....، وفي الطبقة الخفيَّة سيتعامل كُلّ عصبون مع مُتَغَيّرواحدٍ، وينتج من هذا التَّعامل التعرّفي لكلِّ مُتَغَيّرٍقيمةٌ كبيرةٌ، نتيجة ضرب x1 على الوزن w1، ثم تُؤَتْمَت هذه الطَّريقة على بقِيَّة المُتَغَيّرَات، وصولًا إلى طبقة الخرج y، الذي سيَتَضَمَّن قيمةً ناتجةً من عملية x1w1 + x2w2 + x3w3 +.... × b. ولتفادي هذه القيم الضخمةِ؛ يُستعملُ ما يُسمَّى بنماذج التَّنشيطِ (activation functions)، وتُعرف بالـ (RELU)، و(Sigmoid)، و(Softmax)، و(Tanh)، التي تمارسُ دورَ تقليص الأوزانِ، فيجعل النموذج الأول الأوزان بين (0) و(1)، ويجعل النموذج الثاني الأوزان بين (0) و(1) عندما يكون المُتَغَيّرذا صنفٍ واحدٍ، ويجعل النَّموذج الثالث الأوزانَ بين (0) و(1) عندما يكون المُتَغَيّر متعدِّدَ الأصناف، ويجعلُ النموذج الرابع الأوزانَ بين (-1) و(1). يَصِلُ التَّعلُّم في المخرجات إلى أنَّ الآلة قد تعرّفت إلى الأشياء، فصارت تنظرُ وتنطقُ وتسمعُ وتفهمُ مثل الإنسانِ، بعد أن تعلَّمت كُلَّ السماتِ لتلك الأشياء عن طريقِ القيمِ الرَّقَمِيّةِ، وكلُّ هذه الدَّوال

(١) لحلح، محمد. مدخل إلى الذَّكاء الاصْطِنَاعِيّ وتعلم الآلة. أكاديمية حاسوب، المملكة المتحدة، الطبعة الأولى، ٢٠٢٠، صفحة ١٠٨.

التَّنشيطِيَّة أساسِيَّةٌ في تحسين نماذج التعلُّمِ العميق، وتُقيَّم باستخراج نسبة الخطأ عن طريقِ ما يُعرف بدالّة الخسارةِ (Loss Function)، وتُحسَّن عن طريقِ خوارزميات عديدةٍ كـ (Gradient Descent)، و(MiniBatch)، و(ADAM).

يكون هذا التعَلُّم على مستوى اللغةِ النَّصِّيَّة، أما على مستوى مقاطع الفيديو؛ فتُدمج الشبكات العصبِيَّة التلافيفية (Convolutional Neural Networks CNN)، التي تُستخدم لفهمِ الصور ومقاطع الفيديو مع الشَّبكات العصبِيَّة المتكَرّرة (Recurrent Neural Networks RNN)، التي تُستخدم للتعرف إلى الكلامِ في تلك المقاطع. وينتجُ عن ذلك تزييفٌ لمقاطع يتحدثُ فيها أشخاصٌ، وهم في الحقيقة ليسوا هم أنفسهم. تخيّل هنا الحجم الهائل من العملِيَّات العميقة التي بها تتعلم الآلة جزئيات عناصر الأشياءِ؛ لتكتسب معرفةَ ماهيتِها وصورتِها، سواءً كانت صورةً (رؤية الحاسب)، أو كانت كلامًا شفهِيًّا (التعرُّف إلى الكلام)، أو كانت نَصًّا مترجمًا منه، ونصًّا آخر مترجمًا إليه (التَّرجمة الآلية)[1].

وَبالنسبة لمعالجة اللغة الطَّبِيعِيّة التَّطْبِيقِيّةِ؛ فسأقفُ على مرحلتين متتابعتين فيها؛ الأولى: المعالجة الأساسيَّة التي تُعرف باسمِ (CoreNLP)، وهي منطلقُ المعالجة الآلِيَّة للغة الطَّبِيعِيّة، والثَّانية: تطبيقات معالجة اللغة الطَّبِيعِيّة، وهي نِتاجُ المعالجة، وما تُفيد به من مهامّ متعددة.

(١) حول المداخل النظرية لمستويات معالجة اللغة الطَّبِيعِيّة أولاً، والتعلم العميق وتطبيقاته في معالجة اللغة الطَّبِيعِيّة ثانيًا، والترجمة الآلية ثالثًا؛ انظر إلى المراجع الآتية:
الصانع، وليد. طرق ومستويات معالجة اللغة في الذَّكَاء الاصْطِنَاعِيّ. في: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل النَّصّ العَرَبِيّ، ٢٠١٩، الصفحات ١٥-٤٤.
القنيعير، فارس. التعلم العميق وتطبيقاته في معالجة اللغة. في: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل النَّصّ العَرَبِيّ،٢٠١٩، الصفحات ٤٤-٦٨.
الراجح، عبدالله. الترجمة الآلية. في: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل النَّصّ العَرَبِيّ، ٢٠١٩، الصفحات ٦٩-٩٣.

أَوَّلًا: معالجة اللغة الطَّبِيعِيّة (المعالجة الأساسِيَّة)

تعتمد كُلّ المعالجات الأساسِيَّة؛ سواءً التَّابعة لستانفورد أَم غيرها مما تضافرت معًا لأغراض التَّحسين، على قياس أداء أنظمتِها الآليَّة. يُقاس أداء هذه الأنظمة بمقاييس مصفوفة الارتباك (confusion matrix): الصِّحَّة (accuracy)، والدِّقَّة (precision)، والاسترجاع (recall) ومقياس-ف (F-score). تُستخدم هذه المقاييس لمعرفة دقة أداء النِّظام الآلي أو النموذج في بيانات الاختبار (أي: بيانات لم يَرَها النظام بعد التَّدريب) في مَهَامّ مُعَالجةِ اللغة الطَّبيعِيَّةِ المتعَلِّقة بالتَّصنيف. تعملُ مصفوفةُ الارتباك على اختبارِ نماذج تعلم الآلةِ ونماذج التَّعلم العميق، باستخراج نسبة الصِّحةِ، والدِّقة، والاسترجاعِ، ومقياس-ف التي يتم حسابها وفق معادلة كُلّ مقياس: الصحة $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$، والدقة $\left(\frac{TP}{TP+FP}\right)$، والاسترجاع $\left(\frac{TP}{TP+FN}\right)$، ومقياس-ف $\left(2X\frac{\text{الدقة} \times \text{الاسترجاع}}{\text{الدقة} + \text{الاسترجاع}}\right)$، حيث يشير (TP) إلى الإيجابيَّاتِ الصحيحة true positives و(TN) إلى السلبِيَّاتِ الصحيحة true negatives و(FP) إلى الإيجابيَّاتِ الخطأ false positives و(FN) إلى السلبِيَّاتِ الخطأ false negatives. على سبيل المثال، لو كان لدينا مئةُ كلمة، ثُمَّ وسَّمنا يدويًّا ٥٠ كلمةً منها على أنها أسماء؛ لأنها في الحقيقة أسماء، ثم وسَّمنا يدويًّا الخمسين الأخرى على أنها أفعالٌ؛ لأنها في الحقيقة أفعال، ثم درّبنا نموذجَ النِّظَامِ الآليِّ على بيانات الاختبار، وذلك بعد تطبيق أشهر تقنية من تقنيات تقسيم البيانات (بيانات تدريب 80%- بيانات اختبار 20%)، ثم اختبرنا أداء فهم النموذج بعد التدريب على بيانات الاختبار، ستكون مُحصّلة الكلمات في بيانات الاختبار ٢٠ كلمة، أي: ٢٠٪ من الكلمات المئة (١٠ أسماء و١٠ أفعال). إذن، يتكَوَّنُ جدولُ مصفوفة الارتباكِ من القيمِ الحقيقِيَّةِ للتَّصنيف الحقيقي ومن القيم التي تَنَبَّأ نموذج النظام الآليّ بها في نتيجةِ تصنيفه لعشرة أسماء وعشرة أفعالٍ في بيانات الاختبار (الجدول ٢٦). وباستخدام معادلات مقاييس الارتباك الأربعة المذكورة سابقًا، ستكون نسبة الصحة $\left(\frac{8+6}{8+6+4+2}\right)$: 0.7، وبمعدّل خطأ (0.3)، وستكون نسبة الدقة في فئة الأسماء $\left(\frac{6}{8+4}\right)$: 0.67، وستكون نسبة الاسترجاع في فئة الأسماء

$\left(\frac{6}{8+2}\right)$: 0.8، وستكون نسبة مقياس-ف $\left(2X\frac{0.536}{1.47}\right)$: 0.67. ويُعطينا مقياس-ف مُتوسِّطَ التوافق لنسبةِ الدِّقَّة والاسترجاع. أما نسبة الدِّقة والاسترجاع للتنبؤ بفئة الأفعال فهي (0.75) و(0.6)، ونسبة مقياس-ف: (0.67). ولا يُكتفى بنسبة الصحة؛ لأنَّ نسبةَ الدِّقةِ والاسترجاعِ مُهمّةٌ لكشف جودة أداء تصنيف النِّظام الآليّ في بَيَاناتِ الاختبار. ومهما كَثُرتْ حالاتُ الأصناف والفئات، فخوارزمياتُ الآلة في قياس أداء النَّموذج الآليّ في التَّصنيف تعمل آليًّا على استخراجِ قيمةٍ واحدةٍ لكلِّ مقياسٍ من مقاييس مصفوفة الارتباك.

تُفيدُ الدّقة بنسبة حالات الإيجابيَّات الخطأ (FP) إلى حالات الإيجابيّات الصحيحة (TP). المثال على ذلك: أسماء صنَّفها النّظام الآلي في بيانات الاختبار على أنَّها أفعالٌ (التصنيف: خطأ False، والفئةُ الحقيقيَّةُ ظهرت سلبيّةً [أفعال] negative لكنَّ النموذج تنبّأ على أنها إيجابيَّة [اسم] positive). أمَّا الاسترجاع فيُفيدُ بنسبة حالات السلبيَّاتِ الخطأ (FN) إلى حالات الإيجابيَّات الصحيحة (TP). والمثال على ذلك: أسماء صنَّفها النظامُ الآليُّ في بيانات الاختبار على أنَّها أفعالٌ (التصنيف: خطأ False، والفئة الحقيقة ظهرت إيجابيّة [أسماء] positive لكنَّ النموذج تَنبَّأَ على أنها سلبيّة [أفعال] negative).

جدول ٢٦ : جدول مصفوفة الارتباك للقيم الحقيقِيَّة والقيم التنَبُّؤِيَّة

		التَّصنيف التَّنَبُؤيّ (نتيجة تصنيف نَموذج النّظام الآليّ)		
		أداء النموذج لتصنيف الأسماء	أداء النموذج لتصنيف الأفعال	الاسترجاع
التَّصنيف الحقيقيّ	عشرة أسماء	(PT) 8	(NF) 2	10 أسماء تَنَبَّا منها بـ 8 صحيحة (0.8)
	عشرة أفعال	(PF) 4	(NT) 6	10 أفعال تَنَبَّا منها بـ 6 صحيحة (0.6)
	الدِّقَّة	12 اسمًا، تنبَّأ بـ 8 أسماء صحيحة (0.67)	8 أفعال، تنبَّأ بـ 6 أفعال صحيحة (0.75)	

تقـومُ المعالجـةُ الآليَّـة الأساسِـيَّة للغـة الطَّبِيعِيّـة علـى مَهَـامٍّ أوّليّـة، وهـي مهـام ضروريَّـةٌ، تُمكّـن المعالجـة مـن فهـم اللغـة، وتُسـمّى هـذه المعالجـة الأساسـيَّة بالمعالجـة القبلِيَّـة (preprocessing)، ويُعتمـد علـى معالجـات سـتانفورد، الـتي تتطـوَّر في بنائهـا منـذ بدايـة تصميمِهـا في عـام ٢٠١٠، وصـدور النّسـخة الأولى 1.0، وصـولًا إلى آخـر إصـدار 4.4.0 في ٢٠ ينايـر ٢٠٢٢م. وتعمـل معالجـاتُ سـتانفورد علـى بيئـاتٍ بَرْمَجيَّـة متعـددةٍ؛ مثـل البايثـون، والآر، والجافـا، وجافـا سـكريبت، وبـيرل، وسـكالا، وغيرهـا، كمـا أنهـا تغطـي اللغـات الأساسـية، ومـن ضمنهـا اللغـة العَرَبِيَّـة[1]. ويُعتمـد كثـيًرا علـى (البايثـون) في تعلُّـم الآلـة للنصـوص. إذا اخـترت غـير البايثـون؛ فلـن تجـد كُلّ المكتبـات المهمـة لمعالجـة اللغـة، ولنُركّـز هنـا علـى التوكيـد بـ (كلّ). تُعتـبر لغـة (البايثـون) مـن أفضـل لغـات البرمجـة لمعالجـة اللغـة الطَّبِيعِيّـةِ؛ نظَـرًا لتوفـر مكتبـات عـدّة للمعالجـة، ومـن أهمهـا[2]: مكتبـة (Spacy)، ومكتبـة (nltk)، ومكتبـة (AllenNLP)، ومكتبـات المعالجـة الأساسـيَّة لسـتانفورد، ومكتبـة المحـوّلات (Huggingface)، ومكتبـة جينسـم (Gensim) للتضمينـاتِ (Embeddings)، ومكتبـة (Rasa) لفهـم اللغـة الطَّبِيعِيّـة (Natural Language Understanding) ولروبوتـات المحادثـةِ (chatbots).

---

(١) يُمكـن اسـتعمال معالجـة اللغـة الطَّبِيعِيّـة الأساسـية لسـتانفورد في سـبع عشـرة لغـة بَرْمَجيَّـة. انظـر إلى الرابـط الآتي:

Stanford NLP Group. (2020). Using CoreNLP with other programming languages and packages. Retrieved October 24, 2022, from https://stanfordnlp.github.io/CoreNLP/other-languages.html.

(٢) هذه مقدمة مفيدة في معالجة اللغة الطَّبِيعِيّة بلغة البايثون:

Duque, Tiago. (2020). Peeling Natural Language Processing Layer by Layer: An Introductory Overview. In Medium [web source]. Retrieved October 24, 2022, from https://tfduque.medium.com/dissecting-natural-language-processing-layer-by-layer-an-introductory-overview.

لتعلـم (البايثون) من الصـفـر ففـي الموقـع الآتـية مقدمـة أسـاسـيـة مـختصـرة وسـهلـة:

https://www.kaggle.com/learn/python (تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢).

ممَّ تتكون معالجات ستانفورد للغة الطَّبِيعِيّة؟[1] سأُشِيرُ في الآتي إلى مكونات معالجة اللغة الطَّبِيعِيّة التي تُعرف بـ (Stanford CoreNLP). بعض ما سيُشار إليه ليسَ من ضمن حزمة معالجة اللغة العَرَبِيَّة الطَّبِيعِيّة الأساسِيَّة لستانفورد، كالمحلّل الصوتي، والدّلالي، والمقامِيّ، وما تتكوَّن منه هذه الحزمة هو: المقطّع (segmenter)، والموَسِّم النحويّ (postagger)، والمحلّل النحويّ (parser)، والتعرف على أسماء الكياناتِ (Named–Entity Recognition). وفيما يلي عرضٌ تفصيليّ لمكوِّنات المعالجة الآلية للُّغات الطَّبيعيَّة.

المكون الأوَّل: الترميز الذي يُستخدم لتقسيم وحدات النَّصّ إلى وحداتٍ صغرى. تتفرَّع وحداتُ النَّصّ من أصغر وحدةٍ متمثّلةٍ من الوحدة الصَّوتيَّة، إلى أكبر وحدةٍ متمثلةٍ من الوحدة التَّداوليّة. والسؤال الذي يترتَّبُ طرحُه هو: ما الوحدة اللُّغوية التي يُمكنُ معالَجَتُها آليًّا؟ من خلال النَّظَرِ في المحلِّلات اللُّغويَّة الآليَّة؛ سنجدُ أنَّ التَّطْبِيقَات التي تُصمَّمُ لها منقسمةٌ إلى ما هو آت:

- مُحَلِّلٌ صرفِيٌّ ونحويٌّ، ونجد أنَّهما في بعض الأنظمةِ يُعاملان على مُستوى واحدٍ، وفي بعضها يُعاملان على أنَّهما مستويان منفصِلان. يعمل المحلِّلُ الصرفِيُّ على معالجة الكلمة بالتَّوليد والتَّحليلِ، فبالتَّوليد يستخرجُ المحَلّل كُلّ اشتقاقاتِ الكلمة، وبالتَّحليل يستخرجُ المحلِّل جذرَ المشتَقَّات. مثل هذا العملِ يكون لمحللاتٍ صرفيَّةٍ منفصلةٍ عن التَّحليل أو التَّوسيم النَّحْوِيّ. أمَّا ما يكون مدمجًا

(١) لا تندهش من كثرة أدوات المعالجة الأساسية، أو من كثرة مكتباتها، أو من كثرة نماذجها التي ستجد منها ما قد ظهر بمنهج القواعد، وما قد ظهر بمنهج تعَلُّم الآلة، وما قد ظهر بمنهج التَّعلم العميق. تأكّد من تَتَبُّعِ الجديد (novelty) في كُلّ مرحلة يُعلن فيها عن أداء أعلى للنماذج، وتُتاح للاستخدام. كثرةُ النماذج الآليةِ المطوَّرةِ للُّغةِ هي العامل الرئيس الذي يُحقّقُ التَّطوير المستمرِّ في البناء والتحسين، والكثرة هنا مع عواملِ التحسين سمةٌ ملتصقةٌ بكلِّ ما يتعلق بتطور التقنياتِ اللغويّة. تجنّب في البحث العِلْمِيّ وفي العلوم اللُّغَوِيّة والإنْسَانِيّة والاجْتِمَاعِيّة النماذج القديمة ذات الأداء الأقل، واكتشف النماذج الجديدة ذات الأداء الأعلى، والمنشورة مع بياناتها اللُّغَوِيّة.

بينهما فنجده في بعضِ الأنظمة[1] التي تعملُ على استخراج السِّمات المورفيميّة (morphemes) من الكَلِمَاتِ، ثم تعمل على التَّفريق ثم التَّوسيم النَّحْوِيّ حسب حالتها النَّحْوِيّة (اسم، فعل، ضمير).

- مُحلّلات للمعالجة الصَّوتيَّة والدِّلاليَّة والتَّداوليّة لبعض الأنظمة الآليّة للُّغات الطَّبيعيَّة، كالمحلِّل الصوتِيّ الذي يعملُ على تحليل الوحداتِ الصَّوتيَّة (الفونيمية) للكلمة، ثم يعملُ على تحويل الكَلِمَات إلى توصيفٍ صوتِيٍّ خاصٍّ بها. يليه المحَلّل الدلالي الذي يعملُ على بناء العلاقات الدّلالِيَّة (الوجوديَّة ontological) بين الكَلِمَاتِ كعلاقة التَّرادف[2]، ثُمَّ المحلِّل المقامّي الذي يَعمَلُ على استكشاف المعاني السياقِيَّة، وما يعتَوِرُ هذه المعاني من نموٍّ، أو تغَيُّرٍ، أو تخصيصٍ، بناء على التَّأثير الفرديّ-النفسيّ أو البيئيِّ أو الاجْتِمَاعِيّ أو الثَّقافِيِّ على اللّغة.

المكون الثَّاني: التَّفريق (tokenization)، والتَّقْطِيع (segmentation)، والتَّعجيم (lemmatization)، فالتَّفريق يعملُ بوظيفة الفاصل بين المكَوّنات النَّصِّيَّة التي تتطَلَّبُ الفصل والتَّسوية

---

(١) (كلمة-ستار): مولّد ومحَلِّلٌ صرفِيٌّ للغة العَرَبِيَّة، ويعمل على استخراجِ السِّمات المورفيميّة القائمة على الصِّيغة والوظيفةِ، بالإضافة إلى التفريق بين الوحدات المورفيميَّة ووحدات الصِّيغةِ. وللمولّد الصرفِيِّ محرِّكٌ متينٌ، يُمكن دمجه في أنظمةٍ عَديدةٍ، بالإضافة إلى واجهةٍ شبكِيَّةٍ، وواجهة برمجة التَّطْبِيقَات API:

Taji, D., Khalifa, S., Obeid, O., Eryani, F., Habash, N. (2018). An Arabic Morphological Analyzer and Generator with Copious Features. Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, October (pp. 140150-). Brussels (Belgium): Association of Computational Linguistics.

(٢) من المشاريع التي ظهرت بالصورة المطلوبة لبناء محَلِّلٍ دلالِيٍّ للغة العَرَبِيَّة هو مشروع (الأنطولوجيا العَرَبِيَّة) من جامعة بيرزيت، وهو مشروعٌ لشبكة الكَلِمَات العَرَبِيَّة، وصنفت الكَلِمَات فيها حسب المعاني والعلاقاتِ الدّلالية. تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢، من: https://ontology.birzeit.edu. وانظر أيضًا:
Jarrar, M. (2021). The Arabic Ontology – An Arabic Wordnet with Ontologically Clean Content. Applied Ontology 16(1), 1-26.

الهجائيَّـةِ؛ مثـل فصـل علامـات التَّرقيـمِ، والأرقـامِ، والرمـوزِ الـتي تتطَلَّـب أن تكـون في حَـيِّزٍ خـاصٍّ بهـا بمسـافةٍ بينهـا وبـين الكلمـة، أو فصـل العبـارات أو الجمـل إن كان التَّحليـل سـيُطَبَّق علـى مسـتوى العبـاراتِ أو الجمـل، وتتحـدَّدُ مهمّـة التَّفريـق بتحديـد مهمَّـة المعالجـةِ. أمـا التَّقْطِيـع فيعمـلُ علـى فصـل المبـاني النَّحْوِيّـة والصرفِيَّـةِ؛ والمبـنى النَّحْـوِيّ مثـل: حـروف العطـف (و، فـ، ثـم، بـل)؛ لأن لمبانيهـا حالـةً نَحْوِيّـةً خاصَّـةً بهـا، والمبـنى الصـرفِيّ مثـل: (أل) التعريـف؛ لأنَّ للتَّعريـف والتَّنكيـر حالَـةً صرفِيَّـةً خاصَّـةً بهـا. هنـاك نمـاذج أو أنظمـة للتَّوسـيم النَّحْـوِيّ تعمـل علـى فصـل المبـاني الصرفيَّـة، ولا أجـد ذلـك مفيـدًا؛ لأن المهـمَّ هـو المبـنى النَّحْـوِيّ، أمـا المبـنى الصـرفِيُّ فهـو متعلّـقٌ بالكلمـة، وقـد صيغـت الكلمـةُ منـهُ، وبنـاءً عليـه، وفصلُـه خطـأ؛ لأنَّنـا بفصلـه نكـونُ قـد نزعنـا مـن الكلمـة المعالَجـة حالـة التَّصريـف الخاصَّـة بهـا، والتَّقْطِيـع مهـمٌّ علـى مسـتوى المبـاني النَّحْوِيَّـة؛ مـن أجـل دقـة المعجّـم، والمعجِّـم يعـني أن تُـردّ كُلُّ كلمـةٍ في النَّـصّ إلى مادَّتِهـا المعجَمِيَّـة (وليـس الجذريَّـة)، فمثَـلًا؛ لـو عجّمنـا كلمـةَ (تُطعمـون)، فالدَّقيـق أن تكـون المـادة المعجَمِيّـة (أَطعَـمَ)، قـد يتسـاءل أحـدٌ ويقـول لمـاذا لـم تكـن المـادة المعجَمِيّـة هـي (تُطعـم)، والجـوابُ: أنَّ أصـل المـادة المعجَمِيّـة للأفعـال هـي المـاضي، وتحـت هـذه المـادة تـردُ كلُّ صيـغ الفعـل المضـارع الأربعَـة الـتي تُبتَـدَأ بأحـد حـروف المضارعـةِ (أنيـت)، ولا يصلـحُ أن نضـع مـادَّةً معجمِيَّـةً مسـتقِلَّةً للفعـل المضـارعِ (تُطعـم). كذلـك الحـالُ مثَـلًا مـع كلمـةِ (أسـتاذة)، فمادَّتهـا المعجَمِيّـة هـي (أسـتاذ)، فـتُرد صيـغ المؤنَّـث إلى الأصـل المذَكَّـر.

المكـوِّن الثَّالِـث: التَّوسـيم النَّحْـوِيّ، وبـه يعمـلُ النِّظـامُ علـى توسـيم النَّـصِّ كلمـةً كلمـةً، بعـد تهيئتهـا بالتَّفريـق والتَّقْطِيـع، وتختلـف مجموعـةُ التَّوسـيم مـن حيـثُ الوصـف وعـدد الواسـمات النَّحْوِيّـة مـن نظـامٍ إلى نظـامٍ. في نظـام سـتانفورد ثلاثـة وثلاثـون وسـمًا نَحْوِيًّـا، أُبَيِّنهـا في الجـدول (٢٧).

جدول ٢٧ : وسوم ستانفورد النَّحْوِيّة للغة العَرَبِيَّة

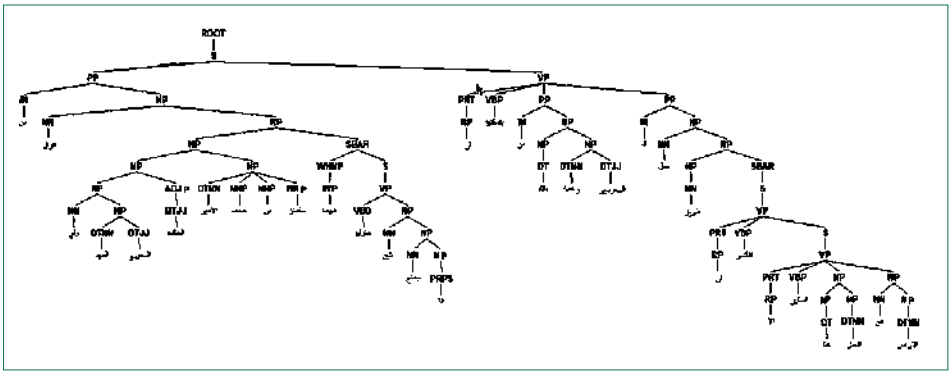
	الواسم	التعريف به	أمثلته
1	FW	كلمات أجنبية	كلمات أجنبية
2	CC	أدوات الربط	و، ف، أو، ثم، بل، لكن، أم، كما، لا، أما، كيما
3	RB	الظروف	هناك، ثمة، هنا
4	WRB	أدوات الاستفهام	كيف، لماذا، أين، متى
5	CD	الأعداد	الأعداد
6	DT	أسماء الإشارة	هذا و أولئك، تلك
7	PRP	الضمائر المتصلة والمنفصلة	هو، ي هـ، نا، كم، نحن
8	$PRP	الضمائر الملكية المتصلة	ها، هم، نا، كم، كن
9	WP	الأسماء الموصولة	التي، الذي، الذين
10	IN	حروف الجر	ل، ب، في، من، عن، إلى
11	RP	الأدوات	لا، قد، لم، س، هل، يا، لقد
12	UH	التعابير العاطفية	نعم، اللهم، كلا، أجل
13	PUNC	علامات الترقيم	علامات الترقيم
14	VBG	المصدر العامل	قول، اعتبار، منح
15	VBD	الفعل التام	قال، صلى، سلّم
16	VBN	الفعل المبني للمجهول	قيل، يقال، ولد
17	VBP	الفعل المضارع	يكون، يسكن، يقول، يجب
18	VB	فعل الأمر	انظر، قم، أضف، خذ
19	VN	أسماء تشبه الأفعال	مشريا، مؤكدا، مضيفا، موضحا
20	NN	أسماء شائعة	سلام، كلام، خلال

	الواسم	التعريف به	أمثلته
21	NNS	المثنى جمع المؤنث السالم	سنتان/سنت رن ي، خدمات
22	DTNN	اسم معرف مفرد	الناس، العمل، اليوم
23	DTNNS	اسم معرف مثنى أو جمع	المسلمين، المعلومات، الولايات
24	NOUN	أسماء التوكيد	جميع، نصف، نفس
25	NNP	اسم علم مفرد	محمد، أحمد، على
26	NNPS	اسم علم جمع	محمدون، مؤمنون، منظرين
27	DTNNP	اسم علم معرف مفرد	الإن بينت، السعودية، المغرب، القاهرة
28	DTNNPS	ألفاظ الجموع	الروحانيات، ال ريازيلي رن ي، القاصرات
29	JJ	الصفات	آخر، خاصة، واحدة، كبيرة
30	ADJ	الصفات العددية	الأول، الثاني، الثالث
31	DTJJ	الصفات المعرفة	العربية، العامة، السياسية، الوطنية
32	DTJJR	الصفات المعرفة للمقارنة	الاقل، الأوسط، الأكبر
33	JJR	صفات المقارنة	أفضل، أكبر، أقل، أعظم

المكوِّن الرَّابع: التَّحليل النَّحْوِيّ[1]، ويعملُ على تحديد العلاقةِ بين كلمةٍ وكلمةٍ أُخرى في الجملة، ويُبنى بتصميم نماذج شَجَريّةِ الاعتمادات (dependencies) للمكوِّنات النحوّية والتركيبيَّة في الجمل؛ لتمثيل شجرةِ الاعتماد والعلائق النَّحْوِيّة. ومن أمثلة ما يقومُ به

(١) سيُلحظ أن بعض الأنظمة تدمج بين المحلّلات الصَّرفية والنَّحْوِيّة، وبعضها تدمج بين المحلّلات النَّحْوِيّة والتَّركيبيَّة، وبعضها تنفرد بنظامٍ واحدٍ، إما للتَّحليل الصَّرفي، أو للتَّوسيم النَّحْوِيّ، أو للتَّحليل التركيبيّ، وأحيل القارئ الذي يرغب في مزيدٍ من التفاصيل عن البنوك الشَّجرية النَّحْوِيّة إلى كتاب: روبي، أحمد محمد. البنك الشجري النَّحْوِيّ بناؤه وتوظيفه في إطار تقنيات الذَّكَاء الاصْطِنَاعِيّ، مركز الملك عبدالله الدولي لخدمة اللغة العَرَبِيَّة (مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللُّغَوِيّة حاليًا)، الرياض، الطبعة الأولى، ٢٠١٧.

المعالـج في التَّحليـل النَّحْـوِيّ للنُّصـوص مـا أجريتـهُ بمعالجـات سـتانفورد في (البايثـون) علـى الجملـة التَّاليـة: (مـن أقـوال ولي العهـد السـعودي الخالـدة الأمـير محمـد بـن سـلمان مهمـا حاولـوا كبـح جماحنـا لـن يتمكنـوا مـن ذلـك وهمـة السـعوديين كجبـل طويـق لـن تنكسـر إلا إذا تسـاوى هـذا الجبـل مـع الأرض)، فظهـرت نتيجـة التَّحليـل كمـا في الشَّـكْل (71).



شكل ٧١ : نتيجة تشجير النحو الاعتمادي

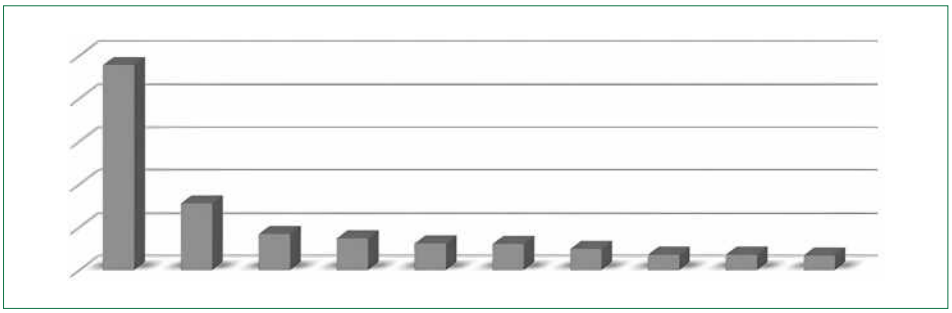
المكـوّن الخامـس: كلمـات الإيقـاف، وتعمـل علـى إبعـاد الكَلِمَـات المضمّنـة في قائمـةِ الإيقـاف، وحـرِيٌّ أن ينتبـه المحَلِّـلُ اللُّغَـوِيُّ إلى محتـوى هـذه القائمـةِ المدمجـة في بعـض المعالجـاتِ، إذ قـد يجـد فيهـا مـا يرغـب في إزالتِـه، أو لا يجـد فيهـا مـا يرغـبُ في تضمينِـه. تَتَضَمَّن هـذه القائمـةُ عـادةً الكَلِمَـات الوظيفِيَّـةَ (حـروف المعـاني) الـتي يُـراد اسـتبعادُها لتحقيـق نتـائج أكـثَرِدِقَّـةً لبعـض التَّطْبِيقَـاتِ؛ مثـل: اسـتخراج الموضوعـات. لننظـر إلى الشَّـكْل (٧٢) الـذي يَتَضَمَّـن قائمـةَ كلمـات الإيقـاف المدمجـة في المعالجـة الأساسِـيَّة لسـتانفورد، سـنلحظ غيـاب بعـض الكَلِمَـات الـتي قـد يحتـاج المحلّـل اللُّغَـوِيّ إلى تغييبِهـا، علـى الرغـم مـن إيـراد ٢٤٣ كلمـةً إيقافِيَّـةً؛ مثـل حـرف العطـف (و) بعـد تقطيـع النَّـصِّ، وفصـل المبـاني النَّحْـوِيّـة الملتصقـة بالكَلِمَـات، والـذي يُلحـظ هـو أن الـواوَ لـم تفصـلْ؛ نظَـرًا لمجيئِهـا ملتصقـةً مـع اسـمِ الموصـولِ «الذيـن» كمـا في «والذيـن»، وقـد يُـوحي ذلـك أنَّ هـذه القائمـةَ مُعتمِـدةٌ علـى النُّصُـوص الـتي لـم تُعالـج بالتَّقْطِيـعِ. ونلحـظُ -أيضًـا- غرابـةَ بعـض الكَلِمَـات مثـل (اللتيـا)، الـتي تـجيء في

التَّعبير الاصطلاحِيِّ النّادر: (بعد اللتيا والتي) في العَرَبِيَّة القديمةِ، ويعني: بعد الخصام والجدال. من المهم أن يتنبّه المحَلّلُ إلى محتوى قوائم الإيقافِ، فحرفُ العطفِ (و) مثَلًا من أكثر الكَلِمَات ورودًا في مُدَوّنَة القرآن الكريم (الشكل ٧٣)، إذ وردت ٩٥٥٢ مرةً، وبنسبة تشكل ١٠,٩٪ من مجملِ القرآن.

```
import nltk
from nltk.corpus import stopwords
", ".join(set(stopwords.words('arabic')))
```

‹والذين، نحو، لستم، أوه، هؤلاء، ومن، ذاك، كما، لئن، إذما، بما، إما، هما، بمن، فيم، هذي، اللاتي، لا، سوى، فلا، لو، فيما، أيها، اللتيا، لاسيما، وما، لكي، إنما، اللتان، كلا، ليست، لن، له، لكن، لوما، اللذين، لك، ذلك، ذو، كأنما، مما، فإن، ولكن، ذات، عسى، كليهما، وإذ، ليت، بل، إلا، كذا، هاته، خلا، ليسا، حيث، بي، ذي، أولئك، حتى، هذه، كيفما، إن، ولو، كلتا، أنتم، من، هذا، كليكما، سوف، هكذا، هذان، ولا، غير، بخ، أولاء، آه، أنتما، هذين، كأين، هيهات، الذين، ذلكما، هنا، إنه، اللائي، لم، منه، عند، أما، على، ليسوا، بهما، كلما، بنا، ذينك، هاتين، هيت، لسن، متى، ذه، تي، إنا، هنالك، ذان، حيثما، لست، لكنما، اللتين، ثمة، حين، إي، ليس، فإذا، أكثر، كي، ماذا، وهو، ممن، لستن، عليك، أين، أنى، وإذا، لهن، كيت، بماذا، لهم، حبذا، منذ، مه، إذا، التي، بعض، إليكما، حاشا، لنا، بين، الذي، عل، لكما، إليك، بهن، قد، عدا، إيه، شتان، ذانك، ذلكم، هي، كل، إذ، بكما، وإن، لعل، تلكما، ألا، يا، هلا، بك، دون، والذي، اللذان، إليكم، أو، أي، تلك، لي، نحن، عما، كأي، ذواتي، أف، ذا، تينك، كلاهما، لدى، لكم، لما، بعد، ليستا، كأن، آها، بهم، عليه، هاهنا، ذوا، بلى، ما، هيا، هن، ذين، إذن، أن، ها، هم، بيد، فمن، لكيلا، هو، نعم، هاتان، ته، فيه، هناك، لولا، كيف، في، كم، أنا، لها، به، أنت، أقل، كذلك، أينما، إليكن، تلكم، ثم، مع، اللواتي، أنتن، إلى، أم، بس، لستما، فيها، هاتي، مذ، لسنا، بها، هاك، ذلكن، عن، آي، منها، مهما، ذواتا، بكم، بكن، تين، لهما، هل، ريث›

شكل ٧٢: قائمة كلمات الإيقاف في مكتبة nltk



شكل ٧٣ : أكثر كلمات الإيقاف تكرارا

ثَانِيًا: تطبيقات معالجة اللغة العَرَبِيَّة الطَّبِيعِيَّة

تـأتي مرحلـةُ بنـاء النمـوذج بعـد معالجـة اللغـة معالجـةً أساسِـيَّةً، وتَتَضَمَّـن اسـتخراجَ السِّـمات والنَّمذجـة والاسـتدلالَ بالمقارنـة مـع الأعمـالِ الأخـرى، وتقييـم الأداءِ لتطبيقـاتٍ عديـدةٍ نذكرُهـا مـع التَّعريـفِ الموجـزِ لـكُلِّ تطبيـقٍ.

لاسـتخراج النُّصُـوص عـدّة تطبيقـات، تعمـلُ علـى اسـتخراج معلومـاتٍ محـدَّدةٍ، منهـا تلخيـصُ النَّـصِّ إلى الموضوعـات الأساسِـيَّة فيـه، ومنهـا اسـتخراجُ الكَلِمَـات المفتاحِيَّـة للنـص، ومنهـا -أيضًـا- التعَـرُّف إلى الكيانـاتِ مـن أسـماء الأشـخاصِ والأماكـن والمنظَّمـات والشَّـركات في النَّـصِّ، أمـا تصنيـف النُّصُـوص فيعمـل تطبيقُـه علـى نمـاذج التَّصنيـف، وتُسـتخدم هـذه النمـاذج في تحليـل المشـاعرِ، واسـتخراج الموضوعـاتِ، وكشـف النِّيَّـةِ، والأخبارِ الزَّائفـةِ، وتطبيقـات التَّصنيـف بشـكلٍ عـامٍّ. في تحليـل المشـاعرِ؛ يُحـدّد النَّـصّ أو الفقـرة أو الجملـة فيمـا إن كانـت إيجابِيَّـةَ المشـاعر أم سـلبِيَّةَ المشـاعر، ويُعتمـد في هـذا التَّصنيـف علـى تدريـب الآلـة علـى عشـراتِ الآلاف مـن التَّصنيفـاتِ مسـبقة التَّعْلِيـم، الـتي حُـدِّدَ فيهـا مـا هـو إيجـابِيٌّ ومـا هـو سـلبِيٌّ، أو علـى معجـم المشـاعرِ الـذي يَتَضَمَّـن عشـراتِ الآلاف مـن الكَلِمَـات الفصيحـةِ واللهْجِيَّـةِ المصَنَّفـة إلى مـا هـو إيجـابِيٌّ ومـا هـو سـلبيٌّ، ويُمكـن تطبيقُهـا دون ذلـكَ كُلِّـه في أداة (أورانج)، عـن طريـق أداةِ (sentiment analysis)، واختيـار اللغـات المتعَـدِّدةِ (multilingual)، ثـم اختيـار العَرَبِيَّـة (Arabic)، ولـن يُطبّـق تحليـل المشـاعر علـى مُدَوّنَـة القـرآنِ

الكريـم، تعـالى كلامُ الله عـن ذلِـكَ. ويُسـتعمل تصنيـفُ النُّصُـوص -أَيضًـا- في تطبيقـاتِ تصنيـف الموضوعـاتِ، وفـق خوارزميَّـة التَّشـابهِ الـدّلاليِّ، أو تقليـل الأبعـادِ، الـتي تسـتخرج الموضوعـات مـن النَّـصّ المتشـابهة، ومـن تطبيقـاتِ تصنيـف النُّصُـوصِ -أيضًـا- مـا يُعـرف بكشـف النِّيَّـةِ (intent detection)، الـذي يسـاعد علـى تحديـد الهـدفِ مـن وراء النَّـصِّ، والـذي يُسـتعمل كثـيرًا في التَّسـويقِ، والمبيعـاتِ، والاشـتراكاتِ البريديَّـة، وكذلـك كشـف الأخبـارِ الزَّائفَـة، أو كشـف البريـدِ الإلكـترونِيّ غـير المرغـوب فيـهِ، أو كشـف خطابـاتِ الكراهيـةِ، الـتي تفيـدُ كلُّهـا العديـدَ مـن الجهـات الإعلاميَّـة والمؤسَّسـيَّة.

مـن تطبيقـات معالجـة اللغـة العَرَبِيَّـة الطَّبِيعِيّـة مـا يُعـرف بروبوتـات المحادثـةِ والدَّعـم الافـتراضِيّ، الـذي يعتمـد علـى الأسـئلة والأجوبـةِ، وتعمـلُ هـذه التَّطْبِيقَـاتُ علـى الإجابـة الآليَّـة عَـن أسـئلة المسـتفيدين، ويُبـنى هـذا التطبيـقُ بواسـطة إدخـال جميـع الأسـئلة المحتملـة مـع إدخـال إجاباتهـا.

مـن تطبيقـاتِ معالجـة اللغـة العَرَبِيَّـة الطَّبِيعِيّـة التَّصحيـحُ الآلِيُّ، والتَّشـكيل الآلِيُّ، ويعتمـدان في البنـاء علـى بيانـات تُـدرّب، وعلـى معـارف الآلـة الـتي عولجَـت معالجـةً أساسِـيَّةً في فهـمِ المكونـات التَّحْلِيليـة الصَّرفِيَّـة والنَّحْوِيّـة والتَّركيبيـة للجمـلِ العَرَبِيَّـة، ومعظـم المصَحّحـات الآلِيَّـة حـتى يومنـا هـذا مفيـدةٌ، ولكـن يبـدو لي أنهـا مـا زالـت لا تعمـلُ علـى تصحيـحٍ كامـلٍ لهمـزات الوصـلِ والقطـعِ في بدايـة الكَلِمَـات العَرَبِيَّـة، الـتي أجدهـا أكـثر أنـواع الأخطـاء شـيوعًا في الكتابـاتِ العَرَبِيَّـة المعاصـرةِ.

تُـدمـج الشَّـبَـكاتُ العصبِـيَّةُ التَّلافيفيـةُ (Convolutional Neural Networks CNN)، الـتي تُـستخـدم لـفهـم الصُّـور ومـقاطـع الفيـديو مـع الشَّـبـكاتِ العصـبِيَّةِ المتكَـرِّرةِ (Recurrent Neural Networks RNN)، في التعَـرُّفِ إلى الـكلامِ في تلـك المقاطـعِ، وقـد ينتـجُ مـن ذلـك وجـود تزييـفٍ لمقاطـع يتحـدَّث فيهـا أشـخاصٌ، وهـم في الحقيقـة ليسـوا هـم أنفسـهم. تخيَّـل هنـا الحجـم الهائـل مـن العمليَّـات العميقـة الـتي بهـا تتعَلَّـمُ الآلـةُ جزئيـات عناصـر الأشـياء؛ لتكتسـبَ معرفـة ماهيتهـا وصورتهـا، سـواءً كانـت صـورة (رؤيـة الحاسـب)،

أو كانت كلامًا شفهيًّا (التَّعَرُّف إلى الكلام)، أو نَصًّا مترجمًا منهُ ونصًّا آخر مترجمًا إليهِ (الترجمة الآلِيَّة). بدأت تطبيقاتُ التّرجمة الآلِيَّة بالاعتماد على منهج القواعد الذي يُحدّد أداء الآلة بقواعد معينةٍ لا يخرج منها، ثم على منهج تعلم الآلة الإِحْصَائِيّ وتدريب النَّماذج على تحذية الكلمة بالكلمة، والعبارة بالعبارة، والجملة بالجملة، وجميعُنا يذكرُ حال التَّرجمة الآلِيَّة غير الدَّقيقة للنُّصوص الطَّويلة. طُوّرت تطبيقاتُ التَّرجمةِ الآلِيَّة باستخدام نماذج التَّعَلّم العميق، فأصبح أداؤُها في الوقت الراهن أفضلَ بكثيرٍ. كذلك الحالُ في تعَلّم الآلة العميق بخوارزميّات الشَّبكات العصبِيَّة المتكَرّرة (RNN) في التعَرُّفِ إلى الكلام، الذي ساعدَ كثيرًا تلك التَّطْبِيقَات التي نعيشُ معها الآن في التعَرُّف إلى اللغة المنطوقة. وصَلَت هذه التَّطْبِيقَاتُ في الوقِت الراهنِ إلى مراحل متقَدِّمةٍ من التَّعَلُّم الآلِيِّ في السَّنتين الماضيتين، وبصورةٍ ضخمةٍ دأبت على السَّعيِ في تحسين جودة حياة الإنسانِ في كثيرٍ من شؤونه الحياتِيَّة؛ في التَّعْلِيم، والاقتصادِ، وذكاء الأعمالِ، ونجد ذلك بين مجتمعات النماذج المدرّبة مسبقًا (pre-trained models) بالشَّبكات العصبِيَّة على اللغة والصُّور والفيديوهات، وفي كُلّ نموذج مجموعةٌ من المطَوِّرين، الذين يعمَلُون على تدريبِ الآلة بالتّعَلُّم العميقِ على بياناتٍ كبيرةٍ. ومن أفضل هذه النَّماذج حتى الآن ما يتعَلّق بتضَمُّن الكَلِمَاتِ مثل (ELMo)، و(Flair)، وما يتعَلَّق بنماذج متعددة الأغراض مثل (ULMFiT)، و(Transformer)، و(Google's BERT وTransformer-XL)، و(OpenAI's GPT3)، بالإضافة إلى معالجات ستانفورد للغاتِ الطَّبِيعِيّة[1].

سيُعرض في الفصول القادمة تطبيقات أدوات تنقيب النُّصُوص والبَيَانَات (أورانج) على مُدَوّنَة القرآن الكريم، وكما أسلفتُ القولَ؛ فالتآلفُ مع هذه الأداةِ بواجهةِ استخدامها السهل في الأبحاث اللُّغَوِيّة والإِنْسَانِيّة والاجْتِمَاعِيّة ليس ترَفًا، بل ضرورَة، وحاجة مُلِحَّة للخروج من الأساليب التَّقليدِيَّة في تلك الأبحاث، وسيُحقّق هذا التآلف

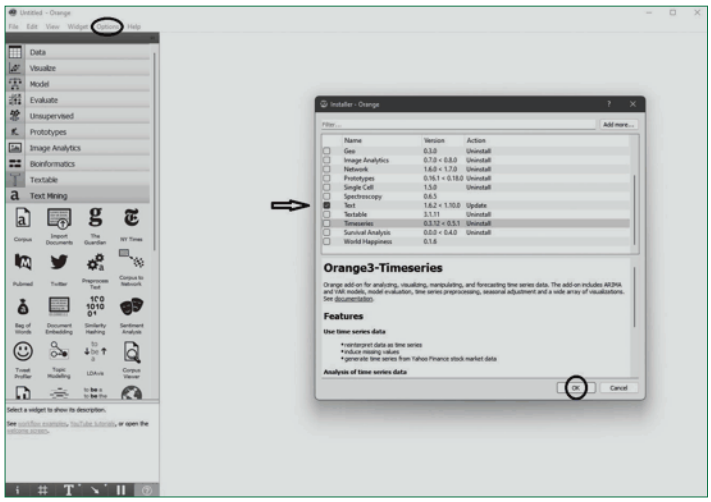
(1) Qui, X., Sun, T., Xu, Y., Shao, Y., Daj, N., Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences 63, 1872-1897.

تمهيـدَ الطريـق إلى معرفـة عمـل تطبيقـاتِ الـذَّكَاء الاصْطِنَاعِـيّ في النَّـصّ اللُّغَـوِيّ أَوَّلًا، ثـم إتاحـة سُـبُلِ إدراك مـا يحـدثُ في عالـم الثَّـورة الرَّقْمِيَّـة ثانِيًـا، ثـم دخـول المهتَمّـين بالعلـوم اللُّغَوِيّـة والإِنْسَـانِيّة والاجْتِمَاعِيّـةِ إلى هـذا العالـم، أسـوةً بأقرانهـم في المراكـز البحثِيَّـة العالمِيَّـة ثالثًـا، فالعلـوم الإِنْسَـانِيّة والاجْتِمَاعِيّـة اليـوَم قـد باتـت تتمحـورُ في مـا يُعـرف بالإِنْسَـانِيّاتِ الرَّقَمِيّـةِ (digital humanities)، والـمُدَوّنَات اللُّغَوِيّـةِ (corpora)، وتقنيـات اللغةِ (language technologies) في معظـم أهـم المراكـز والمؤسسـاتِ والجامعـاتِ في العالـم.

الفصل الأَوَّل
تنقيب النُّصُوص

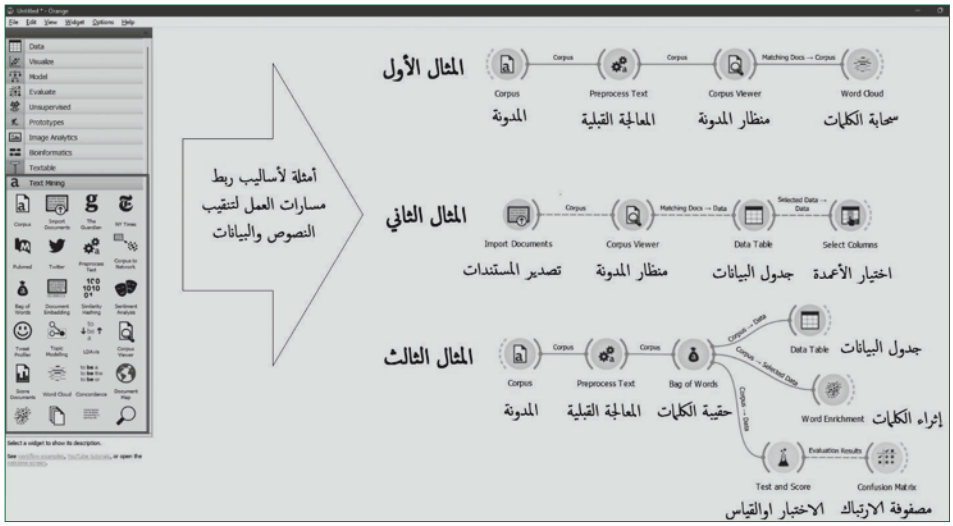
هنـاكَ أربَعـةُ مصطلحـاتٍ تطبيقِيَّـةٍ يجـبُ معرفتُها عندما نقـرأ في معالجـة اللغـة الطَّبِيعِيّـة في منصـات البرمجـة في البايثـون؛ وهـي: المكتبـة (library)، والحزمـة (package)، والوحـدة (module)، والإطـار (framework)، والنمـوذج (model). تَتَضَمَّـن المكتبـةُ مجموعـةً من الحزم، وتَتَضَمَّـن الحزمـةُ الواحـدة مجموعـةً مـن الوَحـدات. هنـاك كَثـيرٌمـن المكتبـاتِ لمعالجـةِ اللغـة الطَّبِيعِيّـة، مثـل مكتبـةِ (nltk). بعـد تنصيـب هـذه المكتبـة في البايثـون؛ يُمكـن تنصيـب حزمـةٍ مـن الحـزم الـتي بداخلِهـا؛ مثـل (nltk.corpus)، وهـي الـمُدَوّنَات اللُّغَوِيّـة، ثـم يُمكـن مثَـلًا أن نسـتدعي وحـدة كلمـات الإيقـافِ (stopwords) مـن الحزمـة. أمـا الإطـارُ فيَتَضَمَّـن مجموعـةً مـن الحـزم والوحـداتِ الـتي تعمـل في مسـارِ بنـاءِ التَّطْبِيقَـاتِ في (البايثـون)، مثـل جانغـو (Django)، وفلاسـك (Flask)، وبوتـل (Bottle). أمـا النمـوذج فهـو مجموعـةُ خوارزمِيَّـات رياضِيَّـةٍ مـن القواعـدِ، والإحصـاءِ، والشَّـبكاتِ العصبِيَّـةِ الـتي تعمـلُ في معالجـة اللغـة الطَّبِيعِيّـةِ للاسـتخراجِ، والتَّصنيـفِ، والفهـمِ، والاسـتيعابِ، والاختبـارِ، والتَّوليـدِ.

ليـس هنـاك واجهـةٌ عربِيَّـةٌ لأداةِ أورانج، ولكنهـا سـهلة الاسـتخدام، ويُمكـن الدخـولُ إليهـا عـن طريـق تنزيلِهـا مباشـرةً، أو عـن طريـق بيئـةِ (anaconda)، ويكـون تنزيـل الأداة مباشـرة مـن خـلال الرابـط https://orangedatamining.com، ويُختـار نظام تشـغيل الجهاز الشَّـخصِيّ؛ سـواءً كان ويـنـدوز، أو مـاك، أو لينكـس. وبعـد فتـحِ الأداةِ؛ لا بُـدَّ مـن تنصيـبِ مكتبـةِ (Text)، وذلـكَ عـن طريـق الضَّغـطِ علـى (خيـارات Options)، ثـمَّ اختيـار (الإضافـات Add-ons) كمـا في الشَّـكْل (٧٤). بعـد تنصيـب المكتبـة، وإعـادة فتـح الأداةِ؛ يتـمُّ التأَكُّـد مـن تنصيـبِ هـذه المكتبـةِ بظهـور صنـدوق أدواتِ تنقيـب النُّصُـوصِ (Text Mining).

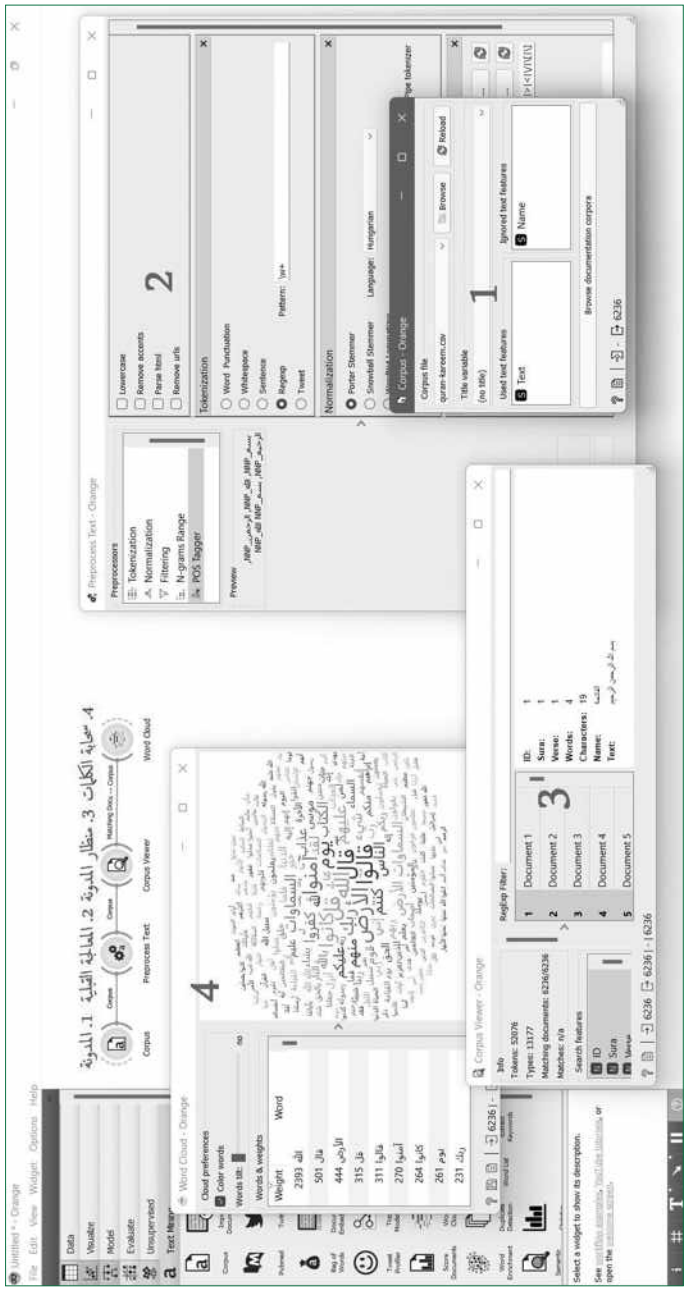


شكل ٧٤ : واجهة أداة (أورانج) وخيارات ونافذة تنصيب مكتبة Text

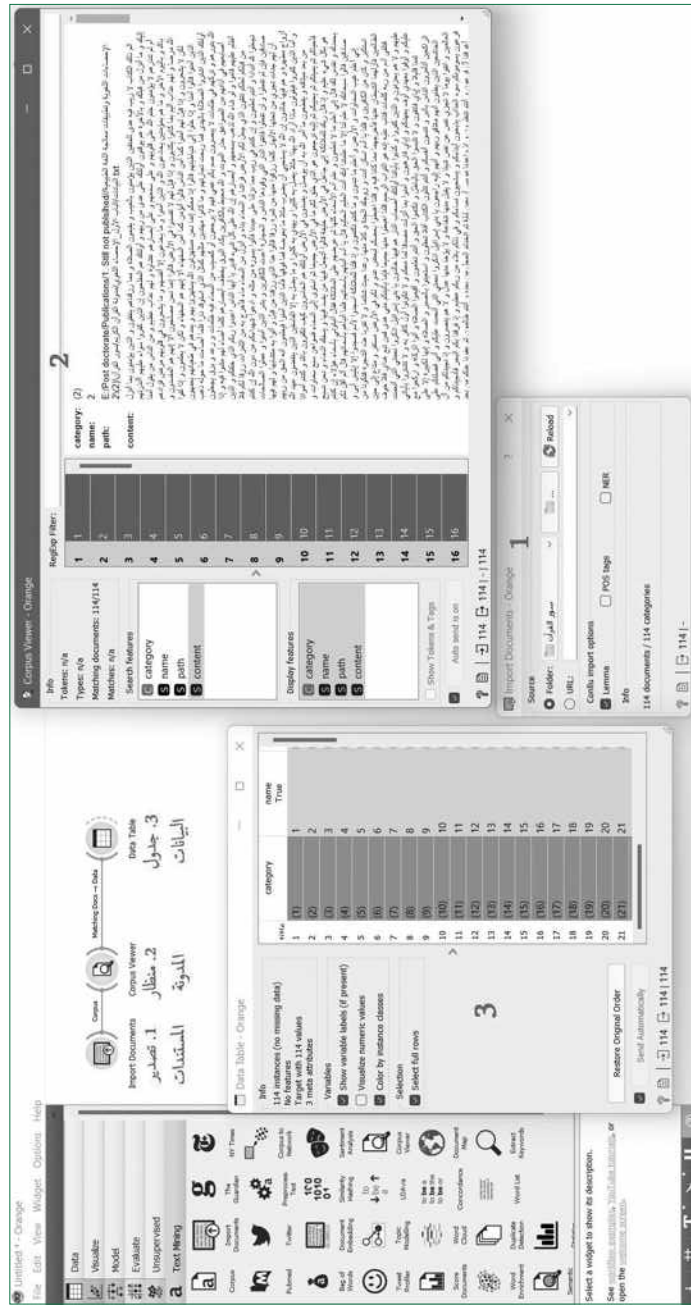
أشـيرُ في الشَّـكْل (٧٥) إلى ثلاثـة أمثلـة لربـط الأدواتِ (widgets) في مسـار عمـل المعالجـة مـن صنـدوقِ تنقيـب النُّصُـوصِ (Text Mining)، وكُلّ مسـار يُعـدُّ اسـتخدامًا معيَّنًـا لمَهَمَّـةٍ مـن مهـامِّ المعالجـةِ.



شكل ٧٥ : أمثلة لأساليب ربط مسارات العملِ للتَّحليل



شكل ٧٦ : مثال ترابط الأدوات وتمثيل خَرْجها داخل أداة (أورانج) للمعالجة القبلية



شكل ٧٧ : مثال ترابط الأدوات وتمثيل خَرْجها داخل أداة (أورانج) لتصدير البَيَانَات واستعراض هيكلتها

يُبدَأ العملُ بسحب أداة من أدواتِ تصدير النصوص من صندوق تنقيب النُّصُوصِ (Text Mining): كسحب الـمُدَوّنَة، أو تصدير المستندات/المجلدات، أو ويكيبيديا، أو تويتر. أما بقية أدوات التَّصدير (صحيفة الجارديان، وصحيفة نيويورك تايمز، وقواعد بياناتِ مُلَخَّصاتِ الأبحاثِ الطِّبِّيَّة) فلا تتعلق باللُّغة العربية، وأما بقِيَّةُ الأدواتِ فهي الأدوات التي تُستخدم لتنقيبِ النُّصُوص ومعالجتها، وهي الأدوات التي ستُستخدم لمعالجة نصوص القرآن الكريم. ولا تعمل كُلُّ تلكَ الأدواتِ إلا بعد ربطِها بالأداة الأُخرى عن طريقِ إخراج مسار بزر الفأرة من أداة الـمُدخَلِ (input) إلى أداة الـمُخرَجِ (output). ولتجَنُّبِ الإكثار من الأشكال التَّوضيحِيّةِ؛ سيُكتفى بصورتين توضيحيّتين (الشكل ٧٦ والشكل ٧٧). تُبَيّن صورة الشكل (٧٦) هيئة ترابط الأدواتِ الآتية: (الـمُدَوّنَة، والمعالجة القبلِيَّة للنَّصِّ، ومنظار الـمُدَوّنَة، وسحابة الكَلِمَات)، وتُبَيّن صورة الشكل (٧٧) هيئة ترابط الأدواتِ الآتية: (تصدير المستندات/المجلدات، ومنظار الـمُدَوّنَة، وجدول البَيَانَات)، ويتعاملُ المستخدم بهذا الأسلوبِ الذي يُعرف بمسار العملِ (workflow) مع جميع الأدواتِ حسب الاستخداماتِ المحَدَّدة لكل مَهَمّة من مَهَامّ المعالجة.

تَتَضَمَّن المعالجة القبلِيَّة للنُّصوص ستَّ وظائف؛ وهي: التَّحويل (Transformation)، والتَّفريق (Tokenization)، والتَّسوية (Normalization)، والتَّصفية (Filtering)، ومدى التَّتابعات اللَّفْظِيّة أو النغرامات (N-grams Range)، والوسوم النَّحْوِيّة (POS Tagger)، وسيُشار في الآتي إلى الاستخداماتِ المفيدة فقط للنَّصِّ العَرَبِيّ من تلك الوظائف.

التَّحويل: يُمكن استخدام التَّحويل في معالجةِ النُّصُوص العَرَبِيَّة معالجةً قبليَّةً، كأن يتم إزالة الروابط والرموز في حال كان النَّصّ العَرَبِيّ من المواقع الشَّبكيَّة.

التَّفريق: وبه يُمكنُ تفريقُ ما يردُ من النَّصّ من علاماتِ ترقيمٍ، فيعمل التَّفريقُ هنا على فصل تلك العلاماتِ عن الكَلِمَات، كما يُمكنُ إجراءُ التَّفريقِ على مستوى الجمل، فيعمَل التَّفريقُ على فصلِ الجملة عن الجملةِ بمعيار وجودِ النُّقطة.

التَّسْوِيَة: وفيه يختار المستخدم إما المقطّع بمنهج (Porter)، أو بمنهج (Snowball)، أو بمنهج مُعجّمِ شبكةِ الكَلِمَات (WordNet Lemmatizer) الذي يعمل على إرجاع الكلمة إلى مادَّتها المعجَمِيّة الأصلِيَّة، والذي يعتمد على المعرفةِ اللغوية-الآلية التي بُنيت في قواعد البَيَانَاتِ المعجَمِيّة، أو بمنهج مُعجّم النحوِ الاعتمادي الكُلِّيّ (Universal Dependency Pipe (UDPipe) Lemmatizer). يعمل المنهجُ الأَوَّلُ والثَّاني على إرجاع الكلمة إلى جذعِها، وتكمن مشكلةُ هذين المنهجين في التَّعاملِ مع بعض الكَلِمَات التي تُعَدُّ في صيغتِها المستعملة أصلِيَّةً، فيتعامل معَها لشبهِها بأسماء أُخرى؛ بسبب التشابه في الصِّيغة، مثل: «اللذان» التي يتعامل معها بوصفِها مثنى، فيُجري لها تسوية بحذف الألف والنون على اعتبارِهما من علامات المثنى، وتنبع هذه المشكلةُ من بيانات التَّدريب والاختبار التي اعتُمد عليها في هذين المنهجين. أما منهج معجّم شبكة الكَلِمَات فليس شموليًّا؛ لأنه يعتمد على تسويةٍ محدودةٍ. وأفضل ما يُستعمل للعربيَّةِ هو معجّم النَّحوِ الاعتمادِيّ الكُلِّيّ للعربيَّةِ؛ لأنه يَتَضَمَّن التَّقْطِيعَ والتَّوسيمَ والتَّحْلِيلَ النَّحْوِيّ معًا.

التَّصفية: وبه يُضيفُ المستخدمُ كلماتِ الإيقاف، أو يُضيفُ معجمَه الخاصّ الذي يرغبُ في تصفيته، أو يعتمدُ على التَّعبيراتِ النَّمطِيَّة. ومن أكثر التَّعبيرات النَّمطية المفيدة للنُّصوص العَرَبِيَّة استعمال \bword\b واستبدال b الأولى بكلمةٍ معينةٍ؛ ليعمل هذا التعبير المنطقي على استخراجها دون غيرها، واستعمال \w+/ لاستخراج الكَلِمَاتِ فقط في حالِ كان النَّصّ يَتَضَمَّنُ علامات ترقيمٍ أو رموز، واستعمال \b(B|b)w+\b، واستبدال b داخل القوسين بأي حرفٍ ليعمل هذا التعبير المنطقيُّ على استخراج كُلّ الكَلِمَاتِ التي تبدأ بهذا الحرفِ المحَدَّد، واستعمال \w{5,}| لاستخراج الكَلِمَات التي تتكوَّن حروفُها مِن أَكثر من خمسة أحرف، واستعمال \b\w+(Y|y)\b واستبدال y داخل القوسين بحرف ما؛ ليستخرج كُلّ الكَلِمَات التي تنتهي بهذا الحرفِ المحدد.

المدى: اختيار مدى التتابع اللَّفْظِيّ (يم-يس).

الوسـوم النَّحْوِيّـة: وبهـا تُوسَّـمُ الكَلِمَـاتُ في النَّـصِّ، ويعتمـدُ علـى منهجـين؛ الأول واسـم بيرسـبتـون المتـوسـط (Averaged Perceptron Tagger)، وواسـم البنـك الشجـريّ (Treebank POS Tagger)، وأفضـل النَّتـائج يُعطيـه المنهـجُ الثَّـاني.

المبحث الأَوَّل: إعداد البَيَانَات اللُّغَوِيّة في جدولةِ بيانات

يتطَلَّـب مـن المسـتخدم إعـداد البَيَانَـات في ملـف الأكسـل بتهيئـةٍ محـدَّدةٍ؛ لكـي تتعامـلَ أداةُ (أورانج) معهـا علـى الوجـه الصَّحِيـحِ، أُشـيرُ هنـا إلى أسـلوبِ إعـداد البَيَانَـات اللُّغَوِيّـة للنُّصـوصِ العَرَبِيَّـة بالصُّـورة الأَوَّلِيَّـة المطلوبـةِ، وسـأقف علـى إعـدادِ البَيَانَـات اللُّغَوِيّـة بمُتَغَيِّرَاتهـا المتعَـدّدة مـن غـير النُّصُـوصِ في الفصـل الخامـس. لإعـداد البَيَانَـات اللُّغَوِيّـة أسـاليبُ محـدَّدةٌ، لـكُلٍّ منهـا مَهَمَّـةٌ، وفي أداةِ أورانج؛ المطلـوبُ هـو عنونـة خصائص الـمُتَغَيِّرَات فيمـا إن كانـت نَصًّـا أو كانـت فئويَّـةً. في مثالِنـا عـن تنقيـبِ مُدَوّنَـة القـرآن الكريـم، هيَّـأتُ البَيَانَـات علـى النَّحـو الآتي.

أَوَّلًا: حُـدّدَ عَـدَدُ الصُّفـوفِ بحسـبِ عـددِ السُّـورِ ليكـون ١١٤ صفًّـا فقـط، ويَتَضَمَّـن العمـودُ الأَوَّلُ التَّصنيـفَ الفئـويَّ (أسـماء السُّـوَر القرآنيـة)، ويَتَضَمَّـن العمـودُ الثَّـاني آيـاتِ السُّـورَة (الشَّـكل ٧٨)، وستُسـتخدم هـذه البَيَانَـات في المبحَـثِ الثـاني مـن هـذا الفصـلِ. أمّـا في الشَّـكْل (٧٩) فتَضَمَّـنَ العمـودُ الأولُ -أيضًـا- أسـماء السُّـوَر القرآنيـة، أمـا العمـود الثَّـاني؛ فتضَمَّـنَ كُلّ صـفٍّ فيـه آيـةً قرآنِيَّـةً واحـدةً؛ ليكـونَ عـددُ الصُّفـوف في هـذهِ البَيَانَـات ٦٢٣٦ صفًّـا، وستُسـتخدم هـذه البَيَانَـاتُ في اسـتخراجِ الكَلِمَـات المفتاحِيَّـة (المبحـث الثـاني في الفصـل الثَّالـث). يتـمُّ ذكـر الفئـةِ (category) علـى أنـه نَـصُّ (Text)، ثـم d (المتغَـيِّر الاسـميّ)، الـذي يشـيرُ إلى نـوعِ مـا سـيعالجُ (سلسـلة حرفِيَّـة)، والنُّصُـوص توصـفُ علـى أنَّهـا (string)، ثـم الصّنـف (class) علـى أنَّـه حقيقـيٌّ كمـا هـو (True)، ويعـني أَنَّ جميـعَ أسـماء السُّـوَر صحيحـةٌ في المدخـلِ المكتـوب.

Category	Text
d	string
class	include=True
الفاتحة	بسم الله الرحمن الرحيم الحمد لله رب العالمين الرحمن الرحيم مالك يوم الدين
البقرة	الم ذلك الكتاب لا ريب فيه هدى للمتقين الذين يؤمنون بالغيب ويقيمون الصلاة
آل عمران	الم الله لا إله إلا هو الحى القيوم نزل عليك الكتاب بالحق مصدقا لما بين يديه و أنزل
النساء	يا أيها الناس اتقوا ربكم الذي خلقكم من نفس واحدة وخلق منها زوجها وبث منهما
المائدة	يا أيها الذين آمنوا أوفوا بالعقود أحلت لكم بهيمة الأنعام إلا ما يتلى عليكم غير محلي
الأنعام	الحمد لله الذي خلق السماوات والأرض وجعل الظلمات والنور ثم الذين كفروا
الأعراف	المص كتاب أنزل إليك فلا يكن في صدرك حرج منه لتنذر به و ذكرى للمؤمنين
الأنفال	يسألونك عن الأنفال قل الأنفال لله والرسول فاتقوا الله وأصلحوا ذات بينكم
التوبة	براءة من الله ورسوله إلى الذين عاهدتم من المشركين فسيحوا في الأرض أربعة أشهر
يونس	الرتلك آيات الكتاب الحكيم أكان للناس عجبا أن أوحينا إلى رجل منهم أن أنذر الناس
...	...

شكل ٧٨ : شكل هيكلة البَيَانَات لمعالجتها في أداة أورانج

Category	Text
d	string
class	include=True
الفاتحة	بسم الله الرحمن الرحيم،
الفاتحة	الحمد لله رب العالمين،
الفاتحة	الرحمن الرحيم،
الفاتحة	مالك يوم الدين،
الفاتحة	إياك نعبد وإياك نستعين،
الفاتحة	اهدنا الصراط المستقيم،
الفاتحة	صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين،
البقرة	الم،
البقرة	ذلك الكتاب لا ريب فيه هدى للمتقين،
...	...

شكل ٧٩ : شكل هيكلة البَيَانَات لمعالجتها في أداة أورانج

ثَانِيًا: تهيئـة بيانـاتِ حقيبـة الكَلِمَـات، وتـضمين الكَلِمَـات (الفصـل الثـاني)، واستـخـراج الموضـوعات (المبـحث الأول في الفصـل الثَّالـث). تطلّـب ذلـك تحـديد الصِّنـف (class) -مكِّيَّـة أو مـدنِيَّة- في العمـود الأوَّل، وهـي مـنفصلةٌ (discrete)، ولـيست مُتَّـصلةً (continuous)؛ لأنَّ كُلَّ صنـفٍ يَتَضَمَّـن متغـيرًا ثابتًـا لـه، ثـم العنـوان (Title)، الـذي يَتَضَمَّـن اسـمَ السُّـورَة في العمـود الثـاني، ثـم المحتـوى (Content)، الـذي يَتَضَمَّـن آيـات السُّـورَة، ثـم رقـم ترتيـب السُّـورَةِ (Numerical). واحتجـتُ هنـا إلى أربعـة أعمـدةٍ بـدلا مـن عمودين. وعنـد التَّعْلِيـمِ؛ يُشـارُ إلى عناوينِهـا بشـكلٍ عمـوديٍّ كمـا في الشَّـكْل (٨٠)، وتُدخـلُ كُلّ آيـةٍ في صَـفٍّ واحـدٍ، ليكـون عـددُ الصُّفـوف ٦٢٣٦ صفًّـا. أمـا الشَّـكْل (٨١) فـكُلُّ سـورةٍ كاملـة قـد وُضعـتْ في صـفٍّ واحـدٍ.

Topic	Title	Content
discrete	string	string
class		include=True
مكية	الفاتحة	بسم الله الرحمن الرحيم
مكية	الفاتحة	الحمد لله رب العالمين
مكية	الفاتحة	الرحمن الرحيم
مكية	الفاتحة	مالك يوم الدين
مكية	الفاتحة	إياك نعبد و إياك نستعين
مكية	الفاتحة	اهدنا الصراط المستقيم
مكية	الفاتحة	صراط الذين أنعمت عليهم غير المغضوب عليهم و لا الضالين
مدنية	البقرة	الم
مدنية	البقرة	ذلك الكتاب لا ريب فيه هدى للمتقين
مدنية	البقرة	الذين يؤمنون بالغيب و يقيمون الصلاة و مما رزقناهم ينفقون
مدنية	البقرة	و الذين يؤمنون بما أنزل إليك و ما أنزل من قبلك و بالآخرة هم يوقنون
مدنية	البقرة	أولئك على هدى من ربهم و أولئك هم المفلحون
مدنية	البقرة	إن الذين كفروا سواء عليهم أأنذرتهم أم لم تنذرهم لا يؤمنون
مدنية	البقرة	ختم الله على قلوبهم و على سمعهم و على أبصارهم غشاوة و لهم عذاب عظيم
مدنية	البقرة	و من الناس من يقول آمنا بالله و باليوم الآخر و ما هم بمؤمنين
...	...	...

شكل ٨٠ : شكل هيكلة البَيَانَات لمعالجتها في أداة أورانج

Topic	Title	Content	Numerical
discrete	string	string	string
class		include=True	
مكية	الفاتحة	بسم الله الرحمن الرحيم الحمد لله رب العالمين الرحمن الرحيم ملك يوم الدين	1
مدنية	البقرة	الم ذلك الكتاب لا ريب فيه هدى للمتقين الذين يؤمنون بالغيب ويقيمون الصلاة	2
مدنية	آل عمران	الم الله لا إله إلا هو الحي القيوم نزل عليك الكتاب بالحق مصدقا لما بين يديه وأنزل	3
مدنية	النساء	يا أيها الناس اتقوا ربكم الذي خلقكم من نفس واحدة وخلق منها زوجها وبث	4
مدنية	المائدة	يا أيها الذين آمنوا أوفوا بالعقود أحلت لكم بهيمة الأنعام إلا ما يتلى عليكم غير	5
مكية	الأنعام	الحمد لله الذي خلق السماوات والأرض وجعل الظلمات والنور ثم الذين كفروا	6
مكية	الأعراف	المص كتاب أنزل إليك فلا يكن في صدرك حرج منه لتنذر به و ذكرى للمؤمنين	7
مدنية	الأنفال	يسألونك عن الأنفال قل الأنفال لله والرسول فاتقوا الله وأصلحوا ذات بينكم وأطيعوا	8
مدنية	التوبة	براءة من الله ورسوله إلى الذين عاهدتم من المشركين فسيحوا في الأرض أربعة	9
مكية	يونس	الر تلك آيات الكتاب الحكيم أكان للناس عجبا أن أوحينا إلى رجل منهم أن أنذر	10
...	...	...	...

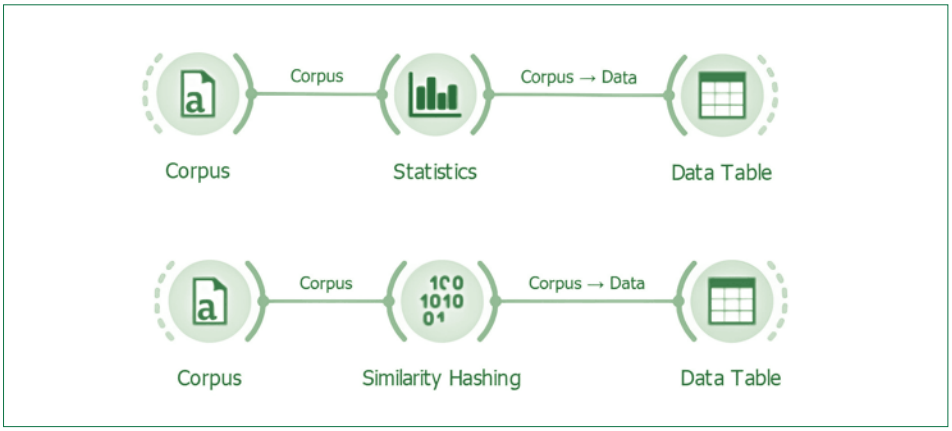
شكل ٨١ : شكل هيكلة البَيَانَات لمعالجتها في أداة أورانج

يُمكن استخدام مُتَغَيِّرَات غير نَصِّيَّةٍ، وفي هذه الحالةِ؛ تُكتب سمة المتغَيِّر، فبدَلًا من (string)، أو اختصاره (s) التي تشيرُ إلى النُّصُوصِ؛ نرمزُ بـ (d) للإشارة إلى المُتَغَيِّرَات الاسميَّة، أو نكتبُ (continuous)، أو اختصارهُ الرَّمز: (c) للمُتَغَيِّرَات المقياسِيَّةِ المتَّصلة، أو نكتبُ (time)، أو اختصارهُ الرَّمز: (t) للتواريخ.

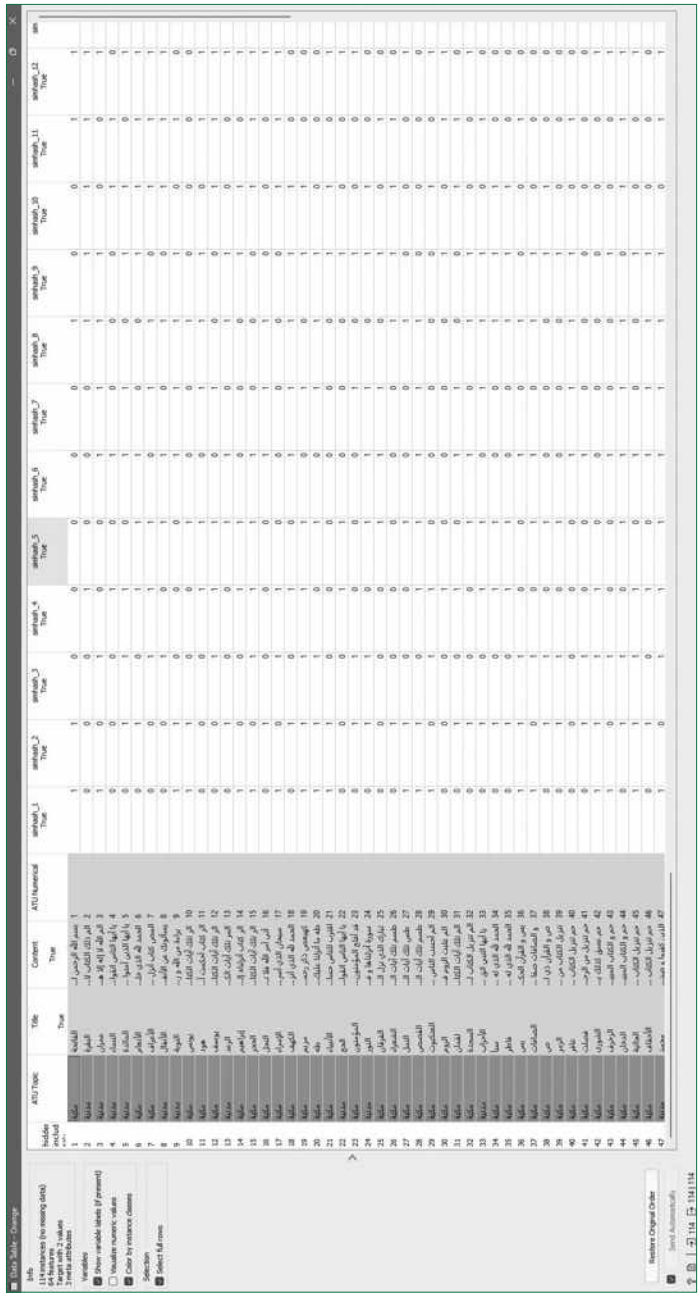
المبحث الثَّانِي: الإحصاءات وتكرار الكلمة-تكرار المستند العكسِيّ

تتوفر إحصاءاتٌ أَوَّلِيَّةٌ للمدَوّنة اللُّغَوِيّة، ويُمكن اختيارها جميعًا، أو اختيار ما يُراد معرفتُه منها، وتتكَوَّن هذه الإحصاءاتُ من عدد الكَلِمَاتِ، وعدد الحروف، وعدد التَّتابعاتِ اللَّفْظِيّة (النغراميات)، ومتوَسّط طولِ الكَلِمَاتِ في المُدَوّنَة، وعدد علاماتِ التَّرقيمِ إن تُرِكت في النَّصِّ دون إزالةٍ في المعالجة القبليَّة للنَّصِّ، والنّسبة المئويّة للكلماتِ الفريدة، وعدد الكَلِمَاتِ التي يُحدّد حروفُها الأولى، كأن تكون مثلًا (أل) التَّعريف، وعدد

الكَلِمَات التي يُحدّد حروفها الأخيرة، كأن تكون مثلًا الألف المقصورة (ى)، وعدد التعبيرات النَّمطِيَّة (regular expressions)، التي تعتمد على رموزٍ تكشفُ أنماطًا محددةً من النَّصِّ، وعدد الوسوم النَّحْوِيّة إذا أضافَ المستخدم في المعالجة القبلِيَّةِ للنَّصِّ خيارَ وسم النَّص. في الشَّكْل (٨٢) مساران للمعالجةِ في أداةِ الـمُدَوّنَة، وفيهما بيانات الشَّكْل (٧٨). المسارُ الأَوَّل للإحصاءِ، والثَّاني لتجزئة التَّشابهِ (similarity hashing). عَرَضْتُ نتائجَ الإِحْصَائِيّاتِ في الجدول (٢٨)؛ لتوضيح أربعة حساباتٍ فقط لبيانات آيات القرآنِ الكريم، ثم عَرَضْتُ نتيجةَ تجزئةِ التَّشابه في الشَّكْل (٨٣).



شكل ٨٢ : مسارا العمل للإحصاءات وتجزئة التشابه



شكل ٨٣ : تجزئة التشابه

جدول ٢٨ : نموذج إحصاءات كُلّ آية من سورة الفاتحة

السورة	الآيـــــة	الكَلِمَات	الحروف	التتابع اللَّفْظِيّ	متوسط طول الكلمة
الفاتحة	بسم الله الرحمن الرحيم	4	19	4	4.75
الفاتحة	الحمد لله رب العالمين	4	18	4	4.5
الفاتحة	الرحمن الرحيم	2	12	2	6
الفاتحة	مالك يوم الدين	3	12	3	4
الفاتحة	إياك نعبد و إياك نستعين	5	19	5	3.8
الفاتحة	اهدنا الصراط المستقيم	3	19	3	6.333
الفاتحة	صراط الذين أنعمت عليهم غير المغضوب عليهم و لا الضالين	10	44	10	4.4

من إحصاءاتِ الـمُدَوّنَات اللُّغَوِيّة في معالجةِ اللغةِ الطَّبِيعِيّة ما يُعرف بتكرارِ الكلمةِ–تكرار المستند العكسِيِّ (Term–Frequency Inverse Document Frequency)، واختصاره (TF–IDF)، وهو مختلفٌ عن التّكْرَار المطلق والتّكْرَار النِّسْبِيّ الذي تناولتُه في المبحث الثَّاني من الفصلِ الأول، فهو لا يَهتَمُّ بَرْتَبِ الكَلِمَات وتكراراتها المطلقة والنسبية، بل بالكَلِمَاتِ المهمَّة في النُّصُوص، فيعطي الكَلِمَات الأكثر تكرارًا بين النُّصُوص قيمةً أقل بين الصِّفر والواحد، ويعطي الكَلِمَات المهمَّة بين النُّصُوص قيمةً أعلى بين الصِّفر والواحد.

تعمل هذه الخوارزمِيَّةُ على إجراءِ حسابٍ بين تكرار كُلِّ كلمةٍ، وعدد ورودِها في كُلِّ نَصٍّ، ومجموع عددِ الكَلِمَات كُلِّها في الـمُدَوّنَة، وذلكَ وفق المعادلةِ الآتية:

- تكرار الكلمة TF = تكرار الكلمة ÷ إجمالي عدد الكَلِمَات في الـمُدَوّنَة
- تكرار المستند العكسي IDF (للكلمة) = $\left(\frac{\text{عدد النُّصُوص}}{\text{عدد النُّصُوص التي في الكلمة}}\right)$،

- تكـرار الكلمة-تكـرار المسـتند العكـسي TF-IDF (للكلمـة) = لـوغ (نسـبة تكـرار الكلمـة × نسـبة تكـرار المسـتند العكـسي)

يتمَـيَّز هـذا الإحصـاءُ للكلمـات بعـدم تركـيزِه علـى مبـدأ زيـف (Zipf) في التَّنـازل الطَّبِيعِـيِّ لـورود الكَلِمَـات في النُّصُـوص اللُّغَوِيّـة، بـل علـى أهَمِّيَّـة الكلمـةِ في النَّـصِّ، والأهَمِّيَّـة هنـا دلالةٌ علـى تـوزُّع الكلمـة توزُّعًـا يحمـلُ أهَمِّيَّـةً معنوِيَّـةً لهـا. ولا تَتَضَمَّـن أداة (أورانج) حسـاب تكـرار الكلمة-تكـرار المسـتند العكـسِيّ، وعليـه يُمكـن تطبيقُهـا باسـتخدامِ الأكـواد[1] في (البايثـون) كمـا في الملحـقِ (د). عَرَضْـتُ أوَّلَ خمسـين كلمـة مـن نتيجـة تكـرار الكلمة-تكـرار المسـتند العكـسِيّ مـن ملـفّ (السُّـوَر المكِّيَّـة) وملَـفّ (السُّـوَر المدَنِيَّـة) كمـا في الملحـق (هـ).

(1) Lavin, M. J. (2019). Analyzing Documents with TF-IDF [Online Lessons]. In: Z. LeBlanc (Ed.), The Programming Historian. Retrieved October 24, 2022, from https://doi.org/10.46430/phen0082.

الفصل الثَّانِي
حقيبة الكَلِمَات وتضمين الكَلِمَات

عندما تتعاملُ الآلة مع النُّصُوصِ؛ أي: عندما تُنمذِجُ الآلةُ النُّصُوصَ؛ فهي تُحَوِّل الكَلِمَات في هذه النُّصُوص إلى أرقامٍ. لُخِّصتْ بعضُ الممارسات التّقنية في نمذجة الكَلِمَات العَرَبِيَّة، وهو تلخيصٌ جديرٌ بالاطّلاع من اللُّغَوِيّين المهتَمِّين[1].

سأُلخِّص فيما يَأتِي فكرةَ تحويل الكَلِمَات إلى أرقام، وفقَ المناهج المطَبَّقَة في معالجة اللغةِ الطَّبِيعِيّة، وذلك من وجهة نظرٍ لُغَوِيَّة. لعملِيَّات التَّحويل ثَلاثُ مراحل أساسِيَّة؛ الأولى: التَّرميز بالرَّقم الواحدِ (One hot encoding)، وهي مرحلةٌ قديمةٌ لا تصلح. والثَّانية: حقيبة الكَلِمَاتِ (Bag of Words BOW)، التي عالجت مشكلةَ المرحلة الأولى. والثَّالثة: تضمين الكلمةِ (Word Embedding)، أو نماذج الكلمةِ إلى متجه (word٢vec)، التي تعدّ الأفضلَ في نمذجة الكَلِمَات العَرَبِيَّة في النُّصُوص.

يعمل الترميز بالرقم الواحد كالآتي؛ لنفترض أنَّ القاموس الذي لدينا يَتَضَمَّن الجملةَ الآتية: (جاء محمد إلى المنزل). يعملُ التَّرميزُ بتمثيل (جاء) بالرقم (١)، وبتمثيل البقية بالرَّقم (٠)، ثم في المرَّة الثانية يعطي الرَّقم (١) لكلمة (محمد) عند التَّمثيل في المرحلة الثَّانية، والرقم (٠) لبقية الكلمات، وهكذا في بقِيَّة الكَلِمَات في المرحلتين الأخيرتين.

جاء	محمد	إلى	المنزل
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

(١) العصيمي، عبدالرحمن. نمذجة الكلمة العَرَبِيَّة: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل الكلمة العَرَبِيَّة لُغَوِيًّا وتوزيعِيًّا. في: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل النَّصّ العَرَبِيّ، ٢٠١٩، الصفحات ٩٥-١٢٤.

لا جـدوى مـن تعَلُّـم الآلـة لتمثيـل الـمُدَوّنَات اللُّغَوِيّـة الضَّخمـة بهـذه الطَّريقـةِ، فلـو افترضنـا أن لدينـا ١٠٠ ألـف كلمـةٍ؛ فذلـك سـيحتاج إلى تمثيـلِ الكلمـة ١٠٠ ألـف مَـرَّةٍ، مَـرَّةً برقـم (1)، و٩٩,٩٩٩ مَرَّة برقـم (0)، وفي ذلـكَ تكلفـةٌ حسـابِيَّةٌ عاليـةٌ، ناهيـك عـن عـدم جدواهـا في تمثيـل الكَلِمَـات ذات العلاقـة الدلالِيَّـة المتشـابهة، وسـتكون هنـاكَ كلمـاتٌ ليـس لهـا علاقـةٌ دلالِيَّـةٌ بينهـا قـد مُثّلـت بأرقـامٍ متقاربـةٍ فيمـا بينهـا.

جـاءت فكـرة حقيبـة الكَلِمَـاتِ (Bag of Words (BoW بسـببِ قصـور التَّرمـيز بالرقـم الواحـد، وتُنـشئ حقيبـةُ الكلمـات أرقامًـا عددِيَّـةً للكلمـاتِ في النُّصُـوص، وتكـون هـذه الأرقـام حقيقيـةً أو ثنائِيّـةً أو لوغارثميـةً. وهـي مُهمَّـةٌ عندمـا نريـد معالجـة إثـراء الكلمـات، ثـم جـاءت فكـرةٌ مطَـوَّرَةٌ أكـثر، تُعـرف بتضمـين الكَلِمَـاتِ (Word Embedding) الـذي يُحلّـلُ التَّتابعـات اللَّفْظِيّـة (النغراميـات) في كُلّ نَـصٍّ في الـمُدَوّنَة، ويُعطـي مضمونًـا عددِيًّـا لـكلِّ نغـرام واحـدٍ، وذلـك باسـتخدام نمـوذج تدريـبِ تعـرّف (pre-trained model)؛ ليكـون لـكلِّ نَـصٍّ مُتَّجـهٌ (vector) يُسـتنبطُ مـن نتيجـةِ التَّضمينـاتِ التَّجْمِيعِيَّـةِ لـكلِّ الكَلِمَـات.

سـنتعَرَّف في المبحـث الأَوَّل إلى نمـوذج حقيبـةِ الكَلِمَـات، وفي المبحـث الثـاني إلى نمـوذج تضمـين الكَلِمَـاتِ، ثـم نسـتعرض تطبيقاتهمـا علـى نصـوصِ مُدَوّنَـة القـرآن الكريـم.

المبحث الأَوَّل: حقيبة الكَلِمَات

تفيدُنـا حقيبـة الكَلِمَـات بإنشـاء أرقـامٍ للكلمـات في الـمُدَوّنَة في كُلّ مسـتندٍ أو نَـصٍّ، ويُعـدّ هـذا النمـوذجُ مـن أبسـط نمـاذج المتّجهـات. ولتبسـيط فكـرةِ النمـوذج، ننظـر إلى هذيـن المسـتندين مـن النُّصُـوص:

المستند الأَوَّل: «الحمد لله رب العالمين».

المستند الثَّانِي: «وقل الحمد لله سيريكم آياته».

يعمل النَّموذج أَوَّلًا على إنشاء قاموس يتكَوَّن من ترقيمِ عدد الكَلِمَات على النَّحوِ الآتي: [1: "الحمد"، 2 "لله"، 3 "رب"، 4 "العالمين"، 5 "وقل"، 6 "سيريكم"، 7 "آياته"]، ولو كان في النَّصِّ كلماتٌ وظيفِيّةٌ مُستبعدةٌ مسبقًا في المعالجة القبليَّة للنَّصِّ فلن تُضمّن في القاموس. ويحتوي هذا القاموسُ على سبع كلماتٍ، ولكُلِّ كلمةٍ رقمٌ مفهرس، ثم يعملُ النّموذج ثانِيًا على تمثيلِ المُتّجِهاتِ لعددِ مَرَّاتِ ظهور الكَلِمَات في كُلّ المستنداتِ، فتتمَثَّل على النَّحو الآتي:

المستند الأَوَّل: [0، 1، 1، 2، 2]

المستند الثَّانِي: [1، 1، 2، 2، 1]

سنلحظ أَنَّ كلمة (الحمد) وكلمة (لله) قد أُعطيتا القيمة (٢)؛ لورودهما مَرَّتين في كلا المستندين. الأمرُ نفسُه عندما نبني نموذجَ حقيبة كلماتٍ لآلاف المستندات أو النُّصُوص. سنلحظُ -أيضًا- أَنَّ النموذج قد وَضَعَ للمستندين في مثالنا خمسةَ مُتَّجهات، وقد وضع متَّجِهًا خامسًا في المستندِ الأَوَّل بقيمة صفر، وذلك أسلوبُ عمل حقيبةِ الكَلِمَات، فالكلمةُ التي تظهر في مستندٍ، ولا تظهر في الآخر؛ تُمثّل في الآخَر بقيمة (0).

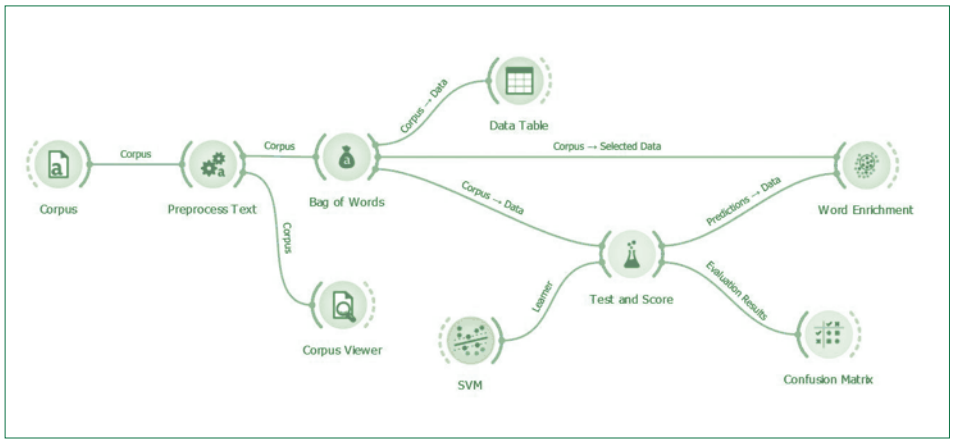
في أداة (أورانج) مقاييسُ نموذج حقيبةِ الكَلِمَاتِ؛ وهي: حساب تكرارِ الكلمة، وحساب تكرار النَّصِّ، والضَّبط (Regularization). يعمل النَّموذج في حسابِ تكرار الكلمة على ثلاثة مناهج يختار المستخدمُ أحدَها، وهي: (التّكْرَار)؛ أي عدد تكرار الكلمة في النَّصِّ، و(الثُّنائِيّ Binary)؛ أي هل الكلمة موجودة في النَّصّ أم لا، و(الخَطّيّ الفرعِيّ Sublinear)؛ أي تحويل التّكْرَارات إلى اللوغارثم. أما في حسابِ تكرار النَّصِّ؛ فيعمل النموذج على ثلاثة مناهج أيضا؛ الأَوَّل: لا شيءَ في حال أراد المستخدم عدم إجراء هذا التّكْرَار، والثَّانِي: تكرار المستند العكسِيّ (inverse document frequency IDF)، والثَّالِث: تكرار المستند العكسِيّ المخَفَّف، الذي يعملُ على استبعاد أَيِّ مستندٍ لا تظهر فيه الكلمة المحسوب تكراراها في المستندات الأُخرى؛ من أجل استبعاد أَيِّ قيمةٍ صفرِيَّةٍ. أمَّا في الضَّبط فيُمكن للمستخدم

أن يختار (لا شيء None)، أو أن يختار (L1)؛ لمعايرة وتوحيدِ طول المتَّجه (الرقم) إن كان بعيدًا جدًّا عن مُتَّجهٍ آخَر، أو لمعايرته بعد استخراجِ المجموع التَّربيعِيِّ للمُتَّجهات.

بعد أن يُنشِئَ النَّموذجُ حقيبةَ الكَلِمَاتِ؛ نَستخدم أداةَ إثراءِ الكَلِمَاتِ (word enrichment) جاكارد وجيب التمام، وتُظهر هذه الأداة قائمة الكَلِمَات بالقيم الاحتماليَّةِ الأعلى دلالةً (أي: الأقل قيمة عادة من 0.05) لمجموعة تُحدَّدُ بوصفِها عَيِّنةً، وتُقارَن مع المُدَوّنَة كُلِّها. تُشير القيمةُ الاحتماليَّة الأقل عادةً من 0.05 إلى احتمالٍ عالٍ بأنَّ الكلمة ليست عشوائيّةً في المُدَوّنَة. تُظهر الأداة -أيضًا- معدل استكشافِ الخطأ (False Discovery Rate FDR)، الذي يُفصح عن النّسبة المتوقّعة من التنَبُّؤ الخطأ في قائمة القيم الاحتماليَّة.

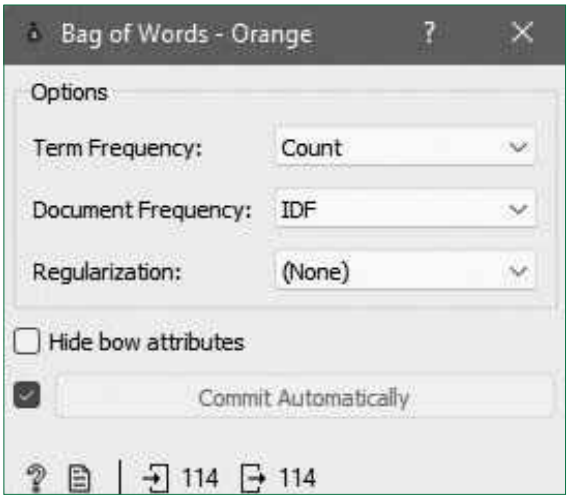
وفيما يلي تطبيق على مُدَوّنَة القرآن الكريم باستخدام أداة تنقيب النُّصُوص أورانج.

في الشَّكْل (٨٤) أسلوب ربطِ أدوات التَّحْلِيل، ففي أداة (مدونة Corpus) تُحمّل البَيَانَات (الشكل ٨٠)، ثم تُربط معها الأدواتُ المتبقِّيَة من المعالجةِ القبليَّة، ثم حقيبة الكَلِمَات، إلى بقيَّة الأدوات التي تعرض النَّتيجةَ، وتختبرُهَا، وتقيسُها.



شكل ٨٤ : مسار عمل ربط الأدوات لمعالجة البَيَانَات وحقيبة الكَلِمَات والاختبار والقياس والتمثيل

ثمَّـة حـالاتٌ عَديـدةٌ في خيـارِ حقيبـة الكَلِمَـاتِ (Bag of Words)، يُمكـن الاختيـارُ مـن بينهـا كمـا في الشَّـكْل (٨٥)، وسـتُعرض هنـاك الحالـة الأولى والحالـة الثانيـة مـن تلـك الحـالاتِ الاثنـتي عشـرة (الجـدول ٢٩)؛ ليتَّضـحَ عمـل نمـوذج حقيبـة الكَلِمَـات في بيانـاتِ مُدَوّنَـة القـرآن الكريـم (السُّـورة ثـم نَـصّ آيـات السُّـورَة الكريمـة).



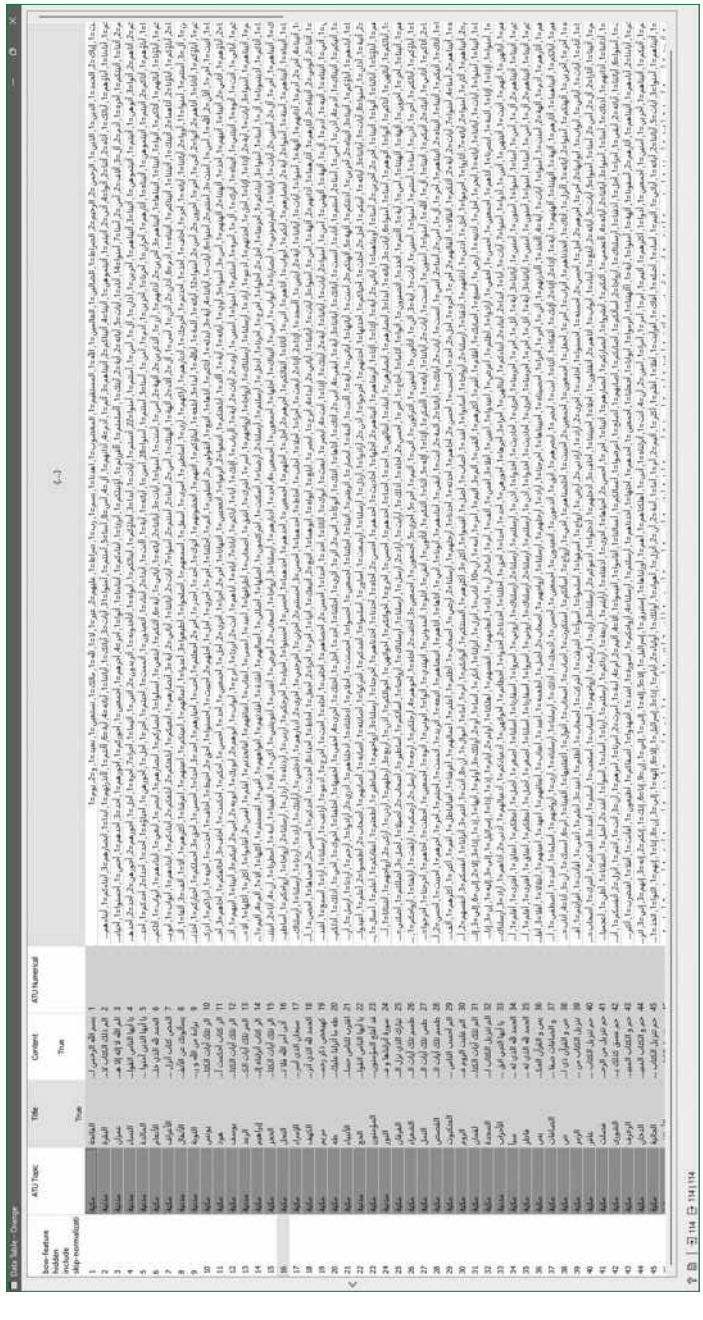
شكل ٨٥ : نافذة أداة حقيبة الكَلِمَات

جدول ٢٩ : حالات نموذج حقيبة الكَلِمَات

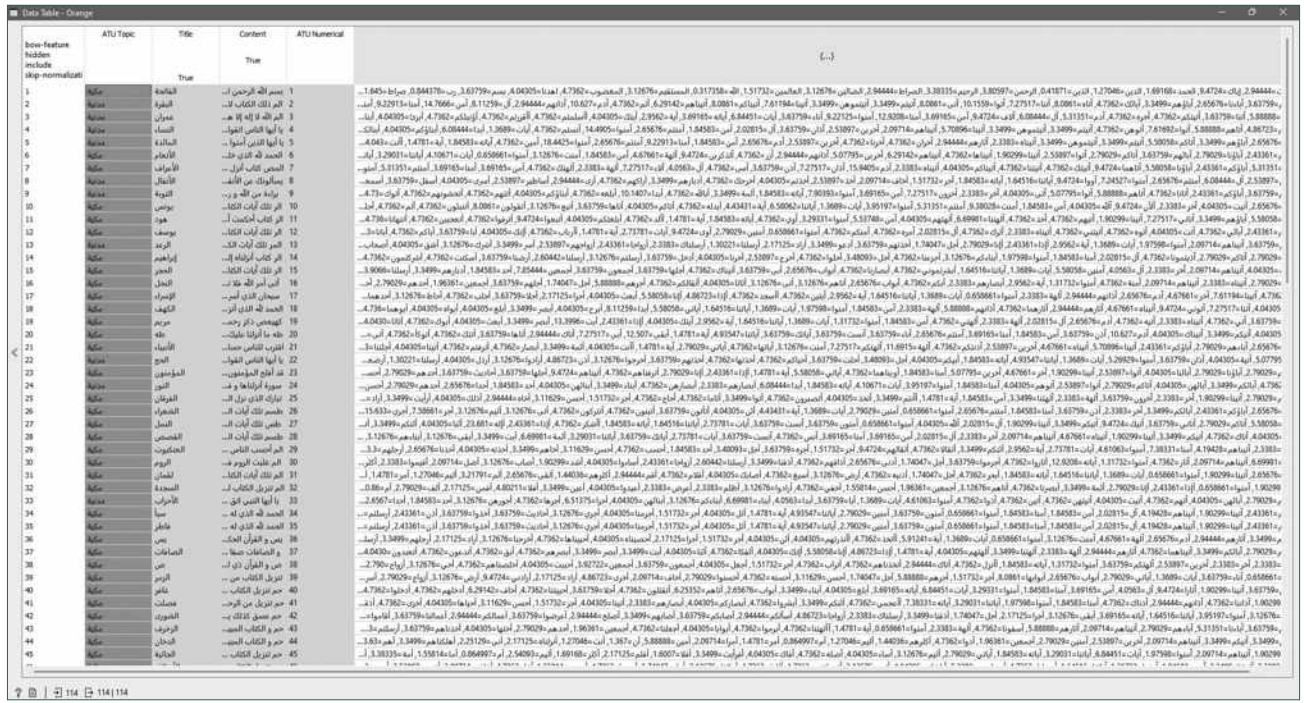
تكرار الكلمة Term Frequency: العدد count، الثنائي binary، الخطي الجزئي sublinear
تكرار المستند Document Frequency: لا شيء none، تكرار المستند العكسي IDF، تكرار المستند العكسي الخفيف
التنظيم Regularization: لا شيء none، المستوى الأول (مجموع العناصر) L1، المستوى الثاني (الإقليدية) L2: Euclidean

الحالة الأولى	الحالة الثانية	الحالة الثالثة	الحالة الرابعة
تكرار الكلمة: العدد تكرار المستند: لا شيء التنظيم: لا شيء	تكرار الكلمة: العدد تكرار المستند: تكرار المستند العكسي التنظيم: لا شيء	تكرار الكلمة: العدد تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الأول	تكرار الكلمة: العدد تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الثاني
الحالة الخامسة	**الحالة السادسة**	**الحالة السابعة**	**الحالة الثامنة**
تكرار الكلمة: الثنائي تكرار المستند: لا شيء التنظيم: لا شيء	تكرار الكلمة: الثنائي تكرار المستند: تكرار المستند العكسي التنظيم: لا شيء	تكرار الكلمة: الثنائي تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الأول	تكرار الكلمة: الثنائي تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الثاني
الحالة التاسعة	**الحالة العاشرة**	**الحالة الحادية عشرة**	**الحالة الثانية عشرة**
تكرار الكلمة: الخطي الجزئي تكرار المستند: لا شيء التنظيم: لا شيء	تكرار الكلمة: الخطي الجزئي تكرار المستند: تكرار المستند العكسي التنظيم: لا شيء	تكرار الكلمة: الخطي الجزئي تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الأول	تكرار الكلمة: الخطي الجزئي تكرار المستند: تكرار المستند العكسي التنظيم: المستوى الثاني

يعمل نموذج حقيبة الكَلِمَات وفق اختيار الحالة الأولى على ترقيم الكَلِمَات ترقيمًا لا يفيد تعلّم الآلة إفادةً تتعلَّق باللفظ والمعنى؛ كونها تعطي الكَلِمَات قيمَ تكراراتها. والأدق هنا هو باستخدام الحالة الثَّانية، التي بها يعملُ النَّموذج على استخلاص قيمة تكرار الكلمة مع تكرار المستند العكسِيِّ (٧٨)، بدلًا من جعل تكرار المستند (لا شيء) كما في الشَّكْل (٨٦).



شكل ٨٦ : تكرار الكَلِمَات (count)؛ تكرار المستند العكسي (لا شيء)



شكل ٨٧ : تكرار الكَلِمَات (count)؛ تكرار المستند (تكرار المستند العكسي IDF)

يعمل نموذجُ حقيبة الكَلِمَات المعتمد على معادلةِ تكرار الكلمةِ وتكرار المستندِ العكسيِّ على محاولة تقريبِ الكَلِمَات ذات الدّلالةِ المتشابهةِ، وهذه المحاولةُ تُعدُّ مرحلةً أوَّلِيَّةً شبه جيّدةٍ في تعلم الآلة الكَلِمَات في النُّصُوصِ، قبل تصميم نموذج حقيبةِ الكَلِمَات المستمِرَّة (continuous bag of words CBOW)، (المبحث الثَّاني: تضمين الكَلِمَات) الذي يُعدّ أفضل. لننظر إلى أسلوبِ النّموذج في محاولةِ تقريبِ الكَلِمَاتِ في مثال كلمة (سنة) وكلمة (عام).

المستند الأَوَّل: "وبلغ أربعين سنة"

المستند الثَّاني: "ثم يأتي من بعد ذلك عام فيه يُغاثُ الناس"

عند النَّظرِ إلى القيمة التي أُعطيت لكلمة (سنة) وكلمة (عام) نجدها متقاربةً فيما بينهما.

السياق	رقم السُّورَة	السورة	القيمة
وبلغ أربعين **سنة** قال رب أوزعني أن أشكر نعمتك التي أنعمت علي وعلى والدي	46.txt	الأحقاف	2.9714
ثم يأتي من بعد ذلك **عام** فيه يغاث الناس وفيه يعصرون	12.txt	يوسف	2.53897

أما نتائج تقييم الاختبار والقياس (Test and Score) لآلة المتَّجه الداعمةِ (Support Vector Machine) لتكرار الكلمة وتكرار المستند العكسي (التدريب-الاختبار ٨٠-٢٠٪)؛ فهي على النَّحو الآتي:

الصحة: 0.791؛ الدقة: 0.626؛ الاسترجاع: 0.791؛ مقياس-ف١: 0.699

في أداة إثراء الكَلِمَاتِ؛ سنلحظُ في النَّتيجة أنَّ كُلَّ الكَلِمَات ذات دلالةٍ إحْصَائيّة قويَّةٍ، وأقلّ من 0.01، ويستحيلُ أن تكون هناكَ مُدَوّنَةٌ لُغَوِيَّةٌ بشرِيَّةٌ ذات إثراءٍ في الكَلِمَات، وتكون كُلّ كلمةٍ فيها ذات دلالةٍ إحْصَائِيّةٍ أقل من 0.01 (الشكل ٨٨).

Word	p-value	FDR
آت	0.0e+00	0.0e+00
آتانا	0.0e+00	0.0e+00
أووا	0.0e+00	0.0e+00
آويناهما	0.0e+00	0.0e+00
آيتين	0.0e+00	0.0e+00
ا أسجد	0.0e+00	0.0e+00
أبا الله	0.0e+00	0.0e+00
أبلغه	0.0e+00	0.0e+00
أبناؤكم	0.0e+00	0.0e+00
أنتهم	0.0e+00	0.0e+00
أتخشونهم	0.0e+00	0.0e+00
أترفناهم	0.0e+00	0.0e+00
أتصبرون	0.0e+00	0.0e+00
أتوك	0.0e+00	0.0e+00
أتيناهم	0.0e+00	0.0e+00
أناما	0.0e+00	0.0e+00
أجاج	0.0e+00	0.0e+00
أجدر	0.0e+00	0.0e+00
أجعلتم	0.0e+00	0.0e+00
أجلب	0.0e+00	0.0e+00
أحبارهم	0.0e+00	0.0e+00
أحسنتم	0.0e+00	0.0e+00
أحملكم	0.0e+00	0.0e+00
أخباركم	0.0e+00	0.0e+00
أخذتم	0.0e+00	0.0e+00
آخرتن	0.0e+00	0.0e+00

شكل ٨٨ : إثراء الكَلِمَات

وعلى أَيِّ حالٍ؛ فمشكلةُ نموذج حقيبة الكَلِمَات تكمنُ في عدم تدقيق الآلة في الكَلِمَات ذات الدلالةِ المتشابهةِ، ووفق المجال الدّلالِيِّ الواحد (السِّمة الدلالة)، إذا كانت معتمدةً على تكرار الكَلِمَات وتكرار المستند العكسيِّ، دون اعتبارٍ لِلُعْبة التَّواردِ (co-occurrences) بين الكَلِمَات الأخرى، وهذا يصعبُ على نموذج حقيبة الكَلِمَاتِ، ولمعالجة هذا العيبِ؛ ابتُكِر تضمينُ الكَلِمَات الذي يعمل في مَهمَّته الأساسِيَّةِ على تقريب أبعاد الكَلِمَات المتشابهةِ

سـياقِيًّا وفـق السِّـمات، وإبعـاد تقريـبِ الكَلِمَـات غـير المتشـابهة سـياقِيًّا، بنـاءً علـى السِّـمات أيضًـا، وذلـكَ علـى مجمـلِ النَّـصِّ، وبلعبـةٍ رَقَمِيّـةٍ ذكِيَّـةٍ مـن الآلـة، تقـوم علـى لعبـة التَّـوارد المسـتمِرِّ للكلمـات المتشـابهة دلاليًّـا بعـد تقريـب أبعادِهـا، ثـم تقـوم علـى إبعـادِ هـذا التَّقريـب لبعـض الكَلِمَـاتِ عندمـا يكـون التـوارد معهـا غـير مسـتمِرٍّ. سـنرى عمـلَ تضمـين الكَلِمَـات وتضمـين المسـتندات في المبحـثِ الآتي، ويجـب أن نـدركَ أنَّ في تضمـين الكَلِمَـات مشـكلةً بسـيطةً، تتعلـق بالمشـترك اللَّفْظِـيّ، وسـتُبيَّن هـذه المشـكلة في مبحـث تضمـين الكَلِمَـات.

المبحث الثَّانِي: تضمين الكَلِمَات

يعمـلُ تضمـينُ الكَلِمَـات علـى تحويـلِ الكَلِمَـات المتباعـدةِ في النُّصُـوص إلى ناقـلاتِ أرقـامٍ ذات مُتَّجهـاتٍ متقاربـةٍ مـن أجـل تدريـبِ سـياقاتها، ويُسـتهدفُ مـن هـذا التَّحويـلِ تصنيـفَ التَّشـابه الـدّلاليِّ بـين الكَلِمَـاتِ، ويتمَـيَّزُ تضمـينُ الكَلِمَـات بـأنَّ المتّجهـات تكـونُ أقَـلَّ عـددًا مـن الكَلِمَـات؛ لأنَّ المتجهـات تعكـس متجهـات أبعـاد السِّـياقات في النُّصُـوص. يُطبّـق نمـاذج الكلمـةِ إلى المتَّجـه (word2vec) -جوجـل- أو نمـوذج المتجـه الكـونِيِّ ((Global Vector (GloVe) -جامعـة سـتانفورد- بمكتبـة جينسـم (Gensim)، أو نمـوذج تمثيـلاتِ المشـفِّر الثنـائيّ الاتِّجـاه مـن المحـوّلاتِ (Bidirectional Encoder Representations from Transformers BERT) في تدريـب البَيَانَـات اللُّغَويّـة. وشـكّل الأخـير ثـورةً هائلـة في فهـم الآلـة اللغـةَ بشـكلٍ أعمـق؛ لأن بـيرت (BERT) نمـوذج عميـق، تتعَلَّـم منـه الآلـة علـى اتجاهـين لسلسـلة الحـروف في النُّصُـوص، وليـس علـى اتجـاهٍ واحـد مثـل النمـاذج السَّـابقة، وذلـك بنمذجـةِ اللغـة المقنّعـةِ (Masked-Language Modeling)، وقـد تطـوّر نمـوذجُ بـيرت بأسـلوبٍ هنـدسيٍّ يَفْهَـمُ بـه تتابـع الكَلِمَـاتِ والعبـارات، ويفهـم هـذا التتابـع وسـياقاته فهمًـا ضمنيًّـا معمّقًـا؛ كونـه يتعلّـم بأسـلوب إخفـاء نسـبةٍ مـن الكَلِمَـات في النَّـصِّ؛ ليتنبـأ باسـتعمالاتها السـياقيّة. ظهـرت نمـاذج للغـة ذات متانـةٍ آليَّـةٍ تعلّميَّـةٍ، عُرفـت بمحـوّلات التَّدريـب التعـرُّفِيِّ التوليـدِيِّ (Generative Pre-trained Transformer GPT)، فظهـرَ GPT-1 (تحسـين فهـم اللغـة بالتدريـب التوليـدِيِّ)، وGPT-2 (نمـاذج متعـددة التعَلُّـم غـير الموَجَّـه)، وGPT-3 (نمـاذج اللغـة بالتعلّـم الأقـلّ).

هـذه النَّمـاذج المطـوَّرة تعمـلُ علـى بلايـين المقاييـس مـن الطَّبقـات الخفيَّـة والعصبونـات المحـددة، كمـا تعمـلُ علـى مـا يُسـمى بالاسـتدلال (inference)ـ الـذي بـه يُقـارِن أداءَ النمـاذج بعـد كُلّ مـرةٍ يُحسّـن فيهـا الأداء مع النَّمـاذج السَّـابقة. هـذه الطفـرة لـم تأخـذ سـنواتٍ طويلـة، بـل ظهـرت في السـنتين الأخيرَتـين فقـط، وبشـكلٍ سـريع ومتراكـم، وبصـورةٍ تُخبرنـا بـأن الـذَّكَاء الاصْطِنَاعِـيّ قـد بـدأ أوج التَّطْبِيقَـاتِ المتقَدِّمـة، وبـأَنَّ شـكلَه بعـد عشـر سـنواتٍ سـيكون بصـورةٍ سـتجعل الآلـة مثـل الإنسـانِ في فهـم اللغـة وإنتاجهـا سـياقِيًّا (وربمـا عاطفِيًّـا).

نعـودُ إلى نمـوذج (الكلمـة إلى متجـه word2vec)، فميزتُـه هـي أنـه يعمـل علـى مـا يُسـمى بتوزيـن الكَلِمَـات (أي: إعطائهـا أوزانهـا)، وفـقَ السـمات (المـيّزات أو الخصائـص features)، وتصـل عـدد السـمات إلى ٣٠٠ أو ٤٠٠ أو ٥٠٠ للنُّصـوص الـتي يبلـغ حجمهـا ٥٠ ألـف كلمـة. وتتمَـيَّزُ هـذه السـماتُ بتحديـد الكَلِمَـات ذات السِّـمة الدلاليَّـة المشـتركة. لـو حددنـا للآلـة عـددًا معَيَّنًـا مـن السِّـماتِ؛ فإنَّهـا في اسـتخلاص كُلّ سـمةٍ تعتمـد علـى لعبـة التَّـواردِ (co-occurrence)، الـتي تكشـف الكَلِمَـات مـن حيـث وضعهـا الـدلالي النَّـصِّيّ. لاحـظ أن السـمة تُحـدَّدُ بالأرقـام، فالآلـةُ لـن تعتبرسـمات كلمـة (الرجـل): [+ذكـر] و[-أنثـى] و[+بالغ]، بـل بقيمـةٍ رَقَمِيّـةٍ تكشـف المعـاني الضِّمنيـة لهـا. لنفـترض هـذا المثـال: (أنجبـتْ الأميـرةُ ابنـةُ الملـكِ ابنًـا). سـتتحول هـذه الكَلِمَـات إلى مُتَّجهـات، كُلّ متجـهٍ يَتَضَمَّـن شـعاعًا رَقَمِيًّـا، وتكـون القيـم الرَّقَمِيّـة للسِّـمة الواحـدة متطابقـةً أو متقاربـةً جِـدًّا، وفي المصفوفـة الآتيـة توضيـحٌ لهـذه النُّقطـة.

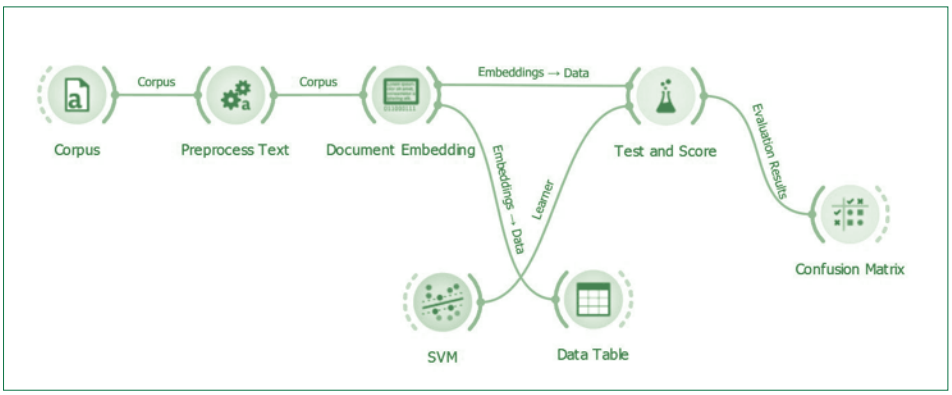
السمات	ابنا	الملك	ابنة	الأميرة	أنجبت
ذكر	0.99	0.99	0.01	0.01	0.01
أنثى	0.01	0.01	0.99	0.99	0.90
العمر	0.20	0.90	0.70	0.70	0.0.60

تعمـل نمـاذج (الكلمـة إلى متجـه word٢vec) لتعلّـم تضمـين الكَلِمَـات علـى نموذجـين أساسـيّين. الأَوَّل: تخطـي الكلمـة (Skip-Gram Model)، والثَّـانِي: نمـوذج حقيبـة الكَلِمَـات

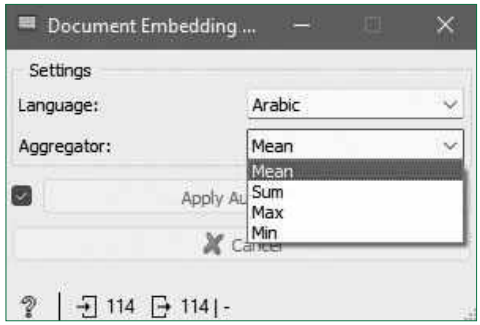
المستمِرَّة (Continuous BOW). والفرق بين هذين النَّموذجين هو أنَّ الأوَّل يتنبأ بالسِّياق من خلال الكلمة، والثاني يتنبأ بالكلمة من خلال السِّياقِ، ويتحدَّدُ سياقُ الكلمة بالكلمة الواردةِ معها في نافذة التَّصَاحُبِ، سواءً كان على مدى كلمةٍ متصاحبةٍ واحدةٍ أو أكثر، وتتحدَّد الكلمة بكلمة السِّياق الواردة معها. لو افترضنا كلمة (أنجبت) فكلمات السِّياق لها تتحدَّدُ بالشُّعاع الرَّقَمِيّ الأعلى، وتكون عندئذ (الأميرة) و(ابنة).

تمثَّلت مشكلة في نماذج (الكلمة إلى متجه word٢vec) عند التعامل مع الكلمة الغنِيَّة بالتَّصريف وبالاشتراك اللَّفْظِيّ، وعليه طوّرت شركةُ فيسبوك نموذجَ (FastText)؛ لمحاولة معالجة مشكلة التَّشابه اللَّفْظِيّ، ثم طُوِّرت شبكات الذَّاكرة القصيرة-الطويلة المدى (Long Short-Term Memory (LSTM))، وكلاهما ليسا مدمجين في أدواتِ أورانج، علمًا بأن هذه الأداة قد تدمج بعض النماذج التي تصل إلى المرحلة الأخيرة من التَّطوير في كُلّ مرحلةٍ مستقبلِيَّةٍ.

وفيما يلي تطبيقٌ على مُدَوّنَة القرآن الكريم باستعمال تضمينِ المستندات في أداة تنقيبِ النُّصُوص (أورانج). ويتَّضح مسارُ العمل كما في الشَّكْل (٨٩)، فتُحمَّل البَيَانَات (الشكل ٨٠) في أداة الـمُدَوّنَةِ (Corpus) أوَّلًا.



شكل ٨٩ مسار العمل لتضمين المستندات



شكل ٩٠ : اختيار اللغة وأسلوب الجامع Aggregator لتضمين المستندات

عدد المخرجات: المتوسط mean، والمجموع sum، والأكبر max، والأقل min

mean

يُمكن اختيار الحساب بالمتوسط (mean)، أو بالمجموعِ (sum)، أو بالأعلى (max)، أو بالأقلّ (min) من أداة تضمين المستنداتِ (Document Embedding) كما في الشَّكْل (٩٠). يُصنِّفُ تضمينُ المستند كُلَّ سورةٍ إلى مستندٍ، ثم يُضمّن كُلّ كلمةٍ (n-gram) باستخدام نموذج التضمين المدرّب والمختبر مسبقًا وفق اللغة، وللغة العَرَبِيَّة نموذج تم اختبارُه مسبقًا، ثم يتضمّنُ كُلّ مستندٍ على متّجهٍ واحدٍ بَعْدَ تجميعِ كُلّ تضمينات الكَلِمَات فيه، وذلك باستخدام المجمّع القائم على المتوسّط، أو المجموع، أو الأعلى، أو الأقَلّ. في الشَّكْل (٩١) عرضٌ لنتيجة استخلاص تضمين المستندات. يَتَضَمَّنُ الشَّكْلُ (٩١) تضمينَ المستندات على ٣٠٠ بُعْد، وبمعامل التجميع aggregator القائم على المتوسّط (mean). تعَلَّمت الآلة تضمين نصوص (مستندات) آياتِ سور القرآن بأسلوبِ التَّعَلُّم العميق، وبنموذج المستند إلى متجهة (doc۲vec)، وعَمَّمت الآلة هذا التضمين لكل مستند بناءً على الكَلِمَات ومتَّجهاتها، لكن ماذا عن الكلمة إلى متَّجه (word۲vec) في داخل كلِّ مستند؟ سأجري تطبيقًا لكلمة (الهدى) وفقَ نتيجة حقيبة الكَلِمَات المستمَرَّة (CBOW)؛ كونها أحَدَ نماذج تضمينِ الكَلِمَات.

Data Table (1) - Orange

	ATU Topic	Title	Content	ATU Numerical	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Dim11	Dim12	Dim13	Dim14
embedding-featu hidden include skip-normalizatio					True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True	True True True
		True	True															
1	مكية	الفاتحة	بسم الله الرحمن ا...	1	-0.0322154	0.00770556	-0.00747347	0.0214129	0.166065	0.00609593	-0.00285352	0.0428541	0.00531364	-0.0845411	0.0255575	-0.125445	0.017596	0.0424457
2	مدنية	البقرة	الم ذلك الكتاب لا...	2	-0.0197762	0.00465462	-0.0305588	0.0242981	0.176814	-0.0187581	0.00322273	0.00346344	-0.0079301	-0.0502327	0.0157738	-0.0925727	0.0234735	0.0393741
3	مدنية	عمران	الم الله لا إله إلا هـ...	3	-0.0190847	0.00871096	-0.0260632	0.0228832	0.179296	-0.0169227	0.00185891	0.0013928	-0.00924583	-0.0526764	0.00910354	-0.095347	0.0226077	0.0413699
4	مدنية	النساء	يا أيها الناس اتقوا...	4	-0.0238352	-0.00363797	-0.0281674	0.0231995	0.177337	-0.0122296	4.41137e-05	-0.00248652	-0.00650883	-0.0530638	0.0187563	-0.092267	0.0292103	0.0479443
5	مدنية	المائدة	يا أيها الذين آمنوا ...	5	-0.0145054	-0.0029846	-0.0347687	0.0254266	0.175078	-0.0153851	0.00613394	-0.00122198	-0.00310161	-0.0539906	0.0138976	-0.0957587	0.028286	0.0420384
6	مكية	الأنعام	الحمد لله الذي خلـ...	6	-0.0227839	0.0085449	-0.0290564	0.0240584	0.176704	-0.0181022	0.00901942	-0.00107279	-0.00350998	-0.0537146	0.0103047	-0.0928097	0.0214727	0.0420531
7	مكية	الأعراف	المص كتاب أنزل ...	7	-0.0192294	0.00514496	-0.0245553	0.0216991	0.17681	-0.0200132	0.00517339	0.00250091	-0.00831522	-0.0566941	0.00815078	-0.0912694	0.0211282	0.0390285
8	مدنية	الأنفال	يسألونك عن الأنفـ...	8	-0.0148057	0.00678119	-0.0254293	0.021509	0.17934	-0.0197867	0.00295419	-0.00140237	-0.00944614	-0.0506779	0.0141988	-0.0933201	0.0199021	0.0443371
9	مدنية	التوبة	براءة من الله و ر...	9	-0.0105868	-0.00134096	-0.0193422	0.0240821	0.166676	-0.0162566	0.00885376	-0.000510728	-0.00284509	-0.0520356	0.0139811	-0.089026	0.0195487	0.0386139
10	مكية	يونس	الر تلك آيات الكتا...	10	-0.0243436	0.00733164	-0.0260478	0.0190508	0.180258	-0.0222508	0.0014692	-1.13832e-05	-0.00873515	-0.0480781	0.00477202	-0.0957791	0.0145417	0.0402231
11	مكية	هود	الر كتاب أحكمت آ...	11	-0.0239246	0.00331008	-0.0244065	0.0273675	0.18788	-0.0180369	0.0057866	0.00755298	-0.000833117	-0.0599474	0.00653778	-0.106882	0.0108262	0.0425269
12	مكية	يوسف	الر تلك آيات الكتا...	12	-0.0147304	0.000993619	-0.0276792	0.0254912	0.182499	-0.0143491	-0.00602517	0.01355	-0.000598914	-0.0531542	0.00712974	-0.109329	0.0177646	0.0324784
13	مدنية	الرعد	المر تلك آيات الكـ...	13	-0.0179995	0.00671868	-0.0232792	0.0279167	0.178485	-0.0205348	0.00197298	-0.00223479	-0.0115296	-0.0526989	0.00713951	-0.0892689	0.0211789	0.0400224
14	مكية	إبراهيم	الر كتاب أنزلناه إلـ...	14	-0.0243504	0.00511156	-0.0242782	0.0231353	0.181407	-0.0206734	0.00582666	0.00426694	-0.00955459	-0.0521134	5.10756e-05	-0.0974812	0.0126373	0.0352896
15	مكية	الحجر	الر تلك آيات الكتا...	15	-0.0207894	0.00626725	-0.0244769	0.0213009	0.174725	-0.00805298	0.00786482	0.00889534	-0.00906837	-0.0562889	0.00690084	-0.0867904	0.00984518	0.0346213
16	مكية	النحل	أتى أمر الله فلا تـ...	16	-0.0216461	0.00133994	-0.0242711	0.0238371	0.175438	-0.0191148	0.00910472	0.00396791	-0.00690312	-0.0537627	0.00819557	-0.0909517	0.015161	0.0425464
17	مكية	الإسراء	سبحان الذي أسرى...	17	-0.0336497	0.00469734	-0.0261737	0.0230243	0.174334	-0.0150787	-0.00530609	0.00334921	-0.00793817	-0.0540578	0.00968549	-0.0897407	0.0285773	0.0418333
18	مكية	الكهف	الحمد لله الذي أنزـ...	18	-0.0227424	0.000498566	-0.01669	0.0234949	0.185105	-0.0180758	-0.00419741	-0.00378203	-0.00923997	-0.0557484	0.0127958	-0.0888333	0.038148	0.0421282
19	مكية	مريم	كهيعص ذكر رحمـ...	19	-0.0303834	-0.000404816	-0.0306535	0.026679	0.201374	-0.0131122	-0.0146264	0.0125913	-0.0023717	-0.0537171	0.00541176	-0.112281	0.022073	0.0421834
20	مكية	طه	طه ما أنزلنا عليك...	20	-0.0186167	0.0150165	-0.0300759	0.0279175	0.191606	-0.0131404	-0.0145327	0.00459298	-0.006666	-0.0609553	0.0108347	-0.100656	0.0248403	0.0352697
21	مكية	الأنبياء	اقترب للناس حسابـ...	21	-0.0211734	0.0114674	-0.0293391	0.0229449	0.188181	-0.0152826	0.00882183	0.00652386	-0.00661933	-0.055032	0.0040802	-0.0853414	0.0162977	0.0405595
22	مدنية	الحج	يا أيها الناس اتقوا...	22	-0.0198884	0.003761	-0.0262992	0.0278111	0.184944	-0.0170362	0.0098932	-0.000628416	-0.0097048	-0.0522845	0.00774017	-0.0932931	0.0170076	0.0431507
23	مكية	المؤمنون	قد أفلح المؤمنون...	23	-0.0192488	0.00955472	-0.0255198	0.0217748	0.171867	-0.0160039	0.00195933	0.0097854	-0.00595844	-0.0540488	0.00710308	-0.0854379	0.00939484	0.0458582
24	مدنية	النور	سورة أنزلناها و فـ...	24	-0.020558	-0.00568716	-0.0225914	0.0255799	0.170112	-0.0145546	-0.000639043	-0.00143439	-0.00696326	-0.0551133	0.00814659	-0.086448	0.029014	0.0405691
25	مكية	الفرقان	تبارك الذي نزل الـ...	25	-0.0302826	0.00424995	-0.0231078	0.0258865	0.182196	-0.0197467	0.00611472	0.00140658	-0.0122758	-0.0649796	0.0091237	-0.0956383	0.027077	0.0414677
26	مكية	الشعراء	طسم تلك آيات الـ...	26	-0.00912551	0.00429706	-0.0289238	0.0239082	0.17072	-0.0110482	-0.000199615	0.0084382	-0.00578799	-0.0573113	0.00974368	-0.0933905	0.0174467	0.0367151
27	مكية	النمل	طس تلك آيات الـ...	27	-0.0207324	0.00458998	-0.0286024	0.0231705	0.183032	-0.0161138	0.00287069	0.00190911	0.000352777	-0.0523496	0.00578953	-0.095025	0.00864866	0.046281
28	مكية	القصص	طسم تلك آيات الـ...	28	-0.0239132	0.0116494	-0.0285943	0.0270314	0.179196	-0.00890563	-0.00832085	0.00560615	-0.00276772	-0.0542343	0.00453473	-0.101039	0.0252023	0.0441368
29	مكية	العنكبوت	الم أحسب الناس ...	29	-0.0235818	0.0059938	-0.0212998	0.0234625	0.163852	-0.0219146	0.00738424	-0.00212782	-0.00705288	-0.0541598	0.00874245	-0.0811747	0.0192855	0.0353164
30	مكية	الروم	الم غلبت الروم فـ...	30	-0.0216659	0.010667	-0.0274116	0.0253087	0.165892	-0.021761	0.00795388	-0.000775464	-0.00995211	-0.0423886	0.00808402	-0.0785255	0.0151895	0.0435557
31	مكية	لقمان	الم تلك آيات الكتا...	31	-0.0306063	0.00353928	-0.0277134	0.0335022	0.196041	-0.011602	-0.00161032	0.0094291	-0.0110735	-0.0542619	0.0052416	-0.106403	0.0157231	0.0401951
32	مكية	السجدة	الم تنزيل الكتاب لـ...	32	-0.0249528	0.0128555	-0.0288407	0.0266945	0.177602	-0.0158242	0.0104986	-0.00176728	-0.0105474	-0.0571745	0.00480448	-0.0907702	0.0147336	0.0404341
33	مدنية	الأحزاب	يا أيها النبي اتق ...	33	-0.028333	-0.00241417	-0.0288374	0.0203778	0.172341	-0.0157797	-0.00632328	-0.0030677	-0.0136286	-0.0523346	0.018875	-0.083814	0.0314996	0.0423017
34	مكية	سبأ	الحمد لله الذي له ...	34	-0.0267157	0.00667481	-0.030881	0.0273881	0.182188	-0.0199967	0.00713983	0.00455014	-0.0103878	-0.0537487	0.00926307	-0.100349	0.023184	0.0467879
35	مكية	فاطر	الحمد لله الذي له ...	35	-0.0267157	0.00667481	-0.030881	0.0273081	0.182188	-0.0199967	0.00713983	0.00455014	-0.0103878	-0.0537487	0.00926307	-0.100349	0.023184	0.0467879
36	مكية	يس	يس و القرآن الحكـ...	36	-0.0268257	0.00633621	-0.0234825	0.0268127	0.171633	-0.0193858	0.00672163	0.00607878	-0.00714002	-0.0554336	0.00315158	-0.0911643	0.0148249	0.0414311
37	مكية	الصافات	و الصافات صفا ...	37	-0.0156823	0.00272557	-0.0202216	0.0261578	0.166194	-0.00991676	0.00152965	0.016525	-0.00304275	-0.0542752	0.00217074	-0.0876219	0.00937046	0.0446157
38	مكية	ص	ص و القرآن ذي ذكـ...	38	-0.0178455	0.0101347	-0.025938	0.0270011	0.187875	-0.0156018	-0.00999914	0.0109127	0.000722503	-0.0544523	-0.00134669	-0.095721	0.0150147	0.032591
39	مكية	الزمر	تنزيل الكتاب من ...	39	-0.0330495	0.00451301	-0.0234082	0.0247547	0.187707	-0.0187444	0.00386395	-0.00527445	-0.00987718	-0.0609562	0.00378972	-0.0847783	0.0197172	0.0358553
40	مكية	غافر	حم تنزيل الكتاب ...	40	-0.0252067	0.00791651	-0.0233938	0.0255703	0.189208	-0.0189233	0.00407634	0.00797087	0.000685536	-0.0506397	0.000844253	-0.103551	0.0152769	0.0413854
41	مكية	فصلت	حم تنزيل من الرحـ...	41	-0.0280947	0.00886901	-0.0273032	0.022731	0.186062	-0.0182647	0.00402767	0.00421103	-0.0133862	-0.0572619	-0.000531108	-0.0822301	0.01731	0.0459326

114 | 114

شكل ٩١ : نتيجة تضمين المستندات

جـاءَت كلمـةُ (هـدى) وكلمـةُ (الهـدى) في القـرآن الكريـم بمعـانٍ عديـدةٍ؛ وهـي: الإيمـان، والطَّريـق، والكتـاب المـنزَّل، ومـن أمثلتهـا مـن الآيـات الكريمـة مـا يـأتي:

قولـه تعـالى ﴿قُلْ أَنَدْعُواْ مِن دُونِ ٱللَّهِ مَا لَا يَنفَعُنَا وَلَا يَضُرُّنَا وَنُرَدُّ عَلَىٰٓ أَعْقَٰبِنَا بَعْدَ إِذْ هَدَىٰنَا ٱللَّهُ كَٱلَّذِى ٱسْتَهْوَتْهُ ٱلشَّيَٰطِينُ فِى ٱلْأَرْضِ حَيْرَانَ لَهُۥٓ أَصْحَٰبٌ يَدْعُونَهُۥٓ إِلَى ٱلْهُدَى ٱئْتِنَا ۗ قُلْ إِنَّ هُدَى ٱللَّهِ هُوَ ٱلْهُدَىٰ ۖ وَأُمِرْنَا لِنُسْلِمَ لِرَبِّ ٱلْعَٰلَمِينَ ﴿٧١﴾﴾ [الأنعـام: ٧١]. المعـنى: الإيمـان.

قولـه تعـالى: ﴿إِذْ رَأَى نَارًا فَقَالَ لِأَهْلِهِ امْكُثُوا إِنِّي آنَسْتُ نَارًا لَعَلِّي آتِيكُم مِّنْهَا بِقَبَسٍ أَوْ أَجِدُ عَلَى النَّارِ هُدًى﴾ [طه: ١٠]. المعـنى: الطَّريـق.

قولـه تعـالى: ﴿إِذْ رَءَا نَارًا فَقَالَ لِأَهْلِهِ ٱمْكُثُوٓاْ إِنِّىٓ ءَانَسْتُ نَارًا لَّعَلِّىٓ ءَاتِيكُم مِّنْهَا بِقَبَسٍ أَوْ أَجِدُ عَلَى ٱلنَّارِ هُدًى ﴿١٠﴾﴾ [السَّـجْدَة: ٢٦]. المعـنى: البيـان والوضـوح.

قولـه تعـالى: ﴿وَلَقَدْ ءَاتَيْنَا مُوسَى ٱلْهُدَىٰ وَأَوْرَثْنَا بَنِىٓ إِسْرَٰٓءِيلَ ٱلْكِتَٰبَ ﴿٥٣﴾﴾ [غافـر: ٥٣]. المعـنى: الكتـاب المـنزّل.

عندمـا دقَّقـتُ النَّظـر في عمـل نمـوذج حقيبـة الكَلِمَـات المسـتمرَّة، والمتجهـات الـتي مثَّلتهـا كلمـة (هـدى) و(الهـدى) مـن سـورة الأنعـام وطـه والسَّـجْدَة وغافـر؛ ظهر أنَّ بـين تلـك المتَّجهات مطابقـةً، وفقًـا للمعـنى في السُّـورَة الواحـدةِ، وتقارُبًـا وفقًـا للمعـنى بـين هـذه السُّـورِ. وفيمـا يلـي عـرضٌ كامـلٌ لنتـائج نمـوذج حقيبـة الكَلِمَـات المسـتمِرَّة لهـذه الكَلِمَـات في تلـك السُّـور.

سياق الكلمة	رقم السُّورَة	السورة	الآية	المعنى = المتجه vectors
لـو شـاء الله لجمعهـم علـى الهـدى فـلا تكونـن مـن الجاهلـين	6.txt	الأنعام	٣٥	الهدى (الإيمان) = ٥,٧٠٨٩٦ هدى (الطريق) = ٣,١٢٦٧٦ هدى* (الكتاب، الفرقان) = ٩,٣٤٨٨٧
قـل إن هـدى الله هـو الهـدى وأمرنـا لنسـلم لـرب العالمين	6.txt	الأنعام	٧١	
قـل إن هـدى الله هـو الهـدى وأمرنـا لنسـلم لـرب العالمين	6.txt	الأنعام	٧١	
ذلك هدى الله يهدي به من يشاء	6.txt	الأنعام	٨٨	
أولئك الذين هدى الله فبهداهم اقتده	6.txt	الأنعام	٩٠	
جاء به موسى نورا وهدى* للناس	6.txt	الأنعام	٩١	
وتفصيـلا لـكل شيء وهـدى* ورحمـة لعلهـم بلقـاء ربهـم يؤمنـون	6.txt	الأنعام	١٥٤	
جاءكم بينة من ربكم وهدى* ورحمة	6.txt	الأنعام	١٥٧	
أو أجد على النار هدى	20.txt	طه	١١	
والسلام على من اتبع الهدى	20.txt	طه	٤٧	الهدى (الكتاب) = ١,٩٠٢٩٩ هدى (الرشاد) = ٧,٧٩٠٧٢
أعطى كُلّ شيء خلقه ثم هدى	20.txt	طه	٥١	
وأضل فرعون قومه وما هدى	20.txt	طه	٨٠	
اجتباه ربه فتاب عليه وهدى	20.txt	طه	١٢٢	
فإما يأتينكم مني هدى فمن اتبع هداي فلا يضل ولا يشقى	20.txt	طه	١٢٣	
أولم يهد لهم كم أهلكنا قبلهم من القرون	32.txt	السَّجْدَة	٢٦	يهد (الهداية) = ٢,٧٩٠٢٩
ولقد آتينا موسى الهدى	40.txt	غافر	٥٣	الهدى (الكتاب) =١,٩٠٢٩٩
هدى وذكرى لأولي الألباب	40.txt	غافر	٥٤	هدى (الهداية) = ١,٥٥٨١٤

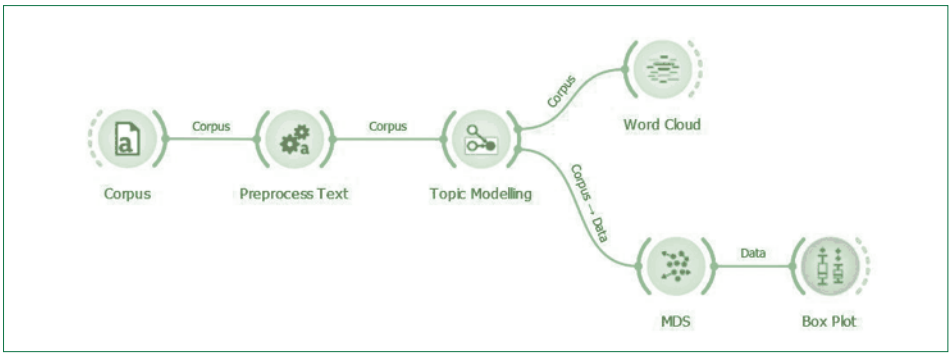
الفصل الثَّالِث
استخراج الموضوعات والكَلِمَات المفتاحِيَّة

المبحث الأَوَّل: نمذجة الموضوعات

في أداة نمذجة أو استخراج الموضوعات في أداة (أورانج) ثلاثُ خوارزمِيَّاتٍ مستخدمَة لإجراء عملية الاستخراجِ؛ وهي: الفهرسة الدلالية الكامنة (Latent Semantic Indexing)، وتخصيص دركليه الكامن (Latent Dirichlet Allocation)، ومعالجة دركليه الهرمِيَّة (Hierarchical Dirichlet Process).

وفيما يَأتِي تطبيقٌ على مُدَوّنَة القرآن الكريم باستخدام أداة تنقيب النُّصُوص (أورانج). يُشير الشَّكْل (٩٢) إلى مسار عمل المعالجة لاستخراج الموضوعاتِ، ويُبدَأ بتحميل البَيَانَات (الشكل ٨٠) في أداة المُدَوّنَةِ (Corpus)، ويحتوي الملفُّ على ٣٠ سطرًا، وفي كُلّ سطرٍ نَصٌّ يُمثّل سورةً قرآنِيَّةً كاملةً[١]. تَضَمّنَ الشَّكْل (٩٣) ٣٠ موضوعًا استُخرج بنموذج الفهرسة الدلالِيّة الكامنة مع عشر كلماتٍ ذات علاقةٍ بكلِّ موضوع، واحتوى الشَّكل (٩٤) على ٣٠ موضعًا استُخرج بنموذجِ تخصيص دركليه الكامن مع عشر كلماتٍ ذات علاقةٍ بكلِّ موضوع. واشتمل الشَّكل (٩٥) على ٣٠ موضوعًا استُخرج بنموذج معالجة دركليه الهرمِيَّة مع عشر كلماتٍ ذات علاقةٍ بكلِّ موضوع.

(١) لو ابتغيتُ تمثيلَ موضوعات القرآن بِدقَّةٍ؛ فالأجدر أن يكون كُلُّ نَصٍّ واحد عبارة عن آيةٍ واحدةٍ، وليس سورةً كاملةً، وتُستخرج الموضوعات على ثلاث مراحِل؛ الأولى: استخراج كُلّ الموضوعاتِ من كُلّ السُّور. والثانية: استخراج كُلّ الموضوعات من كُلّ سورةٍ واحدةٍ. والثَّالثة: استخراج الموضوعاتِ من كُلّ السُّوَر المكِّيَّةِ، وتلك الموضوعات من كُلّ السُّوَرِ المدَنِيَّة.



شكل ٩٢ مسار عمل استخراج الموضوعات وتمثيلها

أُذكّر هُنا أنَّ عدد الموضوعات يتحَدَّدُ بحسب رغبةِ المحلّل، غير أن هناك طريقةً أفضل لتحديد العددِ الأمثل، غير متوفرةٍ في أداةِ أورانج، ويمكنُ الاطّلاع عليها في مصدر آخر[١].

كيف يُمكن استنتاجُ الموضوعات الثَّلاثين من كُلّ نموذجٍ من النَّماذج الثَّلاثة؟ يُعرف ذلكَ عن طريق منهجين؛ **الأوَّل**: أن يُعرف موضوع النَّصّ العام، فالنَّصّ القرآني له موضوعات قد جُمعت في المراجع والمصنّفات التَّفسيريَّة، فمعرفةُ الموضوعات المنَوَّعة وأيهما أكثر وأيهما أقَلّ سيكون معينًا على تحديد موضوعِ كُلّ رقمٍ وفق الكَلِمَات ذات الارتباط الموضوعيّ به. **والثَّاني**: أن يُعتمدَ على تحديد الموضوعاتِ لكُلِّ رقمٍ من ١ إلى ٣٠ وفق الكَلِمَاتِ الواردة في كُلّ موضوعٍ، وهذا يعتمدُ على تفسيريَّة المحَلّل البشري.

سأضربُ مثالًا مُبسَّطًا للمنهج الثاني لتوضيح النُّقطة قبل أن ننظر في الموضوعات والكَلِمَات المستخرجة من النَّصّ القرآني؛ لننظر إلى الموضوعِ الافتراضِيِّ (١) والموضوعِ الافتراضي (٢).

(١) الطلاب، الجامعة، الكُلِّيَّة، القسم.

(٢) الأطفال، الرَّوضة، الحضانة، التَّمهيدي.

---

(١) المجيول، سلطان. الفهرسة الدلالية الكامنة وتخصيص دركليه الكامن لنمذجة موضوعات الشعر العَرَبِيّ عبر العصور، ٥٧-٨٠.

سيتعيَّن أن يكون الموضوعُ الأوَّل متعلقًا بالدراسة الجامعِيَّة، وأن يتعلَّق الموضوع الثاني بدراسة الأطفال. وقد يُخطئ نموذج من نماذج الموضوع الآلِيَّة، فيضيف كلمةً للموضوع (١)، تكون بعيدةً نوعًا ما عن الحقل الموضوعي العامِّ، فيضيف -على سبيل المثال- كلمةَ (القطار). وقد تكون كلمة (قطار) غير بعيدةٍ بحسب السِّياق النَّصِّيِّ الواردة فيه مع الكَلِمَاتِ الأخرى، فتكون مرتبطةً موضوعيًّا بالكَلِمَات الواردة في الموضوع (١)؛ بسبب ارتباطها جميعًا -مثلًا- بسياق ذهابِ الطُّلاب إلى الأقسامِ أو الكُلِّيات بالقطار في كُلّ صباحٍ ومساء. تتطوَّرُ نماذجُ الآلة لاستخراج الموضوعاتِ في كُلّ مرحلةٍ زمنِيَّةٍ لاحقةٍ، وهي تكشف الارتباط الموضوعيَّ بين الكَلِمَاتِ بأساليب آليَّةٍ مختلفةٍ، ومن أدَقّها حتى الآن هو أسلوبُ (الكلمة إلى متّجه word2vec).

Latent Semantic Indexing الفهرسة الدلالية الكامنة **عدد الموضوعات : ٣٠**	
الأرض، آمنوا، الكتاب، الناس، شيء، كفروا، بالله، أنزل، لله، خير	1:
الأرض، آمنوا، بالله، موسى، قوم، فلما، رسوله، ربكم، الكتاب، فرعون	2:
بالله، قوم، رسوله، عظيما، المؤمنين، عليما، سبيلا، كفى، رحيما، خير	3:
تكذبان، فبأي، ربكما، آلاء، إنس، فاكهة، جان، الميزان، خلق، قوم	4:
موسى، رسوله، كفروا، عذاب، الناس، قلوبهم، النساء، خير، عظيما، عليما	5:
رسوله، قوم، الكتاب، موسى، شيء، السماوات، لله، آمنوا، الأرض، قلوبهم	6:
الكتاب، رسوله، أهل، آمنوا، مريم، موسى، اتقوا، عليم، ربنا، قوم	7:
آمنوا، الأرض، أنزل، الكتاب، المؤمنين، اتقوا، ربنا، ابن، خير، قالت	8:
شيء، الأرض، الناس، كذبوا، أنزل، القوم، يؤمنون، الآيات، الكتاب، حرم	9:
يوسف، فلما، أبانا، أخيه، السجن، نفسه، قالت، الملك، المحسنين، الناس	10:
قوم، موسى، عذاب، ظلموا، جعلنا، رسوله، أمر، نوح، ربهم، الناس	11:
العالمين، مؤمنين، الرحيم، لهو، أكثرهم، لآية، أطيعون، فاتقوا، المرسلين، العزيز	12:
قوم، لقوم، الأرض، جعلنا، جاء، ربهم، للذين، يشاء، لله، شيء	13:
موسى، جعلنا، فرعون، الحق، الناس، الدنيا، جهنم، الرحمن، الحياة، سبيلا	14:
كفروا، القيامة، الحق، إله، أمة، لقوم، موسى، الناس، آمنوا، النبى	15:

16:	النبي، كفروا، القرآن، قليلا، كبيرا، ربه، أحدا، جاء، خير، وكيلا
17:	النبي، شيء، الناس، المؤمنين، بيوت، ملكت، يعلم، أعد، سبيل، حرج
18:	كفروا، الرحمن، أحدا، البحر، فهم، ربكم، أرسلنا، مبين، إله، بينهم
19:	النبي، عليم، بيوت، يشاء، الناس، عذاب، العذاب، الرسول، المؤمنين، موسى
20:	كفروا، الرحمن، ربكم، شيء، الكتاب، موسى، أفلا، بيوت، فرعون، سبيل
21:	الرحمن، كفروا، شيئا، الجن، فلما، نبيا، ذكر، قبلهم، صالحا، استضعفوا
22:	الناس، يشاء، السماوات، العزيز، الكتاب، أمة، النار، العذاب، خير، الأرض
23:	عذاب، الأرض، الرحمن، آمنوا، النار، آياتنا، فهم، آيات، العذاب، مكان
24:	الناس، بالحق، جميعا، أفلا، قالت، الحق، عليم، عباد، بينهم، كفروا
25:	الرحمن، نبيا، شيئا، للذين، الأرض، يقول، اذكر، فلما، فرعون، أفلا
26:	يشاء، أمن، سليمان، فهم، لله، الظالمين، إله، يومئذ، القرآن، تسمع
27:	الجحيم، كفروا، الآخرين، أكثرهم، أرسلنا، المخلصين، ذكر، يومئذ، يعلم، المحسنين
28:	القيامة، النار، بالله، بالبينات، المحسنين، قوم، عباد، يقولون، اليوم، أليس
29:	يشاء، آياته، يومئذ، النار، لآيات، ذكر، آمنوا، الليل، العذاب، ربكم
30:	الآخرين، الجحيم، نجزي، المحسنين، المخلصين، لله، المرسلين، يعلم، ألم، تركنا

شكل ٩٣ : نتائج الموضوعات باستخدام الفهرسة الدّلاليَّة الكامنة

في الموضوعاتِ المستخرجة من الفهرسة الدلاليَّةِ الكامنةِ (الشَّكل ٩٣)؛ سنلحظ أنَّ هناك موضوعاتٍ متكرِّرةً تتمحورُ حول قصَّةِ النَّبيِّ موسى مع فرعون، كما في الموضوع (2) و(14) و(20) و(25)، وحول أحوال النَّبيِّ موسى في الموضوع (2) و(5) و(6) و(7) و(11) و(14) و(15) و(19) و(20). لو دقَّقنا النَّظرَ في كُلّ موضوعٍ من غير هذه الموضوعاتِ؛ لألفينا ارتباطَ كثيرٍ من كلماتها بموضوعٍ واحدٍ.

لا تُستعملُ هذه الخوارزميَّةُ لعدم دِقَّتِها، فمن قامَ بتطويرِها (راديم زاهوزيك Radim Řehůřek)؛ قد أكَّد على هذه النُّقطة[1]، إذْ ستُظهِر كلمات لا يمكن تفسير

(1) Retrieved October 24, 2022, from https://orange3-text.readthedocs.io/en/latest/widgets/topicmodelling-widget.html

موضوعاتها؛ بسبب عملِ هذه الخوارزمية التي تلغي دقَّة التماسك في تحديدِ الموضوعات، وقد ثبتَ ذلك من تحليلاتٍ تطبيقِيَّةٍ، ومنها فيما يخصُّ التَّحْلِيلَ التَّطْبِيقِيّ على النُّصُوصِ العَرَبِيَّة، الذي أُجري على بياناتِ القصائد العَرَبِيَّة على مَرِّ العُصورِ[1].

إذن؛ تُستخدم مباشرةً إما الخوارزمِيَّة الثَّانية أو الثَّالثة. تعمل خوارزمِيَّةُ تخصيص دركليه الكامن على تخصيص النُّصُوص الأقرب اتِّجاهًا (الأقرب وَزنًا) للكلمات بين النُّصُوص. أما معالجة دركليه الهرمية فهي منهج بايزي (تصنيفيّ) غير مقياسِيٍّ، ويعمل بأسلوب التَّجْمِيع. سيُلحظ أنَّ معالجة دركليه الهرمِيَّة أفضل من تخصيص دركليه الكامن في نتائج الموضوعات المستخرَجة من مجلدات مُدَوّنَة القرآن الكريم.

Latent Dirichlet Allocation تخصيص دركليه الكامن

عدد الموضوعات: ٣٠

عدد الكلمات في جميع الموضوعات من غير تكرار: ١٤٥ كلمة من مجموع ٣٠٠ كلمة

1:	الأرض، جعلنا، أحدا، شيء، السماء، ربه، بالحق، دونه، رحمة، كفروا
2:	موسى، فرعون، فلما، الأرض، إله، جاء، القيامة، الدنيا، خير، أوتي
3:	الأرض، آمنوا، بالله، السماوات، عذاب، الكتاب، ربهم، العظيم، كثير، فاسقون
4:	آلاء، فبأي، ربكما، تكذبان، الأرض، آمنوا، كفروا، أعمالهم، فلما، السماوات
5:	ربنا، تطع، ربه، فضة، يوما، فهم، يعلمون، يدعون، تحكمون، مصبحين
6:	الأرض، موسى، شيء، فلما، ربنا، إله، فرعون، خير، أمة، للذين
7:	الأرض، أصحاب، كفروا، إله، العظيم، يعلم، أفلا، أنفسهم، إنكم، خير
8:	الأرض، آمنوا، شيء، كفروا، الكتاب، الدنيا، الناس، النار، آمنا، الآخرة
9:	الكتاب، الأرض، الناس، السماوات، كفروا، آمنوا، عذاب، الدنيا، ربنا، النار
10:	لهب، البلاد، فيقول، دكا، صفا، ابتلاه، يومئذ، الإنسان، الأرض، عذاب
11:	الأرض، آمنوا، ربهم، النبي، الكتاب، خير، يقولون، الناس، النار، تعملون
12:	نذير، خلق، الرحمن، البصر، أمن، السماء، السعير، شيء، عذاب، الأرض

(١) المجيول، سلطان. الفهرسة الدلالية الكامنة وتخصيص دركليه الكامن لنمذجة موضوعات الشعر العَرَبِيّ عبر العصور. مجلة اللسانيات العَرَبِيَّة، العدد ١٤، ٢٠٢٢، الصفحات ٥٧-٨٠.

13:	الأرض، ذكر، يعلم، مبين، آلهة، إله، جعلنا، إنكم، السماء، إبراهيم
14:	خلق، ربه، السماء، كذب، اسم، ماله، فسنيسره، كيدا، الطارق، الأعلى
15:	يومئذ، واحدة، رسول، الأرض، الحاقة، بقول، كريم، صيحة، اليوم، الجبال
16:	الأرض، الإنسان، يشاء، السماوات، القرآن، يومئذ، جعلنا، القيامة، الآخرة، كبيرا
17:	العالمين، موسى، قوم، فرعون، الأرض، العزيز، رسول، مؤمنين، أطيعون، فاتقوا
18:	الأرض، السماء، الجبال، جعلنا، عذاب، شيء، السماوات، يقولون، ربه، الجحيم
19:	شيء، الأرض، السماوات، لله، عليم، يشاء، بيوت، ربهم، عذاب، الكتاب
20:	آمنوا، الأرض، بالله، الناس، الكتاب، أنزل، كفروا، شيء، لله، ربكم
21:	الناس، سبيلا، الأرض، جعلنا، نذيرا، ربنا، السماوات، ملك، السماء، يقولون
22:	الأرض، موسى، الكتاب، الرحمن، السماوات، هدى، ربه، أهل، آمنوا، كفروا
23:	عذابي، مدكر، نذر، الأرض، كذبت، القرآن، قوم، شيء، أمر، يسرنا
24:	تعلمون، ألم، اليقين، يجعل، تر، يومئذ، سجيل، الجحيم، أرسل، النعيم
25:	الأرض، كتاب، يومئذ، آمنوا، النار، داوود، مآب، أواب، اذكر، الناس
26:	الأرض، قوم، الناس، عذاب، السماوات، كفروا، آمنوا، ربهم، أمر، جاء
27:	الأرض، آمنوا، شيء، كفروا، عذاب، عليم، السماوات، الناس، لله، الحق
28:	يوسف، فلما، الأرض، الناس، خير، أبانا، الملك، نفسه، السجن، أخيه
29:	يومئذ، المؤمنين، ويل، للمكذبين، الأرض، الآخرين، رسوله، السماء، مبين، المحسنين
30:	النبي، المؤمنين، رسوله، آمنوا، شيء، أعد، بالله، رحيما، غفورا، المنافقين

شكل ٩٤ : نتائج الموضوعات باستخدام تخصيص دركليه الكامن

في الشَّكْل (٩٤) الموضوعاتُ المستخرجةُ باستخدام نموذج تخصيصِ دركليه الكامن، وهو أقل تكرارًا للموضوعات من النَّموذج السابق. لنكتفي بالنظر إلى عدد تكرارات قِصَّة النَّبِيِّ موسى، بواقع أربع مرات، كما في الموضوع (2) و(6) و(17) و(22)، وقصَّة النَّبِيِّ موسى مع فرعون، بواقع ثلاث مرات، كما في الموضوع (2) و(6) و(17). لتفسير بقية الموضوعات وفق إطارٍ ما يُسمى بقابلية التفسير الإِنْسَانِيِّ (human–interpretability)؛ فإني أَرى أَنَّ علماءَ التَّفسير بالقرآنِ الكريم هم من سيجيدون كشفَ ارتباطِ الكَلِمَاتِ من عدم ارتباطِها بكُلِّ موضوعٍ من الناحيةِ الدلالِيَّةِ كما رآها هذا النموذجُ. على سبيل المثالِ؛ ما العلاقة

بـين الكَلِمَـاتِ في الموضـوع (5): رَبّنـا، تُطِـعْ، ربـه، يومًـا، فهـم، يعلمـون، يدعـون، تحكمـون، مصبحـين؟ أمَّـا في نتيجـة نمـوذج معالجـة دركليـه الهرمِيَّـة (الشـكل ٩٥)، فنلحـظ تنوعًـا أكـثر في الموضوعـات بفعـل تنـوّع الكلمـات ذات العلاقـة بـكل موضـوع.

Hierarchical Dirichlet Process معالجة دركليه الهرمية

First level concentration (γ): 1

Second level concentration (α): 1

The topic Dirichlet (α): 0.01

Top level truncation level (T): 30

Second level truncation level (K): 15

Learning rate (κ): 1

Slow down parameter (τ): 64

عدد الكلمات في جميع الموضوعات من غيرتكرار: ٢٤٨ كلمة من مجموع ٣٠٠ كلمة

1:	الأرض، آمنوا، الناس، الكتاب، خير، يشاء، السماوات، أنزل، يعلمون، النار
2:	الأرض، موسى، الحق، آمنوا، عليم، كفروا، عذاب، الدنيا، الزيتون، خير
3:	الأرض، موسى، ربكم، قوم، فلما، بآياتنا، قنوط، القرآن، ربنا، مستبصرين
4:	شيء، الأرض، تخلفه، عذاب، أرجه، أنطقنا، العالمين، الفزع، لهو، يؤمنون
5:	الأرض، يجادلون، يوسف، الناس، فلما، سقيم، أهل، الكتاب، آمنوا، حق
6:	آمنوا، الكتاب، كفروا، الأرض، النار، شيء، عذاب، بالله، الحكيم، ربنا
7:	الكتاب، بالله، الناس، آمنوا، الأرض، عليما، سيقول، ردوه، لتعلن، عظيما
8:	الأرض، السماوات، خطيئته، الرحمن، الوسيلة، الكتاب، بدارا، قمرا، عذاب، نجعلهم
9:	الأرض، تبصرون، أريناه، بالإيمان، بسكارى، أنثى، قارون، يغلب، المصلحين، صبروا
10:	رسوله، روحا، نزعناها، تستفتيان، بالله، إخوانكم، آمنوا، تعتدوا، لصاحبه، كفروا
11:	الناس، مشيك، النبي، آمنوا، بأحسن، دفع، القلم، يلمزون، تشاءون، تجد
12:	الأرض، السماوات، يشاء، أخذناهم، الرحيم، كفروا، أخيه، حصحص، تأثيما، مضغة
13:	آمنوا، الأرض، أنزل، أتحاجوننا، آمنا، اتقوا، الأحلام، معيشتها، الكتاب، فتكونون
14:	قوم، خافون، ليدبروا، تنتصران، عذاب، بأمر، افتريته، أقسموا، أخلق، فاتبعوا
15:	بدأنا، تركبون، فاتبعوا، الشيطان، نجعلهم، مرية، مقمحون، يريكهم، للذين، قست

16:	استكثرتم، الوهاب، قوما، بمستيقنين، يأتيكما، يصدنك، حاسد، روحي، الأرض، اؤتمن
17:	بالبر، فلنأتينهم، القردة، مهيمنا، بكاف، ناديناه، تلفح، ويل، الأرائك، الأرض
18:	تشطط، ذرعا، أحطت، دخل، طغوا، بست، تستفتحوا، عيون، بأيدي، يأتكم
19:	مترفيهم، محاريب، باسطو، البعث، أصحاب، يبطش، رحيق، يضره، منكرة، نجا
20:	يعبدوها، موسى، تريحون، خشوعا، بساطا، آخرنا، الصخرة، يجتنبون، بعوضة، أعمالكم
21:	مدرارا، يرتد، الإشراق، فرجها، ندلكم، الأرض، قربناه، صدقاتهن، يدخلونها، برأسي
22:	فآوى، جثيا، معكوفا، ليتمتعوا، أعناب، أتبعه، رواسي، صارمين، ليبدي، يجاهد
23:	صدقنا، يتجرعه، بشرر، بعولتهن، ذكره، ألقاها، إبليس، فثبتوا، فزده، يصروا
24:	تكذبان، فبأي، ربكما، آلاء، هب، لإبراهيم، أسكنت، لكثير، تأتوني، تفعل
25:	أضلنا، فضلوا، بالشهر، فسترضع، يعبأ، ينشرون، مركوم، سئلت، فأخلفتم، رهينة
26:	تعرف، يعصرون، بأنه، صم، الأنثى، متراكبا، فليأتنا، جهر، يسلبهم، عارض
27:	بيوت، تستفتيان، أشكو، سنكتب، الأوفى، ينادون، أحصنت، اصطنعتك، مفسدون، مأتيا
28:	صواف، خلقها، يورثها، حجوركم، لعملكم، فئتكم، محبة، فخسفنا، يجار، يكشف
29:	تذكيري، صلاته، لأسجد، يتنزل، لساحر، فاستقم، فتطردهم، يعصينك، سيبطله، نورهم
30:	أطوارا، يتولونه، هديناهم، منكر، فاستأذنوك، مطاع، يتخذوا، أبانا، ليقولون، أيدتك

شكل ٩٥ : نتائج الموضوعات باستخدام معالجة دركليه الهرميّة

المبحث الثَّانِي: الكَلِمَات المفتاحية

لا بدّ من التَّمييزِبين المصطلحات الآتية: المصطلحات، والكَلِمَات المفتاحِيَّة، والألفاظ الأكاديميَّة، والألفاظ الشَّائعة. المصطلحات أكثرتمييزًا من الكَلِمَات المفتاحية، فهناك مصطلحاتٌ دقيقة جِدًّا في حقلٍ فرعِيٍّ من حقلٍ واحدٍ من الحقول العِلْمِيّةِ لا تعدُّ كلماتٍ مفتاحيَّةً، وهناك مصطلحاتٌ أقل دِقَّةً تكون كلماتٍ مفتاحِيَّةً في المجال العامِّ. كذلك الحالُ بين الكَلِمَات المفتاحِيَّةِ والألفاظ الأكاديمِيَّة، وبين الألفاظ الأكاديمِيَّة والألفاظ الشَّائعة، وما يُفرّقُ الألفاظ الأكاديمِيَّة عن الكَلِمَات المفتاحِيَّة هو تَضَمُّنها للأفعالِ والصِّفات والعبارات. إذن هناك منطقةٌ ضبابيةٌ نوعًا ما، وصغيرةٌ نسبِيًّا، تَتَضَمَّن كلماتٍ تشترك بين ما ينضوي تحت مفهوم كُلّ مصطلحٍ من هذه المصطلحات الأربعة. هناكَ

أساليبُ إحْصَائِيّةٌ دقيقةٌ، تكشف الحدود المتمايزة بين ما ينضوي تحتَ كُلِّ مصطلحٍ من كلماتٍ وعباراتٍ، وبين ما يكونُ في تلك المنطقة الضَّبابِيَّة[1].

أُجرِيَت دراسةٌ على استخراجِ التباين المعجمي بين المعاجم العَرَبِيَّة القديمة والمعاجم العَرَبِيَّة المعاصرَة[2]، ووفق منهج إحصائي تطبيقي. وفي المقابل، هناك عددٌ من الخوارزميات في تعلّم الآلة، التي تعمل على استخراج الكَلِمَات المرَشَّحَة موضوعِيًّا، والقابلة لأنْ تكونَ كلمةً مميَّزةً نَصِّيًّا. هناك أداة (استخرجْ المصطلحات Extract keywords) في أداة أورانج، وقد دُمج فيها كُلٌّ من أسلوب (تكرار الكلمة-تكرار المستند العكسي Term Frequency–Inverse Document Frequency TF–IDF)، وخوارزمِيَّة (مُستخرج الكلمة المفتاحية الآنية Yet Another Keyword Extractor YAKE)، وخوارزمِيَّة (الاستخراج الآلي السَّريع للكلمات المفتاحِيَّة Rapid Automatic Keyword Extraction RAKE). شَرحتُ مسبقًا عملَ خوارزمية (تكرار الكلمة-تكرار المستند العكسي) في المبحث الثاني (الإحصاءات) من الفصل الأوَّل في هذا الباب.

تعمل خوارزمِيَّةُ (YAKE) على استعمالِ الإحصاءاتِ؛ لفهم أكثر الكَلِمَاتِ صلةً بكلِّ نَصٍّ، ولا تصلح للنَّصِّ العَرَبِيِّ؛ كونها تعتمدُ على تكرار الكَلِمَات المناطةِ بالأنظمة الكتابِيَّة للغاتِ اللاتينية، فعند استخراجِ السِّماتِ (feature extraction)؛ يُفرَّقُ بين الكَلِمَاتِ وفق الاختصاراتِ (acronyms)، ووفق الكَلِمَاتِ التي تبدأ بحروفٍ كبيرةٍ. أما خوارزمِيَّة الاستخراج الآلي السَّريع للكلمات المفتاحِيَّة (RAKE)؛ فتصلح للنُّصوص العَرَبِيَّة، وتعملُ على ترشيح الكلمة المميَّزة في النَّصِّ عن طريق مقياس درجتِها، الذي يُستخلص وفقَ نتيجة: (درجة الكلمة ÷ تكرار الكلمة). نعرف تكرار الكلمةِ، أما درجة الكلمة فتُحدّد بناءً على التَّتابع اللَّفْظِيّ أوَّلًا، وفيما إن كانت درجة التَّتابع اللَّفْظِيّ ١،

(١) لمعرفة أفضل الممارسات حتى اليوم في كشف تلك الحدودِ؛ الاطّلاع على هذه الدراسة: حوراء. الألفاظ الأكاديمية في المُدَوّنَات العَرَبِيَّة العِلْمِيّة: دراسة حاسوبية تحليلية. (رسالة ماجستير) إشراف: د. سلطان المجيول. قسم اللغة العَرَبِيَّة (جامعة الملك سعود)، ٢٠٢٢.

(2) Almujaiwel, S. (2020). An Overview of the Lexical Variation Between Arabic Lexicons and Natural Arabic Language. International Journal of Arabic Linguistics, 6(1 & 2). 91107-.

أم ٢ إلى ٥ على الحدِّ المعمول به في معظم المشاريع. لننظر إلى المثالِ الآتي، ثم نجري توزيعًا يُوَضِّح طريقةَ استخراج درجةِ الكلمة.

المستند الأَوَّل (على اعتبار أنه نصّ): "الحمد لله رب العالمين"

المستند الثَّانِي (على اعتبار أنه نصّ): "[و] قل الحمد لله سيريكم آياته"

في هذا المثال نصَّان خاليان من الكَلِمَات الوظيفِيَّةِ ومن الفواصلِ، وفي حال وجودِهما؛ فإنَّ ذلك سيؤَثِّر، ولذلكَ يجبُ استبعادهما في المعالجة القبلِيَّة للنَّصِّ. استبعدنا حرف العطف (و)؛ لأنَّ هذا النَّص مُقَطَّع. لو حدّدنا التَّتابعَ اللَّفْظِيَّ (النغرام) ليكون (١) فقط؛ فستبدأ الخوارزمِيَّة بتحديدِ قيمة كُلّ كلمةٍ في جدولة مصفوفاتٍ، ثم تُعطي كل قيمة عددًا يُمثّل عدد مرات التّجاور، فإن كانَت لا شيءَ؛ فذلك يعني أَنَّ الكلمةَ لم تكن سابقةً أو لاحقةً للكلمة المجاورة، وإن كانت (١) فذلك يعني مُجاورَتها مرّة واحدة، وإن كانَت (٢) فذلكَ يعني أَنَّ المجاورة بواقع مَرَّتين (الجدول ٣٠). وتزيدُ هذه القيمُ أو تنقصُ بحسب عدد حالات التقاطع المتَجَاور بين صفيفة الكلمات عموديًّا وأفقِيًّا في المصفوفةِ. ويُلحظ بعدَ هذه الخطوة كما في الجدول (٣١)؛ أنَّ المقياسَ الأعلى قد جاءَ لكلمة "الحمد"، و"الله"، ثم "قل"، و"سيريكم"، و"آياته"، ثم "رب"، ثم أخيرًا "العالمين".

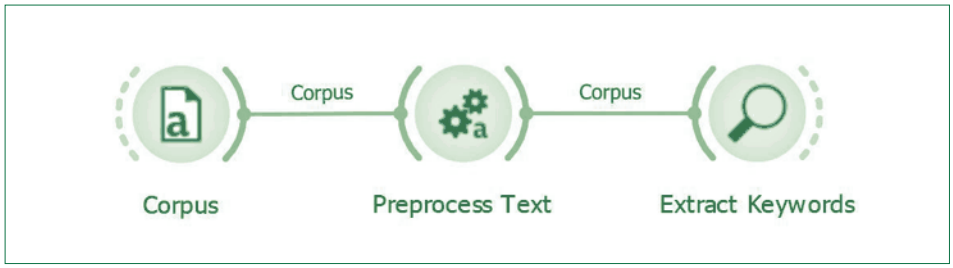
جدول ٣٠ : تمثيل مصفوفة خوارزميَّة الاستخراج الآلِيّ السَّريع للكلمات المفتاحِيَّة

	الحمد	لله	رب	العالمين	قل	الحمد	لله	سيريكم	آياته
الحمد	2	1	1	1		2	1		
لله	1	2	1	1		1	2		
رب	1	1	2	1					
العالمين	1	1	1	1					
قل					2	1	1	1	1
الحمد					1	2	1	1	1
لله					1	1	2	1	1
سيريكم					1	1	1	2	1
آياته					1	1	1	1	2

جدول ٣١ : مقياس درجات الكَلِمَات مع التّكْرَارات

	الحمد	لله	رب	العالمين	قل	سيريكم	آياته
الدرجة	14	14	5	4	6	6	6
التّكْرَار	2	2	1	1	1	1	1
الدرجة ÷ التّكْرَار	7.0	7.0	5.0	4.0	6.0	6.0	6.0

فيما يلي تطبيقٌ على مُدَوّنَة القرآنِ الكريم باستخدام أداة تنقيبِ النُّصُوص (أورانج). البَيَانَات التي حُمّلت في أداة المُدَوّنَة (Corpus) في الشَّكْل (٩٦) تَتَضَمَّن ما عُرِض في الشَّكْل (٧٩)، واستُخلصت النَّتائجُ وعُرضت في الشَّكْل (٩٧).



شكل ٩٦ : مسار عمل استخراج الكَلِمَات المفتاحية

```
('أطعنا سادتنا وكبراءنا فأضلونا السبيلا', 23.5)
('وامرأته قآئمة فضحكت فبشرناها بإسحق', 23.0)
('آبائهم وذرياتهم وإخوانهم واجتبيناهم وهديناهم', 22.833333333333336)
('صلاتي ونسكي ومحياي ومماتي لله', 22.76923076923077)
('الفريقين كالأعمى والأصم والبصير والسميع', 22.73333333333333)
('استحقا إثما فآخران يقومان مقامهما', 22.714285714285715)
('يصدر الرعاء وأبونا شيخ كبير', 22.666666666666668)
('عروشها وبئر معطلة وقصر مشيد', 22.666666666666664)
```

('ثلثي الليل ونصفه وثلثه وطائفة', 22.612244897959183)
('لفريقا يلوون ألسنتهم بالكتاب لتحسبوه', 22.5)
('فتحوا متاعهم وجدوا بضاعتهم ردت', 22.5)
('الساعة ءاتية أكاد أخفيها لتجزى', 22.484848484848484)
('ويدع الإنسان بالشر دعاءه بالخير', 22.4)
('نطوي السماء كطي السجل للكتب', 22.30392156862745)
('موعدهم الصبح أليس الصبح بقريب', 21.923076923076923)
('السراء والضراء والكاظمين الغيظ والعافين', 20.933333333333334)
('قريب نجب دعوتك ونتبع الرسل', 19.964285714285715)
('تأكلوا الربا أضعافا مضاعفة واتقوا', 18.057234432234434)
('تولى ونصله جهنم وساءت مصيرا', 17.90798611111111)
('الكتاب يؤمنون بالجبت والطاغوت ويقولون', 16.981118373275237)
('أتاك حديث ضيف إبراهيم المكرمين', 15.993333333333334)
('يتوكل المتوكلون', 3.333333333333333)
('يصلونها', 3.0)
('آمنا', 2.972972972972973)
('آبائهن', 1.0)

شكل ٩٧ : الكَلِمَات المفتاحِيَّة باستخدام خوارزمِيَّة الاستخراج الآليِّ السَّريع للكلمات المفتاحِيَّة

الفصل الرَّابع
الأنماط النَّحْوِيّة

لمفهـوم النَّمـطِ (pattern) في المعالجـة الآلِيَّـةِ للُّغـة العديـدُ مـن الممارسـاتِ في تنقيـب النُّصُـوص، ومعالجـة اللغـة الطَّبِيعِيّـة، وتعَلُّـم الآلـة. وهنـاك نوعـان مهِمَّـان في معالجـةِ اللغـة الطَّبِيعِيّـةِ؛ وهمـا: التَّعبـيرات النَّمطِيَّـة، والأنمـاط النَّحْوِيّـة المعجَمِيّـة، يُعِـين الأَوَّلُ علـى إجـراء العديـدِ مـن التَّشـذيبِ، والتَّنظيـفِ، والتَّفريـق، والاسـتبدالِ، وكشـف الأنمـاط الخطِّيَّـة[1] للنُّصـوص العَرَبِيَّـة، ويُعِـينُ الثَّـاني علـى فهـم الأنمـاط النَّحْوِيّـة العامـة للنـصِّ العَـرَبِيّ مثـل: [فعل+اسـم]، و[موصوف+صفـة]، و[اسـم+فعل+مفعول بـه]، إلـخ[2]. لـم تُدمـج التَّعبـيرات النمطيـة كلهـا في أداة أورانج، إذْ لـم يُدمـج فيهـا إلا تنميـطُ الكَلِمَـات مـع علامـات التَّرقيـم.

ينبـني مفهـومُ النَّحـو النَّمطِـيّ الـذي انطلقـت فكرتُـه مـن اللُّغَوِيّـات علـى تنميـط الحـالاتِ النَّحْوِيّـةِ التَّتابعيَّـة. واهتـمَّ النحـو العَـرَبِيّ القديـمُ بتحليـل هـذه الأنمـاط في النُّصُـوص العَرَبِيَّـة قبـل اللُّغَوِيّـاتِ الحديثـة. والسـؤال المطـروحُ في هـذا الفصـل هـو: كيـفَ يُمكـن لنـا تطبيـقُ تحليـل الأنمـاط النَّحْوِيّـة في معالجـةِ النَّـصّ العَـرَبِيّ؟ لا يُمكـنُ اسـتخدام هـذا النِّظـام في أداة أورانج، ويتطَلَّـب تطبيقُـه درايـةً عمليَّـة بمكتباتـه في لغـة البرمجـة (البايثـون).

يعتمـد تحليـلُ الأنمـاط النَّحْوِيّـة علـى مجموعـةِ الوسـوم النَّحْوِيّـة والبنـوك الشَّـجريَّة. أشـهر أنظمـة الوسـوم النَّحْوِيّـة للعربيَّـة هـو بـنكُ بـراغ للنحـو الاعتمـادي العَـرَبِيِّ

(١) للاطلاع على قواعد التعبيرات النمطية في لغة البايثون: https://developers.google.com/edu/python/regular-expressions و https://www.w3schools.com/python/python_regex.asp (تاريـخ الاسـترداد ٢٤ أكتوبـر، ٢٠٢٢).

(٢) حـول مزيـدٍ مـن أسـاليب معالجـة النَّحـو النمطـي المعجَـمّي للتَّصاحبـات اللَّفْظِيّـة والتَّعبـيرات الاصطلاحِيَّـة: الغامـدي، أيمـن. تقنيـات الـذَّكَاء الاصْطِنَاعِـيّ والمعالجـة الحَاسُـوبِيَّة للمتلازمـات اللَّفْظِيّـة والتَّراكيـب الاصطلاحِيَّـة. في: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل النَّـصّ العَـرَبِيّ، ٢٠١٩، الصفحـات ١٢٥-١٦٧.

(Prague Arabic Dependency Treebank) بـادت[١]، ومدامـيرا[٢]، وفراسـة[٣]، وسـتانفورد[٤]. وسـأعرض رؤيـةً عامَّـةً عـن هـذه الأنظمـة في المبحـث الأَوَّل، ثـم سـأُطَبّق واحـدًا منهـا في اسـتخراجِ الأنمـاط النَّحْوِيّـة مـن القـرآنِ الكريـم في المبحـث الثَّـاني.

المبحث الأَوَّل: أنظمة الوسوم النَّحْوِيّة واسعة الاستخدام

نعـرضُ هنـا أربعـة أنظمـة وسـومٍ نَحْوِيّـةٍ مشـهورةٍ للغـة العَرَبِيَّـة، وهـي: بـادت (PADT)، ومدامـيرا، وفراسـة، وسـتانفورد. ويتوفـر في كُلِّ نظـامٍ مـن هـذه الأنظمـة المعالجاتُ الأساسِـيَّة، وفي الآتي تعريـفٌ عـن كُلّ نظـامٍ.

بـادت (PADT): ويُعـرفُ ببنك براغ للنَّحو الاعتماديِّ العَـرَبِيّ (Prague Arabic Dependency Treebank)، وسُـميّ بالبنـك الشـجريِّ العَـرَبِيّ Penn Arabic Treebank، وهـو مدمـجٌ مع حُزمـة نَحْـوِ (الاعتماديّـات الكليّـة Universal Dependencies UD) للتَّوسـيم النَّحْـوِيّ والصـرفي والتركيـبي، وتَتَضَمَّـن هـذه الحزمـة قرابـة ٢٠٠ بنـك شـجرِيٍّ لأكـثر مـن مائـة لغـةٍ، وتتبـعُ صيغـةً تنظيمِيَّـةً معَيَّنـةً وفـق صيغـةِ محلـلِ (CoNLL–U Format)، وهـو اختصـارٌ للصِّيغـة الكُلِّيَّـة لفهـم اللغـة الطَّبِيعِيّـة الحَاسُـوبِيَّةِ (Computational Natural Language Learning–Universal). يعمـلُ بـادت عـن طريـق تنصيـبِ (conllu)، أو عـن طريـقِ مكتبـةِ (Stanza) لسـتانفورد.

(1) Taji, D., et al. (2017). Universal Dependencies for Arabic ; Mohamed Maamouri, et al. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus, pp. 102109-.

(2) Pasha, A., et al. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic.

(3) Abdelali, A., et al. (2016). Farasa: A Fast and Furious Segmenter for Arabic; Darwish, K., & Mubarak, H. (2016). Farasa: A New Fast and Accurate Arabic Word Segmenter. LREC-2016; Darwish, K., et al. (2017). Arabic pos tagging: Don't abandon feature engineering just yet; Mubarak, H. (2018). Build Fast and Accurate Lemmatization for Arabic. Paper presented in the International Conference on Language Resources and Evaluation.

(4) Peng Qi, et al. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages; Manning, Christopher D., et al. (2014). The Stanford CoreNLP Natural Language Processing Toolkit.

```
!pip install conllu
!wget https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle11234/1-3105/ud-treebanks-v2.5.tgz
!tar zxf ud-treebanks-v2.5.tgz
#The imports needed to open and parse (interpret) the conllu fil
from io import open
from conllu import parse_incr
#Open the file and load the sentences to a list.
data_file = open("/content/ud-treebanks-v2.5/UD_Arabic-PADT/ar_padt-ud-train.conllu", "r", encoding="utf-8")
```

```
import stanza
# Download an Arabic model into the default directory
print("Downloading Arabic model...")
stanza.download('ar')
# Build an Arabic pipeline, with all processors by default
print("Building an Arabic pipeline...")
ar_nlp = stanza.Pipeline('ar')
```

مداميرا: وقد بُنيت فيه أنظمَةُ المعالجةِ الأساسِيَّةِ من التَّوسيم النَّحْوِيّ، والتَّفريق، والتَّشكيل، والمعجّم، والحالاتِ النَّحْوِيّةِ (base phrases)، واستخراج الكيانات.

فراسة: وبُنيت فيه أنظمةُ المقطّع، والمُعجّم، والموسم النَّحْوِيّ، والمصَحّح الآلِيّ، ومُستخرج أسماء الكيانات، والمُشكِّل الآلِيّ، والمحلّل المكوّني للعبارات، والمحلّل الاعتمادِيّ للنَّحو.

ستانفورد: وبُنيت فيه أنظمة التَّوسيم النَّحْوِيّ، والمُعجّم، واستخراج الكيانات، والمحلّل المكوّني، وتحليل المشاعر، والعلاقات، والإحالات (النَّصِّيَّة).

يُمكن –على سبيل المثال– تجريبُ الاعتماديّات الكليَّة (Universal Dependencies UD) من خلال استخدام الأكواد في الملحق (و) من الكتاب، وللقارئِ أن ينظر في دِقَّة التَّوسيم

لبياناتِ هذا النِّظام، ثم ينظر في أداء توسيمِه لـ (الحمد الله رب العالمين الرحمن الرحيم ملك يوم الدين) باستخدام أكواد الملحقِ (و). سيُلحظ أن توسيمه لهذه الكَلِمَات غير دقيقٍ، ويعود سببُ ذلك إلى أَنَّ البَيَانَات التي تدَرَّب عليها لا تُمثّل نصوصًا عربيَّةً متنوِّعةً بتنَوُّع الأزمنةِ والأوعية والمجالات. نستعرضُ في المبحث الآتي نظرةً عامَّةً عن أنظمة التَّوسيم النَّحْوِيِّ المشهورة في معالجة اللغة العَرَبِيَّة الطَّبِيعِيّة.

المبحث الثَّانِي: نظرة عامَّة عن الأنظمة

أُشيرُ باختصارٍ إلى مكوِّنات أنظمة المعالجة الأساسيّة، وهي: التَّفريقِ (tokenization)، والتَّوسيم النَّحْوِيّ (POS tagging)، والمُعجّم (lemmatizer). يتطلّب التَّفريق أن يعمل النِّظام على فصل التوكن token (الكلمة الفعلية أو الحيّز المفصول عن سابقه ولاحقه بمسافة واحدة) عن بعضه بعضًا في النُّصُوصِ، وهذا الفصل يطال الكلمةَ، ويطال غير الكلمة أيضًا، فغير الكَلِمَات من الأرقامِ، أو علامات التَّرقيم، أو الرُّموزِ؛ تتطَلَّب أن تنفرقَ عن بعضها بعضًا، فجملة (في هذا العام ١٤٤٣هـ) يكون شكل التَّفريق فيها على النَّحو الآتي: ([في]، [هذا]، [ال]، [عام]، [١٤٤٣]، [هـ]). قد تجدُ نظامًا من أنظمة المعالجاتِ الأساسِيَّة لا يفصل (أل) التَّعريف كما هو الحال في ستانفورد، وهذا بحكم تجربتي أفضل؛ لأن فصل المباني الصرفية يُولّد مشكلةً عند تعلم الآلة الكلمات بوصفها وحدات معجمية سياقية. مرحلة التَّفريق هي المرحلة الأولى الأساسِيَّة، يليها معالجةُ التجذيعِ (stemming)، وهو لا يهمُّ كثيرًا إلا في إطار حصرِ الجذور والجذوع. يليها التَّوسيم النَّحْوِيّ الآليُّ لما تمّ تفريقُه، ومرحلة التوسيم مهمَّةٌ في بناء المُعجّم (lemmatizer)؛ أي: رَدّ الكلمةِ إلى أصلها المعجَمِيِّ.

في الشَّكْل (٩٨) نتائج تحليلِ المعالجات المتوفرة في كُلٍّ مِن بادت ومدامِيرا وفراسة وستانفورد على عَيِّنةٍ من القرآن الكريم. الملحوظ من الوهلةِ الأولى أَنَّك ستجدُ معالجةً أساسِيَّةً معينةً في نظامٍ من هذه الأنظمة غير متوفرة في نظامٍ آخر، مثل عدم وجودِ مُعجّم في ستانفورد. أَجريتُ اختبارًا لهذه الأنظمة على سورةِ النَّصر من القرآنِ

الكريـم علـى الرّغـمِ مـن أن هـذه الأنظمـةَ قـد بُنيـت علـى نصـوصٍ عربيَّـةٍ معاصـرةٍ؛ إلا أنَّ العَيِّنـةَ تُمثِّـل سـورةً قصـيرةً لا غرابـة في ألفاظهـا الـتي نسـتعملها في عربيَّتنـا المعاصـرةِ، غـير أنَّ النَّظـمَ فيهـا مختلـفٌ، وليـس كنظـمِ البشـر. بُنيـت للقـرآن الكريـم أنظمـةٌ مبنيـةٌ علـى القواعـد لمسـتوى الصَّـرفِ، والنَّحـوِ، والتَّركيـبِ، والشَّـبكات الدّلاليَّـةِ، وهـي ليسـت أنظمـةً مبنِيَّـةً علـى تعَلُّـمِ الآلـة؛ لأنَّ النَّـصّ القـرآنِيَّ ليـس كمثلـه نَـصٌّ آخـر، ولذلـك نجدُهـا دقيقـة تمامًـا[1]. وَعَـودًا إلى نتـائج المعالجـات في بـادت ومداميـرا وفراسـة وسـتانفورد؛ يُلحـظُ مثـلاً تقطيـعُ (بِحَمْـدِ) إلى ([ب]؛[حمـد]) في نظـام مداميـرا وسـتانفورد، ومـن الطَّبيعِـيِّ نوعًـا مـا أن يُصنِّـف النِّظـام (حَمْـدِ) إلى اسـم علـمٍ، وهـو ليـس بـ(حَمَـد)، لكـن سـبب ذلـك يعـودُ إلى عدم تـدَرّبِ النِّظـام في بياناتـه علـى صيغـةِ (حَمْـد)، فضـلًا عـن غيـابِ التَّشـكيل أساسًـا. في نظـام بـادت؛ لا تتشـارك كلمـة (اِسـتغفرَ) في صيغتِهـا الكتابِيَّـة مـع صيغـةٍ خَطِّيَّـةٍ اسـمِيَّةٍ في حـال غيـاب التَّشـكيل، ومـع ذلـكَ فقـد صنّفهـا النظـامُ علـى أنهـا اسـم، وهـذا غـير صحيـحٍ، ولا يعـني هـذَا أنَّ النظـام لا يصلـح، فنسـبة الدِّقـةِ فيـه عاليـةٌ لبيانـاتٍ لُغَوِيَّـةٍ جديـدةٍ، لكـن كلمـة (اسـتغفر) بصيغـة فعـل الأمـرِ ليسـت مسـتعملةً بكـثرةٍ في النُّصُـوص العَرَبِيَّـة المعاصـرة، فلـم تخضـع الكلمـةُ لتدريـبٍ واختبـارٍ. أمـا بقيـة الملحوظـات فأتركُهـا لواسـع نظـرِ اللُّغَوِيّـين القُـرَّاء. وأُشـيرُ هنـا إلى أنَّ هـذه الأَنْظِمـة الأربعـة للتَّوسِـيمِ النَّحْـوي تختلـف فيمـا بينهـا في اصطـلاح تَسـميةِ الوُسُـوم النَّحويَّـةِ، وفي تعريـفِ كُلِّ وَسْـمٍ (أو: رَمـزٍ) مـن هـذه الوسـوم وفـق الحَالـةِ النَّحويَّـةِ للكَلِمَـةِ في النَّـص. والسـؤال المهـم: هـل اختـلافُ تسـمية الوَاسِـم (أو: الرَّمْـزِ) يؤثـر؟ الجـواب: لا، إلا إذا اختلـف تصنيـف الحـالات النحويـة (الأمثلـة) فيهـا يدويًّـا. أقـترحُ متابعـةَ مـا يتطـور لأنظمـةِ هـذه الوسـوم بعـد ظهـور النَّمـاذجِ اللغويَّـةِ الضَّخمـة Large Language Models LLMs، وأعـني هنـا: نمـاذج تمثيـلاتِ المشـفِّر الثنـائيّ الاتِّجـاه مـن المحـوّلاتِ (Bidirectional Encoder Representations from Transformers BERT) الـتي تُطـوّر

(١) بُنيت معالجة أساسية لمدونة القرآن الكريم، وهي خاصَّةٌ بها، فالتَّصريف والتَّركيب في القرآن الكريم مختلف عن غيره، كما أن النَّصّ القرآني نَصٌّ بلاغِيٌّ مختلفٌ عن نصوص العَرَبِيَّة القديمة في نظمِه المعجز. تميز هذا المشروع ببناء أنظمةٍ شاملةٍ على مستوى الصرف والنحو والتركيب والعلاقات الدلالية. للاطلاع على هذا المشروع: https://corpus.quran.com (تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢).

للغـة العربيـة، والـتي تُحسّـن تحسـينا عاليًا دقـةَ تصنيف الوسـوم النحويـة للعربيـة القديمة، والعربيـة الحديثـة، واللهجـات العربيـة.

سورة النصر	بادت				مدامیرا				فراسة				ستانفورد			
	المعجم	الوسيم النحوي		التفريق	المعجم	الوسيم النحوي		التفريق	المعجم	الوسيم النحوي		التفريق	المعجم	الوسيم النحوي		التفريق
													إذا	إذا	IN	حرف جر
إذا	إذا	إذا	CCONI	إذا	إذا	إذا	أداة	إذا	إذا	PART	أداة	إذا	جاء	جاء	VBD	فعل تام
	جاء	جاء	VERB	جاء	جاء	جاء	فعل	جاء	جاء	V	فعل	جاء	نصر	نصر	NNP	اسم علم مفرد
	نصر	نصر	X	نصر	نصر	نصر	اسم علم	نصر	نصر	NOUN	اسم	نصر	الله	الله	NNP	اسم علم مفرد
الله	الله	الله	X	الله	الله	الله	أداة	الله	الله	PART	أداة	الله	و	و	CC	أداة ربط
جاء	و	و	CCONI	جاء	جاء	جاء	فعل	جاء	جاء	V	فعل	جاء	الفتح	الفتح	DTNN	اسم محدد/مفرد
نصر	الفتح	الفتح	NOUN	نصر	نصر	نصر	اسم علم	نصر	نصر	NOUN	اسم	نصر	و	و	CC	أداة ربط
الله	و	و	CCONI	الله	الله	الله	اسم	الله	الله	NOUN	اسم	الله	رأيت	رأيت	VBD	فعل تام
و	رأيت	رأيت	VERB	و	و	و	أداة	و	و	CONJ	حرف عطف	و	الناس	الناس	DTNN	اسم محدد/مفرد
الفتح	الناس	الناس	NOUN	الفتح	الفتح	الفتح	اسمي	الفتح	الفتح	DET+NOUN	محدد+اسم	الفتح	يدخلون	يدخلون	VBP	فعل تام
و	يدخلون	يدخلون	VERB	و	و	و	أداة	و	و	CONI	حرف عطف	و	في	في	IN	حرف جر
رأيت	في	في	ADP	رأيت	رأيت	رأيت	فعل	رأيت	رأيت	V+ABBREV	فعل+اختصار	رأيت	دين	دين	NN	اسم شائع
الناس	دين	دين	NOUN	الناس	الناس	الناس	اسمي	الناس	الناس	DET+NOUN	محدد+اسم	الناس	الله	الله	NNP	اسم علم مفرد
يدخلون	الله	الله	NOUN	يدخلون	يدخلون	يدخلون	فعل	يدخلون	يدخلون	V+NOUN	فعل+اسم	يدخلون	أفواجا	أفواجا	NN	اسم شائع
في	أفواجا	أفواجا	NOUN	في	في	في	أداة	في	في	PREP	حرف جر	في	في	في	CC	أداة ربط
دين	في	في	CCONI	دين	دين	دين	اسمي	دين	دين	NOUN	اسم	دين	سبح	سبح	VBD	فعل تام
الله	سبح	سبح	NOUN	الله	الله	الله	اسم علم	الله	الله	NOUN	اسم	الله	ب	ب	IN	حرف جر
أفواجا	بحمد	بحمد	X	أفواجا	أفواجا	أفواجا	اسمي	أفواجا	أفواجا	NOUN+NOUN	اسم+اسم	أفواجا	بحمد	بحمد	NNP	اسم علم مفرد
فسبح	ربك	ربك	X	فسبح	فسبح	فسبح	فعل	فسبح	فسبح	ABBREV+V	اختصار+فعل	فسبح	ربك	ربك	NNP	اسم علم مفرد
بحمد	و	و	CCONI	بحمد	بحمد	بحمد	اسم علم	بحمد	بحمد	PREP+NOUN	حرف جر+اسم	بحمد	و	و	CC	أداة ربط
ربك	فاستغفر	فاستغفر	NOUN	ربك	ربك	ربك	اسمي	ربك	ربك	NOUN	اسم	ربك	فاستغفر	فاستغفر	VBD	فعل تام
و	هو	هو	PRON	و	و	و	أداة	و	و	CONI	حرف عطف	و	هو	هو	PRP	ضمير منفصل
استغفره	إنّ	إنّ	PART	استغفره	استغفره	استغفره	فعل	استغفره	استغفره	NOUN+PRON	اسم+ضمير	استغفره	إن	إن	IN	حرف جر
إنه	هو	هو	PRON	إنه	إنه	إنه	أداة	إنه	إنه	PART+PRON	أداة+ضمير	إنه	هو	هو	PRP	ضمير منفصل
كان	كان	كان	AUX	كان	كان	كان	فعل	كان	كان	V	فعل	كان	كان	كان	VBD	فعل تام
توابا	توابا	توابا	X	توابا	توابا	توابا	اسمي	توابا	توابا	NOUN+NOUN	اسم+اسم	توابا	توابا	توابا	NN	اسم شائع

شكل ٩٨ : نتائج تحليل النظام الآلي للمعالجة الأساسية لبادت ومداميرا وفراسة وستانفورد

أُشِيرُ إلى نقطةٍ مهمَّةٍ عن نصوصِ القرآنِ الكريم، وهي منعكسةٌ من النَّظم البديعِ. يكمن النَّظمُ البديع فيه في جملِه القصيرة المكتملةِ معنىً والمتماسكةِ اتّساقًا، حتى في الآيات الطَّويلةِ كآية الدَّين في سورة البقرةِ. وفي المقابل، ستجدُ أيُّها القارئ أنَّ النُّصُوصَ القانونيَّة المعاصِرة نصوصٌ تَتَضَمَّن جملًا قصيرةً؛ لكنَّها ليست مكتملةَ المعنى، ومُتَّسقةَ النَّظم، وفيها من الغموضِ النَّظمي والإحاليّ ما فيها معنىً واتِّساقًا.

نقطةٌ مهمَّةٌ أُخرى تتعَلَّق بطول تراكيب المتمّمات أو الفضلاتِ في الجمل العَرَبِيَّة المعاصرة، ويُقصد بالمتممات complements ما ورد من غير ركني الإسناد (المسند والمسند إليه). نجد المتمّمات في المفعولاتِ، والمنصوباتِ، وشبه الجملِ، والتوابعِ، والإضافة، إلخ. من أمثلةِ الإضافة أنَّ غالبَ أسماء المواضع في العَرَبِيَّة القديمةِ كلمةٌ واحدةٌ، وقليلٌ منها

يكونُ مركَّبًا تركيبًا إضافِيًّا بسيطًا مثل (مكَّة المكَرَّمة). ونجدُ أنّ أسماء المواضعِ في العَرَبِيَّة المعاصرة قد اتّسمت بالطُّول. على سبيل المثالِ: المملكة العَرَبِيَّة السعودِيَّة.

المبحث الثَّالِث: أمثلة على الأنماط النَّحْوِيّة

بعد إجراءِ مقارنةٍ عشوائِيَّةٍ لتكشيفِ الأنماطِ؛ يُرى استعمالُ نظام ستانفورد (الجدول ٢٧) للتوسيم النَّحْوِيّ على تتابع ٢ و٣ متتابعات لفظِيَّة بدَلًا من الأنظمة الأُخرى؛ لوضوحِها ولقِلَّة أخطائِها على البَيَانَات الجديدة. وفيما يَأتِي هذه الأنماط:

١- التَّركيب الإضافِيّ (NN ثم NNP أو DTNN أو DTNNP أو DTNNS). ومن أمثلته ما يَأتِي:
[رب/NN العالمين/DTNNS]: ورد ٣٤ مرة. [حدود/NN الله/NNP]: ورد ١١ مرة. [امرأة/NN العزيز/DTNNP]: ورد مرتين. [زخرف/NN القول/DTNN]: ورد مرة واحدة.

٢- التَّركيب الوصفِيّ، ومن أمثلته ما يَأتِي:
[غفور/JJ رحيم/NNP]: ورد ٤٩ مرة. [الحياة/DTNNP الدنيا/DTNN]: ورد ٩ مرات. [العزيز/DTJJ الحكيم/DTJJ]: ورد ٢٩ مرة. [العزيز/DTNNP الرحيم/DTNNP]: ورد ١٣ مرة. [قول/NNP معروف/JJ]: ورد مرتين. [الروح/DTNN الأمين/DTJJ]: ورد مرة واحدة [مقاما/NNP محمودا/NN]: ورد مرة واحدة، [رجل/NN رشيد/NNP]: ورد مرة واحدة.

٣- التَّركيب الفِعْلِيّ: عدد التَّراكيب الفِعْلِيّة الماضوِيَّة أكثر من المضارعةِ، والأخيرة أكثر من المبنِيِّ للمجهولِ، والأخير أكثر من الأمر، وهذا على عمومِ الأمثلة في مُدَوّنَة القرآن الكريم. ومن أمثلته ما يَأتِي:
فعل تامٌّ: [ملكت/VBD أيمانكم/NN]: ورد ٥ مرات. فعل مضارعٌ: [تؤمنون/VBP ب/IN الله/NNP] ورد ٥ مرات. فعل مبنِيٌّ للمجهول: [و/CC قيل/VBN ل/IN لله/NN]: ورد مرة واحدة. فعل أمر: [ادع/VB ل/IN نا/PRP رب/NN ك/PRP$]: ورد ٥ مرات.

ما الذي ستفيدُنا به قائمة الأنماط للمدَوّنةِ اللُّغَوِيّة؟ الفائدة تكمنُ في استخلاص الأنماط العامَّةِ، وأكثر الأنماط السَّائِدة، وطبيعة الأمثلةِ الطَّبِيعِيّة لكُلِّ نمطٍ من الأنماط، وذلك يُفيد في معرفة البنيةِ النَّحْوِيّة النَّمطِيَّةِ الأكثر استخدامًا، يليها الأقَلُّ استخدامًا، وهي إفادة آلية مهمّة في بناء المعاجم والمصادر المعجميّة.

من فوائدِها أيضا تكشيفُ طبيعةِ أنماط نظمِ القرآنِ البيانِيِّ، وأنماطِ الفواصل القرآنِيَّةِ[1]. يقولُ الباقلاني: «إنَّ نظمَ القرآن على تصرُّفِ وجوهِه، واختلافِ مذاهبِهِ خارجٌ عن المعهود من نظام جميع كلام العربِ، ومباينٌ للمألوفِ من ترتيب خطابِهم، وله أسلوبٌ يختصُّ بهِ، ويتمَيَّز في تصرُّفه عن أساليب الكلامِ المعتاد»[2]. ويقولُ أيضًا: «الفواصل حروفٌ متشاكلةٌ في المقطعِ، يقع بها إفهامُ المعاني»[3]. ومن صور الفواصل ما يكونُ للتَّصديرِ، وما يكون للتَّوشيح في القرآنِ الكريم، فالتَّصديرُ: إشارةٌ لفظيَّةٌ، وهي أن تتقَدَّم لفظة الفاصلة بمادَّتها في أَوَّل صدر الآية، أو في أثنائهِ، أو في آخره، مثل قولِه تعالى: ﴿رَبَّنَا لَا تُزِغْ قُلُوبَنَا بَعْدَ إِذْ هَدَيْتَنَا وَهَبْ لَنَا مِن لَّدُنكَ رَحْمَةً ۚ إِنَّكَ أَنتَ الْوَهَّابُ ﴿٨﴾﴾ [آل عمران: ٨]. والتوشيح: إشارةٌ معنويَّةٌ، وهو أن يكونَ معنى الآية مشيرًا إلى هذه الفاصلةِ، كما ورد في الحديث أنَّ زيدًا بن ثابت كان يكتبُ ما يمْلِي عليه الرسولُ -صلى الله عليه وسلم- فأملى قولَه تعالى: ﴿ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ عِظَامًا فَكَسَوْنَا الْعِظَامَ لَحْمًا ثُمَّ أَنشَأْنَاهُ خَلْقًا...﴾ [المؤمنون: ١٤]، فنهض معاذ بن جبل فقال: «فتبارك الله أحسن الخالقين»، فضحكَ رسولُ الله -صلى اللهُ عليه وسلم-، فقالَ له معاذ: مِمَّ ضحكتَ يا

(١) وقفت على رسالة ماجستير مقَدَّمة في جامعة وهران السانية عن دلالة الخطاب في القرآن الكريم، وقد وقفتُ فيها على نقاط مختصرةٍ ومفيدةٍ عن تميّز نظم القرآن، وعن الفواصل القرآنِيَّة فيها، اعتمدَ الباحثُ فيها على التَّصدي لكثيرٍ من المصادر العَرَبِيَّة القديمة في التّفسير والبلاغة واللغة. انظر: معمر، شباب. دلالة الخطاب في القرآن الكريم (رسالة ماجستير). إشراف: محمد زعراط والجيلاني سلطاني، جامعة وهران السانية، الجزائر، ٢٠٠٦.

(٢) الباقلاني، إعجاز القرآن، صفحة ٣٥.

(٣) المصدر السابق، صفحة ٢٧١.

رسـول الله؟ قـال: "بهـا ختمـت"[1]. ومـن ذلـك قولـه تعـالى: ﴿وَءَايَةٌ لَّهُمُ ٱلَّيۡلُ نَسۡلَخُ مِنۡهُ ٱلنَّهَارَ فَإِذَا هُم مُّظۡلِمُونَ ٣٧﴾ [يس: ٣٧] فنجـد فاصلـة هـذه الآيـة قـد تَوَشَّـحت بـ «مُظلِمـون» نتيجـةَ انسـلاخ النَّهـار.

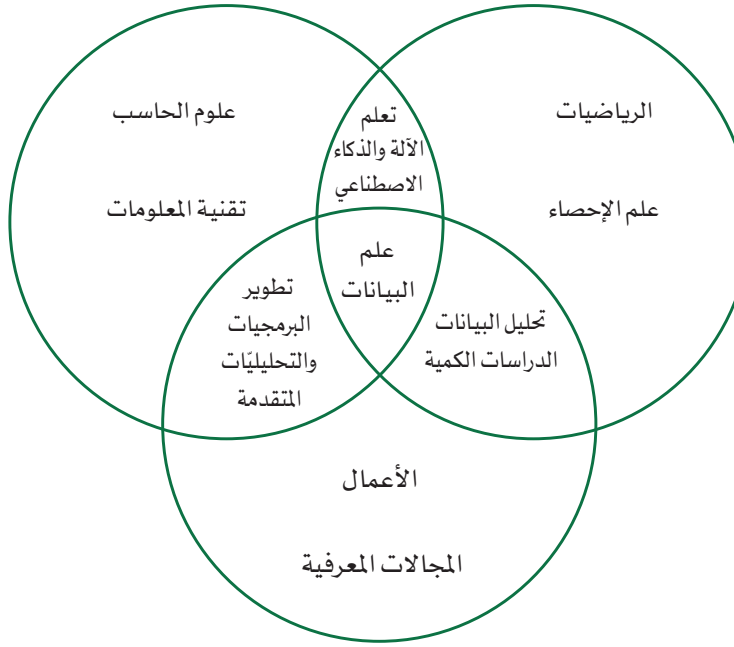
[و/CC جاهـدوا/VBD في/IN الله/NNP حـق/NN جهـاده/NN هـو/PRP اجتباكـم/JJ و/CC مـا/WP جعـل/VBD عليكـم/NN في/IN الديـن/DTNN مـن/IN حـرج/NN ملـة/NN أبيكـم/NN إبراهيـم/JJ هـو/PRP سـماكم/NN المسـلمين/DTNNS مـن/IN قبـل/NN و/CC في/IN هـذا/DT ليكـون/NN الرسـول/DTNN شـهيدا/JJ عليكـم/JJ و/CC تكونـوا/VBP شـهداء/NN علـى/IN النـاس/DTNN فأقيمـوا/VBD الصـلاة/DTNN و/CC آتـوا/VBD الـزكاة/DTNN و/CC اعتصمـوا/VBD بـالله/NNP هـو/PRP مولاكـم/NN فنعـم/NN المـولى/DTNNP و/CC نعـم/UH النصـير/DTNNP]	التصدير: مولاكم... نعم المولى التوشيح: جاهدو... النصير
[سـأصرف/NN عـن/IN آيـاتي/NNS الذيـن/WP يتكـبرون/VBP في/IN الأرض/DTNN بغـير/NNP الحـق/DTNN و/CC إن/VBD يـروا/VBP كل/NOUN آيـة/NN لا/RP يؤمنـوا/VBP بهـا/NN و/CC إن/NN يـروا/VBP سـبيل/NN الرشـد/DTNN لا/RP يتخـذوه/VBP سـبيلا/NN و/CC إن/NN يـروا/VBP سـبيل/NN الغـي/DTNN يتخـذوه/VBP سـبيلا/NN ذلـك/DT بأنهـم/NN كذبـوا/VBD بآياتنـا/NNP و/CC كانـوا/VBD عنهـا/NN غافلـين/JJ]	التوشيح: سبيل الغي... غافلين
[لكـن/VBP الله/NNP يشـهد/VBP بمـا/NN أنـزل/NN إليـك/NNP أنزلـه/NNP بعلمـه/NNP و/CC الملائكـة/DTNN يشـهدون/VBP و/CC كفـى/VBD بـالله/NNP شـهيدا/JJ]	التصدير: يشهدون... شهيد
[إن/VBD عبـادي/NNP ليـس/VBD لـك/DT عليهـم/NN سـلطان/NNP و/CC كفـى/VBD بربـك/NNP وكيـلا/NNS]	التوشيح: سلطان.. وكيلا

(١) أورده السيوطي في الإتقان في علوم القرآن، الجزء الثاني، صفحة ١٠١. انظر أيضا: معمر، شباب. دلالة الخطاب في القرآن الكريم (رسالة ماجستير)، صفحة ١٠٤.

[و/CC تـوكل/VBD علـى/IN الله/NNP و/CC كفـى/VBD بـالله/NNP وكيـلا/NNS]	التصدير: توكل... وكيلا
[و/CC ليخـش/NNP الذيـن/WP لـو/IN تركـوا/VBD مـن/IN خلفهـم/NN ذريـة/JJ ضعافـا/NN خافـوا/VBD عليهـم/NNP فليتقـوا/NNP الله/NNP و/CC ليقولـوا/VBD قولا/NN سـديدا/NN]	التوشيح: ليقولوا قولا سديدا
[و/CC قالـوا/VBD لـن/RP نؤمـن/VBP لـك/DT حتى/IN تفجـر/NN لنـا/VBD من/IN الأرض/DTNN ينبوعا/NN]	التوشيح: تفجر... ينبوعا
[و/CC آت/VBD ذا/DT القـربى/DTNN حقـه/NN و/CC المسـكين/DTNNS و/CC ابـن/NN السـبيل/DTNN و/CC لا/RP تبـذر/VBP تبذيـرا/NN]	التصدير: تبذر تبذيرا

الفصل الخامس
تنقيب البَيَانَات

نسـمعُ بتعلـم الآلـةِ (Machine Learning)، وبتحليـل البَيَانَـاتِ (Data Analysis)، وبتطويـر البَرْمَجيّـاتِ (Software Development)، وهـي حقـولٌ مختلفـةٌ عـن بعضهـا مـن ناحيـة التَّطبيـق والممارسـة. يعتمـدُ تَعَلُّـمُ الآلـةِ علـى علـوم الحاسـبِ، وتقنيـة المعلومـات، مـع الرِّياضيـاتِ (التَّطبيقيَّـة)، وعلـم الإحصـاء. ويعتمـدُ تحليـل البَيَانَـات علـى حقـل الرِّياضيـاتِ، وعلـم الإحصـاءِ، وتطبيقاتهما علـى مجـالات الأعمـالِ، والمعرفـةِ، والعلـوم. ويكـون تطويـرُ البَرْمَجيّات معتمـدًا علـى علـوم الحاسـبِ، وتقنيـة المعلومـات في مجـالات الأعمـالِ، والمعرفـةِ، والعلـوم. أما عِلْـم البَيَانَـاتِ (Data Science)؛ فيعتمـدُ علـى كُلّ مـا سـبقَ ذكـره، شـرطَ أن يبـدأ مـن وجـود مشـكلةٍ محـدَّدة المعالـم، وتتطلَّـب اسـتخدامَ علـوم البَيَانَـات للبحـثِ عـن حـلٍّ لهـا. يُمكـنُ التعبيـرُ عـن نِطَاقـاتِ التَّداخـلاتِ الاختصاصيَّـة بـين كُلّ المجـالات السَّـابقة بالشَّـكْل (٩٩).



شكل ٩٩: تداخلات المجالات التَّطْبِيقِيّة لِعِلْم البَيَانَاتِ

أمَّا ما يتعلّق بإطار مُعالجةِ المدوَّنَاتِ اللُّغويَّة فيُمكن تصنيف التَّداخل على النحو الآتي.

- يَدخل مجالُ معالجة اللغة العَرَبِيَّة الطَّبِيعِيّة ضمن (تعلم الآلة والذَّكَاء الاصْطِنَاعِيّ).
- يَدخل مجالُ الأدوات التي تُبنى وتُصمّمُ على أنها تطبيقاتٌ لمعالجة اللغة العَرَبِيَّة الطَّبِيعِيّة كالمصحح الإمْلَائِيّ، والتعَرُّف إلى الكلام، والتَّشكيل، ضمن (تطوير البَرْمَجيّات والتَّحْلِيليات المتقدمة).
- يَدخل مجالُ لغويَّات المُدَوّنَة ضمن (تحليل البَيَانَات والدِّرَاسَات الكَمِّيَّة).
- يعتمدُ عِلْمُ البَيَانَاتِ (اللُّغَوِيّة) على جميعِ ما سبق، بهدف إيجادِ حلولٍ لمشكلاتٍ واقعِيَّةٍ تتحدّد في سوقِ اللغة الرَّقميّة، شرطَ أن يكون هذا الحلُّ تطبيقًا مفيدًا لمعالجةِ المشكلةِ المحَدَّدة.

سأتطرّق إلى أساليب تعَلُّم الآلة في التَّصنيف والانحدار، وأعيدُ قبل ذلك مثالَ التَّصنيف والانحدار من أجل التَّذكيرِ والتَّوضيح.

في التَّصنيف: لدينا بياناتٌ لُغَوِيَّة فيها مُتَغَيّرَاتٌ فئويّةٌ مصنفة؛ مثل أقسام الكلام.

في الانحدار: لدينا بياناتٌ لُغَوِيَّة فيها مُتَغَيّرَات مقياسِيَّةٌ؛ مثل عدد تكراراتِ أقسام الكلام.

للتَّصنيف مهمَّتان في تعَلُّم الآلة.

الأولى: إيجادُ أفضل نموذج يكونُ فيه المُتَغَيّر المُسْتَقِلّ (المتنبئ) أكثر تأثيرًا في المُتَغَيّر التَّابِع (النَّاتِج)، ويُسمى هذا الإجراء اختيار السِّماتِ (feature selection)؛ على سبيل المثال: [المُتَغَيّرَات المُسْتَقِلَّة: الضَّمائر الشَّخصِيَّة، وضمائر المخاطب، وضمائر الغائِب]، و[المتغَيّر النَّاتِج: كاتب وكاتبه].

سؤال: ما أكثر المُتَغَيّرَات النَّاتِجة استخدامًا لنوع الضَّمائر؟ الجواب: يعمل التَّصنيفُ على البحث عن النَّموذج الأفضل؛ أي: المُتَغَيّر المُسْتَقِلّ الأكثر تأثيرًا في المُتَغَيّر النَّاتِج.

لـو افترَضنـا أن الـمُتَغَيّر (الضمائـر الشَّـخصية) هـو أفضـلُ نمـوذج، فهنـا نسـتنتج بأنـه الأكـثر تأثـيرًا علـى اسـتخداماتِ الجنسَـين لـه.

الثَّانيـة: تَحقيـقُ أفضـل تنَبُّؤ آليّ لأمثلـة التَّصنيـف، وذلـك علـى بيانـاتِ اختبـار (بيانـات لـم تَرَهـا الآلـة). لنفـترض التَّصنيـفَ الكلـمِيَّ الآتي: [محمد/اسـم، خالد/اسـم، يذهب/فعـل]، ولنفـترض أنَّ الآلـة قـد تدرّبـت علـى هـذا التصنيـفَ. لـو أتينـا ببيانـات اختبـارٍ لـم تَرَهـا الآلـة، وفيهـا كلمـة [يزيد/اسـم]، فالمحتمـل أنّ الآلـة سـتُصنِّفها فعـلًا، وليـس اسـمًا، ويتطلّـب عندئـذٍ تدريـب الآلـة مـرة أخـرى علـى أمثلـة صيـغ الأسـماء الـتي تُشـبهُ الأفعـالَ مـن أجـلِ رفـع كفـاءة أداء التصنيـف.

للانحدار مَهمَّتان في تعلُّم الآلة.

الأولى: إيجادُ أفضـل نمـوذج يكـون فيـه الـمُتَغَيّر المُسْـتَقِلّ (المتنبئ) أكـثر تأثـيرًا في الـمُتَغَيّر التَّابِـع (النَّـاتج)، ويُسَـمَّى هـذا الإجـراء اختيـار السِّـماتِ (feature selection). علـى سـبيل المثـال: [الـمُتَغَيّرَات المسـتَقِلَّة: تكـرارات الضَّمائـر الشَّـخصيَّة، وتكـرارات ضمائـر المخاطـب، وتكـرارات ضمائـر الغائـب]، و[المتغـيّر النَّـاتج: كاتـب وكاتبـه]. لا يتعامـل الانحـدار مباشـرةً مـع الـمُتَغَيّرَات الفئويـة، وبالتـالي لابـدّ مـن تحويلهـا إلى مُتَغَـيّرَاتٍ مِقْيَاسِـيَّة. في حالتنـا هنـا: (كاتـب) و(كاتبـة)، يُجـرى في تعلـم الآلـة مـا يُسَـمَّى بترميـز النقطـة السـاخنة (one hot encoding)، ويعمـل ببسـاطة وفـق عـدد الفئـات، فـإن كانـت فئتـيْن؛ تُحـذف الثانيـة؛ لأن الأولى دالَّـةٌ علـى الثَّانيـة (مثـل: إبقـاء كاتـب وحـذف كاتبـة)، أو تُعطـى القيمـة (١) لإحداهمـا والقيمـة (٠) للأخـرى، وإن كانـت أكـثر مـن ذلـك؛ فيتـم وضـع كُلّ فئـة في عمـودٍ مسـتقِلٍّ، وتُعطـى القيمـة (١)، وتُعطـى مـا عداهـا القيمـة (٠) في العمـود نفسـه، ويُطبّـق هـذا المنـوال علـى بقيـة أعمـدةِ الفئـات الأخـرى في جـدول البَيَانـات.

سـؤال: مـا أكـثر الـمُتَغَيّرَات النَّاتِجـة اسـتخدامًا لتكـرارات نـوع الضَّمائـر؟ **الجـواب:** يعمـل التَّصنيـفُ علـى البحـثِ عـن النَّمـوذج الأفضـل؛ أي: قيمـة الـمُتَغَيّر المسْـتَقِلّ الأكـثر تأثـيرًا في

الـمُتَغَيّر النَّـاتج. لـو افترضنـا أن الـمُتَغَيّر (تكـرارات الضَّمائـر الشَّـخصية -أنـثى) هـو أفضـلُ نموذج، فهنـا نسـتنتج بـأن الجنسَ (الأنـثى) هـو الأكـثر اسـتخدامًا لهـذا النـوع مـن الضمائـر.

الثانيـة: تحقيـقُ أفضـل تَنبُّـؤ آليّ للانحـدار الخطّـيّ، وذلـك علـى بيانـاتِ اختبـار (بيانـات لـم تَرَهـا الآلـة). لنفـترِض أن التَّصنيـف لنُصُـوصٍ بـين ثلاثـة أوعيـةٍ مختلفـةٍ، ولنفـترض أنَّنَـا قـد درَّبنـا الآلـةَ علـى التنَبُّـؤ بتكـراراتِ علامـات التَّرقيـمِ في تلـك النُّصُـوص بـين هـذه الأوعيـةِ (الجـدول ٣٢)، وبعـد التَّدريـب؛ تنَبَّـأت الآلـةُ بـأَنَّ تكـرارات علامـات التَّرقيـم في الصُّحـف ١٥، وفي الرَّسـائل العِلْمِيّـة ٩، وفي الرِّوايـات ٣٠. هنـا سـتكون نسـبة قيمـة المنحـدرِ (regressor score) ٧٥٪ مقابـل التّكْـرَارات الحقيقيّـة لعلامـات التَّرقيـم. إذن، فتعلّـم الآلـة لنمـوذج الانحـدار يعتمـدُ علـى تنَبُّـؤ النَّمـوذج بالقِيَـمِ المتنبَّـأة (predicted values) مقابـل القِيَـمِ الحقيقِيَّـةِ (actual value) في البَيَانَـاتِ.

جدول ٣٢ : التنبُّؤ بعلامات التَّرقيم بين الأسماء والأفعال والضَّمائر

	الـمُتَغَيّرَات المتنبِّئة (المُسْتَقِلَّة)			المتغيِّر النَّاتج (التابع)	
	عدد الكلمات			القيمة الحقيقة	القيمة المتنبَّأة
	الأسماء	الأفعال	الضمائر	علامات الترقيم	
الصحف	100	90	66	20	15
الرسائل العِلْمِيّة	90	20	10	12	9
الروايات	150	110	79	40	30

المبحث الأَوَّل: تهيئة المُتَغَيّرَات في البَيَانَات

أُشـيرُ هنـا إلى أبسـط أنـواع الـمُتَغَيّرَات في تعَلّـم الآلـةِ، وإلى اختـلاف المصطلحـات مـع تشـابهِ المفاهيـمِ، بحكـم اختـلافِ مجـال تعَلُّـم الآلـةِ عـن مجـال علـمِ الإحصـاء. في الجـدول الآتي تصنيـفٌ لتلـك الأنـواعِ والاختلافـات.

جدول ٣٣ : ؛أنواع المُتَغَيّرَات في البَيَانَات بين تعَلّم الآلة وعلم الإحصاء

	تعلم الآلة	علم الإحصاء
التَّابِع Y	الـهـدف target، العـلامـة label، النَّاتِج output	التَّـابِـع dependent، الاستجـابـة response، الخرج outcome
المسْتَقِلّ X	الخصيصة feature، السمة attribute، الدخل input	المسْتَقِلّ independent، التأويلـي explanatory، المتنبِّئ predictor

ولشَـرحِ إمكانـاتِ أدوات (أورانج) القائمـة بذاتِهـا في مجـال تنقيـب البَيَانَـات (اللُّغَوِيّـة)؛ أُمَثِّـلُ بثلاثَـةِ أشـكالٍ للبيانـاتِ المهيَّـأة للتَّصنيـف في المبحـث الثـاني (الشَّـكل ١٠٠ و١٠١)، وللانحدار في المبحث الثالث (الشَّكل ١٠٢). يُشـيرd إلى الـمُتَغَيّرَات الاسـمِيَّة (كالزمـن: مكّي MK ومـدنيّ MD) أو الفئوِيَّـة (كالطـول: طويـل long ومتوسـط medium وقصـير short)، ويدلُّ c علـى الـمُتَغَيّرَات المقياسِـيَّة (كعـدد الآيـات)، ويُشـير الصِّنـف (class) إلى الـمُتَغَيّر التَّابِع أو النَّاتِج. سيُجرى التحليل في المبحث الثاني (التصنيف) على بيانات الشكل (١٠٠) و(١٠١)، وفي المبحث الثالث (الانحدار) على بيانات الشكل (١٠٢).

السجدة	الطول	الزمن
d	d	d
class		
no	short	MK
no	long	MD
no	long	MD
no	long	MD
no	long	MD
no	long	MK
yes	long	MK
...	...	...

شكل ١٠٠ : بيانات تحليل للتَّصنيف

الأنبياء	الطول	الزمن
d	d	d
class		
yes	MD	long
yes	MD	long
yes	MD	long
yes	MD	long
yes	MK	long
yes	MK	long
yes	MD	long
...	...	...

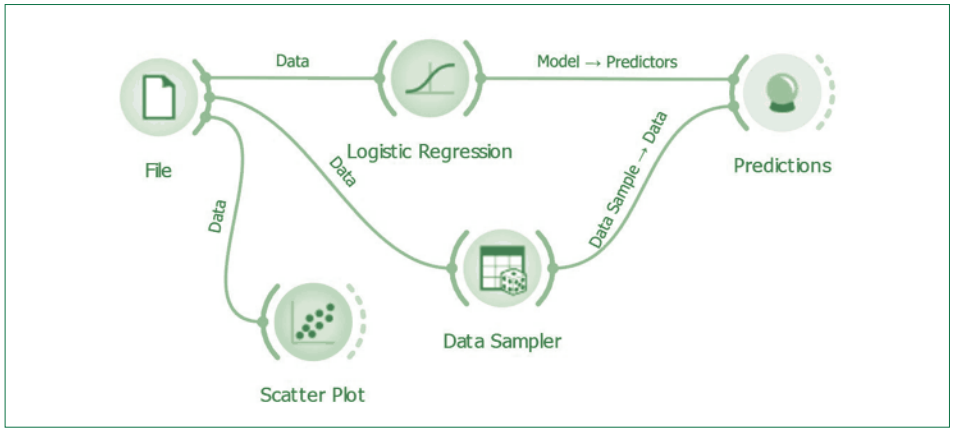
شكل ١٠١ : بيانات تحليل للتَّصنيف

الترتيب الزمني	مدني	مكي	الزمن	السجدة	الأنبياء	عدد الآيات	ترتيب المصحف
c	c	c	d	c	c	c	c
class							
5	0	1	MK	0	1	7	1
87	1	0	MD	0	1	286	2
89	1	0	MD	0	1	200	3
92	1	0	MD	0	1	175	4
112	1	0	MD	0	1	120	5
55	0	1	MK	0	1	156	6
39	0	1	MK	1	1	206	7
88	1	0	MD	0	1	75	8
113	1	0	MD	0	1	129	9
51	0	1	MK	0	1	109	10
...	...	...	...	...	...	...	...

شكل ١٠٢ : بيانات تحليل الانحدار

المبحث الثَّانِي: التَّصنيف

مـن أجـلِ شَـرحِ أُسـلوبِ تَعَلُّـم الآلـة في التَّصنيـف، هُيِّئَـت البَيَانَـاتُ (الشـكل ١٠٠) وفقًـا للمتغـيِّرات الآتيـة: تصنيـف السُّـورَة مـن حيـث مكيَّتهـا أو مدنيَّتهـا، وطـول السُّـورَة (طويـل ومتوسّـط وقصـير). هـذه الـمُتَغَيّرَات فئويَّـة، أمـا المتغَـيّر الهـدف (target) -أو الصِّنـف (class)؛ أي: التَّابـع أو النَّـاتج- فهـو (السَّـجْدَة). لنطـرح هـذا السـؤال: مـا أداء النمـوذج في بيانـات الاختبـار (٢٠٪ مـن مجموعـة البيانـات) لتصنيـف: (وجـود السـجدة مـن عـدم وجودهـا) بنـاءً علـى معطيـات المتغـيرات الفئويـة؟ لبـدء مسـار التَّحْلِيـل وصـولًا إلى تنبـؤات الآلـة بالتَّصنيـفِ؛ يُتّخـذُ مسـار العمـل في أداة أورانج وفـق الشَّـكْل (١٠٣).



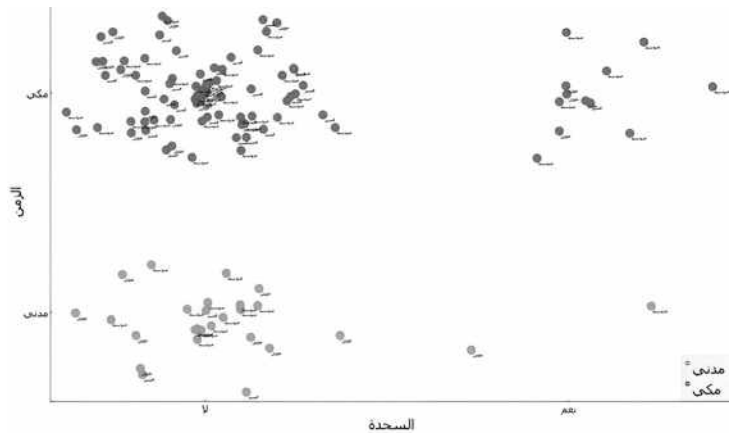
شكل ١٠٣: مسار العمل لتحليل التصنيف بالانْحِدَار اللُّوجسْتِيّ

يُلحـظُ في نتيجـة أداءِ الانْحِـدَار اللُّوجسْـتِيّ للتّصنيـف أن نسـبة (الدقـة) في التَّصنيـف هـي (٠,٧٦٤)، ونسـبة (الاسـترجاع) الـذي يُخبـر عـن نسـبة التَّصنيـف الصحيـح هـي (٠,٨٧٤).

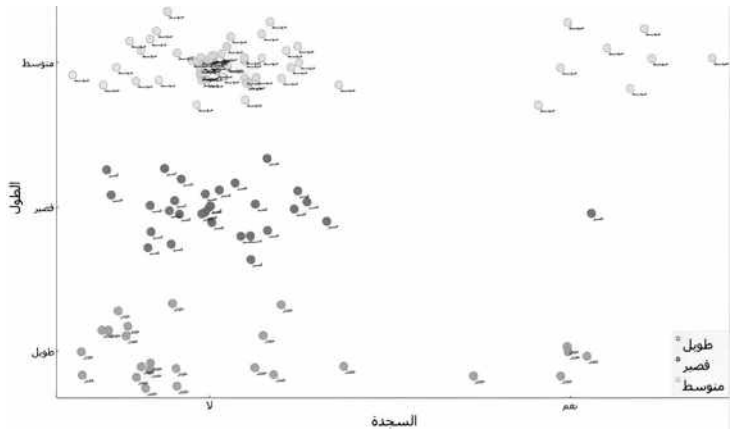
ومـن أسـاليبِ تمثيـل البَيَانَـات المسـتخدمةِ في مسـار هـذا التَّحْلِيـل إجـراءُ مخَطَّـطِ التَّشَـتُّت. أُجـري ذلـكَ لتمثيـل التَّصنيـف بـين السُّـوَر المكِّيَّـة والمدَنِيَّـة علـى محوريـن (المحـور السِّـيني: السَّـجْدَة، والمحـور الصَّـادي: الزَّمـن) كمـا في الشَّـكْل (١٠٤)، ثـم علـى محورَيـن آخَرَيْـن (المحـور السِّـيني: السَّـجْدَة، والمحـور الصَّـادي: الطُّـول) كمـا في الشَّـكْل (١٠٥).

جدول ٣٤ : مصفوفة الارتباك لنموذج الانْحِدَار اللُّوجسْتِيّ

النموذج	الصحة	الدقة	الاسترجاع	مقياس-ف
الانْحِدَار اللُّوجسْتِيّ	٠,٨٧٤	٠,٧٦٤	٠,٨٧٤	٠,٨١٥



شكل ١٠٤ : مخطط التَّشَتُّت لحجم السُّوَر المكِّيَّة والمدَنِيَّة بين السَّجْدَة والزمن

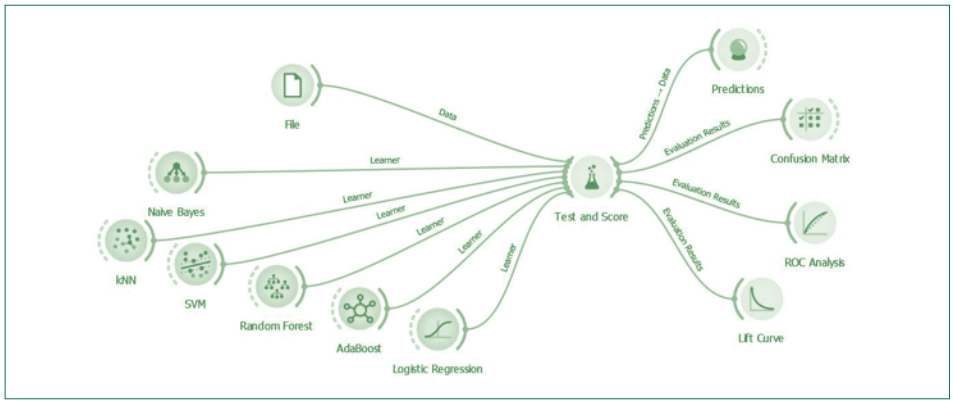


شكل ١٠٥ : مخطط التَّشَتُّت لحجم السُّوَر المكِّيَّة والمدَنِيَّة بين السَّجْدَة والطول

للتَّصنيف في تعَلّم الآلة نماذج أساسِيَّةٌ تُطبّق للتَّدريب والاختبارِ والتَّحقُّق. لمعرفة أداء كل نماذج التصنيف، سنستخدم بيانات (الشكل ١٠١). المتغيرات المستقّلة (المتنبِّئة): السُّوَر من حيثُ الزمن (مَكِّيَّة ومدنِيَّة)، ومن حيث الحجم (طويلة ومتوَسِّطة وقصيرة)، والمتغَيِّر التَّابع أو الناتج: (نعم): وردَ اسمُ نبيّ من الأنبياء؛ و(لا): لم يردْ اسمُ نبيٍّ من الأنبياء. في الشَّكْل (١٠٦) حُمّلت البَيَانَاتُ في أداة (ملف file)، ثم رُبطت مسارات أعمال التَّحْلِيل والمعالجة مع الاختبار والقياسِ (Test and Score)، ثم رُبط الاختبار والقِياس مع سِتَّة نماذج تصنيفيَّة. أما الأدوات التَّمثيلِيَّةُ فَوُضِعتْ بعد أداةِ الاختبار والقياس.

أما النَّماذج فهي ستَّةٌ، وهي على النحو الآتي:

- بايزي البسيط Naïve Bayes
- خوارزمية الجوار الأقرب kNN
- آلة المتجه الداعمة SVM
- الغابة العشوائية Random Forest
- الانْحِدَار اللُّوجسْتِيّ Logistic Regression
- التعزيز التكيّفي AdaBoost



شكل ١٠٦ : مسار التَّحْلِيل لنماذج التصنيف وتمثيل النتائج

في أداة الاختبـار والقيـاس، تُجـرى عـدةُ مناهـج لاختبـارِ أداءِ تصنيـفِ كُلِّ نمـوذجٍ مـن النَّمـاذج السـتة، وهـي علـى النَّحـوِ الآتي:

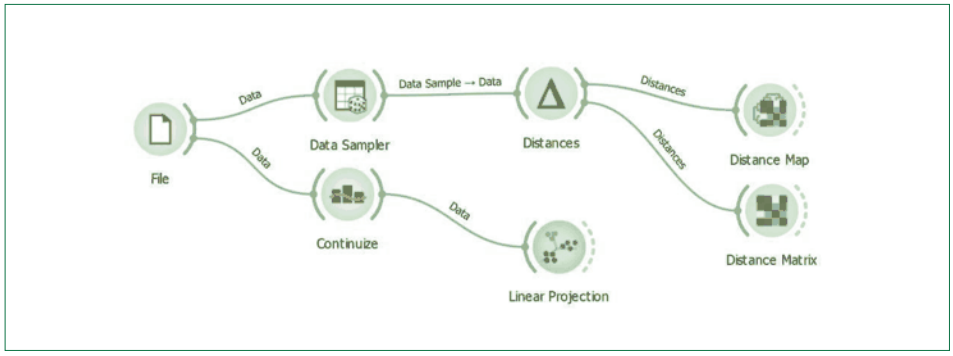
- التَّحقُّـق المتقاطـع (cross validation)، الـذي يُقسّـم البَيَانَـات إلى عـددٍ معـيَّنٍ مـن الطَّيّـات، وعـادةً تكـون ٥ طيّـاتٍ (5-folds) أو ١٠ طيّـاتٍ (10-folds)، ويعمـل علـى اختبـار خوارزميـة النمـوذج عـن طريـق النظـر إلى تصنيـف كُلِّ طيّـةٍ، واسـتنتاج ذلـكَ في الطيّـاتِ الأخـرى، وهكـذا دواليـكَ، إلى أن يُمـرِّرَ كلَّ عـدد الطيّـاتِ المحـدد مسـبقًا.
- العيِّنـةُ العشـوائِيَّة (تدريـب-اختبـار) الـتي تُقَسَّـم فيهـا البَيَانَـات عشـوائِيًّا إلى مجموعـة تدريـبٍ ومجموعـة اختبـارٍ، ويتكـرر هـذا التَّقسـيمُ لعـددٍ محـددٍ في ضبـط التَّدريـب.
- الاختبـار على بيانـاتِ التَّدريـب فقـط، ويُسـتخدم على كامـل البَيَانَـات، ولا يصـل إلا إلى نتـائج خاطئـةٍ، والسَّـببُ يتعَلَّـقُ بـأنَّ النَّمـوذج الـذي يتعَلَّـم على بيانـات التَّدريـبِ، ويختـبر نفسـه فيمـا تعلمـه فقـط؛ لا يحقِّـق هـدف تعلّـم الآلـة في الأسـاس.
- الاختبـارُ على بيانـاتِ الاختبـار فقـط، ويُسـتخدم لكشـفِ أداء الآلـة على التَّصنيـفِ في بَيَانـاتِ الاختبـار (بيانـات لـم تَرَهـا الآلـة).

بعـد اختيـارِ أحـد هـذه المناهـجِ؛ يُتطلَّـب اختيـار صنـفِ الـمُتَغَيّرالهـدف (target) الـذي يُعـدّ في بيانـات ملـف (الأنبيـاء)؛ ثنـائيّ: (نعـم) أو (لا). وهنـاك طريقتـان لاسـتخراجِ مصفوفـة الارتبـاك للتَّصنيـف في مثـل هـذه الحـالاتِ، وهـي أن نسـتخرجَ إمَّـا متوَسّـط الأوزانِ لـكُلِّ الأصنـافِ؛ أي لـ (نعـم) و(لا) في مثالنـا هنـا، أو أن نختـارَ صنفًـا واحـدًا منهمـا. اخـترتُ هنـا التَّحَقُّقَ المتقاطـع ثـم متوَسّـطَ أوزان الأصنـافِ (Average over classes) لـ: (نعم) و(لا). سـنلحظُ أن أفضـلَ أداءٍ مـن هـذه النَّمَـاذجِ قـد عُولـجُ بِنمـوذجِ بايـزي البسـيط ثـم نمـوذج خوارزميـة الجـوار الأقـرب وفـق نسـبِ كلٍّ مـن الدِّقّـة والاسـترجاع (الجـدول ٣٥).

جدول ٣٥ : مقاييس مصفوفة الارتباك لنماذج التَّصنيف

نوع التَّحْلِيل: ٥ طيّات تحقق متقاطع cross-validation الصنف: المُتَغَيّر الهدف (متوسط أوزان الأصناف)					
النموذج		الصحة	الدقة	الاسترجاع	مقياس-ف
بايزي البسيط	Naïve Bayes	0.702	0.707	0.702	0.693
خوارزمية الجوار الأقرب	kNN	0.693	0.761	0.693	0.761
آلة المتجه الداعمة	SVM	0.693	0.699	0.693	0.683
الغابة العشوائية	Random Forest	0.693	0.699	0.693	0.683
الانْحِدَار اللُّوجسْتِيّ	Logistic Regression	0.693	0.699	0.693	0.683
التعزيز التكيّفي	AdaBoost	0.675	0.674	0.675	0.672

لنقـم بإنشـاء مسـارٍ عملِـيٍّ تحليلِـيٍّ كمـا في الشَّـكْل (١٠٧) للبيانـات التَّصنيفِيَّـة للأنبيـاء، لننظـر إلى حسـاب المسـافات ومصفوفتهـا.



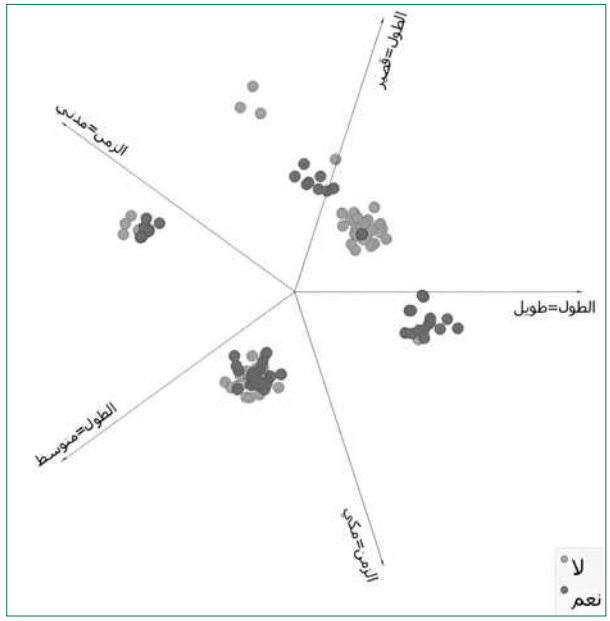
شكل ١٠٧ : مسار العمل لتحليلِ بيانات الأنبياء التصنيفِيَّة

سـأجري حسـابَ المسـافة الإقليدِيَّـة، ويمكـن اختيـار غـيرِه مـن أداة المسـافاتِ (Distances). في الشَّـكْل (١٠٨) أسـلوبُ قيـاس المسـافات بـين الصِّنـفِ (class) -نعـم yes ولا no- لـورودِ أسـماء الأنبيـاء بـين السُّـوَر المكِّيَّـة الطَّويلـة والمتوسّـطة والقصيـرة، وبـين السُّـوَر المدَنِيَّـةِ الطَّويلة

والمتوَسّـطة والقصـيرة، وقـد اسـتخدمتُ هيكلـة البيانـات الـواردة في الشَّـكْل (١٠١). وبلمحـةٍ عامَّـةٍ علـى الشَّـكْل (١٠٨) الـذي يُصـوّرُ المسـافاتِ المقاسـة للمتغـيّر الهـدف؛ فتُشـيرُ المرَبَّعَـاتُ البَيضَـاء إلى عـدم وجـود مسـافة؛ لأنَّ نقطتي الصنـف (نعـم-لا) تمثـلان الصِّنـفَ نفسـه، وتُشـيرُ المربَّعَـاتُ الخضـراء إلى قيمـةِ بُعْـدِ المسـافة بـين كلِّ نُقطَتَـيْن. وأفضـل تمثيـل لفهـم المسـافاتِ هـو مـا يُمثَّـل بالإقحـام الخَطِّـيِّ (linear projection) كمـا في الشَّـكْل (١٠٩)، وخلاصـة النتيجـة بالنظـر إليـه هـي أنَّ أكـثر السُّـوَر الـتي وردت فيهـا أسـماء الأنبيـاء هـي السُّـوَر المكِّيَّـة المتوسّـطة، تليهـا السُّـوَر المكِّيَّـة الطَّويلـة، ثـم السُّـوَر المدَنِيَّـة القصـيرة، ثـم السُّـوَر المدَنِيَّـة المتوسـطة.

	نعم	نعم	لا	لا	نعم	نعم	لا	لا	لا	لا	لا	لا	لا	لا	لا	نعم	نعم	نعم	نعم	لا	لا	نعم	نعم	نعم	نعم	لا	لا	نعم	لا	نعم	نعم
نعم		0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000	1.414	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	0.000
نعم	0.000		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000	1.414	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	0.000
لا	1.000	1.000		1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
لا	1.000	1.000	1.000		1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
نعم	1.000	1.000	1.414	1.414		0.000	1.414	1.414	0.000	1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.414	1.000	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.000
نعم	1.000	1.000	1.414	1.414	0.000		1.414	1.414	0.000	1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.414	1.000	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.000
لا	1.000	1.000	1.000	0.000	1.414	1.414		1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000		1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
لا	1.000	1.000	1.414	1.414	0.000	0.000	1.414	1.414		1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.414	1.000	1.414	1.414	1.414	1.414	1.414	1.000	1.000	1.414	1.000
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414		0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000		1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000		0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000		1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000		1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000		1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000		1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
نعم	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000		1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414		0.000	1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000	0.000	0.000	1.000	1.414
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000		1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000	0.000	0.000	1.000	1.414
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000		1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000		1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000	0.000	1.000	1.414	1.000	1.000	1.414	1.000	1.000	1.000	0.000	0.000	1.000	1.414
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000		1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000		1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000	0.000	1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000	0.000	0.000	1.000	1.414
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000		1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000	0.000	1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000	0.000	0.000	1.000	1.414
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000		1.000	0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	1.000	0.000	1.414	1.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.414	1.414	1.000	0.000	1.414	1.000	0.000		1.000	1.000	1.414	1.414	1.000	1.000
نعم	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000		0.000	0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	0.000	1.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000		0.000	1.000	1.000	0.000	1.000
لا	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000		1.000	1.000	0.000	1.000
نعم	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000	0.000	1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000		0.000	1.000	1.414
لا	1.414	1.414	1.414	1.000	1.000	1.000	1.000	1.414	1.000	1.000	1.000	1.414	1.414	1.000	1.414	1.000	1.414	0.000	0.000	1.000	1.414	0.000	1.000	1.414	1.000	1.000	1.000	0.000		1.000	1.414
نعم	1.000	1.000	1.000	0.000	1.414	1.414	0.000	1.000	1.414	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000		1.000
نعم	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	1.000	1.414	1.000	1.000	1.000	1.000	1.000	1.414	1.414	1.000	

شكل ١٠٨ : قياس المسافات بين الصنف class (نعم ولا)



شكل ١٠٩ : تمثيل الإقحام الخطي

المبحث الثَّالِث: الانْحِدَار الخَطِّيّ

يُسمّى الانْحِدَار الخَطِّيّ بالتَّربيعات الصُّغرى المعتادة (ordinary least squares)، ويُسمّى -أيضًا- بالتَّربيعات الصُّغرى الخَطِّيَّة (linear least squares). في تعلّم الآلة؛ يتعلّم نموذج الانحدار ليُستخدم على بياناتٍ جديدةٍ، على عكس نموذج الانحدارِ في الإحصاء (المبحث الأول في الفصل الرابع من البَاب الأَوَّل) الذي يُكتفى به في الإخبارِ عن أداء الانحدار الخطّي.

ينقسمُ الانحِدارُ الخَطّي إلى نوعين. الأوّل: الانحِدارُ الخَطّي البسيط الذي يُستخدمُ في حال أردنا إيجادَ علاقةٍ خطِّيَّةٍ بين متغيِّرٍ مستقلّ أو مُتَنَبِّئ predictor واحدٍ ومُتغيِّرناتج (تابع). الثَّاني: الانحِدارُ الخَطّي المتعدِّد الذي يُستخدم في حال أردنا إيجادَ علاقةٍ خطِّيَّةٍ بين متغَيِّراتٍ مستقلّة ومُتغيِّرناتج (تابع).

يَفترضُ نموذجُ الانْحِدَار الخَطّيّ وجودَ علاقةٍ بين الـمُتَغَيّرَ المتنبئ (المسْتَقِلّ) والمتغَيِّر النَّاتج (التَّابع). يجبُ أن تكون الـمُتَغَيّرَات كُلُّها في نموذج الانْحِدَار الخَطّيّ مقياسِيَّة (أرقام)، ولا تكون فئوِيَّة. إن كانت فئويةً؛ فيُجرى بالضَّرورة أسلوبُ ترميز النُّقطة السَّاخنة (one-hot encoding)، ومثاله سنراه في الشَّكْل (١١٠). يتدرَّب نموذج الانحدار في تعلم الآلة من القيمِ الحقيقيَّة للمتغَيِّر النَّاتج وفق فهمِها لعلاقتها مع قِيَمِ الـمُتَغَيّرَ النَّاتج (التَّابع)، ثم يُختبر النَّموذج على بيانات اختبار، ثم يتنبّأ بقيمِ المتغَيِّر النَّاتج (قيم متنبّأ بها predicted values) حسب ما تعلّمته الآلةُ من الخصائص والمعطيات في المتَغَيِّرِ المسْتَقِلّ. تُقاس دقة تعلم هذا النموذج بالمقارنةِ بين القيم الحقيقيّة في البيانات والقيم التي تنَبَّأت الآلة بها، فإن كانت القيمُ التي تنبأت بها بعيدة، فأداء التعلّم عندئذٍ يكون ضعيفًا.

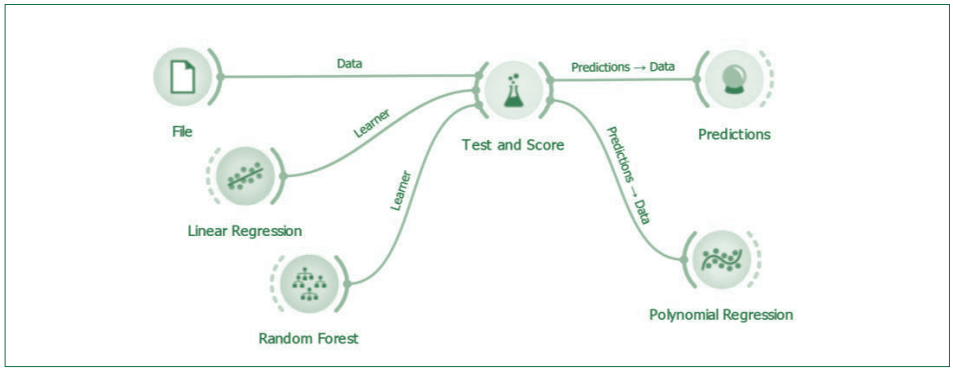
لتَطبيقِ نموذجِ الانحِدَارِ على مُتَغَيّرَات الشَّكْل (١٠٢)، نُجري تحويل الـمُتَغَيّرَات الفئوِيَّة (أسماء السُّور) إلى متغَيِّراتٍ رَقَمِيَّة. وهناك طريقةٌ في عِلْم البَيَانَاتِ تعملُ على تحويل الـمُتَغَيّرَات الفئوِيَّة إلى مُتَغَيّرَاتٍ رَقَمِيّة ثنائِيَّةٍ dummy باستخدام ترميز النُّقطة السَّاخنة[1] one-hot encoding. وفي هذه الحالةِ؛ لا بدَّ من تحويل عمود الزَّمن الذي يَتَضَمَّن مُتَغَيّرَاتٍ فئوِيَّةً إلى مُتَغَيّرَاتٍ مقياسِيَّةٍ، عن طريق فصلها بترميز النقطةِ السَّاخنة، وجعل المكِّيِّ في عمودٍ مستقِلٍّ، والمدنِيِّ في عمودٍ مستقِلٍّ، ثم إزالة عمود الزمن كما في الشَّكْل (١١٠). الـمُتَغَيّر الهدف (أي: النَّاتج أو التَّابِع) في بيانات هذا الشَّكْل هو (التَّرتيب الزمنِيّ). والسُّؤال المطروح: هل يُمكن التنَبُّؤ بالتَّرتيب الزمنِيِّ وفق الـمُتَغَيّرَات المستَقِلَّة: (ترتيب المصحف، وعدد الآيات، وورود أسماء الأنبياء، ووجود السَّجْدَة، وفيما إن كانت السُّورةُ مَكِّيَّةً أو كانت مدنِيَّةً)؟ يكون الجوابُ إما بالإمكان مع أداء عالٍ، أو بعدم الإمكانِ مع أداء ضعيفٍ، وتتبَيَّن جودةُ التنَبُّؤ بارتكاز أو قرب نقاط الـمُتَغَيّرَات على خَطِّ الانحدار. سننظر إلى النتيجة بعد تحميل الملفِّ، وإعداد مسار التَّحْلِيل وفق الشَّكْل (١١١).

---

(١) لمزيد من التدريب على أسلوب التعامل مع الـمُتَغَيّرَات الفئوِيَّة في البَيَانَات الاطلاع على تطبيقات الانْحِدَار الخَطّيّ Linear Regression Tutorial في الرابط الآتي:
https://www.kaggle.com/sudhirnl7/linear-regression-tutorial (تاريخ الاسترداد ٢٤ أكتوبر، ٢٠٢٢).

الترتيب الزمني	مدني	مكي	الزمن	السجدة	الأنبياء	عدد الآيات	ترتيب المصحف
c	c	c	d	c	c	c	c
class							
5	0	1	MK	0	1	7	1
87	1	0	MD	0	1	286	2
89	1	0	MD	0	1	200	3
92	1	0	MD	0	1	175	4
112	1	0	MD	0	1	120	5
55	0	1	MK	0	1	156	6
39	0	1	MK	1	1	206	7
...	...	...	...	...	...	...	...

شكل ١١٠ : مثال لبيانات تحليل الانْحِدَار الخَطِّيّ



شكل ١١١ : مسار العمل لتحليل الانْحِدَار الخَطِّيّ

قبل النَّظرِ في أداء نموذجِ الانْحِدَار الخَطِّيِّ، لابدّ من إجراء تحليلِ الارتباط بين المتَغَيِّرات. أشرتُ في الجدولِ (٣٦) إلى قيم الارتباط بين المُتَغَيِّرَات، وجميعُها ضعيفةٌ، ويُقاس ذلك بمنطقِ +١ و-١، ويتمثَّلُ الضّعف بالجنوحِ إلى القيمة (-1) التي تدُلُّ على العلاقَة السَّلبِيَّةِ، أو القيمة (0) التي لا تدُلُّ على أيِّ علاقةٍ، على عكس الجنوح إلى القيمة (+1) التي تدلُّ على

وجود علاقةٍ إيجابيَّةٍ. ليس بالضرورة أن تكون هناك علاقةٌ خطِّيَّةٌ عندما يكون هناك ارتباط من عدمه بين المتَغَيِّرات، فالارتباطُ إذن ليس بالضرورةِ أن يكون سَبَبًا في وجودِ علاقةٍ خطِّيَّة.

تُؤكِّد الأبحاثُ التَّجريبيَّةُ إلى أنَّ تفسيرَ قيمةَ الارتباط من البَيَانَاتِ غير اللغوية مختلفٌ عن تفسيرها من البَيَانَاتِ اللغويّة، ويعتمدُ ذلك على عددِ الملاحظات وعددِ عناصر البَيَانَاتِ ومعلوماتها (المشاهدات observations). نحتاجُ إلى ارتباطٍ كبيرٍ بقيمةِ (0.63) فأعلى عندما يكون في البَيَانَاتِ عشرُ ملاحظات، وقد تكونُ قيمةُ الارتباطِ الأقل من (0.2) ذات دلالةٍ وحجمِ تأثيرٍ عالٍ للملاحظات التي تزيد عن المئة حالة، كما هو حالُ عددِ ترتيب سورِ القرآن زمنيًّا (١١٤)، كما قد تكون ذاتَ دلالة إحصائية لعدد الملاحظات التي تزيد عن الألف عندما تكون قيمة الارتباط (0.06) فأعلى. يجدرُ التأكّدُ من عدد الملاحظات في البَيَانَاتِ اللغوية، والإفصاحُ عن مُعاملِ الارتباطِ النَّاتجةِ عن التحليل، نظرًا لأهميةِ هذا العدد في كشفِ طبيعةِ الارتباطات بين المتغيراتِ اللغويَّةِ[١].

جدول ٣٦ : قيم الارتباط بين مُتَغَيّرَات بياناتِ نموذج الانْحِدَار الخَطِّيّ

التسلسل	المتَغَيِّر	المتَغَيِّر	قيمة الارتباط
1	الترتيب الزمني	مكي	-0.736
2	الترتيب الزمني	ترتيب-المصحف	-0.419
5	الترتيب الزمني	السَّجْدَة	+0.031
4	الترتيب الزمني	عدد الآيات	+0.118
3	الأنبياء	الترتيب الزمني	+0.277

بعد تحليلِ الارتباط، يُجرى الاختبار والقياس (Test and Score) لأداء نموذج الانْحِدَار الخَطِّيّ بين المتغيِّرات في أداة أورانج، وذلك بمتوسّط الخطأ التَّربيعيّ (Mean Square Error MSE)

(١) بريزينا، فاكلاف. علم الإحصاء في لُغَوِيّات المُدَوّنَة الحَاسُوبِيَّة. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، الطبعة الأولى، ٢٠٢١، الصفحات ١٨٨-١٨٩.

الـذي يُظهـر معدّلَ الخطأ الذي يكون مربّعه هو متوسّـط مربَّعاتِ كُلِّ الأخطاء، وبجذر متوسّـط الخطـأ التَّربيعـيّ (Root Mean Square Error RMSE) الـذي يقيـس مـدى تباعـد القيـم المتنبَّـأ بهـا عـن القيـم الحقيقيـة في البَيَانات، وبمتوسّـطِ الخطـأ المطلـق (Mean Absolute Error MAE) الـذي يقيـسُ إمكان التنَبُّؤ مـن عدمـه بالنَّتـائجُ النهائيَّـة، وبمُعامـل ر$(R^2)^2$ الـذي يقيـس نسـبة التبايـن في المتغـير التابـع، وتكـون قيمـة النسـبة بـين (١) و(٠). وعنـد تقييـمِ مـدى مُلاءَمَـةِ نمـوذج الانْحِـدَارِ الخَطِّـيّ، يُسـتعان بقيـاس كُلٍّ مـن جـذر متوسّـط الخطـأ التَّربيعـيّ (RMSE) ونسـبة مُعامـل ر$(R^2)^2$، وذلـك وِفْـقَ المعطَـى الاطِّـراديِّ الآتي:

- متوسّـط الخطـأ التَّربيعـيّ (RMSE): كُلَّمَـا انخفضـتْ القيمـة، كان النمـوذج أدعـى إلى التَّـلاؤم مـع مجموعـة البيانـات.
- مُعامـل ر$(R^2)^2$: كُلَّمَـا زادتْ قيمـةُ النِّسـبة إلى (1)، كان النَّمـوذجُ أدعـى إلى التَّـلاؤم مـع مجموعـة البيانـات.

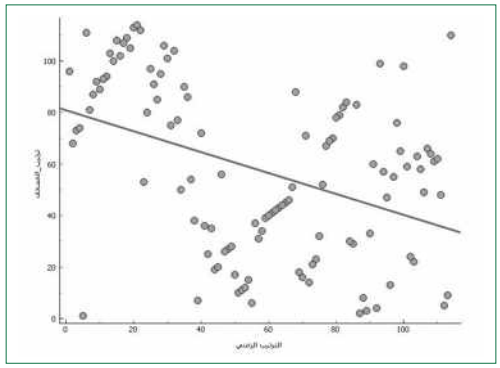
نجـدُ في الجـدول (٣٧) مُدخـلاتِ الاختبـار القيـاس، الـتي اخترنـا فيهـا اختبـار النَّمُـوذجِ بأسـلوبِ العَيِّنـة العشـوائيَّة (تدريـب-اختبـار)، ثـم بأسـلوب التَّحقُّـق المتقاطـع (٣ طيّـات 3 folds). جـاءت معـدلاتُ كُلّ القِيـم عاليـةً، وبنسـبة تبايـن تجـاوزت النصـف (ر 0.610 =2 ثـم ر0.578 =2)، ولا يُمكـن أن يكـون التَّنبُّـؤ الآليِّ ذا كفـاءة؛ لأنـه يصعـب علـى الآلـةِ تَحقِيقَهـا لحالـة هـذه البَيَانَـات اللغويـة الـتي تَتَّسـمُ بفـرط التخصيـص[1] (overfitting) للمتغـيرات اللغويـة (أي: أنّ النمـوذج قـد حفـظ البَيَانَـات دون فهـمٍ).

(١) عَكسُ «فـرط التَّخْصِيـص» هـو «فـرط التَّعميـمِ» (underfitting) في البَيَانـاتِ اللُّغويّـة، ويشـيرإلى أن النمـوذج لـم يتعلّـم شـيئًا. تؤكـد حالـة «فـرط التخصيـص» أو حالـة «فـرط التَّعميـمِ» علـى أهميـة العنايـة بتهيئـة البيانـات في تعلّـم الآلـة، وأنهـا هـي الأسـاس لتحقيـق التعلّـم، ويترتّـب في حـالات «فـرط التخصيـص» التفريـطُ في بعـض الخصائـص الشـاذة في البيانـات اللغويـة، كمـا يترتَّـبُ في حـالات «فـرط التَّعميـمِ» زيـادة التخصيـص مـن خصائـص البيانـات، بغيـةَ تحقيـق هـدف التعلّـم الآلي. هـذا كلـه علـى خـلاف مـا أُلاحظُـه في متَغَـيِّراتِ البيانـات اللغويـة في علـم الإحصـاء، الـتي تُقـاس وتُحلَّـل وتُسـتخرج منهـا المعلومـات والـدلالات الإحصائيـة، ويُفصـح عـن معنويتهـا مـن عدمهـا وعـن تأثيرهـا مـن عدمـه، دون الحاجـة إلى الاعتنـاء بحـالات الفـرط في التَّخْصِيـص أو التَّعميـم.

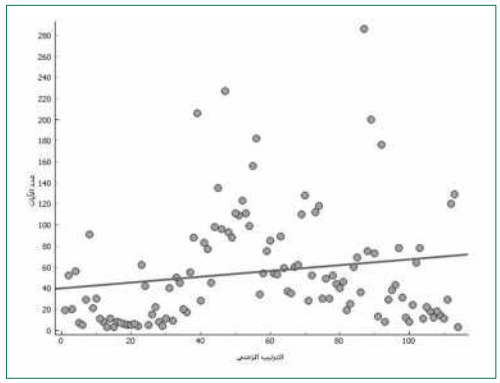
جدول ٣٧ : الاختبار والقياس (الانحدار الخطّي)

	العَيِّنة العشوائيَّة Random sampling (تدريب-80% اختبار 20%)؛ عدد مرات التعلم: ١٠ مرات				التَّحققُ المتقاطع Cross validation (عدد الطيّات: ٣) Regularization			
	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
الانْحِدَار الخَطِّيّ	420.235	20.500	16.059	0.610	457.015	21.378	16.431	0.578

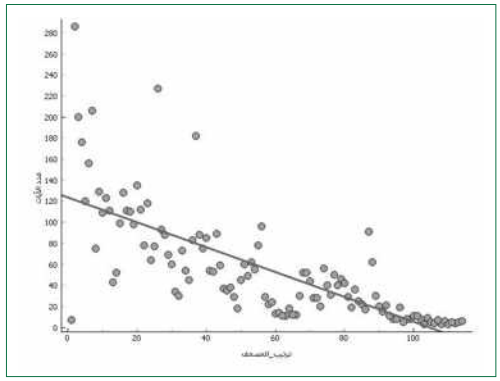
في الشَّكْل (١١٢) تمثيلٌ لخطِّ الانحدار لثلاث حالات؛ وهي (أ): التَّرتيب الزَّمني [مستقلّ] وترتيب المصحف [تابع]؛ و(ب): التّرتيب الزَّمنيّ [مستقل] وعدد الآيات [تابع]؛ و(ج): ترتيب المصحف [مستقل] وعدد الآيات [تابع]، وقد مُثّلَ خطُّ الانحدار بمتعدّد الحدِّيَّة (polynomial)، والانحدار غير متلائم فيها جميعًا مع نقاط البَيَانَات للمُتَغَيِّرَات بين المتغيّر المستقلّ والمغيّر التَّابع (النَّاتج).



(أ) الترتيب الزمني وترتيب المصحف



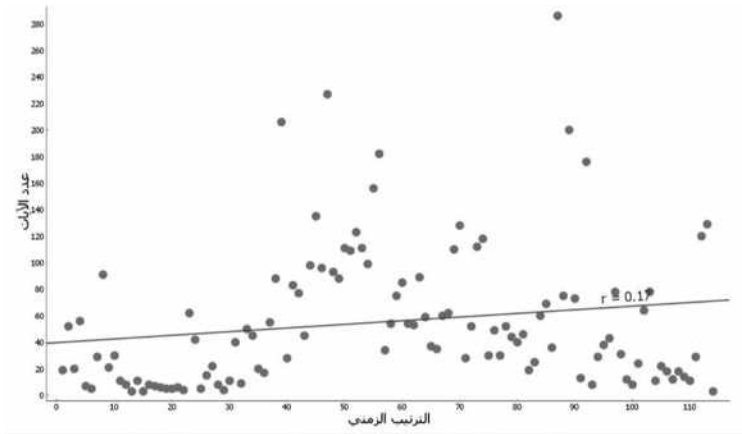
(ب) الترتيب الزمني وعدد الآيات



(جـ) ترتيب المصحف وعدد الآيات

شكل ١١٢ : خط الانحدار البسيط (متغير مستقلّ واحد مقابل متغير تابع واحد)

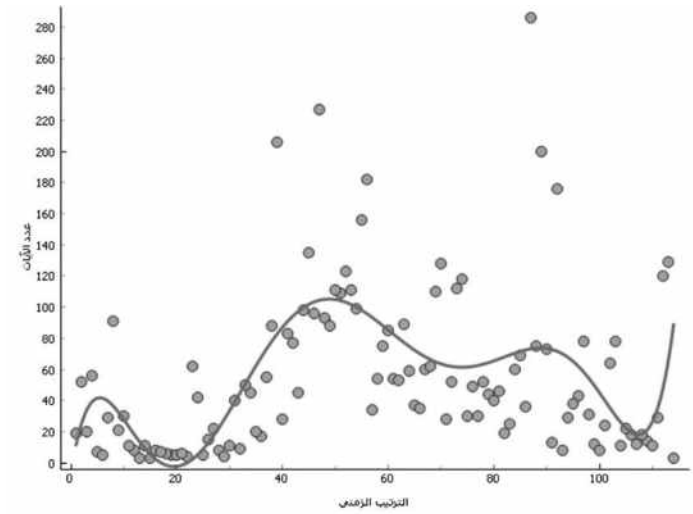
تتضمن أداة أورانج للانحدار الخطي خياراتٍ لمزيد من التحسين، وذلك بإيجَادِ أفضل نموذج خَطِّيٍّ للتنَبُّؤ. ولطرح بعض الحلولِ؛ طُبِّق تنظيمُ انحدارِ ريدج (Ridge regression) بين التَّرتيب الزَّمنِيِّ بصفته المتَغَيِّر المستَقِلّ وبين عدد الآيات بصفته المتَغَيِّر التَّابع (الشكل ١١٣: ب)، وبمعدّل خطأ: ٣٤,٩١٢. لابدّ من الإفصاح عن جذر متوسط الخطأ التربيعي (Root Mean Square Error RMSE) ونسبة التباين لـ (ر٢) إن لم يتحقق تحسينُ التعلّم بأي حلٍّ من حلول التنظيمِ (regularization)؛ كاستخدامِ انحدار ريدج (Ridge regression) الذي يُساعدُ في تقليل التباين الكبير، أو انحدار لاسّو (Lasso regression) الذي يُساعد على التَّقليل من متغيّرات (التَّابع) وعلى التَّقليل من فَرْطِ التَّخصيص.



شكل ١١٣ : مخطط التَّشَتُّت بخط الانحدار بين الترتيب الزمني وعدد الآيات

يُمكن في مثل هذه الحالة استخدامُ نماذج الخَطِّ المتناسبةِ (fitted line models)، وبنموذج الانحدار التَّدريجِيِّ (stepwise regression). أُجري هذا النموذج بأفضل تمثيل على المتغيّرَين: (التَّرتيب الزَّمني وعدد الآياتِ). نلحظُ أنَّ أقربَ تمثيلٍ للخطِّ بين التَّرتيب الزمنيِّ وعدد آيات السُّوَر قد ظَهَرَ كَما في الشَّكْل (١١٣)، وبقيمة ارتباط تُساوي (0.17)،

وهـي قِيمـةٌ قليلـةٌ تـدلُّ علـى وجـود ارتبـاط ضعيـف، ولكنّهـا قيمـةٌ قـد تكـون ذات ارتبـاطٍ مُهـمٍّ في البَيَانـاتِ اللُّغويَّـة.



شكل ١١٤ : الانحِدَارُ المتَعَدِّد الحُدُود بين الترتيب الزمني وعدد الآيات

ثمَّـةَ عـدةُ أسـاليب لتعدُّديّـة الحـدود، أي: حـدود الدَّرجـاتِ المتعـدِّدَة لنقـاطِ البَيَانـاتِ بـين المتغَـيِّرِ المسـتقلّ (الترتيـب الزمـني) والمتغَـيِّرِ التَّابـعِ (عـدد الآيـات)، ولنَمْذَجَـةِ العلاقـة بـين المتغـيِّراتِ مُتعـدّدةِ الحُـدود، يتعـيّن اسـتخدامُ نمـوذجِ الانحـدار اللَّاخَطّـي، ومـن أنواعـه مـا نجـده في أداة الانحـدار الخطّـي linear regression في أُورانج، ويُسـمّى الانحِـدَار المتعـدِّد الحُـدُود (polynomial regression)، الـذي يُمثّـل الشـكل المنحـني الأكـثر ملاءمـةً لمجموعـة البَيَانـات عندما تكـون العلاقـة بـين المتغَـيِّر المسـتقلّ وبـين المتَغَـيِّر التَّابـع غـير خطيّـة. وفي الشـكل (١١٤) تمثيـلٌ للمنـحى الأفضـل الـذي أوجدتُـه باختيـار سـبع دَرَجَـاتٍ مـن دَرَجَـةِ تعدُّدِيَّـةِ الحُـدُودِ (polynomial degree)، وهـو المنحـنى الـذي يُعـدُّ الأكـثَرُ مُلاءَمَـةً لنُقـاطِ البيانـات بـين المتغـيِّرِ المسـتقلِّ (التَّرتيـبِ الزَّمَـنيِّ) والمتغـيِّرِ التَّابِـع (عـدد الآيـات).

الخاتمة

جاء هذا الكتاب في بابين. الأول: الإحصاء اللُّغوي، والثاني: تنقيبُ النُّصُوصِ والبَيَاناتِ العربيَّة. أجريتُ تطبيقات الإحصاء اللغوي وتطبيقات تنقيب النصوص والبيانات على مدونة القرآن الكريم، واحتوى كل فصل من فصول البابين على تطبيقات محددة في مباحثه. اخترتُ أدوات لانكستر الشبكية لتطبيقات التحليل الإحصائي لمدونة القرآن الكريم في الفصول الخمسة من الباب الأول، نظرًا لشمولية أدوات مُعالجتها التحليلية للمدونة اللغوية، كما اخترتُ أداةَ أورانج لتنقيب النصوص والبيانات في معالجة مدونة القرآن الكريم وتحليلها في الفصول الخمسة من الباب الثاني، ذلك أن هذه الأداة الجاهزة قد صُمِّمَت من نماذج آليَّةٍ مُطَوَّرةٍ في البايثون.

ومع تدفُّقِ النُّصُوصِ والمعلوماتِ المنضَويةِ فيها، وضَخَامةِ حَجمِ هذا التَّدفُّقِ من المدوّناتِ اللُّغويةِ، باتَ تطبيقُ الممارساتِ التحليليّةِ مُلحًّا، سواءً كان بالنِّماذجِ الإحصائيَّةِ في البَيَاناتَ اللُّغويَّةِ كما في فُصُولِ الباب الأوَّل، أو كانَ بالمعالجةِ الآليَّةِ للبَيَاناتَ اللُّغويةِ كما في فُصُولِ الباب الثَّاني. والذي أَقتَرِحُهُ عند إجراءِ تطبيقاتِ فُصُولِ الباب الأوَّل، وتطبيقاتِ فُصُولِ الباب الثَّاني، أنْ يُجريَ المتدربُ بعد قراءة كل مبحث تطبيقًا على بيانات هذا الكتاب، المرفوعةِ في موقع Kaggle؛ لكي يتأكّدَ من تَطابُقِ النَّتائجِ في كُلِّ مبحث أو عندَ كلِّ تطبيقٍ تحليليٍّ واحدٍ من مباحث البَابَيْن. يَتَبقّى على المتدرِّبِ بعد صياغة أسئلة بحثِهِ أو بعد صياغة فرضيات تحليله الكمي (في: الإحصاء اللُّغويّ) أو بعد تحديده مهمةً من مَهَامِّ معالجة النصوصِ العربية (في: تنقيب النُّصُوص والبَيَانات اللُّغَويَّة) أن يُهيِّئَ البيانات الخاصة به بالهيكلِ أو الشَّكلِ الذي صَمَّمْتُ به شَكْلَ البَيَاناتِ في كُلِّ تطبيقٍ تحليليٍّ مُحدَّدٍ في مباحث فصول البابَين من هذا الكتاب.

وأختم في الآتي بنتيجتين عامّتَين، وبوجهَتَيْن أساسيَّتين، كُلُّ واحدةٍ منهما تخصّ بابًا من البابين، مع ذِكْر التَّوصِيَات التي تخصُّ كلَّ بابٍ على حدة.

أَوَّلًا: عن الإحصاء اللُّغَوِيّ في البَاب الأَوَّل

النتيجة

من ناحية ما ظهر من نتائج من تحليلِ مُدَوّنَة القرآن الكريمِ؛ فأهَمُّها أَنَّ الكثافة المعجَمِيّة في القرآنِ عاليةٌ، وقد قاربت ٧٠٪، وهذا ممَّا لم أقف عليه في النُّصُوص البشريَّةِ، وأن مقاييس التَّشَتُّت قد أكَّدت على أنه لا يوجد تباينٌ بين أجزاء القرآن، فكانت النِّسبة متدَنِّيَةً جدًّا (١٣٪)، كما أنَّ التَّوْزِيع المتَّزن قد بلغ نسبةً عاليةً، تزيد عن ٩٨٪، ولم أقفْ -أيضًا- على مثل ما يُقارب هذه النسبَ في مُدَوّنَات اللغة العَرَبِيَّة، إذ يَغلبُ عليها ارتفاعُ نِسَبِ التَّبَايُنِ عن هذه النِّسْبَة. ومن أهَمّ النتائج -أيضًا- ما أجريتُه في مقاييس التَّصَاحُب اللَّفْظِيّ، وما تُفيدُ به قِيَمُ هذه المقاييسُ من تحليلٍ إحصَائيٍّ معيّن في تحليل الخطاب، سواءً لاستنتاج ندرةِ وحصرِيَّة التَّصَاحُب مَعًا، أو لاستنتاج حصرِيَّة التَّصَاحُب فقط، أو لاستنتاجِ أثر التَّصَاحُب وكثرتِه، أو لاستنتاج معنوِيَّة التَّصَاحُب، أو لاستنتاجِ حجم زيادة نمُوِّ التَّصَاحُب، أو لاستنتاج ثنائية الاستعمال التبادلِيِّ للتَّصاحب. ومثالُ الأخير: استخدام مقياس دلتا-ب، وما أعطاه من نتيجةٍ مكونةٍ من قيمتين؛ بَرْهَنَتْ لنا أنَّ الاستخدام التبادليَّ أو الثُّنائيّ الحدّ (binomial) لـ ("غفور رحيم"، و"رحيم غفور": 0.803 ;0.844) ممكنٌ، وفق ما أفاد به هذا المقياسُ من مُدَوّنَة القرآن الكريم، وأنَّ الاستخدامَ التبادلِيَّ لـ["الباقيات الصالحات" و"الصالحات الباقيات": 0.032 ;0,999] غيرُواردٍ في السِّيَاقِ الاستعماليّ، وفق ما أفاد به هذا القياس من مُدَوّنَة القرآنِ الكريم؛ لأن القيمتَين متباعدتان جدًّا. ومن أهَمّ النَّتائج -أيضًا- ما جاء في الارتباط بين أقسام الكلامِ في مُدَوّنَة القرآن، فقد جاءَت أقوى الارتباطات إيجابيَّةً بين الأسماء والصِّفات ثم بين الضَّمائر والأدوات، ثم بين الأفعالِ والضَّمائر، وجاءت أقوى الارتباطاتِ سلبِيَّةً (أي انعدامًا) بين الأسماء والضَّمائر، ثم بين الأسماء والأدوات، ثم بين الصِّفات والأدوات. وفي عرض تحليل التَّنَاظُر بين السُّوَر السَّبع الطّوالِ؛ ظهرت السُّورَةُ السابعةُ (الأنفال والتوبة) متميزةً في استخدام الأفعالِ والأدوات، وتميّزت السُّورَة الثالثة والرَّابعة والخامسة بالصِّفات، وتميّزت كُلّ السُّوَر عدا السُّورَة الثَّالثة بالأسماء.

الوِجهة

أَوَّلُ ما يَرُومُه القارئُ ويَطلبُه في الإحصاءِ اللُّغويِّ كي يَتَغَيَّا تَطبيقاتِه ويَجدَّ في استِخدَاماتِه هو أن يتآلفَ مع كُلّ أُسْلوبٍ من الأساليب الإِحْصَائِيَّة. اتَّجهْ إلى التَّفكير الإِحْصَائِيِّ؛ ذلك لأنّه داعمٌ قويٌّ، ورادفٌ متينٌ لما يُراد استنتاجه في التفكير النَّوْعِيِّ، فالتَّحْلِيل النَّوْعِيّ ووسائل التَّفكير لم يعودا كافِيَين بأدواتهما، ولن تكونَ مقبولةً إن لم ترتكز على تفكيرٍ إِحْصَائِيّ في مستقبل الأبحاثِ اللُّغَوِيّة والإِنْسَانِيّة التي تهتَمُّ بالمادة اللُّغَوِيّة. أَمَّا مَن يقولُ بأَنَّ لغةَ الأرقام قد تخدعُ؛ فإنَّ الممارسين في علوم البَيَانَات يُدركون ما هو صحيحٌ وما هو غير صحيحٍ، فعِلْم البَيَانَاتِ بالنّسبةِ إليهم ليس تسويقًا يُمكن أن يُمارَس فيه التَّضليلُ.

التَّوصيات

المعرفةُ النَّظرية للمُعادلات والخوارزميَّات ليست مطلوبةً. المطلوبُ معرفةُ كيفِيَّة عملِ هذه المعادلات والخوارزميات معرفةً سرديَّةً، والقدرة على تفسير نتائجِها. من الأهمية بمكان أن تَتَناولَ الأبحاثُ اللُّغويّة الكيفيَّة في العلوم اللُّغويةِ والإنسانيَّةِ والاجتماعيَّة قدرًا من متانة التطبيقات الإحصائية والمعالجة الآلية، وأن يَجريَ التآلفُ بالممارسة المستمرة مع استخدامات التطبيقات والمعالجات الواردة في هذا الكتاب، إذْ باتت عَصَبَ الأبحاثَ والمشاريعِ في وقتنا الحاضر وفي الغد من المستقبل على المستوى العالميّ. أما المقترحاتُ التطبيقيَّةُ العَلميَّةُ المستقبليَّةُ فهي مفتوحة، مادامت مادةُ التَّحليلِ والمعالجةِ هي (المدوَّنة اللُّغوية) التي تُمثّل اللغة العربية الطبيعية المنتَجَة، الشفهي منها والمكتوب، والماضي منها والحاضر، والفصيح منها واللهجي، والمحليّ منها والدوليّ.

وَمِن أهمِّ التَّوصِياتِ حول ما عُرض من تطبيقاتٍ إحصائيَّةٍ في الباب الأول ما يلي:

أولًا: أهميّةُ تَطبيقاتِ الإحصاءات الوصفيّة قبل الإحصاءات الاستدلاليّة في البَيَاناتِ اللُّغوِيَّة، إذْ تكمنُ أهميَّتُها في التَّكشيفِ الأوّليّ للمعلومات اللُّغويةِ المهِمَّةِ حول التّكرارات والتَّوزيع والتَّنوّع والكثافةِ والتَّشتّت والقِيمِ المتطرِّفة.

ثانيًا: أهميّةُ الإفصاحِ عن أحجامِ التَّأثيرِ المتعلِّقةِ بالإحصاءاتِ الاستدلالية، فهي تُخبرُ عمَّا إنْ كانَت نتائجُ البَيَاناتِ اللُّغويةِ المستخدمة في التحليل ذاتَ تأثيرٍ أم لا.

ثالثًا: أهميّةُ الإفصاحِ عن مجالات الثِّقة؛ كونها تُفيدُ بمدى الثِّقة بنتائج تحليل عيِّنةِ البَيَانات اللغوية، وبمدى الثِّقة بتعميمِها على المجتمعِ اللُّغويّ.

رابعًا: توظيفُ مَقَاييس حِسابِ التَّصَاحبِ اللفظي المكوّنة من اثني عشر مُعاملًا بحسبِ الغرضِ من حالاتِ الاستخدام، والاعتمادُ على تِقْنيةِ مجموعة مَقاييسِ التَّصَاحُبِ (collocation parameters notation) التي تُساعد على دقة ضبطِ مُخرَجات نَتَائِج شبكات التَّصَاحبِ اللَّفظيّ من البَيَاناتِ اللُّغويَّة الكبيرةِ حجمًا.

خامسًا: التَّفكيرُ الإحصائيُّ الدَّقيقُ بنوعيّةِ المتغيرات اللُّغوية والمتغيرات التأويليّة، وما يتفرّع منهما من متغيراتٍ اسميّةٍ وترتيبيَّة ومقياسيّة، سواء كان هذا التكفيرُ منطلقًا من بياناتٍ لغويَّةٍ مُتوفرةٍ أو من أسئلةٍ بحثيةٍ تَتَطلَّبُ إعدادًا وجمعًا وتهيئةً للبَيَاناتِ اللُّغويَّةِ الخاصَّةِ بها. وسَوَاءٌ كان التحليلُ تحليلًا للارتباط، أو للتَّجميعِ، أو لمقارنةِ المجموعات، أو للتَّنَاظُرِ، أو للانحدارِ، أو للبَيَاناتِ اللُّغويَّة التَّاريخيَّةِ والمعجميَّةِ، فالمهمُّ أن يكون جَمْعُ البَيَاناتِ اللُّغويَّة وتهيئتُها وتصميمُ هَيكَلَتِها قائمًا على استيعاب المعلومات اللُّغويَّة في أَعمدةِ البَيَاناتِ وما يَتَمَثَّلُ فيها من أنواعِ المتغيّرات.

ثَانِيًا: عن تنقيب النُّصُوص والبَيَانَات العَرَبِيَّة في البَاب الثَّانِي

النَّتيجة

أَوَّل ما يُرام تحقيقُه -مَرَّةً أُخرى- هو أَن يتآلَفَ اللُّغَوِيُّ مع أساليب تعَلُّم الآلة في النُّصُوص اللُّغَوِيّة العَرَبِيَّة. ومن أهم ما ظهر من نتائج من خلال التَّحْلِيل على مُدَوّنَة القرآنِ؛ سهولةُ تعلّم الآلة للنَّصِّ القرآني في إنشاء المتَّجهاتِ (vectors) للكلماتِ، وقد لُوحظ ذلك على الكَلِمَات ذات التَّشابه الدّلاليّ؛ مثل كلمة (هدى)، وقد يُعزَى سببُ ذلك إلى الكثافة المعجَمِيّة العاليةِ، التي يحظى بها النَّصُّ القرآني. من النتائج -أيضًا- عدم تفاوتِ الموضوعات في القرآن الكريم، فهي مترابطةٌ ترابطًا تامًّا بالقصصِ، والعِبرِ،

والتَّشريعِ، والجزاءِ، والعقابِ، بالإضافة إلى غلبة استعمال التَّركيب الماضي على التَّركيب المضارع. من النَّتائج -أيضًا- فيما يتعلَّق بأنظمة التَّوسيم النَّحْوِيّ التَّفاوت بين الأنظمة، والحاجة إلى بناء معالجاتٍ أساسِيَّةٍ أكثر دِقَّةً؛ للتَّفريق، والتَّوسيم النَّحْوِيّ، والتَّأصِيلِ المعجميّ، للنّصوص العَرَبِيَّة القديمة، والنُّصُوص العَرَبِيَّة المعاصرة. ومن النتائج في سياق التَّصنيف والانْحِدَار الخَطّيّ: أنَّ نماذج التَّصنيف للسُّورِ المكِّيَّة والمدَنِيَّة من حيث الزمن، والسُّوَر المكِّيَّة والمدَنِيَّة من حيثُ الحجم (الطويلة والمتوسطة والقصيرة)، مع ذكرٍ لأسماء الأنبياء من عدم ذكرٍ لها؛ قد جاءت متقاربةً نوعًا ما في الدِّقَّةِ، بعد اختبارها على جزءٍ لم تتعلم منه من هذه البَيَانَاتِ، وأكثر النماذج دقَّةً كانت من آلة المتجهة الدَّاعمة، والانْحِدَار اللُّوجسْتِيّ، والتَّعزيز التكيّفي. أما نموذجُ الانْحِدَار الخَطِّيِّ؛ فبعد تطبيقِه على بياناتٍ مقياسِيَّةٍ لترتيبِ السُّوَر الزمنيِّ، وترتيبِها في المصحف، وذكرِ اسمِ النَّبِيِّ من عدم ذِكرِهِ، ووجود سجدةٍ من عدمِ وجودها؛ فقد اختُبر النموذجُ على بياناتِ الاختبار، وظهر معيارُ خطأ عالٍ، وعدم وجودِ علاقةٍ خطِّيَّةٍ. ولتحقيق العلاقة الخَطِّيَّة؛ يجبُ أن يُؤخذ في الحساب مسألةُ فرط التَّعميم (بيانات قليلة)، وفرط التَّخصيص (بيانات متباينةٌ، وكثيرةٌ، وبها قيمٌ متطرفة)، أو: يُطبّق الانْحِدَار اللّاخطّي الذي يُعطي شكل المنحنى الخَطّي الأَكْثر مُلاءمةً مع نُقاطِ البَيَانات. ومهما كانت النَّتيجةُ؛ لا بدَّ من الإفصاح عنها؛ لمعرفة طبيعة البَيَانَات اللُّغويةِ المستخدمة للمُعالجة، ومعرفة إمكانات البحثِ عن طرائقِ معالجة فرط التَّعميم أو التَّخصيصِ.

الوِجهة

اتَّجه إلى معرفةِ الأسس التي بها تتعَلَّمُ الآلةُ اللغةَ، وتفهمُها، وتُنتجُها، فهذه آخر محَطَّةٍ يتطلّب منك معرفتها اليوم. لا يحتاج اللُّغَوِيّ إلى معرفةِ بناء تطبيقات تعلّم الآلة؛ لأنَّ ذلك من العملِ المشترك بين المطوّرين من مهندسي البَرْمَجيّاتِ؛ ومهندسي البَيَانَاتِ، وعلماء البَيَانَاتِ، واللُّغَوِيّين الحاسوبيين، والممارسين في الـمُدَوّنَات اللُّغَوِيّة. يكفي أن يتعرّفَ اللُّغَوِيُّ إلى كُلّ الأنظمةِ والتَّطْبِيقَات الجديدةِ التي تظهر، فتخدم البحث اللُّغَوِيَّ، وترفَع من جودة المشاريعِ اللُّغَوِيّة.

التَّوصيات

الحياة المستقبلة قائمةٌ على الذَّكَاء الاصْطِنَاعِيّ، الذي يدخل في شؤون الحياةِ؛ ومنها اللغات، ومن هذِه اللغاتِ لغتُنا العَرَبِيَّة. من الضروريِّ أن يكتسبَ غير المتخَصِّصِ بالحاسب الآليِّ الأدواتِ التي تُعينه على إدراك أساليب تعلم الآلةِ، وفهمها، وإنتاجها للأشياءِ، كما نفهمها نحن وننتجُها. لقد أصبحَ التَّاريخِيّ يهتَمُّ بهذه الأساليبِ وتقنياتِها ليُدير المعلوماتِ الأرشيفية، وأصبح مدير الأعمال يهتمُّ بها في تطوير الأعمال، وستجد الاهتمامَ بها في كُلّ حقلٍ من الحقول العِلْمِيّة. ومن الضروريِّ أن يتعرَّف اللُّغَوِيُّ إلى معالجة اللغة الطَّبيعِيّة، وإلى الأدوات البحثِيَّةِ؛ لمعالجة ما يُراد معالجته عن الظواهر اللُّغَوِيّة.

وَمِن أهمِّ التَّوصِياتِ حول ما عُرض من تطبيقاتِ المعالجةِ الآليَّةِ في الباب الثَّاني ما يلي:

أولًا: لابدَّ من تَصميمِ البَيَانات اللُّغويةِ وفق الهيكل المحدّد في المبحث الأَوَّلِ من الفصل الأوَّل (إعداد البَيَانات اللُّغوية في جدولةِ بيانات) كي تَعمَلَ أدواتُ تنقيبَ النُّصُوصِ في أداة أُورَانج.

ثانيًا: عند التَّعَرُّفِ على وظيفة كل أداة في صندوق تنقيبِ النُّصُوصِ والبَيَاناتِ في أداة أورانج، تَتَحَدَّد مهمةُ تحديدِ مسارِ العمل التحليلي وربطِ كلِّ أداةٍ بأخرى بتحديدِ مسار الغرضِ من المعالجة. وما قدّمتُه من تطبيقات لمعالجة النص العربي آليًّا يُشَكِّلُ خمسةَ مساراتٍ أساسيَّة. الأول: مسارُ استخراجِ الكلماتِ المهمّة باستخدام قياس (تكرار الكلمة-تكرار المستندِ العَكسيِّ)، ومسارُ استخراجِ الكلماتِ المفتاحيَّةِ باستخدام نموذج الاستخراج الآلي السَّريع للكلمات المفتاحِيَّة RAKE المدمج في أداة أورانج. والثاني: مسارُ (تضمينِ الكلماتِ) الأكثردقَّة من (حَقيبة الكَلِمات) في نمذجة الكلماتِ ذاتِ العلاقةِ دلاليًّا. والثالث: مسارُ استخراجِ الموضوعاتِ التي تَتَمثَّلُ في الكلماتِ أو العبارات ذاتِ المحتوى من بين النُّصُوص والملفات النَّصيّة اللُّغوية. والرابع: التَّوسيم النحوي للكلمات الذي يُفيدُ بتحديد القسم الكلامي لكُلِّ كلمةٍ في النُّصُوص، واستكشافِ الأنماطِ النَّحويَّة للعبارات والجمل القصيرة. والخامس: خوارزميات (التَّصنيف) التي تكشف مدى دقة

تعلم الآلة التَّصنيفيّ لأصنافِ المتغيِّرات، وخوارزميات الانحدار التي تُقيّم مدى كفاءة الآلة في التَّنَبُؤ بقيمِ المتَغَيِّرات المقياسيَّة للقيمِ الحقيقيّة لها في البيانات.

ثالثًا: يعتمدُ تنقيب النُّصُوص والبيانات اللغوية على نماذجَ إحصائيةٍ مثل نموذج (تكرار الكلمة-تكرار المستند العكسيّ) أو على نماذج اللغة language models مثل (حقيبة الكلمات، وتضمين الكلمات، ونمذجة الموضوعات). لابدّ من إدراكِ مسألةٍ مهمّة تتعلق بهذه النماذج، وهي في كونها نماذج غيرنهائية، فنماذج اللغة المدمجة في أداةِ أورانج تُمثّل آخر النماذج المطوّرة، وتتطورُ هذه النماذجُ في الأداة نفسها في كل نسخة محدّثة بأداء أكثردقة للمُعالجة.

رابعًا: بُنيت مُكوِّناتُ أداةِ أورانج لتقنيبِ النُّصُوص والبَيَاناتِ بالاعتماد على تعلم الآلة، وهذا يعني أن الآلة ستعمل على مَهَامٍّ يصعبُ على الإنسانِ القيام بها، كمَهَمَّةِ توسيم الكلمات النَّحويَّة أو مَهمّةِ استخراجِ الموضوعات النَّصِّية في نُصُوصٍ لغويَّةٍ تَتَضمَّنُ عشرات الآلاف أو مئات الآلاف من الكلمات، غيرأن ذلك يتطلّبُ من المستخدم أن يكونَ مُتفاعلًا مع نتائج هذه الأنظمة الآليَّة، وذلك بتفسيرمُخرجات المعالجة الآلية، وتمحيصها، وتحديد الأخطاء، كَأَنْ تُخطئَ الآلةُ في توسيم كلمةٍ ما نحويًّا، وهي أخطاءٌ معدودةٌ، يتمُّ تقييمُها وإيجادُ حلولٍ لغويَّةٍ/آليَّةٍ لها، نظرًا للتَّطوير الدَّائم والتَّحسينِ المستمرِّ لأداءِ هذه الأنظمةِ في المعالجة الآليَّةِ للُّغة.

خامسًا: استثمارُ هذه الأداة في الأبحاث العِلميَّةِ المتعلِّقةِ بتحليل النُّصُوص ومعالجتها، كون هذه النُّصُوص جزءًا من البَيَاناتِ التي تُجمعُ لأبحاث العلوم الاجتماعية والإنسانية واللُّغوية، وبخاصةٍ في المنهج الكيفي (النّوعي) الذي سَادَت فيه مُعالجةُ الخَصائصِ النَّصيَّة آليًّا، إذْ أَحْدَثَ العصرُ الرقمي نموًّا متزايدًا وهائلًا في النصوص المنتَجَة. وفي الوقت ذاته، تَطَوَّرتْ أكثرُ الأساليب ذكاءً للتحليل الآلي للنصوص، وإن لم تُوظّف مثلُ هذه التِّقنياتِ لمعالجة النُّصُوص في الممارسات البحثيَّة في تلك العُلُومِ، فما نَستوعِبُه فيها سَيَتَضاءل تضاؤلًا كبيرًا مع مرور الزَّمن.

المصادر والمراجع

المراجع العَرَبِيَّة والأجنبية

المراجع العَرَبِيَّة

- بريزينا، فاكلاف. علم الإحصاء في لُغَوِيّات الـمُدَوّنَة الحَاسُوبِيَّة. ترجمة: سلطان المجيول. دار جامعة الملك سعود للنشر، الرياض، الطبعة الأولى، ٢٠٢١.
- الجاسم، حوراء. الألفاظ الأكاديمية في الـمُدَوّنَات العَرَبِيَّة العِلْمِيّة: دراسة حاسوبية تحليلية. (رسالة ماجستير) إشراف: د. سلطان المجيول. قسم اللغة العَرَبِيَّة (جامعة الملك سعود)، ٢٠٢٢.
- الجاسم، حوراء، والمجيول، سلطان. المدونة الأكاديمية العربية (أواك) Academic Written Arabic Corpus AWAC (قائمة الجاسم والمجيول للألفاظ الأكاديمية وقوائم الجاسم والمجيول للمصطلحات العربية العلمية، ٢٠٢٢). الرابط: https://www.arabic-ai.info/awac
- الراجح، عبدالله. الترجمة الآلية. في: خوارزميات الذَّكَاء الاصْطِنَاعِيّ في تحليل النَّصّ العَرَبِيّ. مركز الملك عبدالله الدولي لخدمة اللغة العَرَبِيَّة (مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللُّغَوِيّة حاليًّا)، الرياض، الطبعة الأولى، ٢٠١٩، الصفحات ٦٩-٩٣.
- روبي، أحمد محمد. البنك الشجري النَّحْوِيّ بناؤه وتوظيفه في إطار تقنيات الذَّكَاء الاصْطِنَاعِيّ، مركز الملك عبدالله الدولي لخدمة اللغة العَرَبِيَّة (مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللُّغَوِيّة حَالِيًّا)، الرياض، الطبعة الأولى، ٢٠١٧.
- السيوطي، جلال الدين عبدالرحمن بن أبي بكر. الإتقان في علوم القرآن. تحقيق: شعيب الأرنؤوط، مؤسسة الرسالة، دمشق، الجزء الأول، ٢٠٠٨.

- صالـح، محمـود إسـماعيل. (٢٠١٥). الـمُدَوّنَات اللُّغَوِيّـة وكيفيـة الإفـادة منهـا. في: الـمُدَوّنَات اللُّغَوِيّـة العَرَبِيَّـة بناؤها وطـرق الإفـادة منها. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٥.

- الصانـع، وليـد. طـرق ومسـتويات معالجـة اللغـة في الـذَّكَاء الاصْطِنَاعِـيّ. في: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل النَّـصّ العَـرَبِيّ. مركـز الملـك عبـدالله الدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٩، الصفحـات ١٥-٤٤.

- العصيـمي، عبدالرحمـن. نمذجـة الكلمـة العَرَبِيَّـة: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل الكلمـة العَرَبِيَّـة لُغَوِيًّـا وتوزيعيًّـا. في: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل النَّـصّ العَـرَبِيّ. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حَالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٩، الصفحـات ٩٥-١٢٤.

- الغامـدي، أيمـن. تقنيـات الـذَّكَاء الاصْطِنَاعِـيّ والمعالجـة الحَاسُـوبِيَّة للمتلازمـات اللَّفْظِيّـة والتراكيـب الاصطلاحيـة. في: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل النَّـصّ العَـرَبِيّ. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حَالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٩، الصفحـات ١٢٥-١٦٧.

- الغـزالي، أبـو حامـد. إحيـاء علـوم الديـن. صححـه: محمـد بـن مسـعود الأحمـدي. عالـم الكتـب للطباعـة والنشـر، بـيروت، الطبعـة الأولى، ٢٠٠٥.

- القنيعـير، فـارس. التعلـم العميـق وتطبيقاتـه في معالجـة اللغـة. في: خوارزميـات الـذَّكَاء الاصْطِنَاعِـيّ في تحليـل النَّـصّ العَـرَبِيّ. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حَالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٩، الصفحـات ٤٤-٦٨.

- لحلـح، محمـد. مدخـل إلى الـذَّكَاء الاصْطِنَاعِـيّ وتعلـم الآلـة. أكاديميـة حاسـوب، المملكـة المتحـدة، الطبعـة الأولى، ٢٠٢٠.
- المجيـول، سـلطان. البحـث اللُّغَـوِيّ في الـمُدَوّنَات العَرَبِيَّـة الحَاسُـوبِيَّة بـين الممكـن والمحتمـل والمأمـول. في: الـمُدَوّنَات اللُّغَوِيّـة العَرَبِيَّـة بناؤهـا وطـرق الإفـادة منهـا. مركـز الملـك عبـدالله الـدولي لخدمـة اللغـة العَرَبِيَّـة (مركـز الملـك عبـدالله بـن عبدالعزيـز للتخطيـط والسياسـات اللُّغَوِيّـة حَالِيًّـا)، الريـاض، الطبعـة الأولى، ٢٠١٥.
- المجيـول، سـلطان. الفهرسـة الدلاليـة الكامنـة وتخصيـص دركليـه الكامـن لنمذجـة موضوعـات الشـعر العَـرَبِيّ عـبر العصـور. مجلـة اللسـانيات العَرَبِيَّـة، العـدد ١٤، ٢٠٢٢، الصفحـات ٥٧-٨٠.
- مكانـري، تـوني، وأنـدرو هـاردي. لُغَوِيّـات الـمُدَوّنَة الحَاسُـوبِيَّة: المنهـج والنظريـة والتطبيـق. ترجمـة: سـلطان المجيـول. الريـاض: دار جامعـة الملـك سـعود للنشـر، ٢٠١٦.

المراجع الإنجليزية

- Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. NAACL.
- Almujaiwel, S. (2020). An Overview of the Lexical Variation Between Arabic Lexicons and Natural Arabic Language. International Journal of Arabic Linguistics, 6(1 & 2). 91-107.
- Almujaiwel, S. (2020). A Comparative Evaluation of POS Tagging and N-Gram Measures in Arabic Corpus Resources and Tools. International Journal of Computational Linguistics, 11(1). 1-17.
- Almujaiwel, S. (2019). Grammatical construction of function words between old and modern written Arabic: A corpus-based analysis. Corpus Linguistics and Linguistic Theory, 15(2). 267-296.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139-173.
- Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.
- Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.
- Darwish, K., & Mubarak, H. (2016). Farasa: A New Fast and Accurate Arabic Word Segmenter. LREC-2016.
- Darwish, K., Mubarak, H., Abdelali, A., & Eldesouki, M. (2017). Arabic pos tagging: Don’t abandon feature engineering just yet. In Proceedings of the Third Arabic Natural Language Processing Workshop, (pp. 130-137). Association for Computational Linguistics.
- Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. Corpus Linguistics and Linguistic Theory, 2(1), 23–60.

- Dukes, K. (2009-2017). The Quranic Arabic Corpus [An Open-Source Project]. Availabe at: https://corpus.quran.com.
- Duque, Tiago. (2020). Peeling Natural Language Processing Layer by Layer: An Introductory Overview. In Medium [web source]. Available at: https://tfduque.medium.com/dissecting-natural-language-processing-layer-by-layer-an-introductory-overview.
- Gries, S. Th., Newman, J., Shaoul, C. & Dilts, P. (2009). N-grams and the clustering of genres. Presented at workshop on Corpus, Colligation, Register Variation at the 31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft, March.
- Haugen, E. (1966). Dialect, language, nation. American Anthropologist,68, 922–935.
- Haugen, E. (1972). Language planning, theory and practice. In E. Haugen (Ed.), The ecology of language: Essays (pp. 287–295). Stanford: Stanford University Press.
- Haugen, E. (1987). Blessings of Babel. Bilingualism and language planning. Problems and pleasures. Berlin, New York: Mouton de Gruyter.
- Jarrar, M. (2021). The Arabic Ontology – An Arabic Wordnet with Ontologically Clean Content. Applied Ontology 16(1), 1-26.
- Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788-8790.
- Lancaster Stats Tools Online (Toolbox). (2018). Available at: http://corpora.lancs.ac.uk/stats/toolbox.php.
- Lavin, M. J. (2019). Analyzing Documents with TF-IDF [Online Lessons]. In: Z. LeBlanc (Ed.), The Programming Historian. https://doi.org/10.46430/phen0082.
- Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. Paper presented in the Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools, (pp. 102-109). Cairo, Egypt.
- Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, (pp. 55-60). Baltimore.

- Miner, G., Elder, J., Fast, A., Hill, Th., Nisbet, R., & Delen, D. (2012). Practical Text Mining and Statistical Analysis for non-Structed Text Data Applications. Elsevier, pp. 3-28.
- Mitkov, R. (ed.), (2022). The Oxford Handbook of Computational Linguistics. Oxford University Press, 2nd edition.
- Mubarak, H. (2018). Build Fast and Accurate Lemmatization for Arabic. Paper presented in the International Conference on Language Resources and Evaluation LREC-2018, Miyazaki.
- Parkinson, D., 2001. arabiCorpus. Available at: https://arabicorpus.byu.edu/.
- Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 1094–1101.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. Online.
- Qui, X., Sun, T., Xu, Y., Shao, Y., Daj, N., Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63, 1872-1897.
- Scott, M. (2020). WordSmith Tools version 8 [Lexical Analysis Software]. Available at: https://lexically.net/wordsmith.
- Sinclair, S., & Rockwell, G. (2022). Voyant Tools [v. 2.5.3]. Available at: https://voyant-tools.org.
- Sketch Engine: Create and Search a Text Corpus. (2004-2021). Available at: https://www.sketchengine.eu/.
- Stanford NLP Group. (2020). Stanford CoreNLP (Arabic) [Model Jar]. Available at https://stanfordnlp.github.io/CoreNLP/.
- Stanford NLP Group. (2020). Using CoreNLP with other programming languages and packages. Available at https://stanfordnlp.github.io/CoreNLP/other-languages.html

- Taji, D., Habash, N., & Zeman, D. (2017). Universal Dependency for Arabic. In Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP), pp. 166-176, Association for Computational Linguistics, Stroudsburg, PA, USA.
- Taji, D., Khalifa, S., Obeid, O., Eryani, F., Habash, N. (2018). An Arabic Morphological Analyzer and Generator with Copious Features. Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, October (pp. 140-150). Brussels (Belgium): Association of Computational Linguistics.
- Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purporses: Towards the Individualziation of Needs Satisfaction. In P. A. Fuertes-Olivera & H. Bergenholtz (Eds.), e-Lexicography: The Internet, Digital Initiatives and Lexicography, (pp. 54-79). London, New York: Continuum.
- Tsujii, J. (2021). Natural Language Processing and Computational Linguistics. Computational Linguistics, 47(4), 707-727.
- Veltri, G. A. (2020). Digital Social Research. Medford, MA: Polity Press.

الملاحـق

الملحق (أ): إحصاءات القرآن الكريم

اسم السُّورَة	ترتيب النُّزول	ترتيب المصحف	عدد الآيات	زمن النُّزول	السَّجْدَة	الكَلِمَات	الحروف
الفاتحة	5	1	7	مكية مدنية	0	29	139
البقرة	87	2	286	مدنية	0	6144	25613
آل عمران	89	3	200	مدنية	0	3503	14605
النساء	92	4	176	مدنية	0	3712	15937
المائدة	112	5	120	مدنية	0	2837	11892
الأنعام	55	6	156	مكية	0	3055	12418
الأعراف	39	7	206	مكية	1	3344	14071
الأنفال	88	8	75	مدنية	0	1243	5299
التوبة	113	9	129	مدنية	0	2506	10873
يونس	51	10	109	مكية	0	1841	7425
هود	52	11	123	مكية	0	1947	7633
يوسف	53	12	111	مكية	0	1795	7125
الرعد	96	13	43	مدنية	1	854	3450
إبراهيم	72	14	52	مكية	0	831	3461
الحجر	54	15	99	مكية	0	658	2797
النحل	70	16	128	مكية	1	1845	7642
الإسراء	50	17	111	مكية	1	1559	6480
الكهف	69	18	110	مكية	0	1583	6425
مريم	44	19	98	مكية	1	972	3835
طه	45	20	135	مكية	0	1354	5288
الأنبياء	73	21	112	مكية	0	1174	4925
الحج	103	22	78	مدنية	1	1279	5196
المؤمنون	74	23	118	مكية	0	1051	4354
النور	102	24	64	مدنية	0	1317	5596
الفرقان	42	25	77	مكية	1	896	3786

اسم السُّورَة	ترتيب النُّزول	ترتيب المصحف	عدد الآيات	زمن النُّزول	السَّجْدَة	الكَلِمَات	الحروف
الشعراء	47	26	227	مكية	0	1322	5517
النمل	48	27	93	مكية	1	1165	4679
القصص	49	28	88	مكية	0	1441	5791
العنكبوت	85	29	69	مكية	0	982	4200
الروم	84	30	60	مكية	0	818	3388
لقمان	57	31	34	مكية	0	550	2121
السَّجْدَة	75	32	30	مكية	1	374	1523
الأحزاب	90	33	73	مدنية	0	1303	5618
سبأ	58	34	54	مكية	0	884	3510
فاطر	43	35	45	مكية	0	780	3159
يس	41	36	83	مكية عدا آية	0	733	2988
الصافات	56	37	182	مكية	0	865	3790
ص	38	38	88	مكية	1	735	2991
الزمر	59	39	75	مكية	0	1177	4741
غافر	60	40	85	مكية	0	1228	4984
فصلت	61	41	54	مكية	1	796	3282
الشورى	62	42	53	مكية	0	860	3431
الزخرف	63	43	89	مكية	0	837	3508
الدخان	64	44	59	مكية	0	346	1439
الجاثية	65	45	37	مكية	0	488	2014
الأحقاف	66	46	35	مكية	0	646	2602
محمد	95	47	38	مدنية	0	542	2360
الفتح	111	48	29	مدنية	0	560	2456
الحجرات	106	49	18	مدنية	0	353	1493
ق	34	50	45	مكية	0	373	1473
الذاريات	67	51	60	مكية	0	360	1510
الطور	76	52	49	مكية	0	312	1293

اسم السُّورَة	ترتيب النُّزول	ترتيب المصحف	عدد الآيات	زمن النُّزول	السَّجْدَة	الكَلِمَات	الحروف
الط	23	53	62	مكية	1	359	1405
القمر	37	54	55	مكية	0	342	1438
الرحمن	97	55	78	مدنية	0	352	1585
الواقعة	46	56	96	مكية	0	379	1692
الحديد	94	57	29	مدنية	0	575	2475
المجادلة	105	58	22	مدنية	0	475	1991
الحشر	101	59	24	مدنية	0	447	1913
الممتحنة	91	60	13	مدنية	0	352	1519
الصف	109	61	14	مدنية	0	226	936
الجمعة	110	62	11	مدنية	0	177	749
المنافقون	104	63	11	مدنية	0	180	780
التغابن	108	64	18	مدنية	0	242	1066
الطلاق	99	65	12	مدنية	0	279	1170
التحريم	107	66	12	مدنية	0	254	1067
الملك	77	67	30	مكية	0	337	1316
القلم	2	68	52	مكية	0	301	1258
الحاقة	78	69	52	مكية	0	261	1107
المعارج	79	70	44	مكية	0	217	947
نوح	71	71	28	مكية	0	227	947
الجن	40	72	28	مكية	0	286	1089
المزمل	3	73	20	مكية	0	200	840
المدثر	4	74	56	مكية	0	256	1015
القيامة	31	75	40	مكية	0	164	664
الإنسان	98	76	31	مدنية	0	243	1065
المرسلات	33	77	50	مكية	0	181	815
النبأ	80	78	40	مكية	0	174	766
النازعات	81	79	46	مكية	0	179	762

اسم السُّورَة	ترتيب النُّزول	ترتيب المصحف	عدد الآيات	زمن النُّزول	السَّجْدَة	الكَلِمَات	الحروف
عبس	24	80	42	مكية	0	133	538
التكوير	7	81	29	مكية	0	104	425
الانفطار	82	82	19	مكية	0	81	326
المطففين	86	83	36	مكية	0	169	740
الانشقاق	83	84	25	مكية	1	108	436
البروج	27	85	22	مكية	0	109	459
الطارق	36	86	17	مكية	0	61	249
الأعلى	8	87	91	مكية	0	72	293
الغاشية	68	88	62	مكية	0	92	378
الفجر	10	89	30	مكية	0	139	573
البلد	35	90	20	مكية	0	82	335
الشمس	26	91	15	مكية	0	54	249
الليل	9	92	21	مختلف فيها	0	71	312
الضحى	11	93	11	مكية	0	40	164
الشرح	12	94	8	مكية	0	27	102
التين	28	95	8	مكية	0	34	156
العلق	1	96	19	مكية	1	72	281
القدر	25	97	5	مكية	0	30	112
البينة	100	98	8	مدنية	0	94	394
الزلزلة	93	99	8	مدنية	0	36	156
العاديات	14	100	11	مكية	0	40	164
القارعة	30	101	11	مكية	0	36	158
التكاثر	16	102	8	مكية	0	28	122
العصر	13	103	3	مكية	0	14	70
الهمزة	32	104	9	مكية	0	33	133
الفيل	19	105	5	مكية	0	23	96
قريش	29	106	4	مكية	0	17	73

اسم السُّورَة	ترتيب النُّزول	ترتيب المصحف	عدد الآيات	زمن النُّزول	السَّجْدَة	الكَلِمَات	الحروف
الماعون	17	107	7	مكية	0	25	112
الكوثر	15	108	3	مكية	0	10	42
الكافرون	18	109	6	مكية	0	27	95
النصر	114	110	3	مدنية	0	19	79
المسد	6	111	5	مكية	0	29	81
الإخلاص	22	112	4	مكية	0	15	47
الفلق	20	113	5	مكية	0	23	71
الناس	21	114	6	مكية	0	20	80

الملحق (ب): أقسام الكلام في القرآن الكريم

	الوسم	الوصف	التكرار	النسبة المئوية
1	NN	أسماء شائعة	22559	25.79
2	P	الأدوات (أدوات الربط وحروف الجر وأسماء الإشارة والضمائر والظروف وأدوات النصب والجزم)	20777	23.76
3	Noun	اسم علم مفرد واسم معرف وأسماء التوكيد	17614	20.14
4	VBD	فعل تام	8911	10.19
5	VBP	فعل مضارع	6619	7.57
6	JJ	الصفات	2953	3.38
7	WP	الأسماء الموصولة	2885	3.30
8	DTNNS	اسم معرف مثنى أو جمع	1370	1.57
9	NNS	اسم مثنى أو جمع	1275	1.46
10	DTJJ	الصفات المعرفة	609	0.70
11	DTNNP	اسم علم معرف مفرد	500	0.57
12	VBN	فعل مبني للمجهول	421	0.48
13	VBG	مصدر عامل / أداة فعل	375	0.43

	الوسم	الوصف	التكرار	النسبة المئوية
14	VB	فعل أمر	364	0.42
15	WRB	أدوات الاستفهام	123	0.14
16	JJR	صفات المقارنة	93	0.11
17	DTJJR	الصفات المعرفة للمقارنة	9	0.01

الملحق (ج): أكواد التَّحليل الإحصائي للبَيَانَات اللُّغَوِيّة

بيئة البرمجة (آر)

الفصل الأول الأشكال البيانية

- **المخَطّط الدّرجِيّ histogram**

```
#Histogram
hist(x, breaks="Sturges", col = "gray" , xlab = "linguistic variable", main=
"Histogram")
```

- **مخطَّط الصّندوق (أو رسم المربع) boxplot**

```
#Boxplot
boxplot(myData, ylab = "linguistic variable",
      xlab="(sub)corpora", outline = FALSE, ylim=c(0, max(myData, na.r
m=TRUE)*1.05)); I = 1;while(I <= ncol(myData))
        {
        for(v in myData[,i]){points(jitter(I,3/i),v, col = "blue", pch
=1, cex = 1)};
        points(I, mean(myData[,i],trim = 0, na.rm = TRUE), col = "red"
, pch="_", cex = 4) i= i+1;
```

- **أعمدة الخطأ error bars: مجال الثِّقة ٩٥٪**

```
#Error bars
error.bars(myData,stats=NULL, ylab = “linguistic variable”,xlab=”(sub)corpora”,
main=NULL,eyes=FALSE, ylim = NULL,
        xlim=NULL,alpha=.05,sd=FALSE, labels = NULL, pos = NULL, arrow.len =
         0.05,arrow.col=”red”, add = FALSE,bars=FALSE,within=FALSE, col=”red”)
```

- **مخطَّط التَّشَتُّت scatter plot**

```
#Scatter plot with regression line
plot(myData); fitline <- lm(myData[,2] ~ myData[,1]); abline(fitline,col=”red”)
```

- **مخطَّط خَطّيّ line chart**

```
#Line chart
matplot(myData, type = c(“myData”),pch=1, ylab=”linguistic variable(s)”,
      xaxt = “n”, col = 1:4); axis(1, at=x, labels=v)
```

- **الخريطة الجغرافِيَّة geomapping**

```
#Geomapping
library(maps); library(mapdata)#load libraries
map(‘worldHires’,xlim=c((min(myData[,1])-5),(max(myData[,1])+5)),
ylim=c((min(myData[,2])-5),(max(myData[,2])+5)));
i=1;while(i<=length(myData[,1])){points(myData[I,1],myData[I,2],
col=2,pch=19, cex=(4*(myData[I,3]/x)+0.5)); i<-i+1;};”)
```

- **مخطَّط الأعمدة المرصوفة stacked barchart**

```
#Stacked barchart
matplot(myData, type = c(“myData”),pch=1, ylab=”linguistic variable(s)”,
      xaxt = “n”, col = 1:4); axis(1, at=x, labels=v)
```

- **خطوط المؤَشِّر sparklines**

```
#Sparklines
par(mrow=c(4,1),mar=c(5,7,4,2),omi=c(0.2,2,0.2,2)); for(I in 1:4){x=round(mean(b[,i]),2)
;plot(b[,i],ann=FALSE,axes=FALSE,type="l",col="gray",lwd=2);
mtext(side=2,at=x,names(b[i]),las=2,col="black");
mtext(side=4,at=x,x,las=2,col="black");
points(which.min(b[,i]),min(b[,i]),pch=19,col="blue");
points(which.max(b[,i]),max(b[,i]),pch=19,col="red");}
```

- **مخطَّط الشُّموع candlestick plot**

```
#Candlestick plot
#Prepare data
min<-apply(s, 2, min)
min<-as.vector(min)
max<-apply(s, 2, max)
max<-as.vector(max)
first<-head(s,1)
first<-as.numeric(first[1,])
last<-tail(s,1)
last<-as.numeric(last[1,])
int_1<-ifelse(first<=last,first,ast)
int_2<-ifelse(last>=first,last,first)
dir<-ifelse(first<=last,1,0)
dir<-as.vector(dir)
item<-colnames(s)
item<-as.vector(item)
order<-seq_along(item)
data <- data.frame(order, item, min, int_1, int_2, max, dir)
#Creat plot
with(data,symbols(order, (int_1+ int_2)/2, boxplots=cbind(.25,int_2-
```

```
int_1, int_1-min,max-int_2,0),inches=F,ylim=range(data[,-
c(1:2)]),xaxt="n",ylab="Relative frequency", xlab="Linguistic
variables", bg = ifelse(dir==0, "black", "white"))); data<-
data[with(data, order(order)),];
axis(1,seq_along(data$item),labels=FALSE);
text(x=seq_along(data$item), y=par()$usr[3]-0.05*(par()$usr[4]-
par()$usr[3]), labels=data$item, srt=15, adj=1, xpd=TRUE;
```

الفصل الثاني

- **حساب موافقة المقَدّر البيني**

```
gwet.ac1.raw(myData)
myData1<-table(myData); kappa2.table(myData1)
#Nominal 3 + raters
fleiss.kappa.raw(myData)
#Ordinal
gwet.ac1.raw(myData, weights="ordinal")
#Scale
ICC(myData)
```

الفصل الثالث

- **الارتباط**

```
#Libraries correlation
library(Hmisc)
library(corrplot)
library(stats)
#Pearson's correlation:
cor.test(mydata1, mydata2, method="pearson")
#Spearman's correlation
```

```
cor.test(mydata1, mydata2, method="spearman")
#Correlation matrix
rcorr(mydata, type="pearson")
#Scatter plot
plot(mydata, col ="blue"); fitline <lm(mydata1 ~ mydata2); abline(fitline,col="red")
#Correlation matrix
corrplot(m, method ="color", type = "full", diag = TRUE, addCoef.col="black", add
CoefasPercent=FALSE, addgrid.col="grey", tl.pos = NULL, tl.cex = 1, tl.srt = 45, tl.
col = "black")
```

- **التَّجْمِيع**

```
#Optional z-score transformation
mydata <- scale(mydata)
#Distance matrix
d <- dist(mydata, method = "manhattan")
#Cluster analysis
fit <- hclust(d, method="ward.D")
#Plot dendrogram
plot(fit, xlab="", ylab="Height", main="")
#Draw cluster groups
rect.hclust(fit, k=5, border="red")
```

- **مقارنة المجموعات**

```
#t-test
t.test(data[ ,1], data[ ,2], paired=FALSE)
#t-test: repeated measures
t.test(data[ ,1], data[ ,2], paired=TRUE)
#Mann-Whitney-wilcoxon rank sum test
wilcox.test(data[ ,1], data[ ,2], paired=FALSE)
#Mann-Whitney-Wilcoxon rank sum test: repeated measures
```

```
wilcox.test(data[ ,1], data[ ,2], paired=TRUE)
#One-way ANOVA
aov(measurement ~ group, data = data)
#Kruskal-Wallis test
kruskal.test(data)
```

- **تحليل التَّنَاظُر**

```
library(languageR);
x = corres.fnc(data);
plot(x, ccex = 0.6, rcex = 0.6)
```

الفصل الرابع

- **الانْحِدَار الخَطِّيّ**

```
#Libraries and data
install.packages('tidyverse')
library(tidyverse)
mydata <- read.csv("C:\\Users\\snmuj\\Downloads\\ mydata.csv")
data <- mydata %>% select(characters, verses, words, chronology)
install.packages('Ggally')
library(Ggally)
ggpairs(data)
model <-  lm(data=ar, characters ~ verses + words + chronology)
summary(model)
model$coefficients
model$residuals[0:10]
model$fitted.values[0:10]
ggplot(data = ar, aes(x = verses, y = characters)) + geom_point()
ggplot(data = ar, aes(x = words, y = characters)) + geom_point()
```

```
ggplot(data = ar, aes(x = chronology, y = characters)) + geom_point()
install.packages("ggfortify")
library(ggfortify)
autoplot(model, which = 1:2)
install.packages("stargazer")
library(stargazer)
stargazer(model, type = "html", out = "my model.doc")
stargazer(model, type = "html", out = "my model.doc")
```

الفصل الخامس

- **القمم والقعور**

```
library(ggplot2)
library(mgcv)
p<-ggplot(data, aes(x = data[,1], y =data[,2])) + geom_point() + xladata("Time") + ylada
ta("Linguistic variadatale"); p + stat_smooth(method = "gam", formula = y ~ s(x, datas
= "cr", fx=FALSE, k =15), size = 1, fill="#707070", level = 0.95 )+ stat_smooth(method
= "gam", formula = y ~ s(x, datas = "cr", fx=FALSE, k =15), size = 1, fill="#FFFF00", lev
el = 0.99)
```

- **التَّأرجح الاستعماليّ**

```
#Calculate Gwet's AC1; b…input data frame
I = 1; v <- c(); while(i+1 < ncol(b)) {n=(gwet.ac1.raw(b[,I:(i+2)])[3]);v<- c(v, n); i= i+1; }
#Prepare data frame
h<-seq(from, to, by = 1); g<-data.frame(h,v) #Produce graph
p<-ggplot(g, aes(x = g[,1], y =g[,2])) + xlim(from, to)+ scale_x_continuous(breaks = seq(
from, to, by = 10)) + geom_point() + xlab("Time") + ylab("AC1"); p + stat_smooth(metho
d = "gam", formula = y ~ s(x, bs = "cr", fx=FALSE, k =10), size = 1, fill="#707070", level
= 0.95 )+ stat_smooth(method = "gam", formula = y ~ s(x, bs = "cr", fx=FALSE, k =10),
size = 1, fill="#FFFF00", level = 0.99)
```

الملحق (د): أكواد حساب تكرار الكلمة-تكرار المستند العكسي

بيئة (بايثون)

```
from pathlib import Path

all_txt_files =[]
for file in Path("…FolderPath").rglob("*.txt"):
    all_txt_files.append(file.parent / file.name)
# Counts the length of the list
n_files = len(all_txt_files)
print(n_files)

all_txt_files.sort()
all_txt_files[0]

all_docs = []
for txt_file in all_txt_files:
   with open(txt_file) as f:
      txt_file_as_string = f.read()
   all_docs.append(txt_file_as_string)

vectorizer = TfidfVectorizer(max_df=.65, min_df=1, stop_words=None, use_
idf=True, norm=None)
transformed_documents = vectorizer.fit_transform(all_docs)

transformed_documents_as_array = transformed_documents.toarray()

# represent the numpy array
len(transformed_documents_as_array)
```

```
# Construct a list of output file paths
output_filenames = [str(txt_file).replace(".txt", ".csv").replace("txt/", "tf_idf_
output/") for txt_file in all_txt_files]

# Loop each item in transformed_documents_as_array
for counter, doc in enumerate(transformed_documents_as_array):
    # construct a dataframe
    tf_idf_tuples = list(zip(vectorizer.get_feature_names(), doc))
    one_doc_as_df = pd.DataFrame.from_records(tf_idf_tuples, columns=['term', 'score']
).sort_values(by='score', ascending=False).reset_index(drop=True)
    # Output to a csv
    one_doc_as_df.to_csv(output_filenames[counter])
```

الملحق (هـ): أول خمسين كلمة من حساب تكرار الكلمة-تكرار المستند العكسي

مكية			مدنية		
التسلسل	الكلمة	النتيجة	التسلسل	الكلمة	النتيجة
0	قومه	60.43499965	0	تكذبان	43.56941835
1	خلقنا	30.92023238	1	اعلموا	29.51476727
2	يوسف	30.92023238	2	حدود	16.8655813
3	المجرمين	26.70383705	3	اليتامى	16.8655813
4	نجزي	25.29837195	4	المسيح	15.46011619
5	آلهة	23.89290684	5	الأعراب	14.05465108
6	سلطان	23.89290684	6	يجدون	14.05465108
7	أهلكنا	23.89290684	7	بيوت	14.05465108
8	الملأ	23.89290684	8	إثما	14.05465108

مكية			مدنية		
التسلسل	الكلمة	النتيجة	التسلسل	الكلمة	النتيجة
9	كذبا	22.48744173	9	نساء	14.05465108
10	عباد	21.08197662	10	تتقوا	14.05465108
11	الآخرين	19.67651151	11	قاتلوا	12.64918597
12	يستهزئون	19.67651151	12	لبئس	12.64918597
13	جاءت	18.27104641	13	المساكين	12.64918597
14	أرأيتم	18.27104641	14	النصارى	12.64918597
15	أليس	18.27104641	15	أحل	12.64918597
16	ذكرى	18.27104641	16	تفعلوا	12.64918597
17	أدراك	18.27104641	17	الحج	11.24372086
18	نادى	18.27104641	18	تنفقوا	11.24372086
19	توعدون	18.27104641	19	أطعنا	11.24372086
20	بشرا	16.8655813	20	بأموالهم	11.24372086
21	اليمين	16.8655813	21	رضوان	11.24372086
22	نجعل	16.8655813	22	اليهود	11.24372086
23	الغفور	16.8655813	23	الفتنة	11.24372086
24	للمكذبين	16.8655813	24	القتال	11.24372086
25	أمين	15.46011619	25	سورة	9.838255757
26	تكذبون	15.46011619	26	يحلفون	9.838255757
27	عين	15.46011619	27	أذى	9.838255757
28	فاصبر	15.46011619	28	اشتروا	9.838255757
29	أسألكم	15.46011619	29	لعنهم	9.838255757

مكية			مدنية		
التسلسل	الكلمة	النتيجة	التسلسل	الكلمة	النتيجة
30	أوحي	15.46011619	30	يأمرون	9.838255757
31	السمع	15.46011619	31	المحصنات	9.838255757
32	تنزيل	14.05465108	32	خفتم	9.838255757
33	آياتي	14.05465108	33	المنافقون	9.838255757
34	قومك	14.05465108	34	عزيزا	9.838255757
35	يوعدون	14.05465108	35	ماتوا	9.838255757
36	آباؤنا	14.05465108	36	الخبيث	9.838255757
37	أجري	14.05465108	37	الربا	9.838255757
38	تعالى	14.05465108	38	فاسقون	9.838255757
39	الصور	14.05465108	39	التواب	8.432790649
40	الوعد	14.05465108	40	يؤتيه	8.432790649
41	لوطا	14.05465108	41	يأمركم	8.432790649
42	بأسنا	14.05465108	42	إيمانكم	8.432790649
43	فعل	14.05465108	43	الصدقات	8.432790649
44	يغني	14.05465108	44	الهدي	8.432790649
45	لقاء	14.05465108	45	وصية	8.432790649
46	القرون	14.05465108	46	يعذبهم	8.432790649
47	عذابي	12.64918597	47	يكتمون	8.432790649
48	تبصرون	12.64918597	48	مؤمنة	8.432790649
49	عظاما	12.64918597	49	يطع	8.432790649

الملحق (و): تـدريـب واختـبـار التَّوسيـم النَّحْوِيّ العَرَبِيّ للاعتماديّات الكُلِّيَّة

الواسمات النَّحْوِيّة: https://universaldependencies.org/u/pos.

بيئة البرمجةِ (بايثون): كُلّ مربعٍ يمثِّل مدخلًا مستقِلًّا.

```
import nltk
nltk.download('all')
from nltk.corpus import treeban
```

```
penn_treebank = []
for fileid in treebank.fileids():
  tokens = []
  tags = []
  for word, tag in treebank.tagged_words(fileid):
    tokens.append(word)
    tags.append(tag)
  penn_treebank.append((tokens, tags)
```

```
#Install conllu package
!pip install conllu
!wget https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3105/ud-treebanks-v2.5.tgz
!tar zxf ud-treebanks-v2.5.tgz

#Imports needed to open and parse the conllu file
from io import open
from conllu import parse_incr
```

```
#Open the file
data_file = open("/content/ud-treebanks-v2.5/UD_Arabic-PADT/ar_padt-ud-train.
conllu", "r", encoding="utf-8")
ud_files = []
for tokenlist in parse_incr(data_file):
    ud_files.append(tokenlist)

#List of tuples with a list of words and a list of their PoS tags
ud_treebank = []
for sentence in ud_files:
  tokens = []
  tags = []
  for token in sentence:
    tokens.append(token['form'])
    tags.append(token['upostag'])
  ud_treebank.append((tokens, tags)
```

```
import re
def extract_features(sentence, index):
  return {
      'word':sentence[index],
      'is_first':index==0,
      'is_last':index ==len(sentence)-1,
      'is_capitalized':sentence[index][0].upper() == sentence[index][0],
      'is_all_caps': sentence[index].upper() == sentence[index],
      'is_all_lower': sentence[index].lower() == sentence[index],
      'is_alphanumeric': int(bool((re.match('^(?=.*[0-9]$)(?=.*[a-zA-
Z])',sentence[index])))),
      'prefix-1':sentence[index][0],
      'prefix-2':sentence[index][:2],
```

```
    'prefix-3':sentence[index][:3],
    'prefix-3':sentence[index][:4],
    'suffix-1':sentence[index][-1],
    'suffix-2':sentence[index][-2:],
    'suffix-3':sentence[index][-3:],
    'suffix-3':sentence[index][-4:],
    'prev_word':'' if index == 0 else sentence[index-1],
    'next_word':'' if index < len(sentence) else sentence[index+1],
    'has_hyphen': '-' in sentence[index],
    'is_numeric': sentence[index].isdigit(),
    'capitals_inside': sentence[index][1:].lower() != sentence[index][1:]
  }
#returning the data as two lists
def transform_to_dataset(tagged_sentences):
  X, y = [], []
  for sentence, tags in tagged_sentences:
    sent_word_features, sent_tags = [],[]
    for index in range(len(sentence)):
        sent_word_features.append(extract_features(sentence, index)),
        sent_tags.append(tags[index])
    X.append(sent_word_features)
    y.append(sent_tags)
  return X,
```

```
#for UD Treebank
ud_train_size = int(0.8*len(ud_treebank))
ud_training = ud_treebank[:ud_train_size]
ud_testing = ud_treebank[ud_train_size:]
X_ud_train, y_ud_train = transform_to_dataset(ud_training)
X_ud_test, y_ud_test = transform_to_dataset(ud_testing
```

```
#Ignoring some warnings
import warnings
warnings.filterwarnings('ignore')
```

```
#Install sklearn_crfsuite, as it is not preloaded into Colab
!pip install sklearn_crfsuite
from sklearn_crfsuite import CR
#CRF for UD
ud_crf = CRF(
    algorithm='lbfgs',
    c1=0.01,
    c2=0.1,
    max_iterations=100,
    all_possible_transitions=True
)
print("Started training on UD corpus!")
ud_crf.fit(X_ud_train, y_ud_train)
print("Finished training on UD corpus!"
```

```
#Output for UD
print("## UD ##")
from sklearn_crfsuite import metrics

y_ud_pred=ud_crf.predict(X_ud_test)
print("F1 score on Test Data ")
print(metrics.flat_f1_score(y_ud_test, y_ud_pred,average='weighted',labels=ud_crf.
classes_))

y_ud_pred_train=ud_crf.predict(X_ud_train)
print("F1 score on Training Data ")
```

```
print(metrics.flat_f1_score(y_ud_train, y_ud_pred_train,average='weighted',labels=
ud_crf.classes_))

## UD ##
F1 score on Test Data
0.9370593011076455
F1 score on Training Data
0.9894871290659856
```

```
from sklearn.preprocessing import MultiLabelBinarizer

mlb1 = MultiLabelBinarizer()
a = mlb1.fit_transform([ud_crf.classes_])
y_test_matrix = mlb1.transform(y_ud_test)
y_pred_matrix = mlb1.transform(y_ud_pred
```

```
### Look at class wise score
print("Class wise score:")
from sklearn import metrics
print(metrics.classification_report(y_test_matrix, y_pred_matrix, target_names=ud_
crf.classes_, digits=3))

Class wise score:
              precision    recall  f1-score   support
           X      0.986     0.981     0.983       996
        VERB      0.999     0.997     0.998      1032
        NOUN      0.962     0.948     0.955       134
         ADJ      0.938     0.920     0.928       261
         ADP      1.000     0.995     0.998       884
       PUNCT      0.962     0.993     0.977       436
```

NUM	0.000	0.000	0.000	0
	1.000	0.999	1.000	1202
PRON	0.990	0.987	0.989	479
CCONJ	0.889	0.909	0.899	132
DET	0.991	0.997	0.994	632
PART	1.000	0.444	0.615	18
ADV	1.000	1.000	1.000	992
SYM	0.000	0.000	0.000	4
AUX	0.986	0.994	0.990	985
PROPN	0.928	0.817	0.869	931
INTJ	0.990	1.000	0.995	980
micro avg	0.984	0.973	0.979	10098
macro avg	0.860	0.822	0.835	10098
weighted avg	0.983	0.973	0.978	10098
samples avg	0.982	0.972	0.976	10098

```
sent = "ضع جملةً من عندك"
features = [extract_features(sent.split(), idx) for idx in range(len(sent.split()))]
#Make a prediction on a single input (sentence).
ud_results = ud_crf.predict_single(features)
ud_tups = [(sent.split()[idx], ud_results[idx]) for idx in range(len(sent.split()))]
#The results
print(ud_tups)
```

نبذة عن المؤلِّف

سلطان بن ناصر المجيول، أستاذ اللسانيات الحاسوبية بقسم اللغة العربية وآدابها في كلية العلوم الإنسانية والاجتماعية بجامعة الملك سعود. حصل على درجة الدكتوراه من جامعة إكسيتر (بريطانيا) في لسانيات المدونة الحاسوبية في عام ٢٠١٣م، وعلى دورات احترافية في مجالات الإحصاء اللغوي وتقنيات الذكاء الاصطناعي. نشر ٣٠ بحثًا علميًّا، من بينها أربعة كتب، وأبحاث في مجلات Web of Science. فاز بالمركز الثاني في جائزة الترجمة ضمن الجوائز الثقافية الوطنية (وزارة الثقافة السعودية) في دورتها الأولى ٢٠٢٠-٢٠٢١م. مدرب معتمد، ومحكم لعدد من المشاريع العلمية والتطبيقية في مجالات اللسانيات العربية الحاسوبية، وتطبيقات الذكاء الاصطناعي اللغوي، والإحصاء اللغوي، والسياسة اللغوية والتخطيط.

مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language

نبذة عن الكتاب

يهدف هذا الكتاب إلى ترقية المشاريع البحثية في العلوم اللغوية والإنسانية والاجتماعية، وإلى التشجيع على ممارسة تطبيقات الإحصاء اللغوي وتقنيات تنقيب البيانات اللغوية. وقُدّمت هذه التطبيقات والتقنيات بأسلوب بسيط ومباشر، وأُجريت على مدونة القرآن الكريم، بغيةَ تسهيل استيعاب إجراءات التحليل والتمرّس بها. ينقسم الكتاب إلى بابَيْن. الأول: الإحصاء اللغوي، والثاني: تنقيب النصوص والبيانات العربية. وفي كل باب فصول ومباحث تُفيد القارئ بالخطوات التحليلية بأدوات قائمة بذاتها، وقد ضُمّنت حزم هذه الأدوات ومكتباتها في الملاحق المرتبطة بمباحثها من أجل الرجوع إليها عند الحاجة.

9 786038 413920