مجلة جامعة أم القرى
لعلوم اللغات وآدابها
الموقع الإلكتروني: https://uqu.edu.sa/jll

تقييم النمذجة الموضوعية لنصوص الصحف السعودية باستخدام خوارزمية تخصيص دركليه الكامن LDA: دراسة لغوية حاسوبية

Evaluating Topic Modeling for Saudi Newspapers Texts Using LDA: A Computational Linguistics Study

د. أفراح عبد العزيز التميمي[a*]

[a]أستاذ مساعد، قسم الإعداد اللغوي، معهد تعليم العربية، جامعة الإمام محمد بن سعود الإسلامية، المملكة العربية السعودية

[a]Dr. Afrah Abdulaziz Altamimi[*]

[a]Language Preparation, Arabic Language Teaching Institute, Imam Muhammad bin Saud Islamic University, Saudi

الملخص :تقع هذه الدراسة في مجال معالجة اللغات الطبيعية وتطبق منهج تعلم الآلة غير الموجه في تحديد الموضوعات الكامنة في نصوص الصحف العربية السعودية باستعمال أحد أهم خوارزميات النمذجة الموضوعية غير الموجهة، وهي خوارزمية تخصيص دركليه الكامن للموضوعات. وقد جمعت نصوص الصحف السعودية في مدونة بلغ مجموع نصوصها بعد تهيئتها 4781 نصًّا، تضمنت 649,734 كلمة فعلية. وأظهرت نتائج تدريب 20 نموذجًا عليها بعشر كلمات مميزة أن القيمة المثلى لعدد الموضوعات في تلك النصوص، هي 7 موضوعات، وذلك بدرجة تماسك جيدة بلغت 0.6723. وقد استدل على هذه الموضوعات من خلال كلماتها العشر ذات القيم العليا في كل موضوع. ففسرت الموضوعات على التوالي: الرقابة والتوعية، والتنمية والتطوير، والرياضة، والصحة، والاقتصاد، وشؤون محلية، وسياسة دولية. ثم قيم النموذج ذي الـ7 موضوعات تقييمًا نوعيًّا بفحص تماسك الكلمات في الموضوع الواحد يدويًّا، وفحص الموضوعات بمراجعة النصوص الخمسين الأولى في كل موضوع؛ للتأكد من انتمائها لموضوعها الذي خصصته الخوارزمية لها. وقد ساعد في التقييم النوعي إجراء الخوارزمية مرة أخرى على نصوص كل موضوع من الموضوعات السبعة؛ للوصول إلى تفاصيل أكثر حول كل موضوع على حدة. وعلى الرغم من وجود بعض القصور في نتائج عملية النمذجة الموضوعية لبيانات الدراسة بتلك الخوارزمية، إلا أنه يمكن استكمال أوجه القصور ومعالجتها، واستعمالها في تحليل الخطاب بدلاً من المناهج التقليدية.

الكلمات المفتاحية: نموذج موضوعي، مقياس التماسك، خوارزمية LDA، النموذج الأمثل، معالجة اللغة، تعلم الآلة

Abstract

This paper is in the field of natural language processing. It applied unsupervised machine learning approach to identifying the latent topics in Saudi newspapers using one of the most important unsupervised topic modeling algorithms. This algorithm is called Latent Dirichlet Allocation (LDA). I built a corpus from Saudi newspapers, and it contained 4,781 texts after the preprocessing stage. It consisted of 649,734 tokens. The results of training 20 models with ten words showed that the optimal value for the number of topics in those texts is 7 topics. The 7-topics model got a good coherence degree of 0.6723. These topics were inferred through its ten words that had the highest probabilities on each topic. I interpreted the topics, respectively, according to the following topics: surveillance and awareness, development and improvement, sports, health, economics, domestic affairs, and international politics. The 7-topic model was evaluated qualitatively by manually reviewing the coherence of words in each topic. Also, I reviewed the first fifty texts on each topic to make sure that each of which belongs to the topic that LDA was assigned to it. The qualitative evaluation was supported by the algorithm being conducted again on the texts of each of the seven topics to access more details on each topic separately. Although there are some shortcomings in the results of the topic modeling, they can be optimized and then studied in discourse analysis instead of the traditional approaches.

Keywords: topic model, CV coherence, LDA algorithm, optimal model, machine learning

*بيانات التواصل: أفراح التميمي، جامعة الامام محمد بن سعود الإسلامية
aahaltamimi@imamu.edu.sa

1. المقدمة

يحتاج الباحث إلى أدوات تقرأ البيانات اللغوية المحوسبة، وتستكشفها وتفهم الموضوعات التي تدور حولها، كما يفهمها البشر. وبناء هذا النوع من الأدوات بنماذج تعلم الآلة الموجه supervised يتطلب معرفة مفصلة تتعلق بموضوع كل مقال لتوسيمها، وهذا يستهلك الوقت والجهد. ولذلك يستفاد من النماذج الموضوعية topic models الاحتمالية (الخوارزميات الإحصائية) غير الموجهة unsupervised التي تتعلم من النصوص الخام بالملاحظة وتحليل الكلمات في سياقاتها النصية؛ للكشف عن البنية الموضوعية للبيانات اللغوية مجموعةً، وللنصوص فيها فرادى (Blei et. al., 2003). فهي لا تتطلب أي توصيف أو ترميز أو توسيم يدوي للنصوص قبل تحليلها.

وتشير النمذجة الموضوعية إلى عملية تحديد الموضوعات التي تصف مجموعة من النصوص آليًّا، ولا تظهر هذه الموضوعات إلا عند عملية النمذجة، ولذا يصطلح عليها بأنها كامنة أو خفية latent. وهذا يعني أننا يمكن أن نصف النصوص من خلال توزيع الموضوعات، ويمكن أن نصف الموضوعات من خلال توزيع الكلمات.

إن أبرز مكونين لغويين مؤثرين في النمذجة الموضوعية للنص اللغوي هما: الكلمات الشائعة، والعلاقات بين تلك الكلمات. وإذا كان الخطاب يحدد الموضوعات الصالحة valid objects، فالكلمات التي تشير إلى هذه الموضوعات سترد متكررة في المادة اللغوية التي يؤثر عليها الخطاب (Foucault, 1971). وإذا كان الخطاب يحدد الطريقة المثلى لربط الموضوعات، فستضم البيانات اللغوية العديد من الروابط بين الكلمات التي تعكس هذه العلاقات. وإذا تمكنّا من الوصول للمعلومات (الكلمات الشائعة، والعلاقات بين تلك الكلمات) بدقة، صار من الممكن استنتاج معلومات خطابية من البيانات النصية (Van Dijk, 1993). والموضوع الذي يستكشف في النمذجة الموضوعية يضم هذين العنصرين من المعلومات. وهذا يعني إمكانية الإفادة من النمذجة الموضوعية في تحليل الخطاب.

ولفهم وظيفة النمذجة الموضوعية، ستجري هذه الدراسة وتقيم أكثر الخوارزميات استعمالاً في النمذجة الموضوعية (Gerlach et al. 2018)، وهي خوارزمية تخصيص دركليه الكامن Latent Dirichlet Allocation (LDA) للموضوعات على نصوص الصحف السعودية.

2. خوارزمية تخصيص دركليه الكامن
Latent Dirichlet Allocation (LDA)

تعد خوارزمية LDA نموذجًا من نماذج تعلم الآلة غير الموجه الذي يُزوَّد بمجموعات من النصوص بوصفها مدخلات، ويُوجِد الموضوعات الكامنة فيها بوصفها مخرجات. أي أن الخوارزمية تتعلم من النصوص الخام، وتفترض أن كل نص في المدونة يتضمن عددا من الموضوعات الدلالية المتماسكة التي تتناثر في كامل نصوص المدونة. وأن هذه البنية الموضوعية الدلالية مخفية، ولا يمكننا إلا النظر في النصوص والكلمات وليس في الموضوعات نفسها. وحيث إن البنية الموضوعية كامنة، تسعى الخوارزمية إلى استخلاص هذه البنية الموضوعية في ضوء كلمات ونصوص المدونة. إذ على أساس تصاحب الكلمات داخل النص تعين الموضوعات، فتبحث عن أنماط تصاحبها في النص الواحد. وعندما يغلب على مجموعة من الكلمات التصاحب في كامل النص، وعلى مدى غير محدود من الكلمات، وبانتظام، تفترض الخوارزمية وجود علاقة بين الكلمات[i]. وما إن تُحدّد الكلماتُ المنتظمةُ تصاحبًا، تؤخذ لتُمثِّل موضوعًا متماسكًا مستقلا. فالمخرج إذن سيكون قائمة من عدد معين من مجموعات الكلمات التي ترد متصاحبة في مجموعة من النصوص في المدونة الهدف. وتمثل كل مجموعة كلمات موضوعًا منفصلاً تكون الكلمات الواردة فيه هي الكلمات الأكثر تميزًا في هذا الموضوع. ويمتد استخلاص الموضوعات على كامل نصوص المدونة، فيرتبط النص الواحد بموضوع واحد أو بأكثر من موضوع. وهكذا ينكشف من كل نص مجموعة من الموضوعات المرتبطة به، ودرجة الارتباط به. ولذلك، يصح القول بأن الكلمات التي تندرج في كل موضوع تمثل الكلمات المميزة keywords له، وأن النصوص التي تضم موضوعًا معينًا ذا درجة أو قيمة عالية هي نصوص مميزة keytexts (Murakami et. al., 2017: 245). ولكن هذا لا يعني أن الخوارزمية هي التي ستسمي الموضوعات، فالمحلل البشري بناء على كل مجموعة كلمات مستخرجة بوصفها موضوعًا، هو من سيستنتج بنفسه الموضوعات، وطبيعة الاستنتاج أو الاستكشاف منوطة به. وقبل

أن تبدأ الخوارزمية العمل على النصوص بتحديد الموضوعات وكلماتها المميزة، لا بد من أن نحدد هايبربراميترين hyperparameters في الخوارزمية[ii]:

1- عدد الموضوعات المراد استخلاصها.

2- عدد الكلمات في كل موضوع.

وكلما زاد عدد الموضوعات، زاد تفصيلها، وصعب تفسيرها، وقد يكون من الأفضل دمجها، أما إذا قلت فتزيد عموميتها (Murakami, et. al., 2017: 245). وحيث إن خوارزمية LDA خوارزمية غير قطعية non-deterministic، فهي حتى مع تحديد عدد الموضوعات وعدد الكلمات بقيمة ثابتة، تُخضِع الموضوعات لدرجة من العشوائية تجعلها تختلف في كل عملية تشغيل للخوارزمية على النصوص نفسها (Floyd, 1967).

ولنفترض مثلاً أن لدينا النصوص الخمسة التالية، وسنستعمل تلك الخوارزمية لاستخلاص الموضوعات التي تتضمنها آليًّا بتحديد موضوعين فقط، ولنقل 10 كلمات لكل موضوع.

*نص1: " صعدت **مؤشرات** الأسهم اليابانية في جلسة التعاملات الصباحية ببورصة طوكيو"*

نص2: "المؤشرات الإحصائية الحالية التي تشير الي القيمة السوقية لأسواق المال الخليجية"

نص3: " أُوقظ المريض أثناء العملية لفترة مؤقتة لاستئصال بؤرة الصرع"

نص4: "الفحص المبكر لسرطان القولون واستئصال اللحميات"

نص5: "ما هي مؤشرات هذه العملية الدورية؟"

سيظهر لنا ما يشبه أن:

- النص1 والنص 2 ينتميان بنسبة 100% للموضوع أ.
- النص3 والنص4 ينتميان بنسبة 100% للموضوع ب.
- النص5 ينتمي بنسبة 70% للموضوع أ، وبنسبة 30% للموضوع ب.

والموضوع أ يمثل: صعدت 10%، ومؤشرات 30%، والأسهم 10% ... إلخ

والموضوع ب يمثل: أوقظ 10%، العملية 5%، استئصال 15% ... إلخ

فيمكننا أن نستنتج حينها أن الموضوع أ عن (الاقتصاد)، والموضوع ب عن (الصحة)، حيث لا تحدد الخوارزمية الموضوعات صراحة كما أسلفت، وإنما تقدم لنا احتمالية ارتباط كلمات معينة بموضوع معين يحدده الإنسان.

3. تقييم النموذج الموضوعي

إن اختيار القيمة الصحيحة لعدد الموضوعات في الخوارزمية أحد المهام المهمة في خوارزميات النمذجة الموضوعية، ومنها LDA. ولا بد عند تحديدها أن تراعى عوامل كمية وأخرى نوعية. فمن الناحية الكمية، اقترحت مقاييس متعددة لتحديد العدد الأمثل للموضوعات، ومن ثم تقييم النموذج المناسب. ومن هذه المقاييس مقياس التماسك C_V. وهو مقياس يستعمل لتقدير نموذج موضوعي يتميز بقابلية عالية لفهمه بشريًّا، وقد ثبت أن هذا المقياس هو الأكثر توافقًا مع التفسير البشري (Röder, et. al. 2015). إذ ينظر في تماسك مجموعة الكلمات ذات القيم العليا التي استخلصت في كل موضوع، ويقيم مدى قابلية الموضوع للفهم. ويعتمد هذا المقياس على زيادة تدرجية في عدد الموضوعات باستعمال خوارزمية النافذة المنزلقة sliding window. ويحسب قيمة التماسك لكل خطوة (نافذة). ومن خلال تغيير النافذة تحدد قيم الهايبربراميترات hyperparameters لأفضل الموضوعات. وجوهر ذلك قائم على قياس العلائقيات السياقية (الإحصائية) بين الكلمات في الموضوع الواحد. ويعتمد على دمج قياسين هما: مقياس المعلومات المشتركة المعايَر النقطي normalized pointwise mutual information (NPMI) أحد مقاييس الارتباط القياسية في استخلاص المتلازمات collocation، ومقياس تشابه جيب التمام cosine similarity الذي يقيس جيب التمام للزاوية بين متجهين في فضاء متعدد الأبعاد، وبه يتحدد وجه التشابه بين عنصرين بغض النظر عن حجمهما (Syed and Spruit, 2017). وبهذا يسترجع مقياس التماسك حسابات التصاحب لكلمات معينة باستعمال نافذة منزلقة ذات حجم معين. وتستعمل هذه الحسابات لحساب قيمة NPMI لكل كلمة ذات قيمة عليا بالنسبة لكلمة أخرى ذات قيمة عليا. فينتج عن ذلك مجموعة من المتجهات (متجه لكل كلمة ذات قيمة عليا). ويؤدي اقتطاع المجموعة الواحدة من الكلمات العليا إلى حساب التشابه بين كل

متجه لكلمة ذات قيمة عليا ومجموع المتجهات الأخرى للكلمات العليا باستعمال مقياس تشابه جيب التمام. ومن ثم يكون التماسك هو المتوسط الحسابي لتلك التشابهات. وتقع قيم التماسك الموضوعي بين 0 و1. وكلما زادت القيمة التي يتوصّل لها، زاد التماسك بين الكلمات في الموضوع الواحد، وأصبح الموضوع أو النموذج أفضل.

أما من الناحية النوعية، فلا بد من المعرفة الجيدة بمجال البيانات المراد تحليلها، والقدرة على تخمين تصنيف تقريبي عام للموضوعات التي يراد أن تفصل البيانات على نحوها. وهذا يعني أنه لا بد من أن تكون هناك موضوعات لتمييزها، ولكن ليس بالعدد الذي يفقد به إمكانية تفسيرها. ويكون الحكم البشري على نتائج الخوارزمية قائم على ملاحظة الكلمات الأعلى درجة في كل موضوع، وعلى تحديد الكلمات التي لا تنتمي للموضوعات، والنصوص التي لا تنتمي للموضوعات.

4. الدراسات السابقة

لقد استعملت خورازمية LDA في النمذجة الموضوعية للنصوص العربية على أوعية وأنواع مختلفة من النصوص، كالنصوص الصحفية (Brahmi et. al, 2012) ، والقرآن الكريم (Siddiqui et. al., 2013) ، ومواقع إنترنت (Kelaiaia & Merouani, 2013) ، ولهجات عربية (Zarra et. al. 2017) ، وتغريدات تويتر (Adel & Wang, 2020).

فقد أجرى براهمي وآخرون Brahmi et. al. (2012) دراستهم على نصوص من ثلاث صحف هي: الشروق ورويترز وشينوا. ضمت صحيفة الشروق 11313 نصًّا، نشرت في عامي 2008 و2009، وصنفت تحت 8 تصنيفات. وضمت رويترز 41,251 نصًّا نشرت بين عامي 2007 و2009، وصنفت تحت 6 تصنيفات. أما شينوا فتضمنت 36,696 نصًّا نشرت في عامي 2008 و2009، وصنفت تحت 8 تصنيفات. وهدفت دراستهم إلى إحراز مزيد من التقييم للنمذجة الموضوعية الموجهة في النصوص العربية باستعمال خوارزمية LDA. فبحثوا في أثر أساليب تجذيع الكلمات العربية على أداء النمذجة الموضوعية. وقد استخلصوا الموضوعات بعد تجريب عدد من المجذعات على النصوص الأصلية، ووجدوا أن التجذيع القائم على ما يعرف بالوحدة المعجمية lemma حقق نتائج أفضل من التجذيع القائم على الجذر root. كما بينوا صعوبة تقييم الجوانب الدلالية في النصوص العربية دون معرفة لغوية كافية.

ويحاول صدّيقي وآخرون Siddiqui et. al. (2013) الكشف عن الموضوعات الخفية لسور القرآن الكريم باستعمال خوارزمية LDA. وتمكنوا من الكشف عن الموضوعات المميزة، فضلاً عن الكلمات المميزة التي تصف هذه الموضوعات. وقد كانت منهجيتهم في الكشف عنها على مستويات مختلفة من التفصيل. فعلى المستوى العام، تمكنت الخوارزمية من تحديد موضوعين رئيسين تميزت بهما السور المكية والمدنية. ثم أُسند لكل سورة موضوعًا من بين خمسة موضوعات، ثم موضوعًا من بين عشرة موضوعات، متشاركةً بعض السور الموضوع نفسه. وقد فسر الباحثون في هذه الدراسة النتائج، ولم يقيموا مدى تماسك الكلمات في الموضوع الواحد تقيما آليًّا، ولم يبحثوا عن النموذج الموضوعي الأمثل لسور القرآن الكريم.

وتهدف دراسة كيلايايا ومرواني Kelaiaia& Merouani (2013) إلى إجراء مقارنة بين الخوارزميتين LDA وK-mean في استكشاف الموضوعات من نصوص مواقع إلكترونية عربية تضم نصوصًا في الاقتصاد، والتاريخ، والترفيه، والتعليم والأسرة، والدين، والرياضة، والصحة، والفضاء، والقانون، والقصص، وفنون الطبخ. وكان الهدف من هذه الدراسة مقارنة تأثير الخصائص الصرفية النحوية على أداء الخوارزمية الأولى مقارنة بالثانية. واستعمل مقياسان معروفان هما: مقياس F، ومقياس إنتروبي. وأظهرت النتائج أن LDA تعطي نتائج أفضل من نظيرتها في معظم الأحوال.

وتركز ورقة زارا وآخرين Zarra, et. al. (2017) على اللهجة المغربية، وتطبق منهجًا موجّهًا supervised في تحليل المشاعر على بيانات جمعت من صفحات الفيسبوك تتعلق بآراء الناس حول البرامج التي تبث عبر القنوات الوطنية، وحول الأخبار السياسية، وبعض الآراء والتجارب التي تتعلق ببعض المنتجات التجارية. كما تطبق خوارزمية LDA غير الموجهة unsupervised لاستخلاص الموضوعات الخفية من البيانات. وتقترح منهجًا جديدًا شبه موجه semi-supervised يجمع بين تحليل المشاعر والموضوعات في نموذج واحد يربط كل موضوع بفئة

شعورية معينة. وفيما استعملت الدراسة مصفوفة الإرباك confusion matrix لقياس أداء نموذج تحليل المشاعر، لم تستعمل أي مقياس لقياس النمذجة الموضوعية واكتفت في تقييمها بالاعتماد على رأي خبير.

وجمعت دراسة عادل ووانق Adel & Wang (2020) تغريدات عربية تتعلق بأزمتي المجاعة والكوليرا باليمن، وأزمة اللاجئين في سوريا، تحت الواسمات (الهاشتاقات): كوليرا، ومجاعة، ولاجئ، ومقابلاتها الإنجليزية: cholera وrefugee، وfamine. ثم بني نموذج LDA لاستخراج مواضيع الأزمات من التغريدات باستعمال الكلمات المميزة keywords التي تصف تلك الأزمة. وبعد تحديد العدد الأمثل للموضوعات، أظهر مقياسا الارتباك perplexity والتماسك coherence (u_mass) ضعفًا في الأداء. فقد احتوت النتائج على كلمات ليست ذات علاقة بالأزمة عندما تكون التغريدات قصيرة. كما أن النموذج لم يتعلم الدلالات من الكلمات في الحالة التي تكون فيها الكلمات ذات احتمالية منخفضة. وقد راجع الباحثان العلاقة بين الموضوع والنصوص، ثم جمعوا الكلمات المميزة في كل أزمة وأعادوا الاستعلام عنها عبر تويتر. وقد أنشأوا بذلك مدونة أزمات عربية موسمة بخمس فئات أزمات استخلصت من مكتب الأمم المتحدة (أوتشا). إذ استعمل كل وسم منها مع الكلمات المميزة المولدة من نموذج LDA لتحديد الكلمات الرئيسة لكل صنف من الأزمات. ثم دربت على المدونة لتصنيف الأزمات 3 نماذج هي: آلة الدعم الاتجاهي Support Vector Machine (SVM)، والتصنيف البيزي Naïve Bayes (NB)، وتصنيف الغابة العشوائية Random Forest (RF). وقد قيمت باستعمال مصفوفة الارتباك confusion metrics، وحققت نتائج نموذج تصنيف الغابة العشوائية أعلى النتائج.

وفي هذه الدراسة أجري خوارزمية LDA غير الموجهة على نصوص من الصحف السعودية دون تجذيعها، بهدف الكشف عن الموضوعات الكامنة فيها كشفًا تجريبيًّا يوصل إلى تحديد نموذج بالعدد الأمثل من الموضوعات، وتقييم أدائه تقييمًا آليًّا ونوعيًّا.

5. المنهجية

1.5 *تهيئة البيانات*

تعتمد هذه الدراسة على نصوص إخبارية من الصحف السعودية. وهي 47 صحيفة محلية يفهرس محتوياتها يوميًّا موقع (سعورس). وقد استخلصت نصوص هذه الصحف منه على مدى عشرة أيام من شهر فبراير عام 2021 باستعمال المستخلص العربي[iii]. ويرتب المستخلص العربي النصوص التي استخلصها في مجلدات حسب الأبواب الأكثر ورودًا في الصحف الإلكترونية والورقية. فيصنف النصوص إلى: نصوص رئيسية، واجتماعية، وثقافية، ودينية، وصحية، واقتصادية، وسياسية، ودولية، ورياضية. وكل مجلد مصنف يضم ملفات متعددة بصيغة txt مجموعها 6310 نصا. دمجت النصوص معًا في ملف إكسل واحد حيث يكون كل ملف (نص) في سطر مستقل. ثم روجعت النصوص، وحذف المتكرر منها. وبالاستعانة بالبايثون، حذفت منها علامات الترقيم والتشكيل والرموز والأرقام والحروف الأجنبية والمسافات الزائدة وعلامات التطويل. كما حذفت قائمة الكلمات المستبعدة stop words التي أعددتها وضمنتها جميع الكلمات الوظيفية ذات المباني التقسيمية لا التصريفية، نحو: حروف الجر، والعطف، والضمائر الشخصية، والإشارية والموصولة، ...إلخ؛ بسبب تكراراتها العالية على مستوى النص الواحد، ودورها النحوي الفقير دلاليًّا في ذاتها. كما ضمنتها أيضًا بعض الكلمات الشائعة، كالكلمات الثقافية، ونحوها، مثل: الله ومحمد وبن وكورونا وغيرها؛ لغلبتها على كلمات المحتوى ذات الدلالة الموضوعية العالية في عملية النمذجة. وضمنت أيضا الكلمات المركبة تركيبًا إضافيًّا، نحو: عبد الله، وعبد الرحمن، والكلمات الملازمة للوصف، نحو: البحر الأحمر، ومكة المكرمة؛ لتعاملنا مع الكلمات ذات الوحدة المعجمية الواحدة unigram. وتشير الأرقام في الجدول (1) إلى إحصاءات النصوص التي استخلصت في كل تصنيف، وعدد الكلمات الفعلية والنوعية[iv] قبل التنقيح، ثم إحصاءات ما بعد التنقيح.

2.5 *التطبيق*

لقد أصبح مجموع النصوص بعد تهيئتها 4781 نصًّا، تضمنت 649,734 كلمة فعلية. وتمهيدًا للعمل عليها بالبايثون، فرقت الجمل في النصوص إلى كلمات ذات وحدة معجمية واحدة باستعمال مكتبة جنزم genism. ثم أنشأتُ مدخلين مهمين تتطلبهما الخوارزمية لإنشاء النموذج باستعمال تمثيل حقيبة الكلمات bag of words (BOW)، وهما: رقم تسلسلي لكل

جدول 1:إحصاءات نصوص الصحف السعودية قبل وبعد التنقيح

م	الموضوعات	قبل التهيئة			بعد التهيئة	
		عدد النصوص	عدد الكلمات الفعلية tokens	عدد الكلمات النوعية types	عدد الكلمات الفعلية	عدد النصوص
1	الاجتماعية	1006	149973	22522	84460	734
2	الاقتصادية	925	197280	29636	103422	672
3	الثقافية	244	60626	18665	35748	193
4	الدولية	1033	182755	35581	109890	842
5	الدينية	133	29928	8307	13101	86
6	الرياضية	431	70652	15762	41208	346
7	الرئيسية	1768	336793	52847	191197	1315
8	السياسية	57	14739	2332	5100	42
9	الصحية	713	112354	21192	65608	551
	المجموع	6310	1155100	206844	649734	4781

كلمة في المدونة، وتكرارات كل كلمة. ومن ثم وُلّد لكل كلمة رقم تسلسلي، وربط كل رقم كلمة بتكراراته، لتصبح المدونة بهذه الصورة (رقم تسلسلي، تكرار الكلمة):

[[(0, 1), (1, 2), (2, 1), (3, 1), (4, 1), (5, 1), (6, 5), (7, 1), (8, 1), (9, 2), ...]

ولتشغيل الخوارزمية، استعملت حزمة مالت MALLET الإصدار 2.0 (McCallum, 2002)، والمكتوبة بالجافا من خلال المكتبة السابقة أيضًا، ووجهتها بالعدد المطلوب من الموضوعات. وحيث يوجد لدي معرفة سابقة بموضوعات البيانات منذ جمعها (انظر الجدول 1)، حددت عدد الموضوعات بـ 8 موضوعات، ثم عملت على تحديد القيمة المثلى للموضوعات تحديدًا تجريبيًّا، سعيًا لتحسين أداء الخوارزمية قبل تقديم النتائج. وقد كان منهجي في التحسين هو العمل على إيجاد العدد الأمثل للموضوعات من خلال بناء العديد من نماذج LDA بأعداد مختلفة للموضوعات، ثم اختيار النموذج الذي يعطي أعلى قيمة للتماسك. فمن خلال دالة (compute_coherence_values) في المكتبة دربت 20 نموذجًا من نماذج LDA؛ للوصول للنموذج الأمثل. ثم استعملت نفس الدالة للوصول إلى موضوعات أدق من النموذج الأمثل ذي الموضوعات العامة. حيث أجريت الخوارزمية مرة أخرى على نصوص كل موضوع من موضوعات النموذج الأمثل، بعد تقسيم المدونة إلى سبع مدونات فرعية أخرى ولدتها الخوارزمية في النموذج الأمثل.

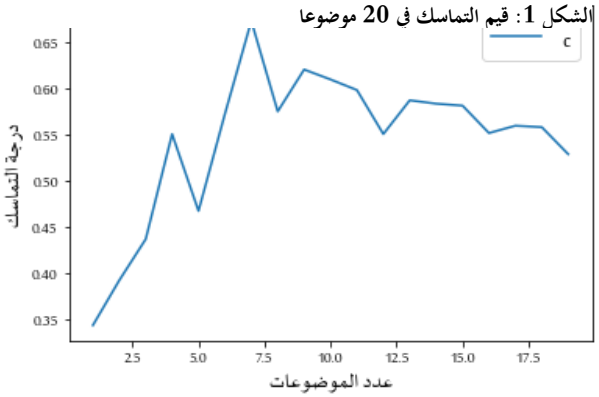
6. التقييم ومناقشة النتائج

إن تقييم النموذج الموضوعي هو جزء مهم من عملية النمذجة الموضوعية. ذلك أن النموذج الموضوعي لا يقدم أي معلومات عن جودة الموضوعات المنتجة. ومن ثم فهو يعيننا على تقييم مدى صلة الموضوعات المنتجة بالمدونة، ومدى فعالية النموذج الموضوعي. ويبين الجدول (2) درجة تماسك النموذج بعد توجيه الخوارزمية لاستخلاص الموضوعات من البيانات بالعدد 8 حدسًا. غير أن النموذج بهذا العدد من الموضوعات – وإن أظهر درجة من التماسك لا بأس بها – يظل نموذجًا حدسيًّا. وقد أظهرت نتائج تحسين اختياري النوعي لعدد الموضوعات بتدريب 20 نموذجًا بعشر كلمات أن القيمة المثلى لعدد الموضوعات في المدونة هي 7 موضوعات، وبدرجة تماسك بلغت 0.6723. ويظهر الشكل (1) درجة التماسك في كل نموذج درب على موضوع واحد إلى 20 موضوعًا بعشر كلمات في كل موضوع. ويتضح في الشكل السابق مدى بلوغ التدريب قمة التماسك في النموذج ذي السبعة موضوعات.

جدول 2: درجات التماسك في نموذجي LDA

الحالة	عدد الموضوعات	درجة التماسك
قبل التجريب	8 موضوعات	0.5664
بعد التجريب (القيمة الأفضل)	7 موضوعات	0.6723

الشكل 1: قيم التماسك في 20 موضوعا



وإذا ما نظرنا في الكلمات العشرة الأكثر احتمالاً أو الأعلى درجة في كل موضوع يمكننا فهم الموضوعات، حيث لا يقدم النموذج تفسيرا لها، ويُتطلَّبُ منا وصف ذلك استنادا على تلك الكلمات (انظر الجدول 3). ويصف الصف الأخير في الجدول السابق المواضيع السبعة التي تدور حولها نصوص المدونة. كما يبين الجدول (4) توزيع هذه الموضوعات على نصوص المدونة. ويلاحظ توازن توزيع الموضوعات، حيث تفاوتت نسبة الموضوعات من كامل المدونة بين درجتين متقاربتين هما 0.124 و0.1558. ويلاحظ أيضًا أن النموذج ولّد موضوعين دقيقين هما: موضوع (1) الرقابة والتوعية، وموضوع (2) التنمية والتطوير. وهما موضوعان لم يكونا ظاهرين في التوزيع التبويبي الذي تنتمي له الصحف في مواقعها على الشبكة، ولربما توزعا بين الأخبار المحلية والمجتمعية. وفي المقابل، لم تظهر الموضوعات الدينية ولا الثقافية في النمذجة، وأعزو ذلك لقلة النصوص في المدونة تحت هذين الموضوعين، حتى وإن بدا أن أقل النصوص هي السياسية فيها. ذلك أن النصوص السياسية يقع غالبها تحت تبويب الرئيسية في الصحف السعودية الذي يضم النصوص الأكثر في مدونتنا.

ولدعم قياس التماسك الدلالي الآلي، فحصتُ الكلمات في كل موضوع فحصًا يدويًّا؛ للتأكد من انتمائها للموضوعات التي فسرتها سابقًا. ولم أعثر على كلمة شاذة واحدة من بين الكلمات في كل مجموعة. إلا أن بعض الموضوعات تتشارك بعض الكلمات باحتماليات متطابقة أو مختلفة، وذلك عائد لطبيعة الموضوع نفسه. فمثلاً في الموضوع (6)، والموضوع (7)، نجد كلمتي (رئيس والرئيس)؛ لأنهما تتشاركان الشأن السياسي عموما سواء كان محليا أو دوليًّا. فيما يتشارك أيضا الموضوع (1) والموضوع (4) في كلمة (وزارة)؛ لتضمن الرقابة والتوعية الشأن الصحي في المنطقة، وهكذا.

ويبين الجدول (5) النص الأعلى تمثيلاً في كل موضوع، ويلاحظ انتماء جميع النصوص الممثلة لموضوعاتها ما عدا النص الأعلى تمثيلاً في موضوع الرياضة (3) الذي سيثبت لنا لاحقًا اختلاط نصوصه بالنصوص الدينية والثقافية، واشتراكه معهما في قلة نصوصه نسبيا مقارنة بغيره من النصوص. وعلى غرار ما فعلته مع الكلمات، استخلصت جميع النصوص الأكثر تمثيلاً في كل موضوع، ثم فحصت الخمسين منها الأكثر تمثيلاً فحصًا يدويًّا، للتأكد من انتمائها للموضوع الذي أسند إليها. وكشفت النتائج في الجدول (6) إلى أن النصوص الخمسين الأعلى تمثيلاً، والواردة في الموضوعات (رقابة وتوعية– تنمية وتطوير– صحة – شؤون محلية) جاءت دقيقة جدًّا في موضوعاتها، وحققت نسبة صحة بلغت 100%. تلتها النصوص في الموضوع (رياضة)، ثم الموضوع (سياسة دولية). فيما حقق الموضوع (اقتصاد) درجة منخفضة تسبب بها الغموض الدلالي بين مفردات النصوص الاقتصادية ونصوص الأحوال الجوية، نحو: بلغت – مستوى – بنسبة، التي يتشاركهما الموضوعان باعتبارهما كلمات عليا تستعمل في سياقات متشابهة بين الموضوعين.

إن حصول بعض الكلمات على احتمالات عالية في نفس الموضوع دلالة على ورودهما معًا في مستند واحد، كما في: (الإجراءات – الاحترازية) و(حالة – الصحة). ولكن قد ترد كلمة ذات احتمالية عالية مع الكلمات في موضوع معين، وليس لها أي علاقة ببقية الكلمات. وهنا تظهر الحاجة إلى استكمال النمذجة الموضوعية من خلال مناهج أخرى، نحو: تحليل العلاقات من خلال الرسوم البيانية الرياضية (التحليل الشبكي network analysis) (Barnes and Harary, 1983). فمن خلال تطبيق هذه المنهجية سنتمكن من الوصول إلى الكلمات المرتبطة بالكلمات الرئيسية، ومعرفة أي الكلمات أهم في علاقتها ببعضها، وأيها أكثر تأثيرًا، فضلاً عن مجموعات الكلمات الأكثر تماسكًا مقارنة بغيرها.

ويبين الجدول (6) نتائج تدريب الخوارزمية مرة أخرى، بعد تصنيف نصوص المدونة في مدونات فرعية حسب موضوعات النموذج الأمثل ذي الـ 7 موضوعات. ويظهر أيضا في الجدول

جدول 3:الموضوعات المولدة من نصوص الصحف السعودية وأوصافها

م	الموضوع 1		الموضوع 2		الموضوع 3		الموضوع 4		الموضوع 5		الموضوع 6		الموضوع 7	
	الكلمة	الدرجة	الكلمة	الدرجة	الكلمة	الدرجة	الكلمة	الدرجة	الكلمة	الدرجة	الكلمة	الدرجة	الكلمة	الدرجة
1	الإجراءات	0.011	العمل	0.009	الأول	0.006	حالة	0.013	الماضي	0.008	رئيس	0.01	المتحدة	0.009
2	الاحترازية	0.011	برنامج	0.004	الاتحاد	0.005	الصحة	0.012	ريال	0.007	مجلس	0.008	الدولي	0.008
3	العامة	0.009	الوطنية	0.004	الفريق	0.005	وزارة	0.007	مليون	0.006	المنطقة	0.006	الخارجية	0.006
4	تطبيق	0.009	إضافة	0.004	المركز	0.004	جديدة	0.007	دولار	0.005	إدارة	0.006	الرئيس	0.006
5	منطقة	0.008	ضمن	0.004	أمام	0.004	الصحية	0.005	بنسبة	0.004	منطقة	0.006	اليمن	0.005
6	وزارة	0.007	القطاع	0.004	فريق	0.004	حالات	0.005	مليار	0.004	التعليم	0.006	الحكومة	0.005
7	الجهات	0.007	الأعمال	0.003	الثاني	0.004	بلغ	0.005	البلاد	0.004	المجلس	0.005	إيران	0.004
8	الوقائية	0.007	الشركة	0.003	كأس	0.003	إصابة	0.005	بلغت	0.003	مدير	0.005	الحوثي	0.004
9	التجارية	0.007	الشركات	0.003	الشباب	0.003	تسجيل	0.005	مستوى	0.003	أمير	0.005	الأمن	0.004
10	الصحية	0.006	الجامعة	0.003	الجولة	0.003	بشكل	0.005	الأربعاء	0.003	التعاون	0.005	الدولية	0.004
الوصف	رقابة وتوعية		تنمية وتطوير		رياضة		صحة		اقتصاد		شؤون محلية		سياسة دولية	

جدول4 : توزيع الموضوعات في المدونة

الموضوع	الوصف	عدد النصوص	نسبة النصوص من المدونة
1	رقابة وتوعية	745	0.1558
2	تنمية وتطوير	651	0.1362
3	رياضة	638	0.1334
4	صحة	740	0.1548
5	اقتصاد	685	0.1433
6	شؤون محلية	593	0.124
7	سياسة دولية	729	0.1525
المجموع		4781	100

جدول 5:النص الأكثر تمثيلاً في كل موضوع

الموضوع	وصفه	النص الأعلى تمثيلا	مدى تمثيله
1	رقابة وتوعية	أعلنت وزارة الداخلية عن مجموعة من القرارات للح...	0.8518
2	تنمية وتطوير	دشن وزير التعليم الدكتور حمد بن محمد آل الشيخ...	0.7671
3	رياضة	أوصى إمام وخطيب المسجد الحرام فضيلة الشيخ الدك...	0.9139
4	صحة	الرياض أعلنت وزارة الصحة اليوم الأربعاء تسجيل...	0.7685
5	اقتصاد	واصل المستثمرون الأجانب تسجيل صافي شراء خلال ت...	0.7266
6	شؤون محلية	عقد مجلس الوزراء جلسته اليوم عبر الاتصال المرئ...	0.8569
7	سياسة دولية	أطلق ولي العهد نائب رئيس مجلس الوزراء وزير الد...	0.819

جدول 6: نسبة الصحة في كل موضوع بفحص 50 نصًّا من النصوص الأكثر تمثيلاً لكل موضوع

الموضوع	الوصف	عدد النصوص	الصحة accuracy
1	رقابة وتوعية	50	%100
2	تنمية وتطوير	50	%100
3	رياضة	50	%98
4	صحة	50	%100
5	اقتصاد	50	%54
6	شؤون محلية	50	%100
7	سياسة دولية	50	%84
المجموع: 350 نصا		**المتوسط**: 91%	

(6) أعلى درجة تماسك حققها النموذج في كل مدونة فرعية تخص كل موضوع بعينه، فضلاً عن تفسيرات الموضوعات الدقيقة في كل موضوع. ففي الموضوع 7 (سياسة دولية)، حقق النموذج المدرب على نصوص السياسة الدولية أعلى درجة تماسك من بين درجات التماسك في الموضوعات الأخرى، وذلك بـ إ8 موضوعات. ورغم جودة النموذج في تحديد تفصيلات موضوع السياسة الدولية في الصحف السعودية، إلا أننا نجد موضوعًا لا يبدو له علاقة بالسياسة الدولية، وهو الموضوع 6: (مشاريع عالمية). وإن كان يظهر انتماؤه الصحيح لموضوع الاقتصاد، إلا أنه قد يُضمن تحت موضوع السياسة الدولية؛ لتحقيق المشاريع الوطنية العالمية مكانة سياسية دولية.

ورغم أن نسبة الصحة في نصوص موضوع (التنمية والتطوير) قد حققت 100%، كما في الجدول (6)، إلا أن تماسك الموضوعات الدقيقة فيها لم يكن جيدًا، والأقل من بين تماسك الموضوعات الأخرى. ورغم ذلك، تمكنا من تفسير موضوعاتها الـ 16 التي حقق النموذج عندها أعلى درجة تماسك ممكنة. ولعل عدم التماسك في موضوعات (التنمية والتطوير) يعود إلى النمو الذي تشهده السعودية في جميع المؤسسات الحكومية والخاصة، وقلة النصوص إذا ما نظرنا في كل مجال من مجالاته على حدة.

الجدول (6) نسبة الصحة في كل موضوع بفحص 50 نصا من النصوص الأكثر تمثيلاً لكل موضوع

وتكشف لنا الموضوعات الدقيقة في الموضوع 5: (الاقتصاد)، أسبابًا أخرى لانخفاض نسبة الصحة في موضوعاته التي قد أشرت إليها سابقًا. إذ لا نجد اختلاط نصوصه بنصوص (الطقس) التي ظهرت فيه في الخمسين الأولى من نصوصه فحسب، بل نجد أيضا نصوصًا في السياسة العربية والعلاقات الدولية تتخلل هذا الموضوع العام. كما يبدو الموضوعان (الاقتصاد المالي والاقتصاد النفطي) فيه منطقيين جدًّا. وقد كنت أشرت إلى عدم ظهور النصوص الدينية والثقافية في الموضوعات العامة، بحكم وجود تبوياتها ضمن تبويبات الصحف المحلية في مواقعها الإلكترونية. وأظهرت النتائج وجود نصوصها في موضوع (3): الرياضة، وتحديدًا في الموضوعات الدقيقة التي فسرت بـ: (الفن– شؤون حياتية — الدين).

ويلاحظ عامة في الموضوعات الدقيقة تكرار ورود كلمات عديدة في أكثر من موضوع دقيق لكل موضوع عام. فنجد مثلا: كلمة (العامة)، وقد وردت في 8 موضوعات، وكلمة (العمل)، وقد وردت 6 مرات. وهذا ما لا يحدث في النماذج الموضوعية المثلى إلا نادرًا. إذ إن ذلك ليس غريبًا في الموضوعات الدقيقة، فمتى ما ارتفع عدد الموضوعات، سيظهر ذلك بوضوح أكثر.

7. الخاتمة

النمذجة الموضوعية مجال متقدم في معالجة اللغات الطبيعية، يعين في تبسيط وتحليل وتصنيف المدونات من خلال تحديد بنيتها الموضوعية الأساسية. وباستعمال خوارزمية LDA التي نفذت بالبايثون، عرضتُ كيفية تطبيقها على نصوص الصحف السعودية. ورأينا كيف يمكن من خلال إجراءات بسيطة نسبيًّا تدريب نموذج موضوعي، باستعمال مدونة ذات نصوص محدودة الزمان والمكان.

إن مهمة تقييم النماذج الموضوعية غير الموجهة أكثر صعوبة من تقييم النماذج الموضوعية الموجهة التي يمكن قياسها. حيث تسعى مقاييس تقييم النماذج الموضوعية غير الموجهة إلى الكشف عن مدى سهولة فهم الموضوعات التي أنتجها النموذج بالنسبة للبشر. وقد كان التقييم للنموذج في هذه الدراسة كميًّا ونوعيًّا، إذ اعتمدت على مقياس التماسك C_V coherence، وركزت في الجانب النوعي على الكشف عن الموضوعات الدلالية العامة والدقيقة التي تضمنتها نصوص الصحف السعودية. وأظهر تدريب النموذج مرة أخرى على نصوص الموضوعات العامة وصولا إلى الموضوعات الدقيقة، ما كان يصعب الوقوف عليه يدويًّا أيضًا. وتميزت الخوارزمية بالوقوف على مواضيع دقيقة جدا يمكن استبعادها لاحقًا لتحسين النموذج.

وقد استعملت أسلوب حقيبة الكلمات BOW لتمثيل كلمات المدونة بالأرقام، باعتبارها مدخلاً أساسيًّا للخوارزمية. وهو الأسلوب العام للخوارزمية. ورغم سهولة هذا التمثيل وبساطته، إلا أنه يمثل الكلمات بمعلومات قليلة تؤثر على النتائج. لذا قد تستعمل أساليب أخرى لتمثيل كلمات المدونة نحو: أسلوب تكرار الكلمة في النص ومعكوسه في النصوص term (TF- frequency-inverse document frequency (IDF، الذي يميز بين الكلمات ذات التكرارات العالية والقيمة الدلالية المنخفضة، والكلمات ذات التكرارات المنخفضة والقيمة

جدول 7:الموضوعات الدقيقة في موضوعات النموذج الأمثل ذي الـ 7 موضوعات

الموضوع	التماسك	الكلمات العليا	الوصف
1 رقابة وتوعية	0.6222	1. العامة، ريال، النيابة، الهيئة، ألف، لمدة، مالية، العقوبات، المخالفة، التجارة 2. الداخلية، الإجراءات، وزارة، الجهات، الاحترازية، بالإجراءات، تقديم، فيروس، الطلبات، الصحية 3. البلدية، ضمن، المخالفين، بالتعاون، العاصمة، الحملة، المقدسة، وأوضح، الجهات، الفرعية 4. الوقائية، الإجراءات، الاحترازية، والتدابير، الالتزام، تطبيق، الصحية، فيروس، المجتمع، انتشار 5. العامة، الأمنية، الرياض، التواصل، النظامية، الإجراءات، النيابة، منطقة، المتحدث، الاجتماعي 6. الوزارة، المصلين، المساجد، الشؤون، بمنطقة، الإسلامية، وزارة، الرياض، الإجراءات، الجهات 7. منطقة، عسير، المنطقة، أمين، أعمال، المخالفات، الحميدي، البيئة، أمير، الخاصة 8. الاحترازية، التجارية، الرقابية، المنشآت، مخالفة، تطبيق، الوقائية، أمانة، الجولات، جولة	تجارية مؤسسية صحية أمنية مجتمعية صحية نظامية دينية بيئية بلدية
2 تنمية وتطوير	0.3884	1. الشركات، الصغيرة، الاقتصاد، السعودي، والمتوسطة، المحلي، الاقتصادي، القرار، ريال، الاستثمار 2. الجامعة، التعليم، الجامعات، جامعة، الوزارة، وفق، الإستراتيجية، التعليمية، وفقا، الوطنية 3. الصحية، برنامج، الطبية، العاملة، البرنامج، الإدارة، الصحي، العامة، الصحة، القوى 4. المالية، الشركات، المالي، العامة، النظام، العالمي، البنك، صندوق، مركز، بشكل 5. وزارة، الوزارة، الطرق، العمل، وفق، طريق، مستوى، إضافة، الخاصة، الخدمة 6. البشرية، الموارد، العمل، برامج، دعم، المركز، البرنامج، الوطنية، عمل، برنامج 7. الطاقة، الجودة، التقني، التدريب، كفاءة، البيئة، اللقاء، المنظمة، والمهني، عدم 8. العمل، الأعمال، الحكومية، الجهات، التعاون، الاتفاقية، رئيس، الغرفة، إطار، المهنية 9. شركة، جازان، قطاع، رؤية، الصناعية، رئيس، المشروع، المنطقة، العالم، الصناعة 10. إدارة، عسير، المنطقة، المعلومات، المبادرة، شركة، الاتصالات، هيئة، مجال، سنوات 11. العالم، الرقمي، التقييم، الإمارات، المريخ، تقييم، منصة، مسبار، أول، دبي 12. الهيئة، الخدمات، خدمات، مجلس، الكهرباء، القطاع، تطوير، الحديدية، الشركة، المهندس 13. المرأة، المتحدة، المجتمع، الاجتماعية، المؤسسة، الوطن، الأمم، وتعزيز، المستقبل، الدولي 14. البرامج، إدارة، مدير، الجمعية، الملتقى، العامة، الإلكتروني، المجتمع، الوظائف، ممثلة، 15. ضمن، ألف، برنامج، الماضي، منطقة، مليون، السكنية، الشركة، مناطق، الحلول، 16. الاستثمار، السعودي، العلمي، البحث، جامعة، تقنية، العلمية، مجالات، إضافة، الورشة	اقتصاد محلي تعليم صحة اقتصاد مالي خدمات الموظفون تدريب شراكات صناعات اتصالات فضاء مرافق عامة المرأة توظيف الإسكان البحث العلمي
3 رياضة	0 .5581	1. الخيل، العالم، سباقات، كأس، وزارة، مختلف، جانب، السباق، مليون، الأولى 2. الدقيقة، الثاني، المباراة، فريق، المركز، الهدف، الدوري، طريق، الأول، دوري 3. الأول، الشوط، جاء، فهد، سنوات، الحالي، دائما، الآن، للممالك، أبو 4. السعودي، البطولة، بطولة، الاتحاد، العالم، منافسات، الأول، الجولف، الدولية، الرياضة 5. الاجتماعي، جدا، المجتمع، التواصل، وسائل، شيء، الرأي، الإعلام، الحديث، الأمر 6. القدم، نادي، لكرة، الاتحاد، النصر، النادي، الشباب، إدارة، الأهلي، اللاعب 7. العمل، علي، مصر، عبدالرحمن، الفنان، قدم، عدة، عمل، أحمد، محمود 8. الحياة، يقول، الإنسان، الآخر، المرأة، الناس، آخر، الحقيقة، أخرى، الكاتب 9. صلى، وسلم، القيامة، الشيخ، رسول، تعالى، الميزان، ابن، لله، أحمد 10. الفريق، الجولة، الهلال، نقطة، أمام، كأس، المركز، للمحترفين، الأهلي، الرياضية	ركوب خيل تحليل رياضي كروي تحليل رياضي فروسي رياضة قولف إعلام رياضي كرة قدم فنون* شؤون حياتية* دين* دوري سعودي
4 صحة	0.5762	1. اللقاح، اللقاحات، بشكل، لقاح، التطبيق، تطبيق، العالم، العالمية، الأمراض، الجديدة 2. حالة، الصحة، جديدة، وزارة، إصابة، حالات، الصحية، وفاة، إجمالي، الطبية	تقنيات إحصاءات
5 اقتصاد	0.584	1. المركز، منطقة، الوطني، الرياح، المياه، الأمطار، الساعة، السطحية، للأرصاد، حالة 2. رئيس، الرئيس، جمهورية، مجلس، بمناسبة، الوزراء، التقدم، برقية، تهنئة، بعث 3. لبنان، قوات، الشرطة، الاحتلال، حزب، وقالت، مصادر، جنوب، مدينة، السبت 4. دولار، ريال، مليون، بنسبة، نقطة، مليار، الماضي، مستوى، بلغت، شركة 5. النفط، البلاد، المتحدة، الوقود، الولايات، يناير، للبرميل، الأمريكية، بسبب، برميل	طقس* تواصل دولي* سياسة عربية* اقتصاد مالي اقتصاد نفطي

6 شؤون محلية	0.50	1. التعاون، رئيس، اللقاء، المشترك، الخارجية، الحرمين، خادم، وجرى، الشريفين، الشيخ 2. الحرمين، العالم، خدمة، العمل، الشريفين، المسجد، جهود، النبوي، الرئيس، الحرام 3. التعليم، التعليمية، الطلاب، الإدارة، للتعليم، والطالبات، مستوى، العربي، اللغة، العملية 4. منطقة، أمير، سموه، صاحب، المنطقة، بالمنطقة، مدير، سلطان، شكره، استقبل 5. المجلس، اللجنة، رئيس، الهيئة، مجلس، التقرير، تطوير، بشأن، الاجتماع، الجهات 6. إدارة، الأعمال، مجلس، الهيئة، الإعلام، المكلف، عسيري، الفترة، التنفيذية، العامة 7. مجلس، الوزراء، عشرة، العامة، وظيفة، بالمرتبة، المجلس، بوزارة، مدير، الرابعة 8. المهندس، الجمعية، العمل، الصحية، البيئة، مركز، العامة، إدارة، برامج، القوات	شؤون الحرمين الشريفين شؤون الحرمين والمسجد النبوي تعليم زيارات مسؤولين اجتماعات مسؤولين تكليفات تعيينات وترقيات أمن بيئي
7 سياسة دولية	0.6762	1. الرئيس، مجلس، ترامب، رئيس، السابق، ترمب، الشيوخ، بايدن، البلاد، أعضاء 2. الدولي، الحوثي، الإرهابية، التحالف، الشرعية، الحوثية، قوات، مطار، دعم، المدنيين 3. إيران، النووي، الإيراني، الأمريكية، دول، الاتفاق، الإيرانية، المتحدة، بايدن، طهران 4. العراق، القوات، الاتحاد، الجيش، الحكومة، الأوروبي، أربيل، اتفاق، العراقية، مقتل 5. نظام، حقوق، التشريعات، الإنسان، العدالة، العهد، الشخصية، المدنية، المجتمع، تطوير 6. العالم، جديدة، الطبيعية، مشروع، الجزيرة، الأولى، جزيرة، مستوى، العالمية، حماية 7. المتحدة، اليمن، الخارجية، اليمنية، اليمني، الولايات، الحوثيين، الأمم، الإنسانية، الأمريكية 8. الخارجية، رئيس، الحكومة، الصحف، الفلسطينية، صحف، الليبي، ليبيا، مصر، الوزراء	سياسة أمريكية حرب اليمن الاتفاق النووي علاقات دولية (عراقية- أوروبية) قانون سياسي مشاريع عالمية قرارات أمريكية علاقات دولية (فلسطينية-أفريقية)

الدلالية العالية. أو استعمال تقنيات التعلم العميق deep learning

نحو تقنية تضمين المعاني في الكلمات word2vec. فحسب ما يذكر بودكار وروجيتش Budhkar & Rudzicz (2019) أن الجمع بين LDA وword2vec يولد خصائص مميزة تعالج إشكالات عدم وجود معلومات سياقية مضمنة في النماذج.

وإذا كانت النمذجة الموضوعية تفكك النصوص إلى عناصرها الخطابية، فيمكن لمستعمليها الاستفادة من نظريات تحليل الخطاب في تفسير نتائجهم. فالموضوع – وهو العنصر الرئيس في النمذجة الموضوعية– يحتوي على معلومتين مهمتين للخطاب هما: الكلمات الشائعة، والعلاقات بين تلك الكلمات. كما أن افتراض وجود موضوعات متعددة كامنة في النمذجة الموضوعية، يمكن من معالجة الخطابات التعددية آليًّا. فضلاً عن أن التقنيات الحديثة في النمذجة الموضوعية يمكنها اكتشاف التغيرات الزمنية في الخطاب، والتفاعلات مع العناصر غير الخطابية. وهو ما يعرف بالنمذجة الديناميكية التي تنظر في التغير الزمني للكلمات المميزة keywords.

ويمكن أيضًا توسيع استعمال النمذجة الموضوعية لتشمل ما يعرف بالنقرامية n-grams التي تُجري النمذجة على كلمتين bigram، أو ثلاث كلمات trigram. كما يمكن أيضًا أن تستعمل للبحث عن موضوعات مخصوصة في نصوص عامة، كالبحث عن الموضوعات الأكاديمية في الصحف، أو الموضوعات الطبية في النصوص الدينية، وغير ذلك.

آمل أن تشجع هذه الدراسة الباحثين في مجال علم اللغة التطبيقي على استعمال النمذجة الموضوعية في أبحاثهم المعتمدة على نصوص عربية ضخمة، لكي نتمكن من البحث في مدى ملاءمة هذه التقنية لأغراض البحث العلمي في مجال علم اللغة التطبيقي.

الهوامش

1. مع ملاحظة أن ما تقوم به الخوارزمية يختلف عن مفهوم التصاحب collocation في المدونات الذي يعمل مع سياقات ضيقة المدى تتفاوت بين ثلاث وخمس كلمات، انظر:
Baker, P. (2006). Using Corpora in Discourse Analysis. London: Continuum, p.95
2. الهايبربراميتر قيمة تستعمل للتحكم في عملية التعلم وتحدد يدويا، على خلاف البراميتر الذي تحدد قيمته خلال عملية التدريب آليا.
3. أداة طورتها مدينة الملك عبد العزيز للعلوم والتقنية يمكن من خلالها إنشاء وتحديث المدونات اللغوية آليًّا باستعمال

RSS feeds، وتتميز بتنظيمها للنصوص وتصنيفها وفق احتياجات الباحث.

4. يقصد بالكلمات الفعلية عدد الكلمات في النص بتكراراتها، ويقصد بالكلمات النوعية عدد الكلمات في النص بدون تكراراتها

المراجع

Adel, G. and Wang, Y. (2020). Detecting and Classifying Humanitarian Crisis in Arabic Tweets. In *3rd International Conference on Artificial Intelligence and Big Data*, Chengdu, China, 28-31/05/2020.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Barnes, J. A. and Harary, F. (1983). Graph theory in network analysis. Social networks, 5(2), 235-244.

Budhkar, A., & Rudzicz, F. (2019). Augmenting word2vec with latent Dirichlet allocation within a clinical application. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, USA, 2-7/5/2019.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993–1022.

Brahmi, A., Ech-Cherif, A. and Benyettou, A. (2012). Arabic texts analysis for topic modeling evaluation. Information retrieval, 15(1), 33-53.

Floyd, R. W. (1967). Non-deterministic algorithms. Journal of the ACM, 14(4), 636–644.

Foucault, M. (1970). The archaeology of knowledge. Social science information, 9(1), 175-185.

Gerlach, M., Peixoto, T. P. and Altmann, E. G. (2018). A Network Approach to Topic Models. Science advances, 4(7), 1-12.

Kelaiaia, A. and Merouani, H. F. (2013). Clustering with Probabilistic Topic Models on Arabic Texts. In: A. Amine, A. Otmane, L. Bellatreche (eds.) Modeling Approaches and Algorithms for Advanced Computer Applications, Cham: Switzerland, Springer.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. Available at: http://mallet.cs.umass.edu (accessed on 18/03/2020)

Murakami, A., Thompsom, P., Hunston, S. and Vajn, D. (2017). What is this corpus about? Using topic modelling to explore a specialised corpus. Corpora, 12(2), 243–277.

Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining, Shanghai, China, 02/2015.

Siddiqui, M. A., Faraz, S. M. and Sattar, S. A. (2013). Discovering the thematic structure of the Quran using probabilistic topic model. In International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, Taibah University, Madinah, Saudi Arabia, 22-25/12/2013.

Zarra, T., Chiheb, R., Moumen, R., Faizi, R. and Afia, A. E. (2017). Topic and sentiment model applied to the colloquial Arabic: a case study of Maghrebi Arabic. In Proceedings of the international conference on smart digital environment, Rabat, Morocco, 07/2017.

Syed, S. and Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19-21/10/2010.

Van Dijk, T. A. (1993). Principles of Discourse Analysis. Discourse & Society, 4 (2), 249–283.