

مجلة اللسانيات العربية، العدد 14، جمادى الآخرة، 1443هـ/يناير، 2022م

الفهرسة الدلالية الكامنة وتخصيص دركليه الكامن لنمذجة موضوعات الشعر العربي عبر العصور

سلطان بن ناصر المجيلول

قسم اللغة العربية، كلية الآداب، جامعة الملك سعود، الرياض، المملكة العربية السعودية

توثيق البحث APA Citation:

المجيلول، سلطان. (2022). الفهرسة الدلالية الكامنة وتخصيص دركليه الكامن لنمذجة موضوعات الشعر العربي عبر العصور. مجلة اللسانيات العربية، 14، 57-80.

Submission Date: 15/05/2021

Acceptance Date: 27/08/2021

تاريخ الإرسال: 1442/10/03

تاريخ القبول: 1443/01/19

Abstract

The current paper entitled “Latent Semantic Indexing and Latent Dirichlet Allocation: Topic modeling for Arabic poetry through the ages” explores the performance of topic modeling algorithms in NLP for texts of Arabic poetry. The latter are available as open-source data covering texts from Arabic poetry, including 58,000 poems composed by 653 poets from five periods ranging from the pre-Islamic, the Islamic, the Abbasid and the Andalusian times to the present day. Latent Semantic Indexing, Latent Dirichlet Allocation and Optimizing Latent Dirichlet Allocation were deployed to extract ten topics from each of the data sets. Utilizing the first and the second models has resulted in an unconvincing performance. Improvements were made to the third model using the Cv measure of coherence score, resulting in topics with maximum human-interpretable capacity of the generated poetic topics.

Keywords: Machine learning; Topic modeling; Latent semantic analysis; Singular value decomposition; Latent Dirichlet allocation.

الملخص

تستكشف هذه الدراسة أداء خوارزميات نمذجة الموضوعات في معالجة اللغة الطبيعية على النصوص الشعرية العربية. تتوافر هذه النصوص في بيانات مفتوحة المصدر، وتتضمن 58 ألف قصيدة لستمئة وثلاثة وخمسين شاعراً من خمسة عصور: الجاهلي، والإسلامي، والعباسي، والأندلسي، والحديث. استخدمت خوارزمية التحليل الدلالي الكامن، وتخصيص دركليه الكامن غير الأمثل، ثم تخصيص دركليه الكامن بالنموذج الأمثل، وذلك لاستخراج عشرة موضوعات من كل مجموعة من مجموعات البيانات. ظهرت نتائج التفسير غير مرضية في النموذج الأول والنموذج الثاني. أُجريت تحسينات في النموذج الثالث بمقياس Cv لتحديد درجة التماسك، ونتج منها توليد موضوعات ذات قابلية تفسيرية دقيقة للموضوعات الشعرية.

الكلمات المفتاحية: تعلم الآلة، نمذجة الموضوعات، التحليل الدلالي الكامن، تجزئة القيمة المفردة، تخصيص دركليه الكامن.

1. المقدمة

أصبح الذكاء الاصطناعي -الذي يُعدّ جزءاً من علوم الحاسب- أساساً لمجالات جودة المجتمعات الإنسانية، ويُعدّ تعلم الآلة جزءاً من الذكاء الاصطناعي، ويُعدّ معالجة اللغة الطبيعية جزءاً من تعلم الآلة لعناصر اللغة وخصائصها وسماتها. ويعمل

الإنسان على تدريب الآلة واختبار قدراتها لفهم اللغة بنماذج models تتضمن خلاصة ما تعلّمت الآلة وفهمته من البيانات اللغوية. يتفرع تعلّم الآلة إلى أربعة أقسام: التعلم التقليدي، والتعلم المعزز، والتجميع، والتعلم العميق (لحلج، 2020). في التعلم التقليدي التعلّم الموجه (بيانات يعمل الإنسان على تصنيفها تصنيفاً واضح الخصائص ثم يُدرّب الآلة على فهمها)، والتعلّم غير الموجه (بيانات تعالجها الآلة معالجة مباشرة دون تدخل الإنسان). وينتهي نموذج التحليل الدلالي الكامن (Landauer & Dumais، 1997)، ونموذج تخصيص دركليه الكامن (Blei et al.، 2003؛ Blei & Lafferty، 2009؛ Blei، 2012) إلى ما يُسمّى بالتعميم generalization الذي تندرج ممارساته البرمجية ضمن التعلم غير الموجه، والتعميم بهذين النموذجين هو ما سيُطبّق في هذه الدراسة على بيانات 58 ألف قصيدة عربية. أما المنهج فسيُستخدم تجزئة القيمة المفردة singular value decomposition للنموذج الأول، وسيُستخدم منهجان للنموذج الثاني: (النموذج غير الأمثل والنموذج الأمثل)¹. يُستعمل هذان النموذجان في عدة أغراض تحليلية، منها: استرداد المعلومات information retrieval ونموذجة الموضوع topic modeling وتلخيص النصوص text summarization ورؤية الحاسب computer vision.

للموضوعات أهمية في الدراسات الخطابية وتحليل الخطاب، فقد أوليت اهتماماً منذ أعمال فان دايك van Dijk (1977a، 1977b، 1980)، لما لثنائية الموضوع topic والتعليق comment من دور في فهم النص فهماً دقيقاً. تجدر الإشارة هنا إلى أن مفهوم الموضوعات topics في تحليل الخطاب، وبحسب ما يتجه إليه فان دايك (1980)، لا ينعكس من الكلمات، بل من خلاصة موضوع النص الذي يتمثل في القضايا المعبرة عن المجمل النصي، وتُستخرج الموضوعات من النصوص بوصفها مفاتيح لفهم ما يتضمنه الخطاب النصي. ويؤكد فان دايك أن موضوع النص يُعبّر عن الموضوعات الافتراضية، غير أن تحديدها باليد والعين ليس ممكناً في النصوص الضخمة، إذ ليس من السهل معرفة موضوع الخطاب من المستوى البنيوي المعقد للخطاب، ولا يكون ذلك سهلاً إلا في المدونة اللغوية الصغيرة، التي فيها يسهل تصنيف التعليقات والموضوعات تصنيفاً واضحاً (فان دايك van Dijk، 1977a). ويؤكد فان دايك أن تحديد الموضوع والتعليق من النص أمر نسبي (فان دايك van Dijk، 1977b: ص 56)، وما دامت النسبية واقعة في مجال تحليل الخطاب تحليلًا يدوياً، فالنسبية باستخدام تعلم الآلة غير الموجه بخوارزمية التحليل الدلالي الكامن وخوارزمية تخصيص دركليه الكامن تكون محتملة، إلا أن أعمال التحسين من جهة أولى، وتطبيق التفسير العقلي الإنساني من جهة ثانية، يُقلّلان كثيراً من هذه النسبية. وفي سياق التفسير العقلي الإنساني، فالمراد منه هو التحليل لمخرجات نموذج الخوارزميتين بالإجابة عن سؤالين مهمين: هل الموضوعات متميزة تمايزاً تاماً أم أن هناك موضوعات مكرّرة؟ وهل للكلمات في كل موضوع علاقة (دلالية) بالموضوع نفسه أم لا؟

وبما أن هذه الدراسة تستكشف الموضوعات الشعرية، فهناك جانب مهم تجدر الإشارة إليه حول المصطلحين الأدبيين: الأغراض الشعرية، والموضوعات الشعرية. استعمل المصطلح الأول من لدن النقاد العرب القدماء، ولم أقف على حدّ علمي على مصطلح الموضوعات الشعرية عندهم، فهو مصطلح على ما يبدو حديث في النقد العربي الحديث، ونقل بهذا المفهوم من وسيط الترجمة، ومن أدبيات تحليل الخطاب الأدبي. والجدير بالإشارة هنا أن الأغراض الشعرية كثيرة، فمنها المدح، والفخر، والحماسة، والثناء، والهجاء، والاعتذار، والوصف، والغزل، والحكمة، والنصح، واللهو والمجون، والطبيعة، والزهد. أما مفهوم الموضوعات فأكثر توسعاً، فقد تصل إلى مئات الموضوعات، وعليه قد تحمل القصيدة غرضاً محدداً أو عدة أغراض كما هو في القصائد الجاهلية، أو قد تحمل غرضاً يتضمن عدة موضوعات متباينة كالغزل والنسيب والمدح، أو قد يكون في داخل الغرض الواحد كالغزل مثلاً عدة موضوعات ذات علاقة بالغزل نفسه كالحب والغيرة والوجد والنأي... إلخ.

تعتمد هذه الورقة إلى تطبيق التحليل الدلالي الكامن وتخصيص دركليه الكامن على بيانات الدراسة (58 ألف قصيدة من العصر الجاهلي، والعصر الإسلامي، والعصر العباسي، والعصر الأندلسي، والعصر الحديث لدول الشرق الأوسط، والعصر الحديث لدول شمال أفريقيا)، وتتضمن المباحث الآتية.

أولاً: التعريف بخوارزمية التحليل الدلالي الكامن وتخصيص دركليه الكامن، والتبسيط المباشر لكيفية عملها على النصوص اللغوية (المبحث 2).

ثانياً: بيانات القصائد العربية المفتوحة، والمعلومات الواصفة لها من حيث أزمنتها وعدد شعرائها (المبحث 3).

ثالثاً: التحليل والنتائج (المبحث 4).

رابعاً: الخاتمة والتوصيات (المبحث 5).

وترمي هذه الدراسة إلى تبيان أداء تعلم الآلة غير الموجه لنمذجة الموضوعات على استخراج الموضوعات من النصوص اللغوية الضخمة، وإلى تقييم هذا الأداء في سياق تحليل الموضوعات العامة من تلك القصائد في تلك الحقب الزمنية لكل تلك العصور.

2. نمذجة الموضوعات

في هذا المبحث إجابة مبسطة عن السؤالين الآتيين: ما دقة الموضوعات التي يستخرجها نموذج التحليل الدلالي الكامن؟ وما دقة الموضوعات التي يستخرجها نموذج تخصيص دركليه الكامن. سنذكر أولاً الأسلوب النظري بطريقة مبسطة لكل نموذج، ثم سنقدّم مثالا على جمل عربية لغرض تبسيط عمل كل نموذج في استخراج الموضوعات من النصوص.

تعمل الخوارزمتان بتحويل الكلمات إلى مصفوفات رقمية، ويختلفان في إجراء ذلك التحويل. يتضمن هذا التحويل إنشاء متجهات vectors لكل ملف (أو نص) وكل كلمة، تحمل أوزاناً تُمثّل قيماً لنسبة تكرار كل كلمة في كل نص إلى المجموع الكلي من الكلمات في كل النصوص. يُجرى حساب التكرارات بالقياس الإحصائي (تكرار المصطلح-تكرار المستند العكسي) term frequency-inverse document frequency TF-IDF، الذي يقوم بقسمة تكرار كل كلمة على المجموع الكلي لعدد الكلمات في النص، ثم تُعطى قيمة أعلى للكلمات الأقل تكراراً، ليكون من ذلك مغزىً دلاليًا عند كشف النصوص.

يعمل نموذج التحليل الدلالي الكامن غير الموجه على الإبقاء على القيم المفردة الأعلى، وذلك بتجزئة القيمة المفردة singular value decomposition الذي يقوم بتحليل عوامل المصفوفة، وهو جبر خطي يُمثّل مصفوفة رقمية للكلمات والنصوص، ويعمل على تحليل التشابه الدلالي بين الكلمات من النصوص. يُعبّر عن هذا الجبر الخطي بـ X (الكلمات والمستندات)، وبـ U (الكلمات والموضوعات)، وبـ V (الموضوعات والمستندات) بالمصفوفة الآتية:

$$M = U\Sigma V^T$$

حيث إن مصفوفة M تشير إلى كل الاحتمالات الرياضية لأوزان النصوص ككل d وأوزان الكلمات ككل، ومصفوفة U تعمل على تجزئة أبعاد الكلمات، ومصفوفة Σ تشير إلى احتمالات الموضوع (مصفوفة الموضوعات)، ومصفوفة V تعمل على تجزئة أبعاد المستندات. عند تحديد عدد الموضوعات المراد استخراجها، يعمل التحليل الدلالي الكامن على تمثيل الكلمات في الصفوف وتمثيل النصوص أو المستندات (d) في الأعمدة، وتحويلها إلى متجهات vectors من أجل قياس التباعد من جهة أولى، ثم تقليص التباعد من جهة ثانية. لنفترض المثال الآتي ليُشرح من خلاله عمل هذا النوع من التحليل في توليد الموضوعات، ولنفرض بأن كل جملة عبارة عن نص أو مستند (d):

سلطان المجلول، الفهرسة الدلالية الكامنة وتخصيص دركليه الكامن لنمذجة موضوعات الشعر العربي عبر العصور

1. d: فهد وسلمان

2. d: القهوة رائعة

3. d: فهد يحب القهوة

4. d: الحياة رائعة

5. d: هل تعلم أن القهوة رائعة؟

بعد إزالة الكلمات الوظيفية من الجمل في المعالجة القبلية (الجدول 1)، تبقى الكلمات الآتية: (فهد)، (سلمان)، (القهوة)، (رائعة)، (يحب)، (الحياة)، (تعلم)، (مفيدة). تعمل تجزئة القيمة المفردة على تحويل هذه الكلمات إلى ثلاث مصفوفات: U ، Σ ، و V (الشكل 1).

جدول 1

مصفوفة IC : المستند-الكلمة $term-document$ حسب ورود الكلمة أو عدم ورودها في المستند

d5	d4	d3	d2	d1	M
0	0	1	0	1	فهد
0	0	0	0	1	سلمان
1	0	1	1	0	القهوة
1	1	0	1	0	رائعة
0	0	1	0	0	يحب
0	1	0	0	0	الحياة
1	0	0	0	0	تعلم

شكل 1

مصفوفة U للمستندات (تشابه الكلمات للموضوعات)، والمصفوفة القطرية Σ لقيم الجذور التربيعية لمصفوفة U و V ، ومصفوفة القيم المفردة V (المتجهات المفردة الصحيحة في الأعمدة لتشابه الموضوعات للمستندات). ملحوظة: الكلمات ذات الأوزان الأكبر في مصفوفة V تكون متشابهة دلاليا وهي التي تُستخرج بوصفها الكلمات للموضوعات

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} 0.236005 & -0.691237 & -0.301982 & 0 & 0.107998 & -0.447214 & -0.40452 \\ 0.048741 & -0.31552 & -0.505087 & 0.5 & -0.171236 & 0.447214 & 0.40452 \\ 0.670733 & -0.131878 & 0.423415 & 0 & -0.284109 & 0.447214 & -0.26968 \\ 0.609306 & 0.44861 & -0.327561 & 0 & -0.217881 & -0.447214 & 0.26968 \\ 0.187264 & -0.375717 & 0.203105 & -0.5 & 0.279234 & 0 & 0.6742 \\ 0.125837 & 0.204771 & -0.547871 & -0.5 & 0.345462 & 0.447214 & -0.26968 \\ 0.26436 & 0.144574 & 0.160322 & 0.5 & 0.795933 & 0 & 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2.41703 & 0 & 0 & 0 & 0 \\ 0 & 1.78628 & 0 & 0 & 0 \\ 0 & 0 & 1.26407 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0.607706 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.117808 & -0.563606 & -0.638467 & 0.5 & -0.104061 \\ 0.529592 & 0.177314 & 0.0758289 & 0 & -0.82604 \\ 0.452623 & -0.671135 & 0.25674 & -0.5 & 0.169692 \\ 0.304152 & 0.365778 & -0.692549 & -0.5 & 0.209939 \\ 0.638966 & 0.25825 & 0.202658 & 0.5 & 0.483693 \end{pmatrix}$$

تم حساب تجزئة القيمة المفردة² (الشكل 1) وبعد اختيار عدد الموضوعات المراد استخلاصها، وقد افترضنا مسبقا عدد الموضوعات ليكون موضوعين اثنين، يتم تعيين العمود الأول والثاني في مصفوفة U ، وتعيين الصف الأول والثاني في مصفوفة V (الجدول 2). تُجري الآلة حسابا رياضياً آلياً يقوم على قياس كل متجهين متجاورين (القيم) في كل صف من مصفوفة U وفي كل عمود من مصفوفة V وذلك بضربها في قيم الجذور التربيعية في مصفوفة Σ ، ثم يعطي كل كلمة في كل صف متجهين (بُعدين)، ويعطي كل مستند في كل عمود متجهين أيضا، ثم يقيس مسافات الأبعاد بجيب التمام cosine distance. في المحصلة النهائية، تخرج متجهات عمودية (متجهين في سياقنا هنا) لكل ملف، ويُنظر إلى الملف الذي تحصل على رتبة أعلى للقيمتين في العمود. سيظهر بأن المستند الثالث d3 هو الأعلى، يليه المستند الأول d1. يعيب على تفريق القيمة المفردة عدم الوقوف بدقة على توزيعات الكلمات المتشابهة دلاليا بين موضوعات متعددة، فاحتمالية عدم ظهور كلمات مهمة كبيرة. كما يعيب على تجزئة القيمة المفردة احتمال عدم تخصيص كلمات موضوعية متنوعة، ويُعوّض بذلك استخدام التحليل الدلالي الاحتمالي الكامن probabilistic latent semantic analysis، غير أن تخصيص دركليه الكامن يُعدّ نسخة مطورة تطويراً

يجعل من أداء التخصيص أكثر دقة (هاي وآخرون. Hai et al.، 2009: ص 5) الأمر الذي دفع إلى تطبيقه أيضا باستخدام نموذج الأمثل في استخراج الموضوعات من بيانات القصائد الشعرية العربية³.

جدول 2

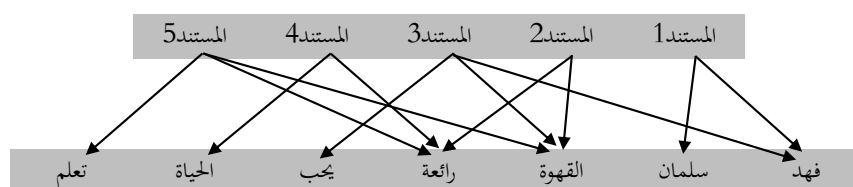
تقليص المتجهات حسب عدد الموضوعات المحددة

V					Σ					U						
d5	d4	d3	d2	d1		5	4	3	2	1		5	4	3	2	1
-0.64	-0.30	-0.45	-0.53	-0.12	1	0.00	0.00	0.00	0.00	2.42	1	0.00	0.00	0.00	0.69	-0.24
-0.26	-0.37	0.67	-0.18	0.56	2	0.00	0.00	0.00	1.79	0.00	2	0.00	0.00	0.00	0.32	-0.05
0.00	0.00	0.00	0.00	0.00	3	0.00	0.00	0.00	0.00	0.00	3	0.00	0.00	0.00	0.13	-0.67
0.00	0.00	0.00	0.00	0.00	4	0.00	0.00	0.00	0.00	0.00	4	0.00	0.00	0.00	-0.45	-0.61
0.00	0.00	0.00	0.00	0.00	5	0.00	0.00	0.00	0.00	0.00	5	0.00	0.00	0.00	0.38	-0.19
0.00	0.00	0.00	0.00	0.00	5	0.00	0.00	0.00	0.00	0.00	5	0.00	0.00	0.00	-0.20	-0.13
												0.00	0.00	0.00	-0.14	-0.26

يعمل تخصيص دركليه الكامن وفقا للجانب الحدسي على توزيع الكلمات من المستندات أو النصوص (الشكل 2). لو حددنا عدد الموضوعات المراد استخراجها ليكون موضوعين فقط (الشكل 3)، فسيعمل تخصيص المستندات الأكثر اتجاها للكلمات، ولو افترضنا تخصيصا محددًا لموضوعين فقط، فسيكون التوزيع لكل موضوع وفق أكثر الكلمات ارتباطا بكل موضوع. هذه الجانب الحدسي لهذا النموذج بسيط في فكرته، ونتائجه ليست مرضية على النصوص اللغوية الكبيرة.

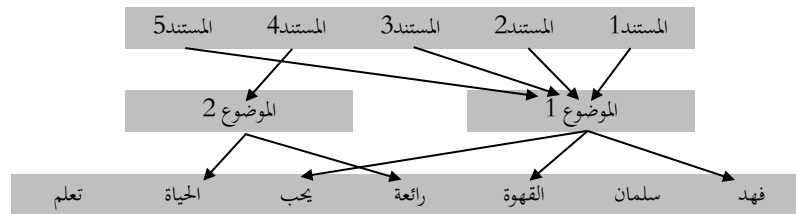
شكل 2

توزيع الكلمات من المستندات



شكل 3

توزيع الكلمات من المستندات بعد تخصيص موضوعين اثنين تخصيصاً افتراضياً



أما نموذج تخصيص دركيه الكامن (النموذج الأمثل)، فيُحوّل النصوص إلى عدة موضوعات محتملة، يتضمن كل موضوع كلمات محتملة، ويتم هذا التحويل وذلك التضمنين مرات عديدة، وبمدخلات ضبط ألفا ($\alpha: 0.1$) وبيتا (β)، والبيتا ($\beta: 0.01$) (الكلمة-الموضوع). يُجرى تخصيص دركيه الكامن بنموذجه المثالي تقييماً آلياً بقياس التماسك coherence measure، ويتم قياس التماسك وفق عدة قياسات، وأشهرها ما يعرف بـ C_v (المبحث 4)، الذي يعمل على إعادة ترشيح الكلمات وموضوعاتها. يعتمد هذا المقياس على المعلومات المتبادلة النقطية المعيارية normalized pointwise mutual information وتشابه جيب التمام cosine similarity، اللذين يحسبان درجات التماسك العالية والمنخفضة لكل عدد محدد من الموضوعات (Röder et al., 2015; Syed & Spruit, 2017). كلما كانت الدرجة أعلى من 0.55 وأقل 0.70 أنتج لنا ذلك توزيعاً متبايناً ودقيقاً للموضوعات (Yeh et al., 2016: 312). تدل الدرجة العليا للتماسك التي تقترب إلى 1.00 على أن الموضوعات المولدة أكثر عمومية في المدونة اللغوية (Jacobs & Tschötschel, 2019)، ويُلبّأ عندئذ إلى انتقاء العدد الأمثل للموضوعات، وذلك فيما يُسعى باستخلاص القيمة المثلى optimal value للتماسك coherence (بين 0.55 و0.70) التي تُخبرنا بعدد الموضوعات المثلى لأجل الدقة في استخراج الموضوعات. لو أخبرتنا بقياس التماسك بـ C_v بأن درجة التماسك لخمسة موضوعات هي 0.50، ودرجة التماسك لعشرة موضوعات هي 0.60، ودرجة التماسك لعشرين موضوعاً هي 0.90، فسيكون تخصيص عشرة موضوعات لأجل استخراجها هو الأفضل.

3. بيانات الدراسة

جُمعت القصائد العربية، ونُشرت في موقع Kaggle، الذي يتشارك فيه البيانيون والحاسوبيون والمطورون بنشر كل ما يتعلق بعلم البيانات، وقد اعتمد من جمعه وهيكله في ملف الإكسل على موقع أدب⁴. تضمن هذا الملف 58.027 سطرًا و9 أعمدة. تضمن السطر الأول البيانات الوافية للقصائد العربية لجميع العصور. نلاحظ من الشكل (4) صورةً للأحد عشر سطرًا من تلك البيانات، وفيها نرى من يسار الشكل الذي تضمن العمود الأول لرقم السطر غير المتسلسل، وتضمن العمود الثاني والثامن الرابط الشبكي للقصيدة والشاعر، وتضمن العمود الثالث مستوى لغة القصيدة (فصحى، عامية)، وتضمن العمود الرابع القصيدة كاملةً، ونجد عنوان القصيدة في العمود الخامس، ثم تصنيف القصيدة وفق العصور القديمة أو الدول في العصر الحديث. بلغ عدد الشعراء في كل العصور وفقًا لهذه البيانات ستمئة وثلاثة وخمسين شاعرًا.

شكل 4

هيكل بيانات القصائد العربية

poem_id	poem_link	poem_style	poem_text	poem_title	poet_cat	poet_id	poet_link	poet_name
21	http://www	عيناك غابتا نخ فصحي	أنشودة المطر	العراق	2	http://www	بدر شاكر السياب	
65546	http://www	انا لا ازال و في فصحي	أقداح و أحلام	العراق	2	http://www	بدر شاكر السياب	
65561	http://www	على مقلتيك ارا فصحي	هوى واحد !	العراق	2	http://www	بدر شاكر السياب	
65554	http://www	اساطير من حشم فصحي	أساطير	العراق	2	http://www	بدر شاكر السياب	
65550	http://www	والثف حولك س فصحي	اللقاء الأخير	العراق	2	http://www	بدر شاكر السياب	
65551	http://www	اطلي على طرف فصحي	أهواء	العراق	2	http://www	بدر شاكر السياب	
65557	http://www	اتبعيني فالضحي فصحي	اتبعيني	العراق	2	http://www	بدر شاكر السياب	
65558	http://www	الداء يثلج راحتي فصحي	رثة تتمزق	العراق	2	http://www	بدر شاكر السياب	
65560	http://www	سوف امضي اس فصحي	سوف أمضي	العراق	2	http://www	بدر شاكر السياب	
65548	http://www	البل والسوق الاز فصحي	في السوق القدي	العراق	2	http://www	بدر شاكر السياب	

يُمكن تطبيق عدة مناهج تحليلية لنمذجة الموضوعات لهذه البيانات، إذ يُمكن - على سبيل المثال - أن نوّلد الموضوعات بالتحليل الدلالي الكامن وتخصيص دركليه الكامن وفقاً للشعراء، أو وفقاً لكل دولة لقصائد العصر الحديث فقط، أو وفقاً للزمن. اعتمدتُ في نمذجة موضوعات هذه القصائد على التصنيف الزمني وفقاً للعصور الخمسة: (الجاهلي، والإسلامي، والعباسي، والأندلسي، والحديث)، ووفقاً لتصنيف أجريته للدول العربية في العصر الحديث بين الشرق الأوسط وشمال أفريقيا. ودول الشرق الأوسط حسب التصنيف في البيانات هي: السعودية، والكويت، والبحرين، وقطر، والإمارات، وعمان، واليمن، والأردن، والعراق، وسوريا، وعمان، وفلسطين، ولبنان، وإيران. أما دول شمال أفريقيا فهي: مصر، والسودان، وليبيا، وتونس، والجزائر، والمغرب، وموريتانيا. وفي مرحلة تهيئة البيانات للتأكد من دقتها أزلتُ أربعاً وتسعين قصيدة لكونها قصائد نبطية عامية، واثنين وعشرين سطرًا لكونها خالية من القصائد، ليكون عدد القصائد الفصيحة فعلياً 57.914 قصيدة. أيضاً نظفت البيانات من علامات الترقيم، وأزلت الكلمات الوظيفية، وحددت النغرام (التتابع اللفظي) n-gram ليكون كلمة واحدة (من الممكن تحديد النغرام ليكون كلمتين bi-grams أو ثلاث كلمات tri-grams). وعند بناء النموذج لتخصيص دركليه الكامن الأمثل حدّدت عدد مرات تكرار المعالجة لتكون 100 مرة لغرض تحسين مدخلات الضبط hyperparameter tuning. يتطلب من أجل الوصول إلى نتائج ذات تفسير عقلي إنساني مرضي أن يعمل المحلل لاختيار مدخلات الضبط للألفا والبيتا، ويكون للألفا بين 0.05 و 50 والبيتا بين 0.01 و 10، إضافة إلى اختيار v (أي: الخصائص features). لاحظت في تجارب التحليلات السابقة أن تحديد الخصائص يعتمد على حجم البيانات، والملاحظة هي الاعتماد على خمس حجم البيانات، فلو كان حجم البيانات 100 ألف كلمة جعلت الخصائص بين 15,000 و 20,000. على أي حال رأيتُ تحديد الأمثلية لعدد الموضوعات ذي الدرجة التماسكية الأعلى، وبناء عليها حددت عدد الخصائص وفق أفضل النتائج المفسرة بعد مراجعة الكلمات والموضوعات من سياقاتها في النصوص الشعرية، وصولاً إلى نتائج مرضية من تباين الموضوعات وتمايز الكلمات.

وفي الجدول (3) أعرض الإحصاءات الأولية لبيانات الدراسة للعصور الخمسة التي هيأتها للتحليل، إضافة إلى عدد القصائد للدول العربية من العصر الحديث في الشكل (5).

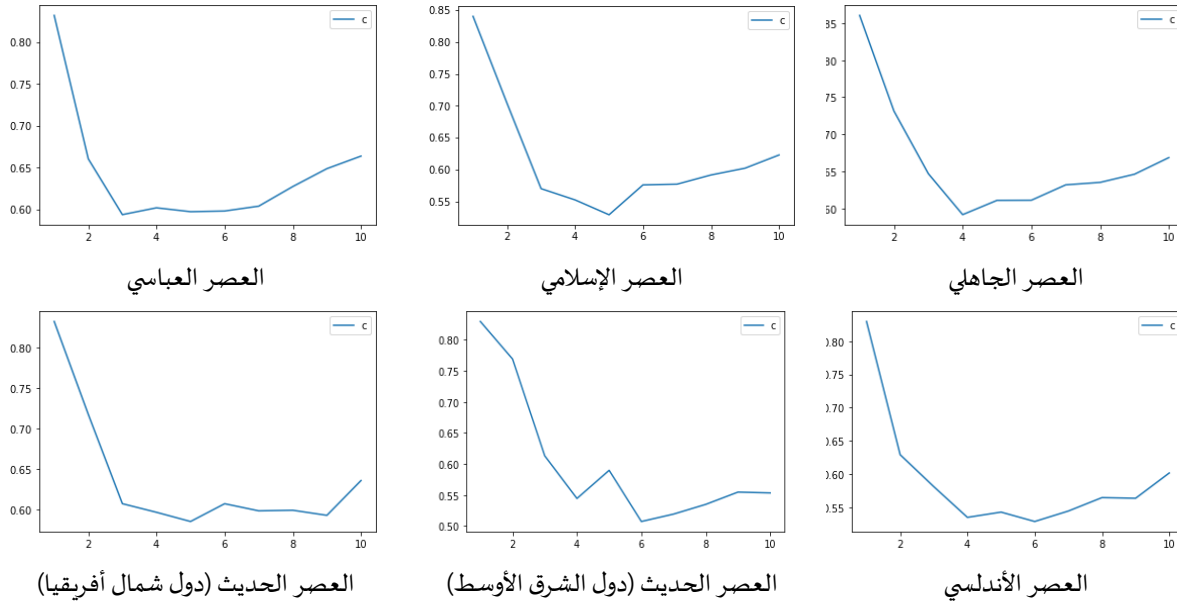
الموضوعات بدقة في النصوص العامة فحسب، بل أيضًا في النصوص أو المدونات المتخصصة، وقد أثبت ذلك في دراسة تحليلية على مدونة متخصصة تتضمن معلومات عن التغير المناخي (Murakami et al., 2017).

في تحليل موضوعات بيانات القصائد العربية، طُبِّقَت ثلاثة نماذج: نموذج للتحليل الدلالي الكامن، ونموذج غير محسن لتخصيص دركليه الكامن، ونموذج مثالي بقياس التماسك coherence score، وأُستُخدمت مكتبة GenSim واستبعدت مكتبة MALLET لأن دقتها أقل من GenSim في التجربة التي أُجريت في هذه الدراسة. مثل هذه التجربة قد أكدته أيضًا عبيد وأرانغو (Ebeid & Arango, 2016) في تقريرهما. وعلى خلاف ذلك، كشفت دراسة اللغوي بروكس ومكانري (Brookes & McEnergy, 2018) لنمذجة موضوعات آراء المرضى حول الخدمات الصحية الوطنية ببريطانيا باستخدام تخصيص دركليه الكامن بمكتبة MALLET عن كلمات غير متعلقة بالموضوعات، فقد حدّدا مسبقًا عدد الموضوعات لتكون عشرين موضوعًا، وبعد تشغيل تخصيص دركليه الكامن بهذه المكتبة أظهر النموذج عشرين موضوعًا، وعشرين كلمة لكل موضوع، ثم حلّلا هذه الكلمات تحليلًا نوعيًا، ووفقا على الكلمات المرتبطة بموضوعاتها (3 موضوعات)، والكلمات القريبة من حيث الارتباط الدلالي الموضوعي (9 موضوعات)، والكلمات المحدودة أو غير المرتبطة بموضوعاتها (8 موضوعات). وقد يكون سبب ذلك عائداً إلى تطبيقات مكتبة MALLET، على عكس GenSim الذي يتيح للمطورين مزيداً من الإضافات لأجل تحسين أداء تخصيص دركليه الكامن. وفي التجربة التي أُجريت باستخدام مكتبة MALLET فإن درجة التماسك التي حققها النموذج لعشرة موضوعات كانت أقل من 0.50، كما أن هناك كلمات قد ظهرت وليس لها علاقة دلالية بالموضوع، الأمر الذي دفعني إلى استبعاد نتائج تخصيص دركليه الكامن باستخدام هذه المكتبة.

وباستخدام مكتبة GenSim، وتخصيص دركليه الكامن الأمثل، اخترتُ لعدد الموضوعات وعدد الكلمات في كل موضوع الرقم (10)، ويشار إليه في ممارسات نمذجة الموضوع بـ k، وسبب هذا الاختيار هو أن تقليل العدد لن يفيدنا بالموضوعات المتنوعة، وزيادة العدد قد ينتج عنه تكرار ملحوظ لكلمات معينة، وهذا ما أكدته أيضًا غرين وآخرون (Greene et al., 2014). ماذا يُقدّم تخصيص دركليه الكامن بالنموذج الأمثل؟ يفيدنا بدرجة التماسك لكل عدد من الموضوعات، وهل هي جيدة أم لا. يتضمن الشكل (6) مخططات بيانات لكل عصر من العصور. يشير المحور السيني x إلى عدد الموضوعات (من 1 إلى 10 موضوعات)، ويشير المحور الصادي y إلى درجات التماسك coherence scores بين الصفر والواحد. نلاحظ في جميع المخططات في الشكل (6) أن درجة التماسك الموضوعي للموضوع الواحد تزيد على 0.80 للعصر الجاهلي والإسلامي والعباسي والأندلسي والحديث (الدول الشرق الأوسطية)، وتبلغ تقريباً 0.75 للعصر الحديث (دول شمال أفريقيا). تقلصت درجة التماسك بعد الموضوع الواحد في جميع العصور، ثم زادت عند عدد الموضوعات (10). يُنبئ الارتفاع بأن درجة التماسك لعدد الموضوعات (10) قد تزيد عندما يكون عدد الموضوعات أكبر، وهذا ما ثبت فعلاً كما في الشكل (7).

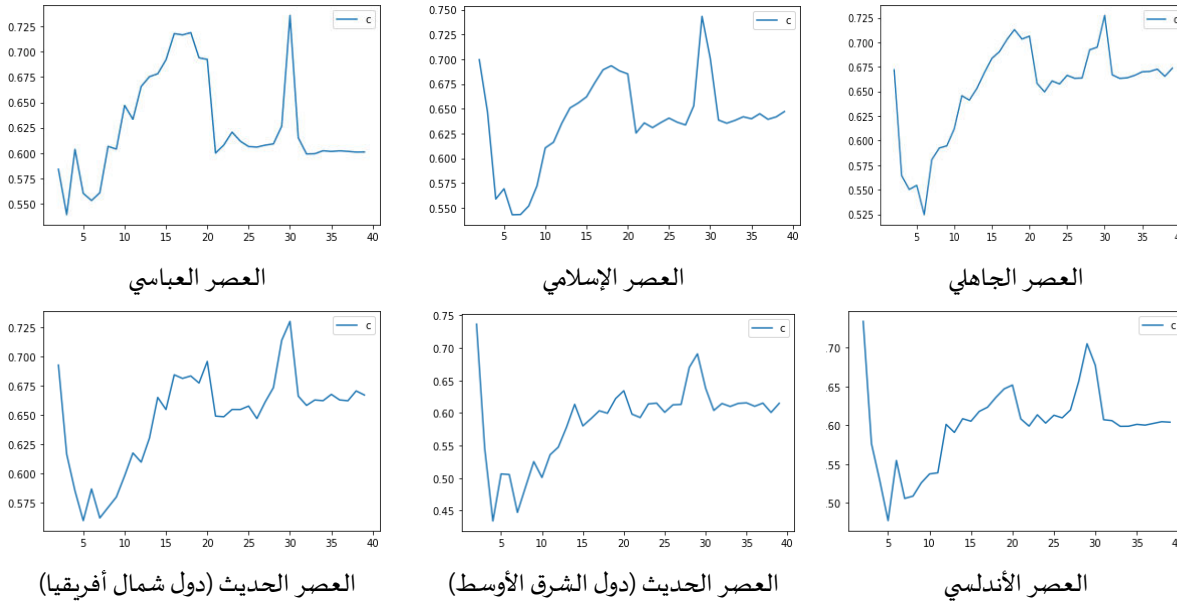
شكل 6

عدد الموضوعات: 10 (المحور السيني) ودرجات التماسك (المحور الصادي)



شكل 7

عدد الموضوعات: 40 (المحور السيني) ودرجات التماسك (المحور الصادي)



لو نظرنا إلى درجة التماسك من حيث دقة استخلاص الموضوعات وكلماتها بتخصيص عدد الموضوعات إلى 40 موضوعاً من أجل مزيدٍ من إمعان النظر، سنلاحظ (كما في الشكل 7) أن الأعداد المثلثية متباينة بين عصور النصوص الشعرية. نلاحظ أن أكثر أعداد الموضوعات تماسكاً في الشكل (7) من كل عصر على النحو الآتي: 30 موضوعاً للعصر الجاهلي، و 29 موضوعاً للعصر الإسلامي، و 18 موضوعاً للعصر العباسي، و 29 موضوعاً للعصر الأندلسي، و 29 موضوعاً للعصر الحديث (دول الشرق الأوسط)، و 30 موضوعاً للعصر الحديث (دول شمال أفريقيا). تقيّم درجات التماسك بالخطأ عند الدرجة 0.90 وأعلى؛ لأن ذلك سيُخرج لنا كلمات معدودة متكررة، وتقيم بالمقبولة عند الدرجة 0.80 وأعلى، وتُقيّم بالجيدة عند الدرجة 0.70 وأعلى. أما ما دون 0.40 فمتدنية وذات أداء غير مقبول.

أُعتمد في هذه الدراسة العدد عشرة لعدد الموضوعات وذلك لمحدودية هذه الدراسة، ودرجة التماسك بعد القيمة 0.55 تُعدّ مقبولة⁹. بعد معالجة القصائد العربية في مجموعة البيانات الستة: الجاهلية، والإسلامية، والعباسية، والأندلسية، والحديثة (دول الشرق الأوسط ودول شمال أفريقيا) جمعتُ الكلمات العشر من كل موضوع من الموضوعات العشرة لكل مجموعة من المجموعات البيانية الستة من النماذج الثلاثة: التحليل الدلالي الكامن، وتخصيص دركليه الكامن، وتخصيص دركليه الكامن (النموذج الأمثل: درجة التماسك). ولأن عدد الموضوعات قد ضُبط مسبقاً بالرقم (10)، أي: عشرة موضوعات، فعدد الكلمات لكل مجموعة من كل نموذج عندئذ هو (100) كلمة، ليصبح المجموع الكلي للكلمات (1800) كلمة، أي: 600 كلمة من كل نموذج من النماذج الثلاثة. رُصفت كل هذه الكلمات وعُرضت في ملحق هذه الدراسة، وظهرت أفضل النتائج من مخرجات تخصيص دركليه الكامن بالنموذج الأمثل، وبواقع أقل تكراراً للكلمات (انظر: الخلايا المضللة في ملحق الدراسة لنتائج تخصيص دركليه الكامن بالنموذج الأمثل).

2.4. قابلية تفسير الموضوعات

أظهرت النتائج بأن الدقة في استخراج الموضوعات من تخصيص دركليه الكامن (مدخلات ضبط ألفا وبيتا وقياس درجة التماسك) أعلى من تجزئة القيمة المفردة في التخصيص الدلالي الكامن، وأعلى أيضاً من تخصيص دركليه الكامن غير الأمثل، وهذا أيضاً يتطابق مع التجربة الرياضية لغيرلتش وآخرين (Gerlach et al., 2018)، ومع التجربة التطبيقية لنافاروا (Navarro, 2018) في مدونة أدبية إسبانية.

يجب أن ندرك أن توليد الموضوعات من النصوص الضخمة ينشأ من نموذج احتمالي. لم أصل إلى ثقة معقولة وفق التفسير العقلي، أي: التحليل النوعي لكل كلمة مولدة لكل موضوع من خوارزمية التحليل الدلالي الكامن، ومن خوارزمية تخصيص دركليه الكامن، إذ لوحظ تكرار لكلمات، ووجود كلمات لا علاقة لها بالموضوع. واستخدام النموذج الأمثل لتخصيص دركليه الكامن، واختيار 10 موضوعات ذات درجات تماسك مقبولة نتج عنه دقة عالية.

بالنظر إلى الموضوعات العشرة للتحليل الدلالي الكامن وتخصيص دركليه الكامن غير الأمثل في ملحق الدراسة نجد تكرار كلمات كثيرة، الأمر الذي دفعني إلى عدم اعتماده في التحليل الموضوعي التتابعي. فمن التحليل الدلالي الكامن نجد كثيراً من الكلمات المكررة كإله، وابن، وبني، يوم، وأيام، ويوماً، والزمان، والناس، والملك، وملك، والأرض، والهوى، والقلب... إلخ. وهناك صعوبة في التنبؤ بطبيعة كل موضوع فيما إن كان مدحاً أو فخراً أو هجاءً، أو ثناءً... إلخ. أما في قوائم الموضوعات وكلماتها بتخصيص دركليه الكامن (النموذج الأمثل) فإن ظاهرة التكرار قليلة، وظاهرة التمايز الموضوعي وفق سياقات الكلمات عالية.

حددت الموضوعات العامة لكل مجموعة من الكلمات من العصور في نتائج تخصيص دركيه الكامن بمدخلات ضبط ألفا وبيتا (النموذج الأمثل)، وبعد تتبع الـ 600 كلمة في سياق قصائدها (المجمل: عشر كلمات مضروبة في عشرة موضوعات، الإجمالي: 100 موضوع لكل عصر من العصر الستة)، فإن هناك القليل من الكلمات المكررة في كل عصر (ملحق الدراسة: تخصيص دركيه الكامن: النموذج الأمثل). تكررت كلمة (الملوك) ثلاث مرات وكلمة (العبد) مرتين في نتائج الشعر الجاهلي، وتكررت الكلمات (لابن)، و(أحمر)، و(ليلة) مرتين في نتائج الشعر الإسلامي، وتكررت كلمة (قلي) ثلاث مرات، و(عثار) مرتين، و(شرقاً) مرتين، و(الشوق) مرتين في نتائج الشعر العباسي، وتكررت كلمة (العيش) فقط مرتين في نتائج الشعر الأندلسي، وتكررت كلمتا (البعيدة) و(قلي) مرتين في نتائج الشعر الحديث لدول الشرق الأوسط، وتكررت كلمة (قلي) خمس مرات، و(الهوى) و(الشمس) و(عمري) و(أغنية) و(سيفي) مرتين في نتائج الشعر الحديث لدول شمال أفريقيا. أما من ناحية الموضوعات العامة فتميز الشعر في العصر الأندلسي بدخول موضوع (العمران)، و(الخلة)، و(الطبيعة) بوصفها موضوعات عامة في النص. أما الموضوع الشعري الجديد في العصر الحديث فنلاحظه في (الثورة) و(الوطنية).

وبالنظر إلى الكلمات الملونة بالداكن لكل نموذج من النماذج الثلاثة في ملحق الدراسة، نلاحظ عددا كبيرا من الكلمات التي تكرر ظهورها في الموضوعات باستخدام التحليل الدلالي الكامن وتخصيص دركيه، وعدداً قليلاً من الكلمات التي تكررت في الموضوعات باستخدام تخصيص دركيه الكامن (النموذج الأمثل). يُستنتج من هذه التكرارات البسيطة مقارنة بنتائج التحليل الدلالي الكامن وتخصيص دركيه الكامن غير الأمثل أفضليته من الناحية التجريبية في استخلاص الموضوعات من النصوص اللغوية العربية.

5. الخاتمة والتوصيات

عولجت البيانات الشعرية العربية الفصيحة باستخدام نمذجة الموضوع، وقد أستخدم التحليل الدلالي الكامن، وتخصيص دركيه الكامن غير الأمثل، وتخصيص دركيه الكامن الأمثل. وحددت عدد الموضوعات وعدد الكلمات في كل موضوع ليكونا عشرة، وأظهرت النتائج دقة استخراج الموضوعات بالنموذج الأخير، إذ ظهرت موضوعات متباينة كما هو حال الموضوعات الشعرية المألوفة في الأدب العربي، وارتبطت الكلمات في كل موضوع وفق النموذج الأمثل لتخصيص دركيه ارتباطاً مباشراً بالموضوعات، إضافة إلى ندرة الكلمات المكررة.

يفيد تخصيص دركيه الكامن بالنموذج الأمثل استخراج الموضوعات من النصوص اللغوية الضخمة، ويُستفاد منه في استخراج الموضوعات التي يتضمنها النص، ويتطلب عمقاً أكثر عند تحليل هذه الموضوعات، وقد أظهر دقة أكبر في ظهور عدد كبير من الكلمات الفريدة للموضوعات.

أما جملة التوصيات للمهتمين باستخدام نموذج دركيه الكامن بالنموذج الأمثل، فهي على النحو الآتي.
أولاً: لضمان الدقة في استخراج موضوعات النصوص، يتعين على المحلل أن يتنبّه إلى طبيعة النصوص من حيث تباين وتقارب الموضوعات، ومن حيث طول كل نص وقصره؛ لأن ذلك سيساعد في قرار تحديد مدخلات ضبط النموذج.
ثانياً: يعتمد تحديد عدد الموضوعات على طبيعة النص، ويُوصى باعتماد درجة التماسك الأعلى عند تحديد الموضوعات من أجل استجلاء موضوعات متباينة عديدة من مستندات النصوص اللغوية.

ثالثاً: كلما زاد عدد الموضوعات كان تتبعها في النص صعباً، فعلى سبيل المثال: استغرق مني تتبع الكلمات الستة ثلاثمائة أسابع، وبجهد فردي قد يكون عرضةً لفهم غير دقيق. ومع أهمية مثل هذه المشروعات في استخراج موضوعات وكلمات كثيرة

تنضوي بدلالاتها السياقية تحت الموضوع فإن تطبيق منهجية التقدير البيني مهم. والتقدير البيني يعتمد على المراجعة من أكثر من قارئ، يعمل لقراءة الكشافات السياقية لكل كلمة، ثم يحدّد بنفسه إن كانت هذه الكلمة تابعة للموضوع أم لا، فإن كانت

تابعة علّمها - على سبيل المثال - ب (تابعة)، وإن لم تكن كذلك علّمها ب (غير تابعة)، والحد الأدنى من عدد المقدّرين ينبغي أن يكون اثنين، ثم تُستخرج النسبة من التي تتطابق مع (تابعة) والتي تتطابق مع (غير تابعة)، ويُوصى بأن تكون نسبة التقدير البيني ثمانين بالمئة وأكثر.

رابعاً: يُطبّق تخصيص دركليه الكامن الأمثل على أي نص، كأن تكون النصوص تغريداتٍ لحدث معين، أو أن تكون تعليقات على منتج من المستهلكين في متجر إلكتروني معين، ... إلخ. لابد من العناية بتهيئة البيانات اللغوية المراد استخراج الموضوعات منها. وأما ما يُوصى به فهو استبعاد الكلمات الوظيفية قدر الممكن.

الهوامش

- 1 هناك مناهج عديدة لنمذجة الموضوع، وقد أختير التحليل الدلالي الكامن بمنهج تجزئة القيمة المفردة وتخصيص دركليه الكامن لكثرة استخداماتهما (انظر: البلوي وآخرون. Albalawi et. al، 2020).
- 2 أُجريت معالجة حساب تجزئة القيمة المفردة للمثال الافتراضي باستخدام الكود الآتي:
<https://machinelearningmastery.com/singular-value-decomposition-for-machine-learning>
- 3 تُسمى مجموعة الخوارزميات المطورة من التحليل الدلالي الاحتمالي الكامن بـ "تحليل عوامل المصفوفة غير السالبة" Non-Negative Matrix Factorization، إذ تعمل على تقليص أبعاد المتجهات في المصفوفة بالتخلص من السالب من أجل دقة أعلى في أداء نموذج استخراج الموضوعات. يُعدّ هذا الاتجاه اتجاهاً حديثاً، وما زال بحاجة إلى مزيد من التطبيقات (انظر: أبوزايد والخليفة Abuzayed & Al-Khalifa، 2021). من خلال التجربة التي أُجريت على بيانات القصائد العربية، ظهر بأن تخصيص دركليه الكامن (النموذج الأمثل) دقيق جداً، كونه متنوعاً من حيث الموضوعات.
- 4 جُمع هذه البيانات عالم البيانات فهد الحازمي، وهي بيانات مفتوحة المصدر في موقع Kaggle، ويمكن الاطلاع عليها وتصديرها من الموقع الشبكي الآتي:
<https://www.kaggle.com/fahd09/arabic-poetry-dataset-478-2017>
- 5 أما مصدر البيانات فهو من موقع أدب: <https://adab.com>
- 6 لا علم لي بمدى اشتغال موقع أدب على كل القصائد العربية الفصيحة لبعض الدول مثل موريتانيا، إذ لم يبلغ عدد القصائد منها في العصر الحديث إلا خمس عشرة قصيدة. وعليه فإن الأعداد المحصورة تُظهر ما قد تضمّنه موقع أدب.
- 6 مستودع الأكواد
 التحليل الدلالي الكامن وتخصيص دركليه الكامن والبيانات:
<https://github.com/salmujaiwel/LSA-LDA-Ar-Poetry>
- تخصيص دركليه الكامن بالنموذج الأمثل والبيانات:
<https://github.com/salmujaiwel/LDA-Coherence-Score-Ar-Poetry>
- 7 أُستخرجت الموضوعات الشعرية بناء على العصور الخمسة، ثم بناء على منطقتي الشرق الأوسط وشمال أفريقيا للعصر الخامس (العصر الحديث). وبحكم محدودية هذه الدراسة، لم يُجرَ استخراج الموضوعات بناء على الدول للقصائد العربية المعاصرة أو بناء على أسماء الشعراء، ويترك ذلك للأبحاث القادمة.
- 8 للاطلاع على قائمة الإيقاف (الكلمات الوظيفية):
<https://github.com/salmujaiwel/LDA-Coherence-Score-Ar-Poetry/blob/main/function-ar>
- 9 يُمكن إجراء تحليلات أوسع باستخدام تخصيص دركليه الكامن (النموذج الأمثل لقياس C_v)، وذلك بتخصيص أكثر عدد ممكن من الموضوعات لغرض مزيد من التحسين للنموذج.

المراجع العربية

لحج، محمد. (2020). مدخل إلى الذكاء الاصطناعي وتعلم الآلة (ط 1). أكاديمية حسوب، المملكة المتحدة.

المراجع الأجنبية

- Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, 189, 191-194.
- Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, Volume 3, Article 42, 1-14. <https://doi.org/10.3389/frai.2020.00042>
- Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.
- Blei, D., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, (pp. 71-94). Chapman & Hall/CRC Press.
- Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022.
- Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1), 3-21.
- Ebeid, I., & Arango, J. (2016). *Mallet vs GenSim: Topic Modelling for 20 news groups report*. Little Rock: University of Arkansas. <http://dx.doi.org/10.13140/RG.2.2.19179.39205/1>
- Gerlach, M., Peixoto, T. P., & Altmann, E., G. (2018). A network approach to topic models. *Science advances*, 4(7), eaaq1360.
- Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (pp. 498-513). Springer.
- Hai, N. C. T., Kim K-I., & Park, H-R. (2009). SVD-LDA: A Combined Model for Text Classification. *Journal of Information Proceeding Systems*, 5(1), 5-10.
- Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 22(5), 469-485.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.
- Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?' Using topic modelling to explore a specialised corpus. *Corpora* 12(2), 243-277.
- Navarro-Colorado, B. (2018). On Poetic Topic Modeling: Extracting Themes and Motifs from Corpus of Spanish Poetry. *Frontiers in Digital Humanities*, 5:15. doi:10.3389/fdigh.2018.00015
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, (pp. 399-408). ACM Press.
- Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In *The 4th IEEE International Conference on Data Science and Advanced Analytics. IEEE*, (pp. 165-174). doi:10.1109/DSAA.2017.61
- Van Dijk T. A. (1977a). Sentence topic and discourse topic. In: Stolz, BA (ed) *Papers in Slavic Philology 1: In Honour of James Ferrell*. Ann Arbor, (pp. 49-61). Department of Slavic Languages and Literature, University of Michigan.

- Van Dijk T. A. (1977b). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Longman.
- Van Dijk T. A. (1980). *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum.
- Yeh, J-F., Tan Y-Sh., & Lee, Ch-H. (2016). Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation. *Neurocomputing 216*: C, 310–318.

بيانات الباحث

AUTHOR BIODATA

Sultan Almujaivel is an Associate Professor of Corpus Linguistics in the Department of Arabic Language, College of Arts, King Saud University. Dr. Almujaivel received his PhD degree in Corpus Linguistics (2013) from the University of Exeter. His research interests include Corpus/Computational Linguistics.

سلطان المجبول، أستاذ لغويات المدونة الحاسوبية المشارك في قسم اللغة العربية بجامعة الملك سعود (السعودية). حاصل على درجة الدكتوراه في لغويات المدونة الحاسوبية من جامعة إكسستر عام 2013م. تدور اهتماماته البحثية حول المعالجة الآلية للغة، والبيانات اللغوية، واللسانيات الحاسوبية.

معرف أوركيد (ORCID): 0000-0001-5155-8096

Email: salmujaivel@ksu.edu.sa

1. موضوعات القصائد في العصر الجاهلي

التحليل الدلالي الكامن

الموضوع	الكلمات									
الأول	الحي	الحرب	ترى	يومًا	ابن	الدهر	إله	الناس	بني	يوم
الثاني	قول	بغافل	دينًا	عادل	قوم	جزاء	عبد	لعمري	بني	إله
الثالث	لعمري	الشرائع	أدان	والعلي	قرين	الأمر	أيام	يومًا	رميم	شئت
الرابع	مقاربًا	السراب	معشر	يميل	تدري	حمامها	ليلة	عرض	العشيرة	الحي
الخامس	الجرفين	قدامة	فظل	صنعا	ضاحية	النفس	الناس	نفعا	يومًا	المراء
السادس	عصارة	المفتل	جذع	لدي	أغتدي	ظبي	يوم	الغبيط	وبما	ترى
السابع	فارس	أول	عبله	رأيت	الكماة	تركت	الوغي	دار	فإني	عبل
الثامن	طويل	صادق	الفتي	الخير	السري	القل	الخل	فعل	طال	شاء
التاسع	لبانة	الوحش	لباتهن	محقق	حمار	تخال	بطنه	خميص	بالأذنان	ويروي
العاشر	البأس	عنا	إلقاء	يومًا	بكر	تعلم	فإذا	سر	القوم	غر

تخصیص در کلیه کامن

الموضوع	الكلمات									
الأول	عين	فإني	ترى	هاشم	غر	بني	بكر	لدى	الناس	ابن
الثاني	فارس	لدى	يومًا	الناس	خير	ابن	ولانت	بني	يوم	إله
الثالث	الأرض	القوم	ابن	الدهر	غداة	يومًا	كليب	إله	بني	يوم
الرابع	جنيًا	الصبا	غصن	يرى	صرمة	يقطرن	يقم	استأثر	طويل	بني
الخامس	عشبة	ترى	أمنًا	لدى	الحرب	عبد	الديار	الحي	الدهر	يوم
السادس	تظن	يرى	ليلة	المآكما	دار	النفوس	فقال	الناس	بيت	يوم
السابع	القوم	الفتى	أبًا	اليوم	بني	الناس	غد	الرجال	يوم	أعادل
الثامن	أدان	القلب	شئت	خير	رميم	الموت	لعمرى	الهوى	الناس	يومًا
التاسع	الحرب	عبل	الأرض	إله	الحي	الخيل	ترى	الناس	بني	يوم
العاشر	الموت	الحرب	ترى	الدهر	يومًا	ابن	الناس	يوم	إله	بني

تخصيص دركليه الكامن (النموذج الأمثل)

الموضوع	الكلمات									
الرزق	أتغصب	الهطل	لياليك	لقاء	ومقبل	جحول	مصاعب	رابت	وغله	نعمة
الحماسة	ورائي	دعوا	ونستلب	أناسي	فلوموا	مثله	وغله	زلفة	أسيافنا	يشتجر
الغزل	كبدني	ترمي	إنثما	شاغل	فؤادي	أتاني	الغرام	مستحب	واستعجمت	سواد
الحكمة	بشور	الملامة	حذاري	أبلغ	بالنور	الخمر	شاعب	الملوك	الحامي	كالطلوح
الثأر	العبد	والخيل	سلوا	أشفي	وصيري	القنا	واللوام	الصبا	كفيت	الخبيت
الحماسة	أفني	الأصفر	المطى	قومها	حسنًا	الملوك	وأعرض	والقنا	السلم	وينو

الحماسة	انتحاهها	تعوق	الدوابل	تسبي	عمرو	الملوك	حاجز	غيرها	الصبي	تحاول
المدح	وآخر	الخوف	لاقيحًا	مشروبة	مجهولة	ندوبًا	اجتنبتنه	سربال	يثني	فاعتاقه
الوصف	أطلت	زعمت	أغوي	تغوي	ضربت	ذراع	تمري	غيبته	العبد	بنصل
الرتاء	بستانه	المحزون	الموت	مقنب	نعامة	العرض	عبيدًا	منقذ	صبحًا	والشمس

2. موضوعات القصائد في العصر الإسلامي

التحليل الدلالي الكامن

الموضوع	الكلمات	الموضوع	الكلمات
الأول	القلب	أبي	أري
الثاني	شعر	الفوالك	الوجد
الثالث	وأفلت	أهل	أبي
الرابع	تلتقي	يوم	جني
الخامس	الضحى	هوى	الأصل
السادس	التغلي	الفؤاد	إله
السابع	الأم	تر	تيمم
الثامن	ماء	أطراف	مالك
التاسع	الفؤاد	ترى	سبيل
العاشر	الحي	الجود	مي

تخصيص دركليه الكامن

الموضوع	الكلمات	الموضوع	الكلمات
الأول	ابن	أبًا	شتم
الثاني	الهوى	ابن	حسن
الثالث	طمحت	نبه	جاءت
الرابع	الزمان	الدنيا	أرى
الخامس	القلب	غيث	حال
السادس	الدهر	يوم	أيها
السابع	قليل	قضاء	حادي
الثامن	إله	الناس	الأرض
التاسع	ابن	قلي	الزمان
العاشر	الإله	قلت	وجه

تخصيص دركليه الكامن (النموذج الأمثل)

الموضوع	الكلمات	الموضوع	الكلمات
المجون	كأنك	الهوى	ضعاف
الغزل	أسعرت	نفسك	بزائل
الرتاء	يقيمها	الضرب	الراقصات
الهجاء	يعرف	عنة	وأستبيح

الوصف	قناها	أحمر	حسورها	ابن	جاءهم	مديرها	ليلة	ذوده	هريرها	بالأكف
المدح	نبيل	فاسلمي	كرائمه	الملك	ليلة	أفاطم	أمامي	يفوز	وأعرض	لسانك
الحرب	قريب	اليأس	إليكم	المروءة	الحماس	صلاة	تسعى	تجزعوا	تكاد	فتحششوا
الوصف	حيثها	ثنية	فأشكرها	وأخذهم	أغيب	أسعى	ويترك	يصبحني	صعود	أنامله
لغزل	والعذر	الفتة	نحني	لملاءة	ابن	لكبد	والشوق	يدرس	حببيه	
لثأر	مساعي	تكذب	محنق	مق	بواحد	وقا	شاحبًا	بد	مصفر	مضيه

3. موضوعات القصائد في العصر العباسي

التحليل الدلالي الكامن

الكلمات

الأول	ابن	لدني	الأرض	الزما	تري	الهوى	ماس	مدهر	يوم	إله
الثاني	غله	خا	تغير	أخا	عين	لسوء	عمري	لشوكي	شق	ابن
الثالث	عرف	ملك	خير	تكونو	وكم	لشعر	مدهر	منكم	ماس	إله
الرابع	الهوى	ماس	لغزال	حقه	فلست	ماءه	إله	ألده	لشباب	الزما
الخامس	تطفيفه	ملك	ناهيك	لمشي	وقف	الهوى	وترى	الوزير	وحق	الزما
سادس	مال	يغلب	يد	لدني	علا	المغرب	ارتجعت	تخضب	يوم	شرف
سابع	الأعداء	الإله	مه	علا	تري	إله	ماس	مكا	لموك	أراك
الثامن	بابي	لكمال	الموالي	ضل	فضل	ملك	قلت	نوال	رجائي	بغير
تاسع	ليلى	وجه	بد	لفقير	تجير	لذة	يل	الحمر	قلبي	لدين
ماشر	قدس	الهوى	الظي	الأرض	كلم	الزما	المهم	إله	ملك	لدين

تخصيص دركليه الكامن

الكلمات

الأول	ابن	أبًا	شنة	باق	الأكف	الأرض	تأتي	الأيام	وكم	لفراق
الثاني	الهوى	ابن	حسن	مدهر	وكم	تري	الزما	يل	يوم	إله
الثالث	طمحت	نبه	جاءت	وجه	عاد	إلا	عين	دار	الم	عيب
الرابع	الزما	لدني	أرى	تري	الهوى	الأرض	يوم	ماس	مدهر	إله
الخامس		غيث	حال	لدني	خرقًا	الأرض	الراح	إله	مدهر	تراه
سادس	مدهر	يوم	أيها	يرى	شق	ماس	ملك	الهوى	قلبي	إله
سابع	قليل	قضاء	حادى	حبه	قلبًا	كل	تري	نفسى	ذكر	إله
الثامن	خير	رب	مدهر	عبد	منكم	لشعر	وكم	الأرض	ماس	إله
تاسع	ابن	قلبي	الزما	الأرض	الهوى	أبي	ماس	يوم	ماس	إله
ماشر	الإله	قلت	وجه	يرى	سوي	الوجود	أهل	يوم	عين	إله

تخصيص دركليه الكامن (النموذج الأمثل)

الكلمات

لغزل	بالعيش	حسد	قلبي	وخن	أيام	أعرف	لبس	ويخل	فيذا
------	--------	-----	------	-----	------	------	-----	------	------

اللهم	يدين	احتمالها	والمقلتين	نصارى	فضوح	السواري	دمتا	وبالشكر	ولهوا	الرزايا
الولاء	الشوق	وفاء	لبست	بالعدل	المشيب	مناهجهم	حميداً	النجح	أراهم	يزيدا
الوجد	أكيار	قصور	ليلة	ديارهم	لتلك	زفرات	حملت	عقار	شموس	حسناً
مدح الذات	المعجم	بقلب	تبلغ	بقلبي	يهوى	العيش	بالأنجم	فقلبي	بالنيزك	شرقاً
الشوق	ألما	افتضاحي	الصبح	الرداح	جراح	الخود	شرقاً	البيض	الوشاح	وثغر
مدح الحبيبة	ولأشكو	هواكم	بالسحر	ويلاه	العقد	قلبي	الجوانح	انظر	حماكم	غريز
المدح	مستحسن	بكنبوش	عثار	الملوك	المسجد	وسطيح	عرتي	يجلوك	بجوده	الجيش
الوجد	الأطيار	أودى	لهفي	أمتع	القمري	مثله	أظنت	عابساً	بالقمري	عثار
الندم	والعقاب	قلبي	فوجدت	عالم	الشوق	إليه	النقاخا	فياخا	جلاله	النيق

4. موضوعات القصائد في العصر الأندلسي

التحليل الدلالي الكامن

الموضوع	الكلمات
الأول	الناس ملك الملك الزمان ابن الهوى الدهر يوم إله
الثاني	ثغر برد النجوم زلت الخلافة بدا الجمال إله الحسن
الثالث	الحب الهوى حالياً بقلبي الجو جاء غدا القلب إله
الرابع	دين خير الدين جاء حاضر قريب الهدى قلبي إله
الخامس	رعا أعبا الأرض الدنيا رد يحيي العليا ابن حياً
السادس	ملك خير شاء أحد حي شيء سماء فإذا يوم
السابع	القلوب أسد بدور العلا الدمع يوم ماء مولاي الزمان الهوى
الثامن	الدهج زلت الدر لخير الدنيا المجد العلا الملك إله
التاسع	حرف البدر البدر خير صدر فعلي بدر الوري قلبي
العاشر	سقيم الجبار الديار الرسول حق وعلى نسب أهل بني

تخصيص دركليه الكامن

الموضوع	الكلمات
الأول	لهلال طبعت مشحنة أمهاتها أودنت شكرًا لطرده عينيك إمائاً أول
الثاني	الناس وكم الأرض الزمان الملك ابن الهوى يوم الدهر إله
الثالث	الندي داعي الهدى ظهور الإله التقى النبي رأيت عامر إله
الرابع	الهوى الحسان القلوب ملوكاً عيون در إله رأيت العقول زمان
الخامس	الحب بني وإله ليل عين قلت سوى بيدق قلبي الفرزان
السادس	الفتى الجد المداد رضوى شاد ألهب الثنا ابن ماء الزمان
السابع	الأرض نفسي الصبا الفرد العبد ما زلت الأول ملك إله الزمان
الثامن	زوراً بحر شكر يك رقة بعثت أكلت القلب نسبت ابن
التاسع	المعالي الدهر ابن الوري الملك الزمان الهوى قلبي يوم إله
العاشر	أطاع وأنجاد رضا أناس ابن أنتم شرف محتاج إله الزمان

تخصيص دركليه الكامن (النموذج الأمثل)

الموضوع	الكلمات									
الرجاء	المرتجى	بجباله	أبدى	المسلمين	الطويل	جوازي	تحيا	علقت	إحسانها	ماسورة
العمران	عادت	ملاءة	الموفق	وعينه	وقفت	صديق	الراح	قميص	أيكة	بيتين
الشوق	العادة	الحلي	نصري	حسنًا	العيش	زارها	المحب	فطرك	عهدي	الحبيب
الوجد	تحمل	وقلبي	فظلمك	باهوى	الهوى	أقول	الشميل	العقيق	الرطب	صداء
الخلة	الجميل	ويذم	الأيام	الدهر	بتبعد	يعتب	صديقًا	نصيحتي	أملأكها	بالزجاج
الوجد	حبكم	فصيتي	نظيره	زفرات	بكاك	بلغت	ذكرت	الصدور	خيامي	لقلبي
الذكرى	راحوا	رقيقًا	المستطيع	طلوع	بنجيع	صاحبه	نسبوا	إليه	الأدمع	أتاني
الطبيعة	مدام	بالعدل	بمحجة	تسبي	منسوبة	واله	طلاب	والجار	يتفرق	النجم
الغزل	الحاليتين	نفني	يصغي	يعرفه	يميز	العيش	شعري	سروًا	باللقاء	أجفانه
المجون	يحمل	مقر	الوشاح	وشاهد	تدخل	قلبي	الحساد	الردف	والكأس	راحته

5. موضوعات القصائد في العصر الحديث (دول الشرق الأوسط)

التحليل الدلالي الكامن

الموضوع	الكلمات									
الأول	أشعر	النفس	الحق	قلبي	بلا	يوم	خير	الليل	عبد	إله
الثاني	خلف	الريح	أين	شيء	بلا	وأنا	قلبي	الأرض	الحب	الليل
الثالث	سرير	ليلة	حلم	نون	أغنية	نونك	الحلم	حب	والموعد	قبلة
الرابع	ورهيف	وأثف	الرواة	موجز	قيطون	امرأة	أحد	أيشبهني	ولنقل	أحسد
الخامس	أهل	الناس	الهدى	حق	الفضل	الدهر	الدنيا	العلم	إله	مصر
السادس	الصحراء	أسمع	الطرق	أزمنة	تلوها	الصباح	ألمح	أيام	الجبال	العالم
السابع	أبي	شعر	فلسطين	إله	يعد	تخرجون	أطلع	لأن	نأتي	باقون
الثامن	الرمل	صندوق	مستوي	الأزرق	يدي	الحب	البحري	الناس	الماء	البحر
التاسع	بمستوى	ونغت	صديقتي	لديك	جواهر	رميت	أساور	عيوني	وبعت	جنوبي
العاشر	البحر	أغنيتي	الحلم	العربي	الحصار	أيها	تقول	دمي	اذهب	تك

تخصيص دركليه الكامن

الموضوع	الكلمات									
الأول	السيدة	الحنين	سوى	سيدي	أخذت	أوريانتيا	الفراغ	الطالع	نقول	تحسدوا
الثاني	الماء	بلا	عد	مرة	الطريق	ابن	سوى	أيها	الزمان	إله
الثالث	آخر	شاهدت	الآن	بلا	العالم	واحدة	الليل	وأنا	أيها	شيء
الرابع	بلا	قلت	الدنيا	الغني	أدري	نفسى	حق	إله	قريبًا	بعيدًا
الخامس	أثمرت	ورب	وقام	نخلاتها	الأمر	أوصال	نجاحها	عاجلة	يومًا	عسى
السادس	المجد	دار	نار	شاء	كسروان	الزمان	وغله	خير	أيها	نفسى
السابع	أيام	نجا	ترى	ليل	البحر	غله	لزائمًا	الماء	الموجة	أيتها

الثامن	بلا	الرياح	القلب	البحر	يوم	قلبي	الحب	الأرض	الليل	إله
التاسع	قلبي	الدنيا	ابن	الهوى	الحب	أهلاً	مصر	الليل	والموعد	القبلة
العاشر	وأنا	سلام	صغارا	فاطمة	القطار	ألف	غودو	سواسبول	العالم	منتظر

تخصيص دركليه الكامن (النموذج الأمثل)

الموضوع	الكلمات
الحنين	متسولاً قطارات وتوقظني قديم يدعي الأخرى قدام تسكنه سيمنحي قلبي
الغربة	الظلام وحدهم تعود أرقّت الهواء راسمو حلبات هواجسهم ملوك قلبي
التجربة	السعادة أمامي الصباح وآخر يسوق مصيره فطوبى طريدة غصناً لمسعى
الوجود	مقوده زهولاً تكاد وربانه والشرع خطاها المدى لفظه الضلوع وطناً
اللهو	الخفريات النجوم رعوية قامة فتألقت الغناء الرقصات أناشدك ذكراك سلمى
الوجد	حزني خفقانه وسرابي الفؤاد ينسل أطيب سبيله وجراحك مفترس مثنوك
السكنى	توجه البيت كأنك كوني ليبدأ حولك تحيا والبيت يعلمك محكوم
الغزل	يذوي أقمت أيسر سبيلك بغرامي ودعية فقال أتاك للمليك إليك
الغزل	دعاك سورية الشفاة الشوق فأوشك الجميل ندفع المحفوظ وتصفي لميدانه
الوطنية	القصد عيداً تنساكم نبضة نجمة ومرحوماً الصبر عوضاً وطني أنساك

6. موضوعات القصائد في العصر الحديث (دول شمال أفريقيا)

التحليل الدلالي الكامن

الموضوع	الكلمات
الأول	رأى الثرى السهل الجيش يوم الماء الملك الترك إله الأرض
الثاني	لؤلؤة القلب مصر ناظم صمناً الصوص صديق مثلما نم حبيبي
الثالث	النوى الليل ألقاك شيء زمان الحياة حبيبي يوماً الهوى قلبي
الرابع	بالعشق وحيداً البلاد صمت بلادي الحياة علم السماء هرميس الآن
الخامس	الدهر الحق الأيام الأمر المرء المال يجد النفس الناس إله
السادس	لأنكم الآن عنكم الأرض مثلما الطريق البحر الشعر الناس الحب
السابع	جميعاً ثوروا حسن القدس شكرًا أكبر الحكام الجبل إله أيها
الثامن	برج الشعراء الصغيرة الليل المدى أغنية وجه بامرأة قادرة أحلم
التاسع	القصر المر وقع المدينة الطرقات الليل قصيدة الهوى الصباح قلت
العاشر	الصبا الشيب ومنيتي والشيب الثرى القنا ألوي هوى الرياح الشباب

تخصيص دركليه الكامن

الموضوع	الكلمات
الأول	يجد بلادي يوماً الآن فؤادي الأرض الطريق الشعر قلبي الحب
الثاني	الحارس الحب الزمان البحر أسير صرت تصطف ليلي الناس ألقاك
الثالث	السماء المدينة ليل العين الطرقات مضى هرميس الأرض قصيدة قلت
الرابع	الأيام تجمعت النوى جميعاً شكرًا أصبنا الماء زمان الجبل أيها

الخامس	ثرنا	الجنون	نكون	باله	ثوروا	أسلافكم	حكامكم	العربي	يوم	الحكام
السادس	وجه	الأخضر	البحر	الحوت	الشعراء	القلب	برج	بامرأة	قادرة	أحلم
السابع	الحق	البحر	الدنيا	داخلي	الهوى	الناس	حببتي	يوماً	قلبي	إله
الثامن	سوى	الحياة	الشعر	القلب	أكبر	الأرض	الحق	يوم	مصر	إله
التاسع	إله	مصر	القلب	نم	الترك	الأرض	الشمس	الهوى	الليل	حببي
العاشر	الظبي	حسناً	القدس	الهوى	الزمان	الأرض	إله	أهم	ملك	الدين

تخصيص دركليه الكامن (النموذج الأمثل)

الموضوع	الكلمات									
الرتاء	الطري	فرحتي	والشمس	نشور	بلاداً	سيفي	جبيه	الرحيل	الروض	موتي
التشاؤم	إليه	ودار	يتبع	قلبي	سبيل	معشر	عاده	ظلامه	فتجنح	قلصا
الشوق	مكان	وتوقظ	حلوه	ستأخذني	الرسالة	قلبي	الزل	محكم	الهوى	الضلوع
الثورة	الشوارع	أطلقت	خلفي	ليلة	الشمس	الدرب	أتيت	والظلام	أواخر	انكشف
الحنين	يعيش	سعوداً	يرضى	المنحنى	أغنية	الدنيا	النافذة	الإقامة	الضنى	وطغى
الحزن	رغبة	اليمني	مسبحتي	أحكمت	الدار	عمري	لرحيل	مجروراً	تنهري	سيفي
الغزل	الأشواق	أسعى	الأيام	أغنية	عمري	قلبي	الركب	لعمري	كفتني	مطالبي
الغزل	أحلم	أنساك	سبيله	وناجاه	الجميل	قلبي	جنبي	العيش	يتضرم	والهوى
الوطنية	شرقاً	تسوح	الهوى	قلبي	الشمس	وطني	قصيدة	روحي	الغبوق	شوقي
الوجد	والتفجع	المفيد	لتلك	السعود	ونقرته	ترجي	باحث	الورقاء	نقرة	يطاع