



مراجعة

## خوارزميات تصنيف النصوص: دراسة استقصائية

1

كينا جعفري ميماندي<sup>1</sup>  
لورا بارنز<sup>1,2,3</sup> ودونالد براون ID 1,3

سانجانا ميندو<sup>1</sup>

قسم هندسة النظم والمعلومات، جامعة فرجينيا، شارلوتسفيل، فرجينيا، الولايات المتحدة الأمريكية؛<sup>1</sup> 22904  
kjm6vd@virginia.edu (KJM); mh4pk@virginia.edu (MH); lb3dp@virginia.edu (SM)

2

مختبر أنظمة الاستشعار للصحة، جامعة فرجينيا، شارلوتسفيل، فرجينيا، الولايات المتحدة الأمريكية<sup>2</sup>  
كلية علوم البيانات، جامعة فرجينيا، شارلوتسفيل، فرجينيا، الولايات المتحدة الأمريكية \* للتواصل:  
+١-٢٠٢-٨١٢-٣٠١٣ kk7nc@virginia.edu

تاريخ الاستلام: ٢٢ مارس ٢٠١٩ تاريخ القبول: ١٧ أبريل ٢٠١٩ تاريخ النشر: ٢٣ أبريل ٢٠١٩



ملخص: شهدت السنوات الأخيرة نمواً هائلاً في عدد الوثائق والنصوص المعقّدة التي تتطلّب فهّماً أعمق لأساليب التعلم الآلي لتصنيف النصوص بدقة في العديد من التطبيقات. وقد حققت العديد من مناهج التعلم الآلي نتائج باهرة في معالجة اللغات الطبيعية. ويعتمد نجاح هذه الخوارزميات على قدرتها على فهم النماذج المعقّدة والعلاقات غير الخطية داخل البيانات. ومع ذلك، يمثّل إيجاد البني والهيكل والتقنيات المناسبة لتصنيف النصوص تحدياً للباحثين. تتناول هذه الورقة نظرة عامة موجزة على خوارزميات تصنيف النصوص، تشمل استخدام خصائص النصوص المختلفة، وأساليب تقليل الأبعاد، والخوارزميات والتقنيات الحالية، وأساليب التقييم. وأخيراً، تناقش قيود كل تقنية وتطبيقاتها في مشاكل العالم الحقيقي.

الكلمات المفتاحية: تصنیف النصوص؛ استخراج البيانات النصية؛ تمثیل النصوص؛ تصنیف النصوص؛ تحلیل النصوص؛ تصنیف الوثائق

### 1. مقدمة.

حطّيت مشاكل تصنیف النصوص باهتمام واسع النطاق، وتم تناولها في العديد من التطبيقات العملية [1-8] على مدى العقود القليلة الماضية. ومع التطورات الحديثة في معالجة اللغات الطبيعية (NLP) واستخراج البيانات النصية، يهتم العديد من الباحثين بتطوير تطبيقات تستفيد من أساليب تصنیف النصوص. يمكن تقسيم معظم أنظمة تصنیف النصوص وتصنیف المستندات إلى المراحل الأربع التالية: استخراج الميزات، وتقليل الأبعاد، واختيار المصنف، والتقييم. في هذه الورقة، ناقش بنية أنظمة تصنیف النصوص وتطبيقاتها التقنية من خلال مسار العمل الموضح في الشكل 1 (يتوفر الكود المصدري والنتائج كأدوات مجانية على الرابط [https://github.com/kk7nc/Text\\_Classification](https://github.com/kk7nc/Text_Classification)).

تكون مدخلات خط المعالجة الأولية من مجموعة بيانات نصية خام، بشكل عام، تحتوي مجموعات البيانات النصية على تسلسلات نصية في مستندات على النحو التالي:  $X_1, X_2, \dots, X_N = D$  حيث يشير  $X_i$  إلى نقطة بيانات (أي مستند، مقطع نصي) تحتوي على جملة، بحيث تتضمن كل جملة  $w$  الكلمة مكونة من  $w$  حرفًا.

يتم تصنیف كل نقطة رقمية فئة من مجموعة من مؤشرات القيم المنفصلة المختلفة [7].

بعد ذلك، ينبغي إنشاء مجموعة بيانات منظمة لأغراض التدريب، ويُطلق على هذا القسم اسم "استخلاص الميزات". تُعد خطوة تقليل الأبعاد جزءاً اختيارياً من مسار المعالجة، ويمكن أن تكون جزءاً من نظام التصنیف (على سبيل المثال، إذا استخدمنا تردد المصطلح - تردد المستند العکسی (TF-IDF) لاستخلاص الميزات، وكان لدينا 200 ألف كلمة فريدة في مجموعة التدريب، فإن وقت الحساب سيكون مرتفعاً للغاية، لذا يمكننا تقليل هذا الخيار عن طريق نقل مساحة الميزات إلى فضاء ذي أبعاد أخرى). تُعد الخطوة الأهم في تصنیف المستندات هي اختيار أفضل خوارزمية تصنیف. أما الجزء الآخر من مسار المعالجة فهو خطوة التقييم، والتي تقسّم إلى جزأين (التنبؤ بمجموعة الاختبار و...).

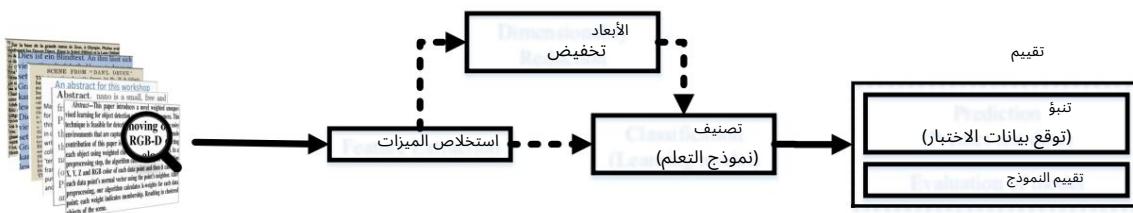
(تقييم النموذج). بشكل عام، يحتوي نظام تصنيف النصوص على أربعة مستويات مختلفة من النطاق يمكن تطبيقها:

1. **مستوى المستند:** على مستوى المستند، تحصل الخوارزمية على الفئات ذات الصلة من كامل وثيقة.

2. **مستوى الفقرة:** في مستوى الفقرة، تحصل الخوارزمية على الفئات ذات الصلة لفقرة واحدة (جزء من المستند).

3. **مستوى الجملة:** على مستوى الجملة، يتم الحصول على الفئات ذات الصلة لجملة واحدة (جزء منها). (فقرة).

4. **مستوى الجملة الفرعية:** في مستوى الجملة الفرعية، تحصل الخوارزمية على الفئات ذات الصلة من التعبيرات الفرعية داخل الجملة (جزء من الجملة).



الشكل 1. نظرة عامة على مسار تصنيف النصوص.

(ا) استخلاص الميزات: تُعتبر النصوص والوثائق عمومًا مجموعات بيانات غير مهيكلة. مع ذلك، يجب تحويل هذه التسلسلات النصية غير المهيكلة إلى فضاء ميزات مهيكل عند استخدام النمذجة الرياضية كجزء من المصنّف. أولًا، يجب تنظيف البيانات لحذف الأحرف والكلمات غير الضرورية. بعد تنظيف البيانات، يمكن تطبيق أساليب استخلاص الميزات الرسمية. من التقنيات الشائعة لاستخلاص الميزات: تردد المصنّد العكسي، TF-IDF [10] وTF-IDF المصطلح، Word2Vec [9] و[11] والمتجهات العالمية لتمثيل الكلمات. في القسم 2، تُصنّف هذه الأساليب إلى تقنيات تضمين الكلمات وتقنيات الكلمات الموزونة، ونُناقش تفاصيل التنفيذ التقني.

(II) تقليل الأبعاد: نظرًا لاحتواء مجموعات بيانات النصوص أو المستندات غالباً على العديد من الكلمات الفريدة، قد تتأخر خطوات المعالجة المسبقة للبيانات بسبب ارتفاع تقييد الوقت والذاكرة. يتمثل الحل الشائع لهذه المشكلة في استخدام خوارزميات منخفضة التكلفة. مع ذلك، في بعض مجموعات البيانات، لا تتحقق هذه الخوارزميات منخفضة التكلفة الأداء المتوقع. لتجنب انخفاض الأداء، قُفضل العديد من الباحثين استخدام تقليل الأبعاد لتقليل تقييد الوقت والذاكرة لتطبيقاتهم. قد يكون استخدام تقليل الأبعاد للمعالجة المسبقة أكثر كفاءة من تطوير مصنفات منخفضة التكلفة.

في القسم 3، نستعرض أكثر تقنيات تقليل الأبعاد شيوعًا، بما في ذلك تحليل المكونات الرئيسية (PCA)، وتحليل التمييز الخطى (LDA)، وتحليل المصفوفات غير السالبة (NMF)، كما نناقش تقنيات جديدة لتقليل أبعاد استخلاص الميزات غير الخاضع للإشراف، مثل الإسقاط العشوائي، والمشفرات التلقائية، وتضمين الجوار العشوائي الموزع (t-SNE).

t

(ثالثًا) تقنيات التصنيف: تُعد اختبار أفضل مُصنّف أهم خطوة في مسار تصنيف النصوص. في بدون فهم مفاهيمي كامل لكل خوارزمية، لا يمكننا تحديد النموذج الأكثر كفاءة لتطبيق تصنيف النصوص. في القسم 4، نناقش أكثر تقنيات تصنيف النصوص شيوعًا. أولاً، نتناول الطرق التقليدية لتصنيف النصوص، مثل تصنّيف روكيو. ثم نتحدث عن تقنيات التعلم القائمة على التجميع، مثل التعزيز والتجميع، والتي استُخدمت بشكل أساسى في استراتيجيات تعلم الاستعلامات وتحليل النصوص [12-14]. ثُمَّ نُعد الانحدار اللوجستي (LR) من أبسط خوارزميات التصنيف، وقد استُخدمت في معظم مجالات استخراج البيانات [15-18]. في بدايات تاريخ استرجاع المعلومات،

نظاماً لكنه تطبيقاً عملياً، حظى مصنف بايز الساذج (NBC) بشعبية واسعة. نقدم هنا لمحة موجزة عن مصنف بايز الساذج، الذي يتميز بانخفاض تكلفته الحسابية واحتياجه إلى ذاكرة قليلة جداً.<sup>[19]</sup>

تمت دراسة التقنيات غير البارامترية واستخدامها في مهام التصنيف، مثل خوارزمية أقرب جار [20]، تقنية SVM [21، 22]، آلية المتجهات الداعمة [KNN] [23]، وتقنية شائعة أخرى تستخدم مصنفًا تمييزياً لتصنيف المستندات. يمكن استخدام هذه التقنية أيضًا في جميع مجالات استخراج البيانات، مثل المعلوماتية الحيوية، والصور، والفيديوهات، وتصنيف الأنشطة البشرية، والسلامة والأمن، وغيرها. كما يستخدم هذا التموزج كأساس للعديد من الباحثين للمقارنة مع أعمالهم الخاصة، بهدف إثارة الجوانب الجديدة والمساهمات.

فيما يتعلق بتصنيف المستندات [23]سيتم تناول كل خوارزمية قائمة على الأشجار في قسم فرعي منفصل. في السنوات الأخيرة، تم اعتبار التصنيفات الرسمية [24] مهمة تصنيف مثل الحقول العشوائية الشرطية (CRFs) ومع ذلك، تُستخدم هذه التقنيات في الغالب لتلخيص المستندات [25] واستخراج الكلمات المفتاحية تلقائياً [26].

في الآونة الأخيرة، حققت أساليب التعلم العميق نتائج متفوقة مقارنة بخوارزميات التعلم الآلي السابقة في مهام مثل تصنیف الصور ومعالجة اللغة الطبيعية والتعرف على الوجوه وما إلى ذلك. يعتمد نجاح خوارزميات التعلم العميق هذه على قدرتها على نمذجة العلاقات المعقدة وغير الخطية داخل البيانات.<sup>[27]</sup>

(رابعاً) التقييم: يُعد التقييم المرحلة الأخيرة من عملية تصنيف النصوص. ويُعد فهم أداء النموذج أمراً أساسياً لاستخدام وتطوير أساليب تصنيف النصوص. توفر العديد من الطرق لتقدير التقنيات الخاصة بالإشراف. يُعد حساب الدقة أبسط طرق التقييم، ولكنه لا يُجدي نفعاً مع مجموعات البيانات غير المتوازنة.<sup>[28]</sup> في القسم 5، يُبين طرق التقييم التالية لخوارزميات ROC (AUC).

في القسم ،نتحدث عن القيود والعيوب للطرق المذكورة أعلاه.

نقارن بإيجاز خطوات مسار المعالجة، بما في ذلك استخلاص الميزات، وتقليل الأبعاد، وتقنيات التصنيف، وأساليب التقييم، ونقارن في هذا القسم أحدث التقنيات وفقاً لمعايير متعددة، مثل تبني النموذج، وجدة العمل، وتقنية استخلاص البيانات المستخدمة، ومقياس التحقق، وقيود كل تقنية. يتطلب إيجاد النظام الأمثل لتطبيق ما اختيار طريقة استخلاص الميزات المناسبة. ويعتمد هذا الاختيار كلياً على هدف التطبيق ومجموعة بياناته، إذ أن بعض تقنيات استخلاص الميزات غير فعالة لتطبيقات محددة. على سبيل المثال، نظراً لأن GloVe مدرب على ويكيبيديا، وعند استخدامه مع الرسائل النصية القصيرة (SMS) لا يؤدي هذا النموذج أداءً جيداً مثل TF-IDF، بالإضافة إلى ذلك، فإن محدودية نقاط البيانات تجعل من الصعب تدريب هذا النموذج بكفاءة التقنيات الأخرى. أما الخطوة التالية في مسار المعالجة، فهي تقنية التصنيف، حيث نتناول بإيجاز قيود وعيوب كل تقنية.

في القسم 7، نصف تطبيقات تصنيف النصوص والوثائق. يُعدّ تصنيف النصوص تحدياً رئيسياً للباحثين في العديد من المجالات. تستخدم أنظمة استرجاع المعلومات [33] ومحركات البحث [34, 35] أساليب تصنيف النصوص بشكل شائع، وانطلاقاً من هذه التطبيقات، يمكن استخدام تصنيف النصوص في تطبيقات أخرى مثل تصفية المعلومات (مثل تصفية الرسائل الإلكترونية والرسائل النصية غير المرغوب فيها). [36] بعد ذلك، تناول اعتماد تصنيف الوثائق في مجال الصحة العامة [37] ودراسة السلوك البشري [38] ومن المجالات الأخرى التي استفادت من تصنيف النصوص تنظيم الوثائق وإدارة المعرفة. أخيراً، سنتناول أنظمة التوصية التي تُستخدم على نطاق واسع في التسويق والإعلان.

## ٢. معالجة النصوص المسبقة

يُعد استخلاص الميزات والمعالجة المسبقة خطوتين أساسيتين لتطبيقات تصنيف النصوص. في هذا القسم، نقدم طرفاً لتنظيف مجموعات بيانات النصوص، وبالتالي إزالة التشويش الضمني والسامح

للحصول على معلومات مفيدة. علاوة على ذلك، نناقش طريقتين شائعتين لاستخراج ميزات النصوص: تقنيات الكلمات الموزونة وتقنيات تضمين الكلمات.

#### 2. تنظيف النصوص ومعالجتها المسيبة

تحتوي معظممجموعات بيانات النصوص والوثائق على العديد من الكلمات غير الضرورية، مثل الكلمات الشائعة، والأخطاء الإملائية، واللغة العامية، وما إلى ذلك. في العديد من الخوارزميات، وخاصة خوارزميات التعلم الإحصائي والاحتمالي، يمكن أن تؤثر الضوضاء والخصائص غير الضرورية سلبيًا على أداء النظام. في هذا القسم، نشرح بإيجاز بعض التقنيات والأساليب لتنظيف النصوص ومعالجةمجموعات بيانات النصوص مسبيًا.

##### 2.1.1. التجزئة

التجزئة هي طريقة معالجة مسيبة تقسم سلسلة النصوص إلى كلمات أو عبارات أو رموز أو عناصر أخرى ذات دلالة تُسمى الرموز. [39, 40] الهدف الرئيسي من هذه الخطوة هو دراسة الكلمات في الجملة. [40] يتطلب كل من تصنيف النصوص واستخراج المعلومات منها محلًا نحوًابالعاجل تجزئة النصوص. على سبيل المثال: [41]:

{ بعد أن نام أربع ساعات، قرر أن ينام أربع ساعات أخرى.

في هذه الحالة، تكون الرموز كما يلي:

{ بعد أن نام أربع ساعات، قرر أن ينام أربع ساعات أخرى.

##### 2.1.2. الكلمات المحظوظة

يشمل تصنيف النصوص والوثائق العديد من الكلمات التي لا تحمل دلالة مهمة لاستخدامها في خوارزميات التصنيف، مثل "...، "again" ، "afterwards" ، "about" ، "above" ، "across" ، "after" ، "a" ، "about" ، "above" ، "across" ، "after" ، [42] وتمثل التقنية الأكثر شيوعاً للتعامل مع هذه الكلمات في إزالتها من النصوص والوثائق.

##### 2.1.3. استخدام الأحرف الكبيرة

تنوع أنماط كتابة الأحرف الكبيرة والصغيرة في النصوص والوثائق لتكون الجملة. ونظرًا لأن الوثائق تتكون من جمل عديدة، فإن هذا التنوع في استخدام الأحرف الكبيرة والصغيرة قد يشكل مشكلة كبيرة عند تصنيف الوثائق الكبيرة. والطريقة الأكثر شيوعاً للتعامل مع عدم اتساق استخدام الأحرف الكبيرة والصغيرة هي تحويل جميع الأحرف إلى أحرف صغيرة. تسقط هذه التقنية جميع الكلمات في النص والوثيقة في نفس فضاء الميزات، ولكنها تُسبب مشكلة كبيرة في تفسير بعض الكلمات (مثل تحويل "US"(الولايات المتحدة الأمريكية) إلى "us"(ضمير)). [43] ويمكن لمحولات المصطلحات العامية والاختصارات أن تساعد في تفسير هذه الاستثناءات. [44]

##### 2.1.4. اللغة العامية والاختصارات

تعد اللغة العامية والاختصارات من أشكال الشذوذ النصي الأخرى التي تعالج في خطوة المعالجة المسيبة. الاختصار [45] هو شكل مختصر لكلمة أو عبارة، ويكون في الغالب من الأحرف الأولى للكلمات، مثل SVM الذي يرمز إلى آلة المتجهات الداعمة (Support Vector Machine).

اللغة العامية هي جزء من اللغة المستخدمة في المحادثات أو الرسائل النصية غير الرسمية، ولها معانٍ مختلفة مثل عبارة "فقدوا السيطرة على أنفسهم"، والتي تعني أساساً أنهم أصابوا بالجنون. [46] ومن الطرق الشائعة للتعامل مع هذه الكلمات تحويلها إلى لغة رسمية. [47]

##### 2.1.5. إزالة الضوضاء

تحتوي معظممجموعات بيانات النصوص والوثائق على العديد من الأحرف غير الضرورية، مثل علامات الترقيم والأحرف الخاصة. تُعد علامات الترقيم والأحرف الخاصة مهمة لفهم الإنسان للوثائق، ولكنها قد تكون ضارة بخوارزميات التصنيف. [48]

### 2.1.6. تصحیح الأخطاء الإملائیة

يُعدّ تصحیح الأخطاء الإملائیة خطوةً اختياریةً في المعالجة المسبقة. وتنتشر الأخطاء المطبعية (اختصاراً للأخطاء المطبعية) في النصوص والوثائق، لا سيما في مجموعات بيانات النصوص على وسائل التواصل الاجتماعي (مثل تويتر). وقد تناولت العديد من الخوارزميات والتقييمات والأساليب هذه المشكلة في معالجة اللغات الطبيعية. [49] وتتوفر للباحثين تقنيات وأساليب عديدة، منها تقنيات تصحيح الأخطاء الإملائية القائمة على التجزئة والحساسة للسياق. [50] بالإضافة إلى تصحيح الأخطاء الإملائية باستخدام شجرة تري وثئيات داميراو-ليفنشتاين. [51]

### 2.1.7. التفرع

في معالجة اللغات الطبيعية، قد تظهر الكلمة الواحدة بأشكال مختلفة (مثلاً صيغة المفرد والجمع) مع بقاء المعنى الدلالي لكل شكل كما هو. [52] إحدى طرق دمج الأشكال المختلفة للكلمة في نفس فضاء السمات هي التجريد. يُعدّ التجريد النصي الكلمات للحصول على أشكال مختلفة باستخدام عمليات لغوية متنوعة، مثل الإلصاق (إضافة الواحد). [53, 54] على سبيل المثال، جذر كلمة "studying" هو "study".

### 2.1.8. التجريد اللغوي

التجريد هو عملية معالجة اللغة الطبيعية التي تستبدل لاحقة الكلمة بلحقة مختلفة أو تزيل لاحقة الكلمة تماماً للحصول على شكل الكلمة الأساسي (الجذر). [54-56]

### 2.2. التمثيل النحوی للكلمات

عمل العديد من الباحثين على تقنية استخلاص خصائص النصوص هذه لمعالجة مشكلة فقدان العلاقة النحوية والدلالية بين الكلمات. وقد تناول العديد منهم تقنيات جديدة لحل هذه المشكلة، إلا أن العديد من هذه التقنيات لا تزال تعاني من بعض القيود. في [57] تم تقديم نموذج يستمد فائدته من النصوص الجينومية التقنية، حيث يُدمج فيه المعرفة النحوية والدلالية في تمثيل النص لاختيار الجمل. أما الحل الآخر للمشكلة النحوية فهو استخدام تقنية n-gram- الاستخلاص الخصائص.

### N-Gram 2.2.1. نموذج

تقنية n-gram هي مجموعة من الكلمات المكونة من  $n$  كلمة والتي تظهر "بهذا الترتيب" في مجموعة نصية. هذا ليس تمثيل نص، ولكن يمكن استخدامه كميزة لتمثيل نص.

BOW هو تمثيل لنص باستخدام كلماته (1-gram) التي تفقد ترتيبها (النحوی).

يسهل الحصول على هذا النموذج، ويمكن تمثيل النص باستخدام متوجه، عادةً ما يكون بحجم مناسب. من ناحية أخرى، تُعدّ n-gram إحدى خصائص BOW لتمثيل النص باستخدام 1-gram و 2-gram و 3-gram. ومن الشائع استخدام 1-gram وبهذه الطريقة، يمكن لخواص النص المستخرجة أن تكشف معلومات أكثر مقارنةً بـ 1-gram.

#### مثال على ثنائي الغرام

بعد أن نام لمدة أربع ساعات، قرر أن ينام لأربع ساعات أخرى.

في هذه الحالة، تكون الرموز كما يلي:

{ "بعد النوم", "النوم لمدة", "النوم لمدة أربع", "أربع ساعات", "أربع ساعات هو", "قرر", "قرر أن", "أن ينام", "النوم لمدة", "المدة أخرى", "أربع ساعات أخرى" }.

#### مثال على 3Gram

بعد أن نام لمدة أربع ساعات، قرر أن ينام لأربع ساعات أخرى.

في هذه الحالة، تكون الرموز كما يلي:

{ "بعد النوم لمدة", "النوم لمدة أربع", "أربع ساعات هو", "ساعات قرر", "قرر أن", "ينام لمدة", "النوم لمدة أخرى", "المدة أربع ساعات أخرى" }.

## 2.2.2 التركيب النحوي N-Gram

في [58]، تمت مناقشة n-grams النحوية التي يتم تعريفها بواسطة المسارات في التبعة النحوية أو الأشجار المكونة بدلاً من البنية الخطية للنص.

### 2.3. الكلمات الموزونة

أبسط أشكال استخلاص ميزات الكلمات الموزونة هو TF، حيث تربط كل كلمة برمز يشير إلى عدد مرات ظهورها في جميع النصوص. أما الطرق التي توسيع نتائج TF فتستخدم عادةً تردد الكلمات كمعيار ترجيح منطقى أو لغاريتمي.

في جميع طرق ترجيح الكلمات، تُترجم كل وثيقة إلى متوجه (بطول مساوٍ لطول الوثيقة) يحتوى على تكرار الكلمات فيها. ورغم أن هذا الأسلوب بديهي، إلا أنه محدودٌ بسبب هيمنة كلمات معينة شائعة الاستخدام في اللغة على هذه التمثيلات.

#### 2.3.1. حقيقة الكلمات (BoW)

نموذج حقيقة الكلمات (نموذج BoW) هو تمثيل مختزل ومبسط للنص وثيقة من أجزاء مختارة من النص، بناءً على معايير محددة، مثل تكرار الكلمات.

تُستخدم تقنية BoW في العديد من المجالات مثل رؤية الحاسوب، ومعالجة اللغات الطبيعية، والبريد العشوائي البایزی المرشحات، بالإضافة إلى تصنیف المستندات واسترجاع المعلومات بواسطة التعلم الآلي.

في نموذج BoW، يُنظر إلى مجموعة النصوص، مثل المستند أو الجملة، على أنها حقيقة من الكلمات. تُنشأ قوائم الكلمات في عملية BoW. هذه الكلمات في المصفوفة ليست جملًا تُشكل بنية الجمل وقواعدها، ويتم تجاهل العلاقة الدلالية بين هذه الكلمات عند جمعها وتكتينها. غالباً ما تُمثل هذه الكلمات محتوى الجملة.

بينما يتم تجاهل القواعد النحوية وترتيب الظهور، يتم احتساب التعدد ويمكن استخدامه لاحقًا لتحديد نقاط التركيز في المستندات.

إليك مثال على BoW:

وثيقة

"باعتبارها موطنًا لبرامج البكالوريوس والدراسات العليا المعترف بها في جامعة فرجينيا في هندسة النظم، فإن طلاب قسم هندسة النظم والمعلومات في جامعة فرجينيا يتعرضون لمجموعة واسعة من المجالات."

#### حقيقة الكلمات (BoW)

```
"in", "Department", "Information", "students", "", "are", "exposed", "wide", "range" 
"undergraduate", "and", "graduate", "degree", "program", "in", "systems", "engineering",
 {"As", "the", "home", "to", "UVA's", "recognized",
```

#### حقيقة الميزات (BoF)

= {1,1,1,3,2,1,2,1,2,3,1,1,2,1,1,1,1,1}

### 2.3.2. قيود نموذج حقيقة الكلمات.

تقوم نماذج "حقيقة الكلمات" بترميز كل كلمة في المفردات كمتوجه مُرفرفً أحداً، فعلى سبيل المثال، بالنسبة لمفردات بحجم  $|A|$  تمثل كل كلمة بمتوجه متفرق ذي  $|A|$  بعد، حيث يكون الرقم 1 عند الفهرس المقابل للكلمة، والرقم 0 عند كل فهرس آخر. ونظراً لأن المفردات قد تصل إلى الملايين، فإن نماذج "حقيقة الكلمات" تواجه تحديات في قابلية التوسيع (على سبيل المثال، "هذا جيد" و"هل هذا جيد؟" لهما نفس التمثيل المتوجه تماماً). كما تُعد المشكلة التقنية لنموذج "حقيقة الكلمات" التحدى الرئيسي لمجتمع علوم الحاسوب وعلوم البيانات.

يُعد تكرار المصطلحات، والذي يطلق عليه أيضًا اسم "حقيقة الكلمات"، أبسط تقنية لاستخراج ميزات النص. تعتمد هذه الطريقة على حساب عدد الكلمات في كل مستند وتعيينها لمساحة الميزات.

### 2.3.3 تردد المصطلح -تردد المستند العكسي

اقترح ك. سبارك جونز [59] طريقة لاستخدامها جنباً إلى جنب من أجل تقليل تأثير الكلمات الشائعة ضمنياً في المدونة.

يُعطى مؤشر IDF وزناً أعلى للكلمات ذات الترددات العالية أو المنخفضة في المستند.

يُعرف هذا المزيج من تردد المصطلح وتردد المستند العكسي باسم تردد المصطلح - تردد المستند العكسي (TF-IDF). ويُعطى التمثيل الرياضي لوزن المصطلح في المستند باستخدام TF-IDF في المعادلة (1).

$$W(d, t) = TE(d, t) \cdot \log(\frac{1}{d}) \cdot fft \quad (1)$$

هنا، يمثل  $N$  عدد المستندات، و  $f(t)$  عدد المستندات التي تحتوي على المصطلح  $t$  في المجموعة. يُحسّن الحد الأول في المعادلة (1) الاستدعاء، بينما يُحسّن الحد الثاني دقة تمثيل الكلمات. [60] على الرغم من أن  $\text{TF-IDF}$  يُحاول التغلب على مشكلة المصطلحات المشتركة في المستند، إلا أنه لا يزال يعني من بعض القيود الوصفية الأخرى. تحديداً، لا يمكن لـ  $\text{TF-IDF}$  مراعاة التشابه بين الكلمات في المستند، حيث تُقدم كل كلمة بشكل مستقل كمؤشر. مع ذلك، ومع تطور نماذج أكثر تعقيداً في السنوات الأخيرة، ظهرت أساليب جديدة، مثل تمثيل الكلمات، التي يمكنها دمج مفاهيم مثل تشابه الكلمات وتصنيف أجزاء الكلام.

## 2.4 تضمين الكلمات

على الرغم من وجود تمثيلات نحوية للكلمات، إلا أن هذا لا يعني أن النموذج يستوعب المعنى الدلالي للكلمات. من جهة أخرى، لا تراعي نماذج "حقيقة الكلمات" دلالات الكلمات. على سبيل المثال، تُستخدم الكلمات "طائرة" وـ"طائنة" وـ"مركبة" بشكل متكرر في السياق نفسه. ومع ذلك، فإن المتوجهات المقابلة لهذه الكلمات متعامدة في نموذج "حقيقة الكلمات". تُشكل هذه المشكلة عائقاً كبيراً أمام فهم الجمل داخل النموذج. كما أن المشكلة الأخرى في نموذج "حقيقة الكلمات" هي عدم مراعاة ترتيب الكلمات في العبارة.

لا يحل نموذج n-gram هذه المشكلة، لذا يجب إيجاد تشابه لكل كلمة في الجملة.

عمل العديد من الباحثين على تضمين الكلمات لحل هذه المشكلة. يقترح نموذج CBOW وحقيقة الكلمات المستمرة Skip-gram [61]، وهي بسيطة وأحادية الطبيعة تعتمد على الضرب الداخلي بين متجهين من الكلمات.

يُعدّ تضمين الكلمات تقنيةً لتعلم الميزات، حيث تحول كل كلمة أو عبارة من المفردات إلى متغير ذي  $N$  بعدد من الأعداد الحقيقية. وقد ظهرت طرق عديدة لتضمين الكلمات بهدف تحويل الكلمات المفردة إلى مدخلات مفهومة لخوارزميات التعلم الآلي. يذكر هذا العمل على `Word2Vec` و `GloVe` و `FastText` وهي ثلاثة من أكثر الطرق شيوعاً والتي استُخدمت بنجاح في تقييمات التعلم العميق. مؤخراً، طرحت تقنية جديدة لتمثيل الكلمات، حيث تعتمد متوجهات الكلمات على سياق الكلمة، وُسمى "تمثيلات الكلمات السياقية" أو "تمثيلات الكلمات السياقية العميق".

#### 2.4.1. Word2Vec

قدم ت. ميكولوف وآخرون [61، 62] تمثيل "الكلمة إلى متوجه" كبنية محسنة لتضمين الكلمات. يستخدم نهج Word2Vec شبكات عصبية سطحية ذات طبقتين مخفيتين، ونموذج حقيبة الكلمات المتصلة (CBOW) ونموذج Skip-gram لإنشاء متوجه عالي الأبعاد لكل كلمة. يحلل نموذج Skip-gram مجموعة من الكلمات  $w$  وسياقها [10]  $c$  والهدف هو تعظيم الاحتمالية:

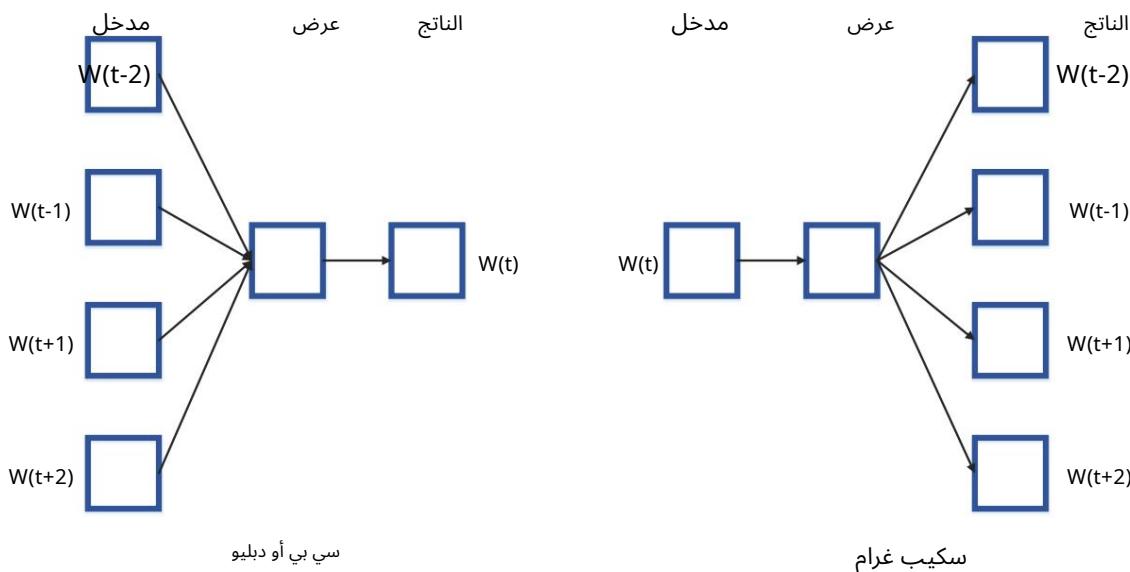
$$\Theta \vdash \Box_{\mathcal{P}}^{\mathcal{C}} \psi \rightarrow \psi \quad (2)$$

حيث يشير  $T$  إلى النص، و  $\theta$  هو معامل.

يوضح الشكل 2 نموذج CBOW بسيطًا يحاول إيجاد الكلمة بناءً على الكلمات السابقة، بينما يحاول نموذج Skip-gram إيجاد الكلمات التي قد تأتي في جوار كل كلمة. تمثل الأوزان بين طبقة الإدخال وطبقة الإخراج مصفوفة  $N \times V$  من  $W$ .

$$h = W^T c = W^T \frac{1}{\sqrt{V}} \quad (3)$$

تُوفّر هذه الطريقة أدلةً فعالةً للغاية لاكتشاف العلاقات في مجموعة النصوص، بالإضافة إلى أوجه التشابه بين الكلمات. فعلى سبيل المثال، يعتبر هذا التضمين كلمتي "كبير" و "أكبر" متداوينتين في فضاء المتجهات الذي يخصّصهما.



الشكل 2. تتبّعاً بنية حقيبة الكلمات المستمرة (CBOW) بالكلمة الحالية بناءً على السياق ، ويتبعها Skip-gram بالكلمات المحيطة بناءً على الكلمة الحالية المعطاة [61].

#### نموذج حقيبة الكلمات المتصلة

يتم تمثيل نموذج حقيبة الكلمات المستمرة بكلمات متعددة لمجموعة معينة من الكلمات المستهدفة. على سبيل المثال، كلتا "طائرة" و "عسكرى" كلمات سياقية لكلمة "القوات الجوية" باعتبارها الكلمة المستهدفة. يتضمن هذا تكرار المدخلات إلى وصلات الطبقة المخفية  $\beta$  مرّة، وهو عدد كلمات السياق [61]. لذا، تُستخدم نموذج "حقيبة الكلمات" في الغالب لتمثيل مجموعة غير مرتبة من الكلمات كمتجه، أول خطوة هي إنشاء مجمّع، أي جميع الكلمات الفريدة في المدونة. سيكون ناتج الشبكة العصبية الضحلة عبارة عن مهمة "التنبؤ بالكلمة بناءً على سياقها". يعتمد عدد الكلمات المستخدمة على إعداد حجم النافذة (الحجم الشائع هو 4-5 كلمات).

#### نموذج سكيب غرام المستمر

يُعد نموذج Skip-gram المستمر بنية نموذجية أخرى مشابهة جدًا للمودع [61] CBOW إلا أن هذا النموذج، بدلاً من التنبؤ بالكلمة الحالية بناءً على السياق، يسعى إلى تعطيم تصنيف كلمة ما بناءً على كلمة أخرى في الجملة نفسها. يستخدم كل من نموذج حقيبة الكلمات المستمر ونموذج Skip-gram المستمر لاحتفاظ بالمعلومات النحوية والدلائلية للجمل لصالح خوارزميات التعلم الآلي.

#### 2.4.2. المتجهات العالمية لتمثيل الكلمات (GloVe)

تُعد تقنية [11] Global Vectors (GloVe) إحدى تقنيات تضمين الكلمات القوية الأخرى المستخدمة في تصنيف النصوص . يشبه هذا النهج إلى حد كبير طريقة Word2Vec، حيث يتم تمثيل كل كلمة بـ

يتم تقديمها بواسطة متوجه عالي الأبعاد وتدربيها بناءً على الكلمات المحيطة بها على مجموعة ضخمة من النصوص. تعتمد تقنية تضمين الكلمات المدربة مسبقاً، والمستخدمة في العديد من الدراسات، على 400,000 معجم تم تدريبيها على ويكيبيديا 2014 ومجموعة بيانات 5. مع 50 بعضاً تمثيل الكلمات. كما توفر GloVe تقنيات أخرى لتمثيل الكلمات المتوجهة المدربة مسبقاً بأبعاد 100 و 200 و 300. والتي تم تدريبيها علىمجموعات بيانات أكبر، بما في ذلك محتوى توپر، يوضح الشكل 3 تمثيلاً مرجئياً لمسافات الكلمات على مجموعة بيانات نموذجية باستخدام تقنية SNE-Self-Tuning. [64] دالة الهدف هي كما يلي:

$$\text{Pik} = \frac{\text{f}(w_i \oplus w_j, w_k)}{\text{Pj}_k} \quad (4)$$

حيث يشير  $w_i$  إلى متوجه الكلمة  $i$ ، و  $\text{Pik}$  يشير إلى احتمال ظهور الكلمة  $k$  في سياق الكلمة  $i$ .



الشكل 3. المتوجهات العالمية لتمثيل الكلمات.

#### العنصر السريع 2.4.3.

تجاهل العديد من طرق تمثيل الكلمات الأخرى بنية الكلمات من خلال تخصيص متوجه مميز لكل كلمة. [65] وقد أطلق مختبر أبحاث الذكاء الاصطناعي في فيسبوك تقنية جديدة لحل هذه المشكلة، وذلك بتقديم طريقة جديدة لتمثيل الكلمات تسمى FastText. على شكل مجموعة من n-gram من الأحرف، على سبيل المثال، عند إدخال الكلمة "introduce" و  $n=3$ ، تُنتج التمثيل التالي المكون من ثلاثيات الأحرف:

`<في، ce، int، ntr، tro، Rod، Odu، duc، uce>`

لاحظ أن التسلسل ،`<int>` المقابل للكلمة هنا، يختلف عن الثلاثي "int". من كلمة "تقديم".

لنفترض أن لدينا قاموساً من n-grams بحجم  $G$ ، ومعطى الكلمة  $w$  مرتبطة بـ دالة التسجيل التي تم الحصول عليها [65] في هذه الحالة هي:

$$s(w, c) = \frac{z^{vc}}{g \oplus gw} \quad (5)$$

حيث  $.gw \in \{1, 2, \dots, G\}$

نشر فيسبوك متجهات كلمات مدرّبة مسبقاً لـ ٢٩٤ اللغة، تم تدريبيها على ويكيبيديا باستخدام FastText القائم على نموذج [60]. استخدم FastText كـ Skip-gram بالمعلمات الافتراضية.

#### 2.4.4. تمثيلات الكلمات السياقية

تُعد تمثيلات الكلمات السياقية تقنية أخرى لتضمين الكلمات، وهي مبنية على تقنية [66] context2vec التي قدمها ب. ماكان وآخرون. تستخدم طريقة context2vec [67] ذاكرة طويلة المدى ثنائية الاتجاه. وقد طور م. بيترز وآخرون [67] هذه التقنية لإنشاء تقنية تمثيلات الكلمات السياقية العميقه. تتضمن هذه التقنية السمتين الرئيسيتين لتمثيل الكلمات: (أ) الخصائص المعقدة لاستخدام الكلمات (مثل النحو والدلالة)، و(ب) كيفية اختلاف هذه الاستخدامات عبر السياقات اللغوية (مثل نمذجة تعدد المعاني) [67].

الفكرة الرئيسية وراء تقنيات تضمين الكلمات هذه هي أن متجهات الكلمات الناتجة يتم تعلمها من نموذج لغة ثنائي الاتجاه، (b) bLM والذي يتكون من نماذج لغة أمامية وخلفية.

فيما يلي نماذج اللغة الأمامية:

$$\text{شمالي} = p(t_k | p(t_1, t_2, \dots, t_{k-1})) \quad (6)$$

فيما يلي نماذج التعلم العكسي:

$$\text{شمالي} = p(t_k | p(t_1, t_2, \dots, t_{k-1})) \quad (7)$$

تعمل هذه الصيغة على زيادة احتمالية اللوغراريتم للاتجاهين الأمامي والخلفي معاً على النحو التالي:

$$\text{شمالي} = \log p(t_1, \dots, t_{k-1}; \Theta_x, \Theta_s) + \log p(t_k | t_1, \dots, t_{k-1}; \Theta_x, \Theta_s) \quad (8)$$

حيث يمثل  $\Theta_x$  تمثيل الرمز المميز، ويشير  $\Theta_s$  إلى طبقة. بعد ذلك، يتم حساب ELMo كوزن خاص بال مهمة لجميع طبقات bLM على النحو التالي:

$$= E(R_k; \Theta_t^{\text{LM}}) \sum_{j=0}^{L_{\text{task}}} \text{ مهمة}_{\text{LM}}^j \quad (9)$$

حيث يتم حساب  $\text{LM}_j$  بواسطة:

$$= \frac{1}{L} \sum_{k=1}^L h_k^j \quad (10)$$

حيث  $\text{LM}_j$  يرمز إلى الأوزان المعيارية باستخدام دالة softmax، وهو المعامل القياسي.

#### 2.5. القيد

على الرغم من استخدام نموذج حقيقة الكلمات المستمرة ونموذج Skip-gram المستمر لاحتفاظ بالمعلومات النحوية والدلالية لكل جملة لخوارزميات التعلم الآلي، إلا أنه لا تزال هناك مشكلة كيفية الحفاظ على المعنى الكامل للوثر المتماسكة للتعلم الآلي.

مثال:

الوثيقة: {"ذهبت مريم إلى باريس في الرابع من يوليو/تموز 2018. فاتها عروض الألعاب النارية والاحتفالات بيوم الاستقلال . هذا اليوم عطلة رسمية في الولايات المتحدة الأمريكية، تُحيي ذكرى إعلان استقلال الولايات المتحدة في الرابع من يوليو/تموز 1776. أعلن الكونغرس القاري أن المستعمرات الأمريكية الثلاث عشرة لم تعد خاضعة لملك بريطانيا، وأنها أصبحت الآن دولة موحدة وحرة ومستقلة. ترحب مريم في البقاء في البلاد العام المقبل والاحتفال مع أصدقائها."}

مستوى الجملة في هذه الوثيقة:

S1: {"ذهبت مريم إلى باريس في 4 يوليو 2018:."

S2: {"لقد فاتها الألعاب النارية والاحتفالات بيوم الاستقلال."}

S3: {"هذا اليوم عطلة رسمية في الولايات المتحدة الأمريكية لإحياء ذكرى إعلان استقلال..."}

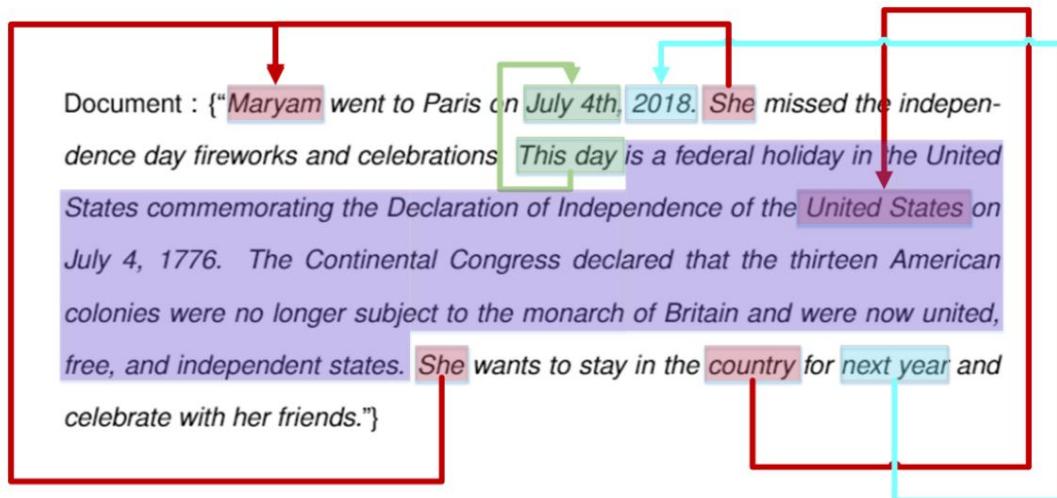
الولايات المتحدة في 4 يوليو ."S4:

{"أعلن المؤتمر القاري أن المستعمرات الأمريكية الثلاث عشرة لم تعد خاضعة لملك بريطانيا، وأنها الآن دول موحدة وحرة ومستقلة."}

{"لديها خطة للعام المقبل للبقاء في البلاد والاحتفال مع أصدقائها."}

القيود:

يوضح الشكل 4 فشل استخلاص الميزات على مستوى الجملة. يشير اللون الأرجواني في الشكل إلى تاريخ عبارة "هذا اليوم". كما تشير عبارة "هذا اليوم" إلى "الرابع من يوليو". في الجملة، تشير كلمة "هي" إلى "ميريم" في الجملة. S1



الشكل 4. قيود استخراج ميزات المستند على مستوى الجملة الواحدة.

### 3. تقليل الأبعاد

تتألف تسلسلات النصوص في نماذج المتجهات القائمة على المصطلحات من العديد من الخصائص. ولذلك، فإن التعقيد الزمني واستهلاك الذاكرة مرتفعان للغاية في هذه الطرق. ولمعالجة هذه المشكلة، يستخدم العديد من الباحثين تقنية تقليل الأبعاد لتقليل حجم فضاء الخصائص. في هذا القسم، نناقش بالتفصيل خوارزميات تقليل الأبعاد المتاحة.

#### 3.1. تحليل المكونات

##### 3.1.1. تحليل المكونات الرئيسية (PCA)

يُعد تحليل المكونات الرئيسية (PCA) التقنية الأكثر شيوعاً في التحليل متعدد المتغيرات وتقليل الأبعاد. PCA هي طريقة لتحديد فضاء فرعي تقع فيه البيانات تقريباً [68]. وهذا يعني إيجاد متغيرات جديدة غير مترابطة وتعطيم البيانات "لحفاظ على أكبر قدر ممكن من البيانات" [69]. نفترض أن لدينا مجموعة بيانات  $\mathbf{X}$  حيث  $\mathbf{X} = \mathbf{A}\mathbf{B}^T$  حيث  $\mathbf{A}$  العمود ز من المصفوفة  $\mathbf{X}$  هو متجه  $\mathbf{a}_j$  وهو يمثل المشاهدات على المتغير  $j$ . يمكن أن يكون التركيب الخطى لـ  $\mathbf{B}$  هو

مكتوبة على النحو التالي:

$$\mathbf{B} = \sum_{j=1}^k \mathbf{a}_j \mathbf{x}_j \mathbf{a}_j^T \quad (11)$$

حيث  $\mathbf{a}_j$  يمثل متجهاً من الثوابت  $a_{1j}, a_{2j}, \dots, a_{mj}$ . ويمكن التعبير عن تبيان هذا التركيب الخطى كما يلي:

$$\text{var}(\mathbf{X}_{\mathbf{a}}) = \mathbf{a}^T \mathbf{T} \mathbf{a} \quad (12)$$

حيث  $\mathbf{T}$  هي مصفوفة البيانات المشتركة للعينة. الهدف هو إيجاد التركيبة الخطية ذات البيانات الأقصى. وهذا يعني تعظيم  $\mathbf{a}^T \mathbf{T} \mathbf{a}$  حيث  $\mathbf{a}$  هو مضاعف لاغرانج.

يمكن استخدام تحليل المكونات الرئيسية (PCA) كأداة معالجة مسبقة لتقليل أبعاد مجموعة البيانات قبل تطبيق خوارزمية التعلم الخاضع للإشراف [70].  
[71] طبقت تحليل المكونات الرئيسية على المكونات الخطية حيث تم تحويل المكونات إلى مكونات خطية بخطىء متحيز، مما يعتمد على المكونات الأصلية.

### 3.1.2. تحليل المكونات المستقلة (ICA)

ظهرت تقنية تحليل المكونات المستقلة (ICA) بواسطة ه. جيني [72]. ثم ظهرت هذه التقنية لاحقاً بواسطة س. جوت و. هيرولت [73]. تُعد طريقة للنمذجة الإحصائية حيث تُعتبر البيانات المرصودة بتحويل خطىء [74] لنفترض أننا نرصدنا  $n$  من الحالات الخطية ( $x_1, x_2, \dots, x_n$ ) حيث المكونات المستقلة:

$$x_j = a_{j1} s_1 + a_{j2} s_2 + \dots + a_{jn} s_n \quad (13)$$

يمكن كتابة التمودج [75] على النحو التالي:

$$X = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \quad (14)$$

ويمكن كتابة التمودج [75] على النحو التالي:

$$S = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \quad (15)$$

### 3.2. تحليل التمييز الخطى (LDA)

تُعد تقنية تحليل التمييز الخطى (LDA) أسلوباً شائعاً لاستخدامه لتصنيف البيانات وتقليل أبعادها [76]. وتنبأ هذه التقنية بشكل خاص عندما تكون ترددات الفئات غير متساوية، وقد تم تقييم أدائها على بيانات اختبار مُولدٌ عشوائياً. ويُعتبر التحويل المعتمد على الفئة والتحويل غير المعتمد على الفئة نهجين لتقنية LDA، حيث يُستخدم فيما على التوالي نسبة التباين بين الفئات إلى التباين داخل الفئات، ونسبة التباين الكلي إلى التباين داخل الفئات [77].

لنفترض أن  $R_d$  هي عينات ذات بعد  $d$ ، وأن  $\{x_1, x_2, \dots, x_c\}$  هي عدد المستندات و  $c$  هو عدد الفئات. يُحسب عدد العينات في كل فئة كما يلى:

$$S_w = \sum_{i=1}^c S_i \quad (16)$$

أين

$$S_i = \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \quad (17)$$

يتم تعريف التعميم بين مصفوفة تشتت الفئات على النحو التالي:

$$S_b = \sum_{i=1}^c n_i(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \quad (18)$$

أين

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^c x_i \quad (19)$$

بالنسبة لمتجه الإسقاط  $w$  الذي يمكن إسقاطه في المصفوفة  $W$ :

$$W = [w_1 | w_2 | \dots | w_c] \quad (20)$$

$$y_i = w^T x_i \quad (21)$$

وبالتالي، فإن متجه  $\mu$ (المتوسط) ومصفوفات  $S$ (مصفوفات التشتت) للإسقاط إلى بعد أقل كما يلي:

$$S_w = \sum_{i=1}^n y_i w_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \quad (22)$$

$$S = \sum_{i=1}^n (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \quad (23)$$

إذا لم يكن الإسقاط قياسياً  $1 \leq c \leq n$ ، فسيكون محدداً مصفوفات التشتت  
يُستخدم على النحو التالي:

$$J(SB) = \frac{|WTSBW|}{|WTSWW|} \quad (24)$$

[جوب عرب]

انطلاقاً من تحليل التمييز لفيشر ، يمكننا إعادة كتابة المعادلة على النحو التالي:

$$\frac{|WTSBW|}{|WTSWW|} \quad (25)$$

### 3.3. تحليل المصفوفات غير السالبة (NMF).

لقد ثبت أن تحليل المصفوفة غير السالبة (NMF) أو تقرير المصفوفة غير السالبة هو أسلوب قوي للغاية للبيانات عالية الأبعاد مثل تحليل

النصوص والتسلسلات [79].

تعُد هذه التقنية طريقة واحدة لتقليل الأبعاد [80]. في هذا القسم، نقدم لمحنة موجزة عن تحليل المصفوفات غير السالبة (NMF) لمجموعات بيانات النصوص والوثائق. بافتراض أن المصفوفة  $V$  غير سالبة من الرتبة  $n \times m$ ، فإنها تُشكل تقريراً لـ:

$$V = WH \quad (26)$$

حيث  $W = Rr \times m$ ،  $H = Rr \times r$  إذا افترضنا أن  $r < nm$ . فيمكن اعتبار حاصل ضرب  $WH$  شكلًا مضغوطاً للبيانات في  $V$ . عندئذ، يكون  $v_i$  و  $h_j$   
هما العمودان المقابلان في  $V$  و  $H$ .  
يمكن إعادة كتابة حساب كل عمود مقابل على النحو التالي:

$$v_i = \sum_j (W_{ij} H_{jj}) \quad (27)$$

يمكن كتابة وقت الحساب لكل تكرار، كما قدمه إس. تسوجي وآخرون [80] على النحو التالي:

$$v_i = \sum_j \frac{(W_{ij} H_{jj})}{(W_{ij} H_{jj})} \quad (28)$$

$$v_i = \sum_j \frac{(V_{ij} H_{jj})}{(V_{ij} H_{jj})} \quad (29)$$

وبالتالي، يتم حساب الحد الأدنى المحلي لدالة الهدف على النحو التالي:

$$(V_{ij} - \sum_k (W_{ik} H_{kj}))^2 \quad (30)$$

يمكن إعادة كتابة عملية تعظيم دالة الهدف على النحو التالي:

$$(V_{ij} \log((W_{ik} H_{kj})) - \sum_k (W_{ik} H_{kj}))^2 \quad (31)$$

يتم تعريف دالة الهدف، المعطاة بواسطة تباعد كولياك-لايبير ، [81,82] على النحو التالي:

$$V_{kj} = \frac{1}{\sqrt{\sum_{i=1}^m w_{ij}^2}} \quad (32)$$

$$w_{ij} = \frac{z_i}{\sqrt{\sum_{k=1}^n v_{kj}^2}} \quad (33)$$

$$z_i = \frac{1}{\sqrt{\sum_{k=1}^n v_{kj}^2}} \quad (34)$$

أمثلة

تتضمن عملية تقليل الأبعاد هذه القائمة على تحليل المصفوفات غير السالية الخطوات الخمس التالية [80] (الخطوة السادسة اختيارية). لكنها شائعة الاستخدام في استرجاع المعلومات:

استخرج مصطلح الفهرس بعد المعالجة المسبيقة مثل استخراج الميزات وتنظيف النص كما هو موضح في القسم 2. عندئذٍ، سيكون لدينا  $m$  مستندًا مع  $n$  ميزة؛ أنشأ  $m$  مستندًا ( $d = \{d_1, d_2, \dots, d_n\}$ ). حيث يشير  $z_i$  إلى الأوزان المحلية للمصطلح  $d_i$  المستند،  $z_i = G_i$  هي الأوزان العامة للمستند (III)؛ أي  $G_i = \text{NMF}(d_i, m)$  على جميع المصطلحات في جميع المستندات واحدًا تلو الآخر؛ (IV)

قم بإسقاط متوجه المستند المدرب في فضاء ذي أبعاد  $m$  باستخدام نفس التحويل، قم برسم مجموعة الاختبار في الفضاء ذي الأبعاد  $m$ ؛ احسب التشابه بين متوجهات المستندات المحولة ومتوجه الاستعلام. (V)

(نحو)

### 3.4. الإسقاط العشوائي

الإسقاط العشوائي تقنية حديثة لتقليل الأبعاد، تُستخدم غالباً معمجموعات البيانات الضخمة أو فضاءات الميزات ذات الأبعاد العالية. تُنتج النصوص والوثائق، وخاصةً مع استخراج الميزات الموزونة، عدداً هائلاً من الميزات. وقد طبق العديد من الباحثين الإسقاط العشوائي على البيانات النصية [83, 84] لاستخراج المعلومات من النصوص وتصنيفها وتقليل أبعادها. في هذا القسم، نستعرض بعض تقنيات الإسقاط العشوائي الأساسية. كما هو موضح في الشكل، يقدم هذا القسم نظرة عامة على الإسقاط العشوائي.

#### 3.4.1. أحواض مطبخ عشوائية

تعتمد الفكرة الأساسية لخوارزمية أحواض المطبخ العشوائية [85] علىأخذ عينات باستخدام تكامل مونت كارلو [86] لنearib النواة كجزء من تقليل الأبعاد. وتعمل هذه التقنية فقط مع النواة الثابتة تحت الإزاحة.

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle = K(x - x') \quad (35)$$

حيث النواة الثابتة تحت الإزاحة، وهي نواة تقريرية لـ

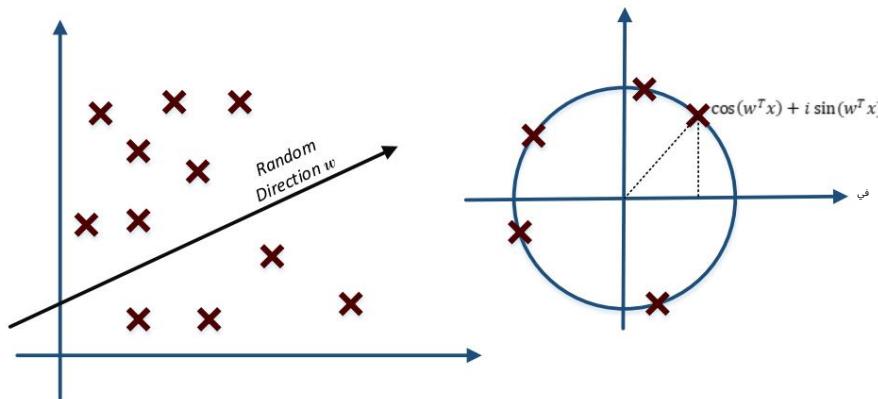
$$K(x - x') = z(x)z(x') \quad (36)$$

$$K(x, x') =$$

$$z(x)z(x')$$

أمثلة

حيث  $D$  هو عدد العينات المستهدفة، و  $(w)$  هو توزيع احتمالي، و  $w$  يرمز إلى الاتجاه العشوائي، و  $R \times D$  حيث  $R$  هو عدد الميزات و  $D$  هو الهدف.



الشكل 5. يوضح الرسم البياني على اليسار كيف نقوم بتوسيع اتجاه عشوائي، ويوضح الرسم البياني على اليمين كيف نقوم بإسقاط مجموعة البيانات في الفضاء الجديد باستخدام الأعداد المركبة.

$$K(x, x') = K(x \otimes x') \otimes \frac{1}{D} \prod_{j=1}^D e^{iw_j^T (x \otimes x')} \quad (38)$$

$$\begin{aligned} \frac{1}{D} \prod_{j=1}^D e^{iw_j^T (x \otimes x')} &= \frac{1}{D} \prod_{j=1}^D e^{iw_j^T j x} = \\ &= \frac{1}{D} \prod_{j=1}^D e^{w_j^T j x} = \frac{1}{D} \prod_{j=1}^D e^{w_j^T x} \end{aligned} \quad (39)$$

$$k(x \otimes x') \otimes \varphi(x)\varphi(x') \quad (40)$$

$$\varphi(x) = \cos(w^T x + bi) \quad (41)$$

حيث يكون المتغير  $w$  متغيراً عشوائياً منتظماً. ( $bi \in [0, \pi]$ )

### 3.4.2. معادلة جونسون ليندنشتراوس

أثبتت ويليام ب. جونسون وجورام ليندنشتراوس [87, 88] أنه لأي فضاء إقليدي ذي  $D$  نقطة، فإن  $\log n \leq R_d \leq R_k$  حيث  $R_d = O(x = u \otimes v)$  ويمكن تقريبها في النهاية إلى الحد الأدنى لاحتمالية النجاح

—

(2) لأي  $u \otimes v \in n$

$$(i \otimes) | | u \otimes v | | 2 \otimes | | f(u) \otimes f(v) | | 2 \otimes (i +) | | u \otimes v | | 2 \quad (42)$$

برهان جونسون ليندنشتراوس ليما:

لأي مجموعة  $n$  من نقاط البيانات من  $R$  حيث  $n \leq m$  ومتغير عشوائي  $w \in \mathbb{R}^D$

$$\frac{1}{4} \leq k(2 \otimes 2) \leq 1 \otimes m \quad (43)$$

$$k = \frac{16 \log n}{\pi^2} :$$

$$\begin{aligned} \frac{1}{e^{\frac{16 \log n}{\pi^2}}} &\leq k(3 \otimes 3) \leq \frac{16 \log n}{\pi^2} \cdot \frac{2m}{4} \\ \frac{16 \log n}{\pi^2} \cdot \frac{2m}{4} &\leq k(3 \otimes 3) \leq \frac{16 \log n}{\pi^2} \cdot \frac{2m}{4} \\ 4 \otimes 2 \cdot 1 \otimes 2 \otimes 0 &> 1 \otimes 2 \otimes 0 \end{aligned} \quad (44)$$

البرهان على اللمة [89]: 1

ليكن  $\Psi$  متغيراً عشوائياً بدرجة حرية  $k$ ، إذن بالنسبة لـ  $[0, 1]$

العلاقة العامة [١ - (ك) ١] (ك) ٢e ١ (45)

نبدأ بمتباينة ماركوف [90]:

$$E[\Psi] \quad \text{العلاقـات العامة} [\Psi] (1 - \kappa) [1 - \kappa] \quad (46)$$

$$\Pr[e^{\lambda \Psi} \geq \lambda(1-\lambda)] \leq \frac{\lambda(1-\lambda)}{E[e^{\lambda \Psi}]}$$

(47)

$$\Pr[\{\Psi \models (1 \sqcap k)\}] \leq 2^{-k} \quad \text{حيث } 0.5 < e \leq 1$$

$\Pr[\{\Psi \models (1 \sqcap k)\}] = \Pr[\{\Psi \models (1 \sqcap k) \wedge \text{متضمنة}\}]$

$$\frac{(1+)_9}{9} \quad \frac{9 - \underline{\quad}^2}{9} = \frac{(2-3)_9}{9(30)4 = e} \quad (48)$$

العلاقات العامة[1 - ك] (1) ك م + ك

$$\frac{\square \Psi \square \Psi}{4} \quad (49)$$

البرهان على اللمة [89]: 2

ليكن  $W$  متغيراً عشوائياً حيث  $W \in [0, 1]^k$  و  $d \times k$  هي نقاط البيانات،  $Rd$  إذن لأي:

$$\boxed{1} \quad \text{---} \quad \boxed{4}$$

$x_0$  هي قيمة التقرير العشوائي في  $\text{النهاية}_{\text{لذ}}(x)$  يمكننا إعادة كتابة

— Br[1, 2] + Br[4, 2] + 2M[1, 1, 2, 1] (150)  $\frac{1}{2} \pm 1.1$

$\zeta_i = ||x_i||$  معاشر

$$\vdash \zeta_i | \mathcal{B}[\alpha_1 \beta_1] * \beta_k] +$$

## العلاقات العامة[1 - 2]

حيث يمكننا إثبات المعادلة (51) باستخدام المعادلة (45):

المشفر التلقائي 3.5.

المشتهر التقليدي هو نوع من الشهادات العصبية تدرب على محاولة نسخ مدخلاته إلى مخرجهاته. [٩١] وقد حقق المشتهر التقليدي نجاحاً كبيراً كطريقة لتقليل الأبعاد بفضل قدرة الشهادات العصبية الفائقة على إعادة إنتاج البيانات. [٩٢] تم تقديم الإصدارات الأولى، من المشتهر التقليدي،

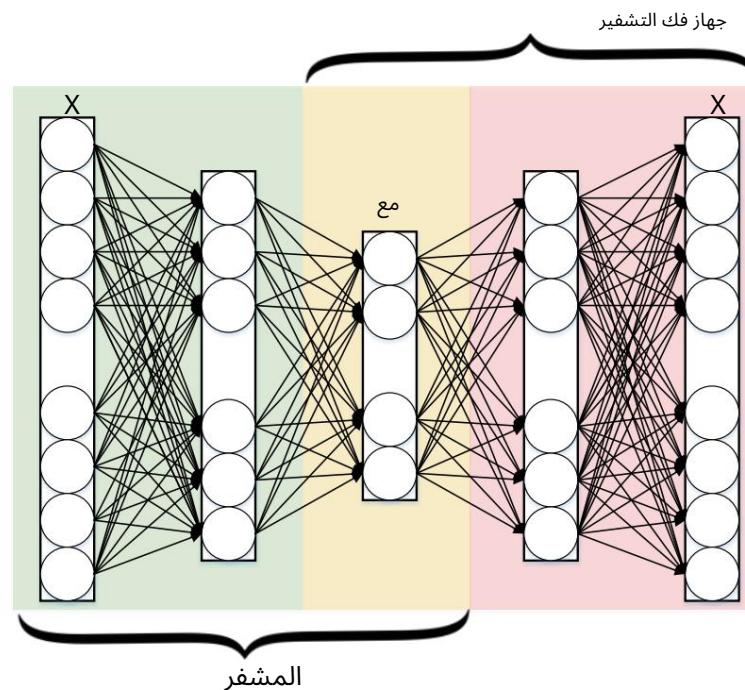
بواسطة [93] DE Rumelhart et al. في عام 1985. الفكرة الرئيسية هي أن طبقة مخفية واحدة بين طبقات الإدخال والإخراج تحتوي على وحدات أقل [94]. وبالتالي يمكن استخدامها لتقليل أبعاد مساحة الميزات. وخاصة بالنسبة للنصوص والوثائق والسلسلات التي تحتوي على العديد من الميزات، فإن استخدام برنامج التشفير التلقائي يمكن أن يساعد في السماح بمعالجة البيانات بشكل أسرع وأكثر كفاءة.

### 3.5.1. الأطر العام

كما هو موضح في الشكل ، تحتوي طبقات الإدخال والإخراج في المشفر التلقائي على  $n$  وحدة حيث ،  $Rn = Z^{|x|}$  بينما تحتوي الطبقة المخفية على  $p$  وحدة حيث .  $n < p$  في هذه التقنية لتقليل الأبعاد، يتم تقليل أبعاد فضاء الميزات النهائي من  $n$  إلى  $p$ . يتضمن تمثيل المشفر مجموع تمثيل جميع الكلمات (في حالة نموذج حقبة الكلمات)، مما يعكس التردد النسبي لكل كلمة [96].

$$a(x) = c + \sum_{i=1}^n w_i x_i, \quad \varphi(x) = h(a(x)) \quad (53)$$

حيث  $(.)$  هي دالة غير خطية على مستوى العناصر مثل الدالة السينية (المعادلة (79)).



الشكل . يوضح هذا الشكل كيفية عمل مشفر تلقائي بسيط. يحتوي النموذج الموضح على الطبقات التالية:  $Z$  هي الشفرة، وتُستخدم طبقتان مخفيتان للتشفير وطبقتان لفك التشفير.

### 3.5.2. بنية التشفير التلقائي التقليدية

يمكن تقسيم المشفر التلقائي القائم على الشبكات العصبية الالتفافية (CNN) إلى قسمين رئيسيين الخطوات [97] (الترميز وفك الترميز).

$$\text{إذا كان } a(i, j) = d \quad \begin{cases} \boxed{\phantom{0}} & \text{ـ 1} \\ \boxed{\phantom{0}} & \text{ـ 2} \\ \boxed{\phantom{0}} & \text{ـ 3} \\ \boxed{\phantom{0}} & \text{ـ 4} \end{cases} \quad \begin{matrix} 2k+1 & 2k+1 \\ \text{ـ 1} & \text{ـ 2} \end{matrix} \quad \begin{matrix} u & v \\ i & j \end{matrix} \quad \begin{matrix} m & \dots & 1 \\ \text{ـ 1} & \text{ـ 2} & \text{ـ 3} \end{matrix} \quad \text{ن ، } \mu(v) \text{ معرف} \quad (54)$$

$$F \leftarrow \{F \cup \{I\} \text{ حيث } I = \{I_1, \dots, I_n\} \text{ ، الذي يتعلم تمثيل المدخلات من خلال دمج الدوال غير الخطية: } \\ \dim(I) = \dim(\text{decode}(\text{encode}(I))) \text{ ، حيث } b \text{ هو الانحياز، وعدد الأصفار التي نريد إضافتها إلى المدخلات هو بحيث: } \\ \sum_{i=1}^n z_i = 1 \text{ حيث } z_i \in \{0, 1\} \text{ .} \quad (55)$$

(1) حيث  $b$  هو الانحياز، وعدد الأصفار التي نريد إضافتها إلى المدخلات هو بحيث:  $\sum_{i=1}^n z_i = 1$  حيث  $z_i \in \{0, 1\}$ . تكون عملية الالتفاف للترميز متساوية لما يلي:

$$Ow = Oh = (Iw + 2(2k + 1) \odot 2) \odot (2k + 1) \quad (56) \\ = Iw + (2k + 1) \odot 1$$

تنتج خطوة فك التشفير الالتفافية  $n$  من خرائط الميزات  $z_m = 1, \dots, n$ . هي نتيجة الالتفاف بين حجم  $F(2)$  [97-99] وحجم مرشحات الالتفاف  $Z = \{z_i = 1\}$ .

$$\tilde{I} = a(Z \odot F^{(2)} \odot (2k + 1) \odot 1) \quad (57)$$

$$Ow = Oh = (Iw + (2k + 1) \odot 1) \odot (2k + 1) + 1 = Iw = Ih \quad (58)$$

حيث توضح المعادلة (58) عملية فك التشفير الالتفافية ذات الأبعاد. أبعاد المدخلات تساوي أبعاد المخرجات.

### 3.5.3. بنية التشفير التلقائي المتكررة.

الشبكة العصبية المتكررة (RNN) هي تعليمي طبقي للشبكات العصبية ذات التغذية الأمامية لتشمل المتتاليات. [100] يوضح الشكل 7 بنية المشفر التلقائي المتكرر. تحسب الشبكة العصبية المتكررة القياسية التشفير كسلسلة من المخرجات من خلال التكرار.

$$ht = \text{sigm}(W_h x_t + W_h h_{t-1}) \quad (59)$$

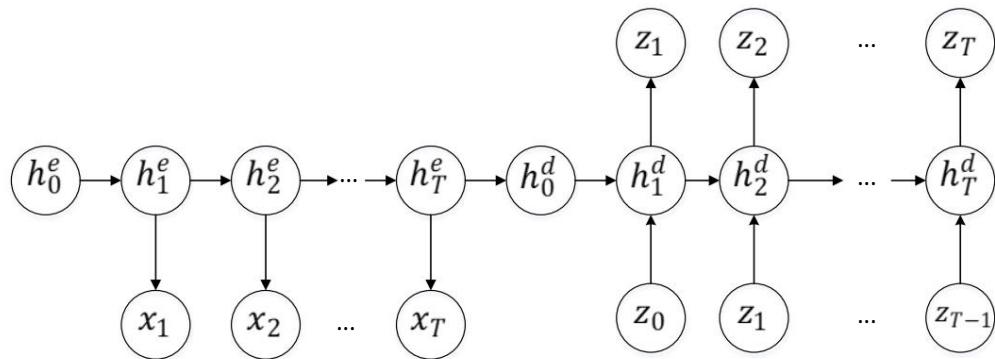
$$y_t = W_y h_t \quad (60)$$

حيث يمثل  $x$  المدخلات  $(x_1, \dots, x_T)$  ويمثل  $y$  المخرجات  $(y_1, \dots, y_T)$  يمكن إخراج توزيع متعدد الحدود (ترميز 1 من  $K$ ) باستخدام دالة التنشيط softmax:

$$\frac{\exp(w_j h_t)}{\sum_k \exp(w_k h_t)} \quad (61)$$

دمج هذه الاحتمالات، يمكننا حساب احتمال المتتالية  $x$  على النحو التالي:

$$\prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}) = \prod_{t=1}^T \exp(w_j h_t) \quad (62)$$



الشكل 7. بنية المشفر التلقائي المتكرر.

## 3.6. تضمين الجوار العشوائي الموزع (t-SNE)

تـSNE هي طريقة غير خطية لتقليل الأبعاد لتضمين البيانات عالية الأبعاد. تُستخدم هذه الطريقة بشكل شائع للتوصير في فضاء الميزات منخفض الأبعاد [64]. كما هو موضح في الشكل 8، يعتمد هذا النهج على G. Hinton [102] وـRoweis [64]. t-SNE تعمـل خوارزمية t-SNE عن طريق تحويل المسافات الإقليدية عالية الأبعاد إلى احتمالات شرطية تمثل أوجه التشابه. يتم حساب الاحتمال الشرطي  $p_{j|i}$  كما يلي:

$$p_{j|i} = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})}{\sum_{k=1}^n \exp(-\frac{\|x_i - x_k\|^2}{2\sigma^2})} \quad (63)$$

حيث  $\sigma$  هو تباين البيانات المتمركزة حول نقطة البيانات  $x_i$ . ويتم حساب تشابه زامع  $z$  كما يلي:

$$\exp(-\frac{\|y_i - y_j\|^2}{2\sigma^2}) \quad (64)$$

دالة التكلفة  $C$  هي كما يلي:

$$C = \sum_{i=1}^n \sum_{j=1}^n p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (65)$$

حيث  $KL(P_i | Q_i)$  هو تباعد كولباك-لايبير [103] والذي يتم حسابه على النحو التالي:

$$KL(P_i | Q_i) = \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (66)$$

يكون تحديث التدرج مع حد الزخم كما يلي:

$$\frac{\partial C}{\partial q_{j|i}} = \frac{(t^{j|i} - q_{j|i})}{q_{j|i}} \quad (67)$$

للشـعـلـمـ (t)، يـكـلـنـ بـأـعـدـلـ لـتـكـرـارـةـ، عـلـىـ كـلـلـمـتـبـاـدـلـاـتـ، فـيـ الـغـفـمـ لـهـمـاـتـ الـلـيـكـلـلـيـجـيـوـتـوزـعـهـ لـلـلـعـلـمـ، وـهـوـ مـفـعـلـ لـلـفـضـاءـ مـنـخـفـصـ الأـبـعـادـ عـلـىـ النـحـوـ التـالـيـ: [64]

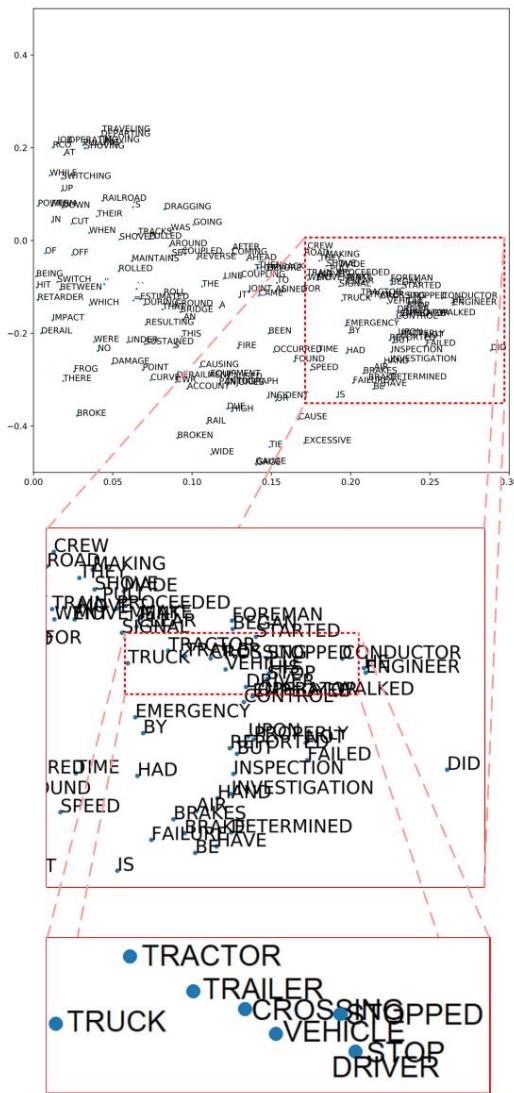
$$C = KL(P || Q) = \sum_{i=1}^n \sum_{j=1}^n p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (68)$$

في الفضاء عالي الأبعاد، يكون  $\pi$  هو:

$$= \frac{2 \exp(-\sigma^2)}{\|x - \bar{x}\|} \quad (69)$$

يكون تدرج  $\Delta$ المتاظر كما يلي:

$$-\sum_j \frac{\delta C}{\zeta} (y_i - y_j) \delta y_i = 4 \square \quad (70)$$



الشكل 8. يعرض هذا الشكل تصور تضمين الجوار العشوائي الموزع (t-SNE) لمجموعة بيانات إدارة السكك الحديدية الفيدرالية (FRA).

#### 4. تقنيات التصنيف الحالية.

في هذا القسم، نستعرض خوارزميات تصنيف النصوص والوثائق الحالية. نبدأ بوصف خوارزمية روكيو المستخدمة لتصنيف النصوص. ثم نتناول تقنيتين شائعتين في خوارزميات التعلم الجماعي: التعزيز والتجميع. بعض الطرق، مثل الانحدار логистي،

تُعد خوارزمية بايز البسيطة وخوارزمية أقرب جار من الخوارزميات التقليدية، لكنها لا تزال شائعة الاستخدام في الأوساط العلمية، كما تُستخدم آلات المتجهات الداعمة، (SVMs) وخاصةً آلات المتجهات الداعمة ذات النواة، على نطاق واسع كتقنية تصنيف، وتتميز خوارزميات التصنيف القائمة على الأشجار، مثل شجرة القرار والغابات العشوائية، بالسرعة والدقة في تصنيف المستندات، وتناول أيضًا خوارزميات الشبكات العصبية، مثل الشبكات العصبية الالتفافية، (CNN) والشبكات العصبية المتكررة، (RNN)، وشبكات الاعتقاد العميق، (DBN) وشبكات الانتباه الهرمية، (HAN) بالإضافة إلى تقنيات الدمج.

#### 4.1. تصنیف روکیو

ظهرت خوارزمية روکیو لأول مرة على يد جيروکی روکیو [104] عام 1971 كطريقة لاستخدام التغذية الراجعة للملاعنة في الاستعلام عن قواعد بيانات النصوص الكاملة. ومنذ ذلك الحين، تناول العديد من الباحثين هذه التقنية وطوروها لتصنيف النصوص والوثائق [105, 106]. تستخدم خوارزمية التصنيف هذه أوزان IDF-IDF لكل كلمة ذات دلالة بدلًا من الميزات المنطقية. باستخدام مجموعة تدريبية من الوثائق، تُنشئ خوارزمية روکیو متجهاً أولياً لكل فئة. هذا المتجه الأولي هو متوسط متجهات وثائق التدريب التي تتبع إلى فئة معينة. ثم تُسند كل وثيقة اختبار إلى الفئة ذات أعلى تشابه بين وثيقة الاختبار وكل متجه من المتجهات الأولية. [107] [بحسب المتجه المتوسط مركز ثقل الفئة = (مركز كتلة عناصرها)].

$$\frac{1}{\sum_{d \in D} |C_d|} \sum_{d \in D} C_d = v_d \quad (71)$$

حيث  $D$  هي مجموعة المستندات في  $\mathcal{D}$  التي تتبع إلى الفئة  $C$ ،  $v_d$  هو التمثيل المتجهي الموزون للمستند  $d$ . التصنيف المتوقع للمستند  $d$  هو التصنيف الذي يحقق أصغر مسافة إقليلية بين المستند  $d$  ومراكزه.

$$C^* = \arg \min_C \mu_C \cdot v_d \quad (72)$$

يمكن تطبيق مراكز الثقل إلى وحدة الطول كما يلي:

$$\mu_C = \frac{v_d \cdot d \cdot v_d}{\|d\|_2^2} \quad (73)$$

وبالتالي، يمكن الحصول على تصنیف وثائق الاختبار على النحو التالي:

$$C^* = \arg \min_C \mu_C \cdot v_d \quad (74)$$

#### 4.2. تعزيز والتجميع

لقد ظهرت بنجاح تقنيات تصنیف التصویت، مثل التجمیع والتعزیز، لتصنیف مجموعات بيانات المستندات والنصوص. [110] في حين أن التعزیز يُغير توزیع مجموعه التدربی بشکل تکیفی بناءً على أداء المصنفات السابقة، فإن التجمیع لا يعتمد على المصنف السابق. [111]

##### 4.2.1. تعزیز

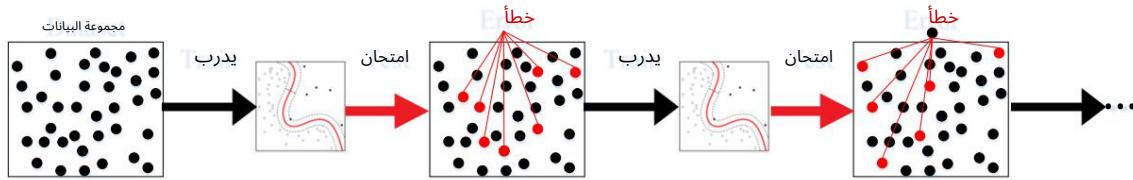
تم تقديم خوارزمية التعزیز لأول مرة بواسطة آر إي شاپیر [112] في عام 1990 كتقنية لتعزیز أداء خوارزمية التعلم الضعیف. وقد تم تطوير هذه التقنية لاحقًا بواسطة فرونڈ. [113, 114]

يوضح الشكل 9 كيفية عمل خوارزمية التعزيز لمجموعات البيانات ثنائية الأبعاد، حيث قمنا بتصنيف البيانات ثم تدريبيها باستخدام بنى متعددة النماذج (التعلم الجماعي). وقد أدت هذه التطورات إلى ظهور خوارزمية AdaBoost (التعزيز التكيفي). [115] لنفترض أننا أنشأنا  $Dt$  بحيث يكون  $h_t = \frac{1}{Z_t} \sum_{i=1}^{D_t} z_i h_t(x_i)$

$$= \frac{\prod_{i=1}^{D_t} \exp(-y_i h_t(x_i))}{Z_t} \quad (75)$$

حيث يشير  $Z_t$  إلى عامل التطبيع، و  $z_i$  كما يلي:

$$z_i = \frac{\ln 2}{1 + e^{-y_i h_t(x_i)}} \quad (76)$$



الشكل 9. هذا الشكل هو بنية تقنية التعزيز.

كما هو موضح في الخوارزمية، نستخدم مجموعة التدريب  $S$  بحجم  $m$ ، والمُحقّق  $C$  والعدد الصحيح  $N$  كمدخلات. ثم تجد هذه الخوارزمية أوزان كل  $x_i$  وأخيراً، يكون الناتج هو المصنف الأفضل  $(C)$ .

#### الخوارزمية 1: طريقة AdaBoost. المدخلات : مجموعة التدريب $S$ بحجم $m$ ، والمُحقّق $C$ والعدد الصحيح $N$ .

```

لكل 1 = إلى N نفذ
Ci = τ(S)

    لو
        = 1
        xj ∈ S ; Ci(xj) ≠ yi (x)
        الوزن(yi(xj)) = الوزن(xj) * Ci(xj)
        < قم بتعيين 'كمينة تمهدية من' Ci(xj) = 1
        الجميع الحالات وانتقل إلى أعلى
    endif
    endfor
    endfor
    قم بتطبيع أوزان الحالات

```

$$C_i = \frac{1}{\sum_j w_j C_i(x_j)}$$

$$\text{الناتج: المصنف } C^* = \arg \max_{C^*} \sum_i y_i C_i(x_i)$$

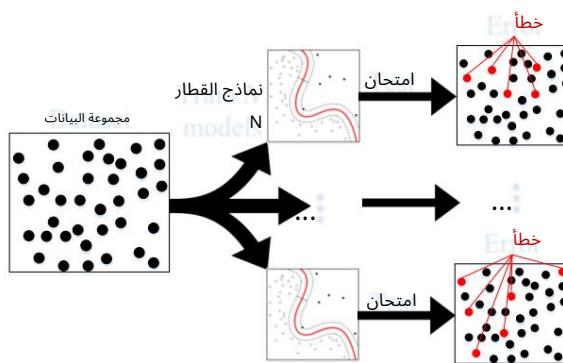
يمكن كتابة الصيغة النهائية للمصنف على النحو التالي:

$$H_f(x) = \text{sign}(\alpha \cdot h_t(x)) \quad (77)$$

#### 4.2.2. التعبئة

تم تقديم خوارزمية التجميع بواسطة [Breiman, 1996] في عام 1996 كطريقة تصنيف تصويبية. تُنشأ الخوارزمية باستخدام عينات بوتستراب مختلفة [111]. تُولد عينة البوتستراب عينة موحدة من مجموعة التدريب. إذا تم توليد  $N$  عينة بوتستراب،  $CN$  ....  $B1, B2, ..., BN$  فسيحصل على  $N$  مصنف  $(C_i)$  حيث يُبني المصنف  $C_i$  من كل عينة بوتستراب  $B_i$ . في النهاية، يحتوي المصنف  $C$  على المصنفات  $C1, C2, ..., CN$  أو يُولد منها ، ويكون ناتجه هو الفئة التي تنبأت بها المصنفات الفرعية بشكل متكرر، مع حل حالات التعادل بشكل عشوائي [111, 116]. يوضح الشكل 10 خوارزمية تجميع بسيطة درجت  $N$  موجهاً.

كما هو موضح في الخوارزمية ، لدينا مجموعة تدريب  $S$  التي تم تدريبيها ونجد أفضل مصنف  $C$ .



الشكل 10. يوضح هذا الشكل نموذجاً بسيطاً لتقنية التعبئة.

---

#### الخوارزمية: 2. التجميع

---

المدخلات : مجموعة التدريب  $S$ ، والمحفز  $\tau$ ، والعدد الصحيح  $N$

لكل  $i = 1$  إلى  $N$  نفذ

عينة بوتستراب من  $S$  =  
 $C_i = \tau(S)$  endfor

$$C(x) = \arg \max_{i=1}^N y_i$$

---

\* الناتج: المصنف  $C$

---

#### 4.2.3. قيود التعزيز والتجميع

كما أن لأساليب التعزيز والتجميع العديد من القيود والعيوب، مثل التعقيد الحسابي وفقدان قابلية التفسير [117]. مما يعني أنه لا يمكن اكتشاف أهمية الميزات بواسطة هذه النماذج.

#### 4.3. الانحدار اللوجستي

يُعد الانحدار اللوجستي (LR) من أقدم أساليب التصنيف. وقد طوره الإحصائي ديفيد كوكس عام 1958 [118]. الانحدار اللوجستي هو مصنف خطى ذو حدّ فاصل  $Tx = 0$ . يتبين الانحدار اللوجستي بالاحتمالات بدلاً من تحديد الفئات [119, 120].

## الإطار الأساسي 4.3.1.

يهدف الانحدار اللوجستي إلى التدريب بناءً على احتمالية أن تكون قيمة المتغير  $z$  إما 0 أو 1 بمعلومية. لنفترض أن لدينا بيانات نصية  $\mathbf{X} \in \mathbb{R}^{n \times d}$  في حالة وجود مسائل تصنيف ثنائية، يجب استخدام دالة نموذج خليط برنولي [121] كما يلي:

$$\begin{aligned} L(\theta | \mathbf{x}) &= p(y | \mathbf{x}; \theta) = \\ &\prod_{i=1}^n \beta(y_i | \text{sigm}(x_i \theta)) = \\ &\prod_{i=1}^n \text{sigm}(y_i(x_i \theta)) = \\ &\prod_{i=1}^n \frac{1}{1 + e^{-x_i \theta}} \end{aligned} \quad (78)$$

(78) هي دالة سigmoid يتم تعريفها كما هو موضح في حيث  $\theta_0 + \sum_{j=1}^d x_j \theta_j$ .

$$\sigma(\eta) = \frac{1}{1 + e^{-\eta}} = \frac{e^\eta}{1 + e^\eta} \quad (79)$$

## 4.3.2. دمج التعلم القائم على الحالات والتعلم الانحداري

يحدد نموذج الانحدار اللوجستي احتمالية المخرج الثنائي  $\{0, 1\} = \{y_i\}$  بالنظر إلى المدخل  $\mathbf{x}_i$ . اعتبار الاحتمال اللاحق كما يلي:

$$\pi_0 = P(y_0 = +1) \quad (80)$$

أين:

$$\frac{\pi_0}{1 - \pi_0} = \frac{+1(0 | \text{ص.} 0)}{+1(0 | \text{ص.} 0)} \cdot \frac{1 - \pi_0}{\pi_0} \quad (81)$$

حيث  $p$  هي نسبة الاحتمال، ويمكن إعادة كتابتها على النحو التالي:

$$\frac{1 - \pi_0}{\pi_0} = \frac{\text{ص.} 0}{\text{ص.} 0} \quad (82)$$

$$\text{سجل } \frac{\pi_0}{1 - \pi_0} = \log(p) + w_0 \quad (83)$$

بالنسبة إلى:

$$w_0 = \log(p_0) - \log(1 - p_0) \quad (84)$$

للمثال للبدأ الأساسي للتعلم القائم على الحالات [122]. يجب أن يكون المصنف دالة المسافة  $\delta$ ، حيث  $y_i = 1$  إذا كانت  $\delta_i$  كبيرة إذا كانت  $0 < \delta_i < 1$ ، وصغيرة عندما  $\delta_i > 1$ . يجب أن تكون قيمة  $w_0$  قريبة من 0 إذا كانت  $\delta_i$  صغيراً، لن تكون القيمة  $+1$  وبالتالي تكون الدالة المعلمقة كما يلي:

$$p = p(\delta) = \exp \left( \frac{1 - \pi_0}{\pi_0} \right) \quad (85)$$

أخيراً،

$$\text{سجل } \frac{1 - \pi_0}{\pi_0} = w_0 + \alpha \cdot \frac{k(x_i) - N(x_i)}{x_i} \quad (86)$$

حيث  $k(x_i)$  هو مقاييس التشابه.

#### 4.3.3. الانحدار اللوجستي متعدد الحدود

يستخدم التصنيف اللوجستي متعدد الحدود (أو متعدد التصنيفات) [123] احتمال انتقاء  $x_i$  إلى الفئة (كما هو محدد في المعادلة (87))

$$(i) = 1 \mid x, \theta = p y \frac{\sum_{j=1}^n \exp(\theta_j)}{\sum_{j=1}^m \exp(\theta_j)} \quad (87)$$

حيث  $(i) \theta$  هو متجه الوزن المقابل للفئة  $i$ .

بالنسبة للتصنيف الثنائي ( $m=2$ ) والذى يُعرف باسم الانحدار اللوجستي الأساسي، ولكن بالنسبة للانحدار اللوجستي متعدد الحدود ( $m > 2$ ) فإنه يستخدم عادة دالة  $f(t)$ ، حيث  $t = \max(x_i)$ . دالة التطبيع هي:

$$(i) = 1 \mid x, \theta = 1 \prod_{j=1}^m p_j \quad (88)$$

في مهمة التصنيف ضمن سياق التعلم الخاضع للإشراف، تُحسب قيمة  $\theta$  من مجموعة فرعية من بيانات التدريب  $D$  التي تنتمي إلى الفئة أحى  $\{1, \dots, n\}$  أو لإجراء تقدير الاحتمال الأقصى (ML)، تحتاج إلى تعطيم دالة الاحتمال اللوغاريتمي كما يلي:

$$\begin{aligned} & \text{أنا} \\ & \text{أنا} \ln(p_i) = \ln(y_i) - \ln(1 - y_i) \\ & = \sum_{j=1}^n \sum_{i=1}^m \ln(p_i) \ln(\exp(\theta_i) / \exp(\theta_j)) x_{ij} \end{aligned} \quad (89)$$

اعتماد تقديرات الاحتمال اللاحق الأقصى (MAP) على النحو التالي:

$$\text{أنا} \ln(p_i) = \ln(p_i) + \ln(\exp(\theta_i)) \quad (90)$$

#### 4.3.4. قيود الانحدار اللوجستي

يعد مصنف الانحدار اللوجستي فعالاً في التنبؤ بالنتائج الفئوية. مع ذلك، يتطلب هذا التنبؤ أن تكون كل نقطة بيانات مستقلة. [124] وهو ما يحاول التنبؤ بالنتائج بناءً على مجموعة من المتغيرات المستقلة. [125]

#### 4.4. مصنف بايز الساذج

يُستخدم تصنيف النصوص باستخدام خوارزمية بايز البسيطة على نطاق واسع في مهام تصنيف المستندات منذ خمسينيات القرن الماضي. [126, 127] واستند هذه الخوارزمية نظرياً إلى نظرية بايز، التي وضعها توماس بايز بين عامي 1701 و 1761. [128, 129] وقد تناولت الدراسات الحديثة هذه التقنية على نطاق واسع في مجال استرجاع المعلومات. [130] تعد هذه التقنية نموذجاً توليدياً، وهي الطريقة الأكثر تقليدية لتصنيف النصوص. نبدأ بأسط نسخة من خوارزمية بايز البسيطة، والتي ظهرت باستخدام نموذج حقيقة الكلمات (TF) وهي تقنية لاستخراج الميزات تعتمد على حساب عدد الكلمات في المستندات.

#### 4.4.1. وصف عالي المستوى لمصنف بايز الساذج

إذا كان عدد المستندات ( $n$ ) يندرج ضمن  $k$  فئة حيث  $\{c_1, c_2, \dots, c_k\}$  فإن الفئة المتوقعة كنتاج هي  $C$ . ويمكن وصف خوارزمية بايز الساذجة على النحو التالي:

$$\frac{P(c_i | d)}{P(c_i)} = \frac{P(d | c_i) P(c_i)}{P(d)}$$

حيث يمثل  $d$  المستند و  $c$  يشير إلى الفئة.

$$\text{CMAP P}(\text{from} | c) = \arg \max_c P(c | x_1, x_2, \dots, x_n) \quad (92)$$

يُستخدم هذا النموذج كأساس للعديد من الأوراق البحثية، وهو مصنف Naïve Bayes على مستوى الكلمات [3, 131] كما يلي:

$$\frac{P(c_j | \theta)}{P(j | \theta)} = \frac{\prod_i P(w_i | c_j; \theta)}{\prod_i P(w_i | \theta)}$$

#### 4.4.2. مصنف بايز الساذج متعدد الحدود

إذا كان عدد المستندات ( $n$ ) يندرج ضمن  $\{c_1, c_2, \dots, c_k\}$  كفأان الفئة المتوقعة هي الناتج هو  $C$ . يمكن كتابة خوارزمية بايز الساذجة على النحو التالي:

$$\frac{P(c) \prod w_i d_i P(d | c) n_{wd}}{P(c | d)} = \frac{P(c) \prod w_i d_i}{P(d)} \quad (94)$$

حيث يشير  $d_i$  إلى عدد مرات ظهور الكلمة  $w_i$  في المستند، و  $P(w_i | c)$  هو احتمال ملاحظة الكلمة  $w_i$  بالنظر إلى الفئة.

$$\begin{aligned} &\text{ يتم حساب } P(w_i | c) \text{ على النحو التالي:} \\ &P(w_i | c) = k + \frac{n_{wd}}{n_w + n_d} \end{aligned} \quad (95)$$

#### 4.4.3. مصنف بايز الساذج للفئات غير المتوازنة

من عيوب خوارزمية NBC ضعف أدائها علىمجموعات البيانات ذات الفئات غير المتوازنة. وقد طور إيببي فرانك وريمكو ر. بوكيرت [132] طريقة لتطبيع كل فئة باستخدام المعادلة ، (96) ثم استخدامها مُصنف المركز [22] في خوارزمية NBC للفئات غير المتوازنة. ويعطى مركز الفئة  $C$  في المعادلة (97).

$$\frac{x}{\sum w_i d_i D_C} = \frac{\text{قمح}}{\sum w_i d_i D_C} \quad (96)$$

$$\begin{aligned} C_C = & \frac{\sum w_i d_i D_C n_{wd2}}{\sum n_{wd2} d_i D_C n_{wd}} \dots \\ & \frac{\sum n_{wid} w_i d_i D_C}{\sum d_i D_C n_{wd2} d_i D_C} \dots \\ & \frac{\sum d_i D_C n_{wd2}}{\sum d_i D_C n_{w1d} w_i} \end{aligned} \quad (97)$$

تُعرف دالة التسجيل على النحو التالي:

$$x_d \cdot c_1 + x_d \cdot c_2 \quad (98)$$

لذا يمكن حساب لوغاريتم مصنف بايز الساذج متعدد الحدود على النحو التالي:

$$\log P(c_1) + \sum_{i=1}^k n_{wid} \log(P(w_i | c_1)) - \log \sum_{i=1}^k n_{wid} \log(P(w_i | c_2)) \quad (99)$$

باستخدام المعادلين (95) و (96) وإذا كانت  $a = 1$  ، فيمكننا إعادة كتابة ما يلي:

$$\frac{\sum d_i D_C}{\sum w_i D_C} = \frac{\text{الناتج}}{\text{الكل}} \quad (100)$$

بالنسبة إلى:

$$\frac{\partial \text{log}(\text{P}(C|x))}{\partial w_i} < 1 \quad (101)$$

بالنسبة لمجموعات البيانات النصية، يكون  $x$  في هذه التقنية من  $\text{NBC}$  التجريبية [22] النتائج مشابهة جدًا لمصنف المركز.

#### 4.4.4. قيود خوارزمية بايز البسيطة

يعاني خوارزمية بايز الساذجة من عدّة قيود. فهي تفترض افتراضًا قويًا حول شكل توزيع البيانات [134, 135]، كما أنها محدودة بندرة البيانات، حيث يتطلب تقدير قيمة الاحتمالية لأي قيمة محتملة في فضاء الميزات استخدام خوارزمية إحصائية تكرارية [136].

#### 4.5. أقرب جار

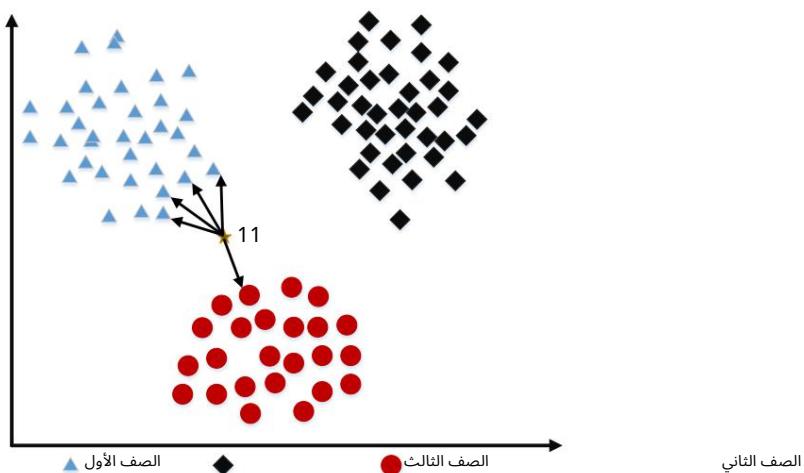
خوارزمية أقرب الجيران (KNN) هي تقنية غير بارامترية تستخدم للتصنيف. تُستخدم هذه الطريقة لتطبيقات تصنیف النصوص في العديد من المجالات البحثية [137] في العقود الماضية.

##### 4.5.1. المفهوم الأساسي لخوارزمية أقرب جار (KNN)

بالنظر إلى مستند اختبار  $x$ ، فإن خوارزمية KNN تجد أقرب جار له من بين جميع المستندات في مجموعة التدريب، وتقوم بتقييم المرشحين للفئة بناءً على فئة جار. يمكن أن تكون درجة تشابه المستند  $x$  مع كل مستند مجاور هي درجة تشابه المستندات المجاورة. قد تنتهي عدة مستندات من خوارزمية KNN إلى نفس الفئة؛ في هذه الحالة، يكون مجموع هذه الدرجات هو درجة تشابه الفئة  $k$  المستند. وبعد ترتيب قيم الدرجات، تُعين الخوارزمية المرشح إلى الفئة الحاصلة على أعلى درجة من المستند [137]. يوضح الشكل 11 بنية خوارزمية KNN، ولكن لنبسيط الأمور، تم تصميم هذا الشكل باستخدام مجموعة بيانات ثنائية الأبعاد (مشابهة لمجموعة بيانات النصوص، ولكن ببعاد أعلى). قاعدة القرار في خوارزمية KNN هي:

$$f(x) = \underset{j}{\operatorname{argmax}} \sum_{i \in KNN} \text{sim}(x, di)y(di, Cj) \quad (102)$$

حيث يشير  $C_j$  إلى قيمة النتيجة بالنسبة إلى  $j$ ، وهي قيمة النتيجة للمرشح إلى فئة  $j$ ، ويكون ناتج  $f(x)$  عبارة عن تسمية لمستند مجموعة الاختبار.



الشكل 11. بنية نموذج أقرب جار (KNN) للمجموعة البيانات ثنائية الأبعاد وثلاث فئات.

#### 4.5.2. تصنيف أقرب جار Kالمعدل حسب الوزن.

يُعد تصنيف الجوار الأقرب kالمعدل بالأوزان (WAKNN) نسخة من تصنيف الجوار الأقرب (KNN) والتي تحاول تعلم متوجهات الأوزان للتصنيف. [138] ويتم حساب مقياس جيب التمام الموزون [139] كما يلي:

$$\cos(x, y, w) = \frac{\sum_{t \in T} (x_t \times w_t) \times (y_t \times w_t)}{\sqrt{\sum_{t \in T} (x_t \times w_t)^2} \times \sqrt{\sum_{t \in T} (y_t \times w_t)^2}} \quad (103)$$

حيث تشير T إلى مجموعة الكلمات، و  $x_t$  و  $y_t$  هما، كما نوشط في القسم 2 بالنسبة لنموذج التدريب. ليكن  $\{d \in D\}$  مجموعه أقرب kJar لـ  $d$ . وبمعرفة  $N_d$  يُعرف مجموع التشابه بين  $d$  جاراً يتبع إلى الفئة  $c$  على النحو التالي:

$$Sc = \frac{\sum_{i=1}^{N_d} \cos(d, n_i, w)}{N_d} \quad (104)$$

يتم حساب التشابه الكلي على النحو التالي:

$$T = \frac{\sum_{c=1}^C Sc}{C} \quad (105)$$

يتم تعريف مساهمة  $d$  بدالة  $Sc$  من الفئتين  $c$  و  $T$  على النحو التالي:

$$Sc = \begin{cases} 1 & \text{إذا كان } d \in C, \text{ فإن } c = \text{class}(d), \\ & \text{الفئة } c(d) \\ 0 & \text{و } c \neq \text{class}(d) \text{ خلاف ذلك} \end{cases} \quad (106)$$

حيث يرمز  $Sc$  إلى المساهمة  $(d)$

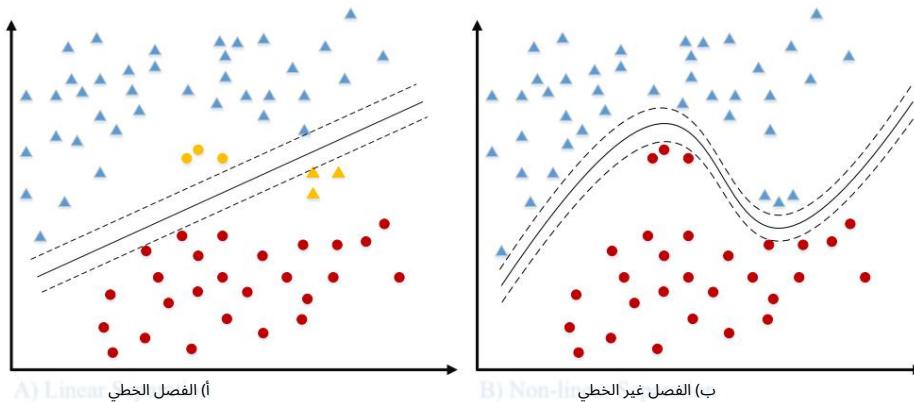
#### 4.5.3. قيود خوارزمية أقرب جار K.

KNN هي طريقة تصنيف سهلة التنفيذ وتتكيف مع أي نوع من مساحات الميزات. يتعامل هذا النموذج بشكل طبيعي مع حالات التصنيف المتعدد. [140, 141] مع ذلك، فإن خوارزمية أقرب الجيران (KNN) محدودة بسبب قيود تخزين البيانات في مسائل البحث الكبيرة للعثور على أقرب الجيران. إضافة إلى ذلك، يعتمد أداء خوارزمية أقرب الجيران على إيجاد دالة مسافة ذات دالة، مما يجعلها خوارزمية تعتمد بشكل كبير على البيانات. [142, 143]

#### 4.6. آلة المتوجهات الداعمة (SVM).

تم تطوير النسخة الأصلية من SVM بواسطة فابنيك وشيرفونينكيس [144] في عام 1963.

قام بي إيه بوسر وأخرون [145] بتطوير هذه النسخة إلى صيغة غير خطية في أوائل التسعينيات. ضممت آلة المتوجهات الداعمة (SVM) في الأصل لمهام التصنيف الثنائي. مع ذلك، يعمل العديد من الباحثين على مشاكل التصنيف المتعدد باستخدام هذه التقنية السائدة. [146] يوضح الشكل 12 المصنف الخطى وغير الخطى المستخدم لمجموعات البيانات ثنائية الأبعاد.



الشكل 12. يوضح هذا الشكل نموذج آلة المتجهات الداعمة (SVM) الخطى وغير الخطى لمجموعة بيانات ثنائية الأبعاد (لديناآلاف الأبعاد لبيانات النصوص). اللون الأحمر يمثل الفئة ، والأزرق يمثل الفئة ، والأصفر يمثل نقاط البيانات المصنفة بشكل خاطئ.

#### آلة المتجهات الداعمة للفئة الثنائية

في سياق تنصيف النصوص، لنفترض أن  $x_1, x_2, \dots, x_n$  هي أمثلة تدريبية تتبع إلى فئة واحدة، حيث  $X$  هي مجموعة جزئية مضغوطة من  $\mathbb{R}^n$ . ثم يمكننا صياغة مصنف ثانوي على النحو التالي:

$$\sum_{i=1}^n w_i x_i + b = 0 \quad (107)$$

رهنًا بما يلي:

$$f(x) = \text{sign}(w \cdot \Phi(x)) - p \quad i = 1, 2, \dots, n \quad (108)$$

إذا كان  $w$  و  $p$  يحلان هذه المشكلة، فإن دالة القرار تُعطى بالصيغة التالية:

$$f(x) = \text{sign}(w \cdot \Phi(x)) - p \quad (109)$$

#### آلة المتجهات الداعمة متعددة الفئات

بما أن آلات المتجهات الداعمة (SVMs) تُستخدم تقليديًا للتصنيف الثنائي، فنحن بحاجة إلى إنشاء آلة متجهات داعمة متعددة [147] (MSVM) لمشاكل التصنيف المتعدد. تُعد تقنية "واحد ضد واحد" (One-vs-One) إحدى تقنيات آلات المتجهات الداعمة متعددة التصنيفات، حيث تُنشئ  $(N \times N)$  مصنفًا على النحو التالي:

$$f(x) = \max_{k=1}^K f_k(x) \quad (110)$$

الطريقة الطبيعية لحل مشكلة الفئات  $K$  هي بناء دالة قرار لجميع الفئات  $K$  عند مرة واحدة [148, 149]. بشكل عام، تُعتبر آلة المتجهات الداعمة متعددة الفئات مشكلة تحسين على النحو التالي:

$$\min_{w_1, w_2, \dots, w_K, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^K \xi_i \quad (111)$$

$$\text{st. } w_k \cdot x_i - b - \xi_i \geq 1, \quad (xi, yi) \in D, k \in \{1, 2, \dots, K\}, k = yi \quad (112)$$

حيث يمثل  $(xi, yi)$  نقاط بيانات التدريب بحيث يكون  $w_k \cdot xi - b$  هو معامل الجزء، و  $\xi_i$  هو معامل التراخي، و  $k$  يرمز إلى الفئة.

تُعد تقنية "الكل مقابل واحد" إحدى تقنيات التصنيف متعدد الفئات باستخدام خوارزمية SVM. ويعتمد استخراج الميزات عبر SVM عمومًا على إحدى طريقتين: استخراج ميزات تسلسلات الكلمات [150] و-IDF-FT-. لكن بالنسبة لـ

في حالة التسلسلات غير المنظمة مثل تسلسلات الحمض النووي الريبي (RNA) والحمض النووي (DNA)، تُستخدم نواة السلسلة. ومع ذلك، يمكن استخدام نواة السلسلة لتصنيف المستندات. [151]

#### نواة السلسلة 4.6.3.

تمت دراسة تصنيف النصوص أيضًا باستخدام نواة السلسلة. [151] الفكرة الأساسية للسلسلة تستخدم النواة ( $\Phi$ ) لرسم السلسلة في مساحة الميزات. تم تطبيق نواة الطيف كجزء من SK في العديد من التطبيقات المختلفة، بما في ذلك تصنیف النصوص والحمض النووي والبروتينات [152]. وتمثل الفكرة الأساسية لنواة الطيف في حساب عدد مرات ظهور الكلمة ما في السلسلة  $\Delta K$  خريطة مميزة، حيث يتم تعريف خرائط الميزات من  $x$  إلى  $R$ .

$$\Phi_k(x) = \Phi_j(x) j \otimes \sum_k \text{أين} \quad (113)$$

$$x = \Phi_j(\text{عدد العناصر زالتي تظهر في}) \quad (114)$$

يتم توليد خريطة الميزات ( $\Phi_i(x)$ ) بواسطة التسلسل  $x$  ويتم تعريف النواة على النحو التالي:

$$F = \Sigma \quad (116)$$

$$K_{ij}(x) = \langle \Phi_i(x), \Phi_j(x) \rangle \quad (115)$$

يتمثل القيد الرئيسي لـ SVM عند تطبيقه على تصنیف تسلسل السلسلة في تعقید الوقت. [154] يتم توليد الميزات باستخدام حجم القاموس،  $F$  هو عدد الميزات، وهو محدود بالمعادلة. [115] وتشابه عملية حساب النواة مع  $SP$ ، حيث تستخدم المعادلة. [116] ثم يتم تطبيق النواة باستخدام المعادلة. [117]

$$K_{ij}(x, y) = \frac{K(x, y)}{\sqrt{K(x, x) K(y, y)}} \quad (117)$$

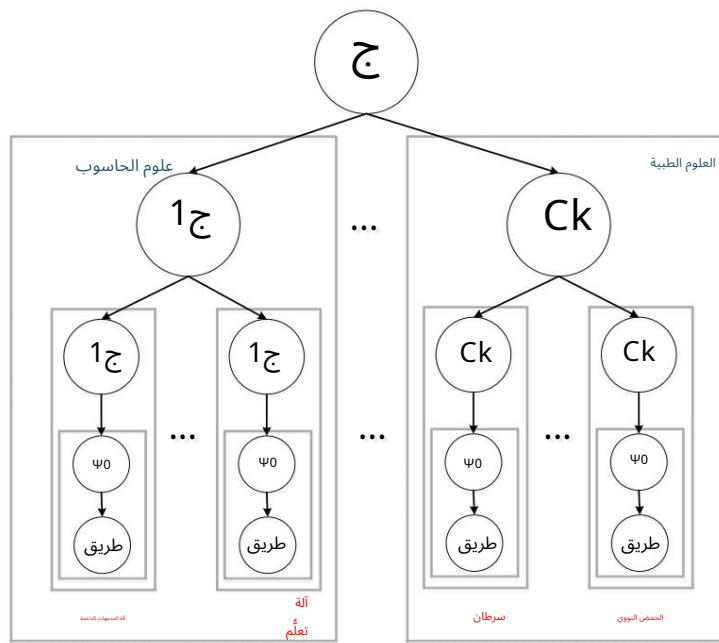
$$\langle f^x | y \rangle = \sum_{i=1}^n \sum_{j=1}^m s_i^1 s_j^2 u_i h_j(u) \quad (118)$$

حيث يوجد متباينان،  $u$

#### آلة المتجهات الداعمة المكدسة (SVM) 4.6.4.

تُعد خوارزمية SVM المكدسة طريقة تصنیف هرمية تُستخدم لهيكل شجرة الفئات، وتعتمد على منهجية تنازيلية قائمة على المستويات. [155] توفر هذه التقنية نموذجًا هرميًّا لمصنفات SVM الفردية، وبالتالي تُنتج عمومًا نتائج أكثر دقة من نماذج SVM الفردية. [156]

كما هو موضح في الشكل 13، يستخدم نموذج التجميع مصنفًا هرميًّا يحتوي على عدة طبقات (في هذا الشكل لدينا مستويات مثل المجال الرئيسي والمجالات الفرعية).



الشكل .13 طريقة التصنيف الهرمي.

#### 4.6.5. التعلم متعدد الحالات (MIL).

يُعد التعلم المتعدد الحالات (MIL) أسلوباً للتعلم الخاضع للإشراف [157] ويتم صياغته عادةً كإحدى طريقتين تعتمدان على خوارزمية SVM (mi-SVM) [158]. تأخذ MIL مجموعة من الأكياس المصنفة كالمدخلات بدلاً من الحالات. تُصنف الحقيقة على أنها موجبة إذا كان هناك حالة واحدة على الأقل فيها ذات قيمة موجبة. يُصنف المفهوم على أنه سلبي، ويُصنف على أنه سلبي إذا كانت جميع حالاته سلبية. بعد ذلك، يحاول المتعلم استنتاج المفهوم، التي تصنف الحالات الفردية بشكل صحيح [157] في التعرف الإحصائي على الأنماط، يفترض أن

تتوفر بطبعها ملحوظة من الأنماط المصنفة حيث كل زوج  $(x_i, y_i)$

توزيع غير معروف بشكل مستقل. الهدف هو إيجاد مصنف من الأنماط إلى التصنيفات، أي

$\hat{y}_i$  في لغة MIL، تفترض الخوارزمية أن المدخلات متاحة كمجموعة من أنماط الإدخال  $x_1, \dots, x_n$ .

تجمع في حفائب  $B_m, \dots, B_1$  حيث  $B_i = \{x_i : y_i = 1\}, \dots, \{x_i : y_i = -1\}$ . كل حقيقة

يرتبط بالعلامة  $Y_I$  حيث  $Y_I = 1$  إذا كانت  $y_i = 1$  أو  $Y_I = -1$  إذا كان هناك واحد على الأقل

المثال  $B_I$  ذو التصنيف الموجب [158] العلاقة بين تصنيفات الأمثلة  $y_i$  وتصنيفات الحقيقة  $Y_I$

يمكن التعبير عنها على النحو التالي:  $y_i = \max_{I \in B_I} Y_I$  أو مجموعة من القيود الخطية:

$$\begin{aligned} & \text{do } \boxed{1} \\ & \quad i \in I \\ & \quad \text{أنا سرت بي } = 1 \\ & \quad y_i = 1, \text{ st } Y_I = 1. \end{aligned} \tag{119}$$

تُسمى دالة التمييز  $f(x)$  دالة فصل متعددة الحالات بالنسبة لحالة متعددة مجموعة البيانات إذا كانت  $y_I = \max_{I \in B_I} f(x_i)$  لجميع الحفائب التي تحملها.

#### 4.6.6. قيود آلة المتجهات الداعمة (SVM).

تعتبر خوارزمية SVM واحدة من أكثر خوارزميات التعلم الآلي كفاءة منذ ظهورها في تسعينيات القرن العشرين [159]. ومع ذلك، فإن خوارزميات آلة المتجهات الداعمة لتصنيف النصوص محدودة بسبب نقص الشفافية. في النتائج الناتجة عن كثرة الأبعاد، ولهذا السبب، لا يمكن عرض تقييم الشركة كـ لا تعتمد الدالة البارامترية على النسب المالية ولا على أي شكل دالي آخر [159] وهذا قيد آخر وهو معدل نسب مالية متغيرة [160].

## 4.7. شجرة القرار.

إحدى خوارزميات التصنيف السابقة لاستخراج النصوص والبيانات هي شجرة القرار. [161] تُستخدم مصنفات شجرة القرار (DTCs) بنجاح في العديد من المجالات المتنوعة للتصنيف. [162] يعتمد هيكل هذه التقنية على التفكير الهرمي لمساحة البيانات. [163] . [7] د. مورغان [164] الفكرة الرئيسية هي إنشاء شجرة بناءً على سمة نقاط البيانات المصنفة، لكن التحدي الرئيسي لشجرة القرار هو تحديد السمة أو الميزة التي يمكن أن تكون في مستوى الأباء وتلك التي يجب أن تكون في مستوى الأبناء. حل هذه المشكلة، قدم دي مانتاراس [165] النمذجة الإحصائية لاختيار الميزات في الشجرة. بالنسبة لمجموعة تدريب تحتوي على  $n$  من القيم الموجبة و  $m$  من القيم السالبة:

$$\text{اختر السمة } A \text{ ذات قيمة مميزة، وقم بتقسيم مجموعة التدريب } E \text{ إلى مجموعات فرعية من } \{E_1, E_2, \dots, E_k\}.$$

$$\text{قيمة الإنتروبيا المتوقعة (EH(A)) بعد تجربة السمة } A \text{ (مع الفروع: } i = 1, 2, \dots, k\text{):}$$

$$EH(A) = \sum_{i=1}^k \frac{p_i \cdot H(E_i)}{p_A}$$

$$(120)$$

اختر السمة ذات قيمة مميزة، وقم بتقسيم مجموعة التدريب  $E$  إلى مجموعات فرعية من  $\{E_1, E_2, \dots, E_k\}$ .  
 $i = 1, 2, \dots, k$ : مع الفروع

$$EH(A) = \sum_{i=1}^k \frac{p_i \cdot H(E_i)}{p_A}$$

$$(121)$$

إن اكتساب المعلومات ( $I$ ) أو انخفاض الإنتروبيا لهذه السمة هو:

$$I(A) = H - \sum_{i=1}^k \frac{p_i \cdot H(E_i)}{p_A}$$

$$(122)$$

اختر السمة ذات أكبر قدر من المعلومات كعقدة أصلية.

## قيود خوارزمية شجرة القرار

تُعد شجرة القرار خوارزمية سريعة جدًا للتعلم والتنبؤ على حد سواء؛ إلا أنها شديدة الحساسية للتغيرات الطفيفة في البيانات. [166] ويمكن أن تتعرض بسهولة للتجاوز. [167] يمكن التغلب على هذه الآثار باستخدام أساليب التحقق والتقليم، ولكن هذا الأمر لا يزال غير واضح المعالم. [166] كما يُعاني هذا النموذج من مشاكل في التنبؤ خارج نطاق العينة. [168]

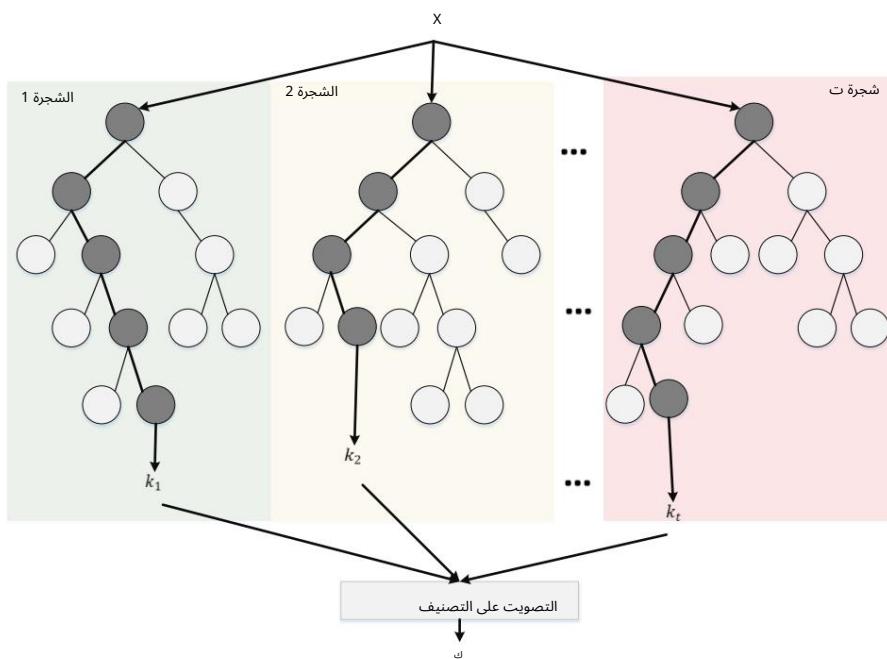
## 4.8. الغابة العشوائية.

تُعد تقنية الغابات العشوائية أو غابات القرار العشوائية أسلوبًا للتعلم الجماعي لتصنيف النصوص. وقد ظهر هذا الأسلوب، الذي يستخدم شجرة  $t$  كنموذج متوازن، بواسطة تي. كام هو [169] في عام 1995. كما هو موضح في الشكل 14، تمثل الفكرة الرئيسية لخوارزمية الغابات العشوائية في توليدأشجار قرار عشوائية. وقد ظهرت هذه التقنية بشكل أكبر في عام 1999 على يد ل. بريمان. [170] الذي وجد تقارباً لخوارزمية الغابات العشوائية كمقاييس هامشية ( $mg(X, Y)$  على النحو التالي):

$$mg(X, Y) = avk I(hk(X) = Y) - avk I(hk(X) \neq Y)$$

$$(123)$$

حيث يشير  $I(\cdot)$  إلى دالة المؤشر.



الشكل .14. الغابة العشوائية.

#### 4.8.1 التصويت

بعد تدريب جميع الأشجار كفالة، يتم تعين التنبؤات بناءً على التصويت [171] على النحو التالي:

$$\delta V = \arg \max_{ij} \sum_j I\{r_{ij} > r_{ji}\} \quad (124)$$

حيث

$$r_{ij} + r_{ji} = 1 \quad (125)$$

#### 4.8.2. قيود الغابات العشوائية

تتميز الغابات العشوائية (أي مجموعات أشجار القرار) بسرعة تدريبيها على مجموعات البيانات التصورية مقارنةً بتقنيات أخرى مثل التعلم العميق، ولكنها بطبيعةِ نسبيةٍ في توليد التنبؤات بعد التدريب [172]. لذا، لتحقيق بنية أسرع، يجب تقليل عدد الأشجار في الغابة، لأن زيادة عدد الأشجار يزيد من تعقيد الوقت في خطوة التنبؤ.

#### 4.9. الحقل العشوائي الشرطي (CRF)

نموذج الحقول العشوائية الشرطية (CRF) هو نموذج بياني غير موجه، كما هو موضح في الشكل .15. يُعد نموذج CRF في جوهره طريقةً تجمع بين مزایا التصنيف والتمندحة البيانية، إذ يجمع بين القدرة على نمذجة البيانات متعددة المتغيرات بإيجاز، والقدرة على الاستفادة من فضاء الميزات عالي الأبعاد للتنبؤ [173]. يُعد هذا النموذج فعالً للغاية مع البيانات التصورية نظرًا لكبر فضاء الميزات). يُحدد نموذج CRF الاحتمال الشرطي لتسلسل تصنيف 2 مُعطى بتسلسل ملاحظات ، أي  $P(Y|X)$ . يُمكن نموذج CRF دمج ميزات معقدة في تسلسل الملاحظات دون الإخلال بفرضية الاستقلال، وذلك من خلال نمذجة الاحتمال الشرطي لتسلسل التصنيف بدلاً من الاحتمال المشترك. [174, 175]. يستخدم جهد الزمرة (أي الرسم البياني الفرعي المتصل بالكامل) لحساب  $P(Y|X)$  بالنسبة لدالة الجهد لكل زمرة في الرسم البياني، فإن احتمال تكون متغير ما يُقابل حاصل ضرب سلسلة من دوال الجهد غير السالبة.

القيمة المحسوبة بواسطة كل دالة كامنة تعادل احتمالية المتغيرات في الزمرة المقابلة لتكوين معين [174] أي:

$$\frac{1}{\prod_{v \in \text{uniques}} P(v)} = \psi(c) \quad (126)$$

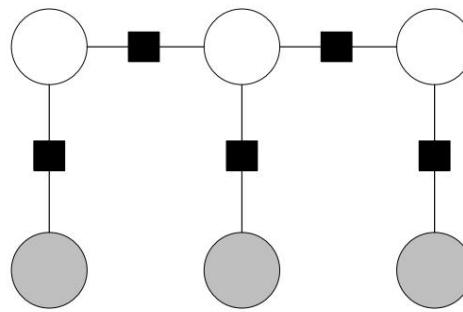
حيث  $Z$  هو حد التوحيد. يمكن صياغة الاحتمال الشرطي  $P(Y|X)$  على النحو التالي:

$$\frac{1}{\prod_{t=1}^n \psi(t, y_t | X)} \quad (127)$$

بالنظر إلى دالة الجهد، فإن الشرطي  $\psi(t, y_t | 1, y_t, X) = \exp(w \cdot f(t, y_t | 1, y_t, X))$ . يمكن إعادة كتابة الاحتمالية على النحو التالي:

$$X = \exp(w \cdot f(t, y_t | 1, \bar{y}_t P(X))) \quad (128)$$

حيث  $w$  هو متوجه الوزن المرتبط بمتوجه الميزات المحسوب بواسطة  $f$ .



الشكل 15. حقل عشوائي شرطي ذو سلسلة خطية (CRF). المربعات السوداء هي زمرة انتقالية

#### قيود نموذج الحقل العشوائي الشرطي (CRF)

فيما يتعلق بخوارزمية الحقول العشوائية الشرطية، فإن أبرز عيوبها هو التعقيد الحسابي العالي لخطوة التدريب، خاصةً مع مجموعات بيانات النصوص نظراً لكبر مساحة الميزات. علاوة على ذلك، لا تعمل هذه الخوارزمية بكفاءة مع الكلمات غير المرئية (أي الكلمات التي لم تكن موجودة في عينة بيانات التدريب) [177].

#### 4.10. التعلم العميق

حققت نماذج التعلم العميق نتائج متميزة في العديد من المجالات، بما في ذلك مجموعة واسعة من تطبيقات معالجة اللغات الطبيعية. يتضمن التعلم العميق لتصنيف النصوص والوثائق ثلاثة بنى أساسية للتعلم العميق تعمل بالتوالي. سنشرح كل نموذج منها بالتفصيل أدناه.

##### 4.10.1. الشبكات العصبية العميقية

صممت الشبكات العصبية العميقية (DNN) للتعلم من خلال اتصال متعدد الطبقات، حيث تستقبل كل طبقة اتصالاً من الطبقة السابقة فقط، وتنشئ اتصالات مع الطبقة التالية فقط في الجزء المخفي [2]. يوضح الشكل 16 بنية شبكة عصبية عميقية قياسية. يتكون المدخل من اتصال فضاء ميزات الإدخال (كما نوقش في القسم 2) مع الطبقة المخفية الأولى للشبكة. يمكن إنشاء طبقة الإدخال باستخدام TF-IDF أو تضمين الكلمات، أو أي طريقة أخرى لاستخراج الميزات. تساوي طبقة الإخراج عدد الفئات في حالة التصنيف متعدد الفئات، أو فئة واحدة فقط في حالة التصنيف الثنائي. في الشبكات العصبية العميقية متعددة الفئات، يتم تحديد كل نموذج تعلم (يتم تحديد عدد العقد في كل طبقة وعدد الطبقات بشكل عشوائي تماماً).

يُعد نموذج الشبكة العصبية العميق نموذجاً مُدرّجاً تميّزاً بـاستخدام خوارزمية الانتشار العكسي القياسي باستخدام دالة سigmoid (المعادلة 129) ودالة f<sub>tmax</sub> (المعادلة 130) كدالة تشبيط. يجب أن تكون طبقة الإخراج للتصنيف متعدد الفئات دالة softmax (المعادلة 131) كما هو موضح في المعادلة.

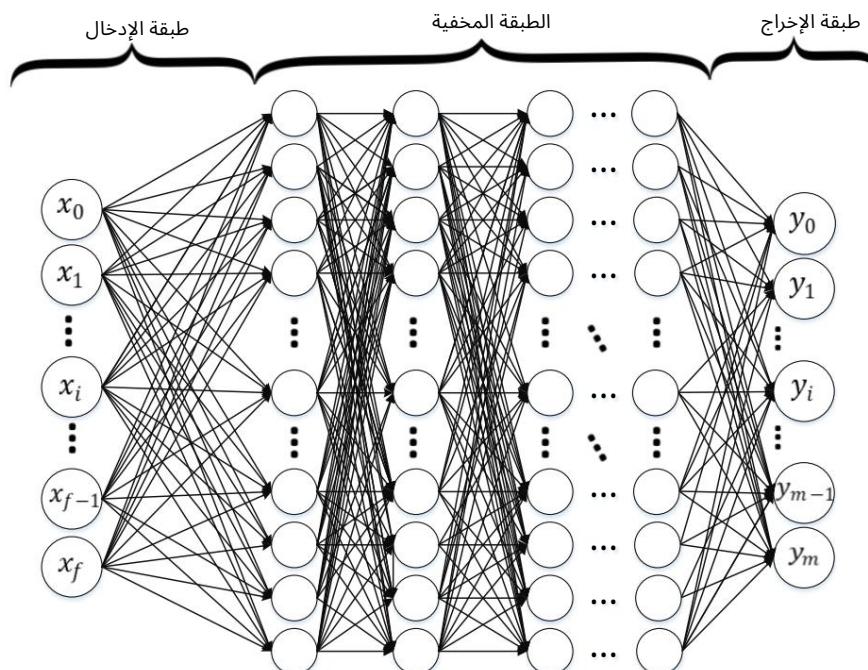
$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (129)$$

$$(130)$$

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad (131)$$

$$j \in \{1, \dots, K\}$$

بفرض وجود مجموعة من أزواج الأمثلة (س، ص)، حيث س، ص يُمثلان الهدف هو تعلم العلاقة بين فضاءات الإدخال والهدف باستخدام الطبقات المخفية. في تطبيقات تصنيف النصوص، يكون الإدخال عبارة عن سلسلة نصية يتم إنشاؤها من خلال تحويل بيانات النص الخام إلى متجهات.



الشكل .16. شبكة عصبية عميق قياسية متصلة بالكامل. (DNN).

#### 4.10.2. الشبكة العصبية المتكررة (RNN).

من بين بنى الشبكات العصبية الأخرى التي استخدمها الباحثون في استخراج البيانات النصية وتصنيفها ، الشبكة العصبية المتكررة. [179، 180]. تُعطي RNN أو زاناً أكبر لنقاط البيانات السابقة في التسلسل. ولذلك، تُعد هذه التقنية طريقة فعالة لتصنيف النصوص والسلسل والبيانات التسلسلية. تأخذ RNN في الاعتبار معلومات العقد السابقة بطريقة متطورة للغاية، مما يسمح بتحليل دلالي أفضل لبنية مجموعة البيانات. تعمل RNN في الغالب باستخدام LSTM أو GRU لتصنيف النصوص، كما هو موضح في الشكل ، 17 والذي يتكون من طبقة إدخال (تضمين الكلمات)، وطبقات مخفية، وطبقة إخراج. يمكن صياغة هذه الطريقة على النحو التالي:

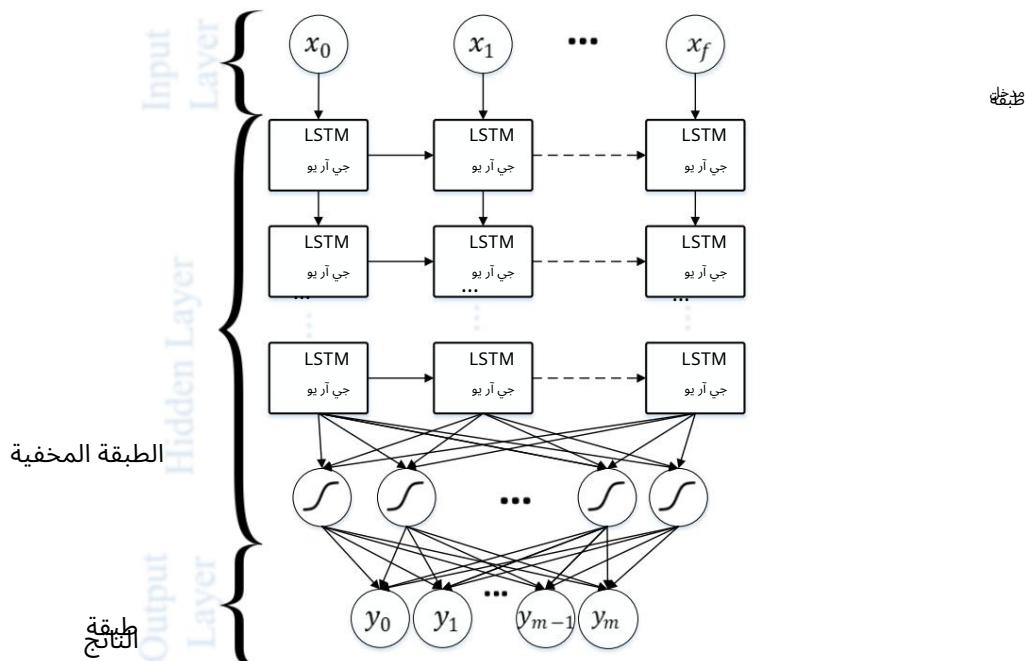
$$x_t = F(x_{t-1}, u_t, \theta) \quad (132)$$

حيث يمثل  $x_t$  الحالة عند الزمن  $t$ ، ويشير  $u_t$  إلى المدخلات عند الخطوة  $t$ . وبشكل أكثر تحديداً، يمكننا استخدام الأوزان لصياغة المعادلة. (132) التي تم تحديدها بالمعاملات التالية:

$$xt = W_{\text{rec}} \sigma(xt \odot 1) + W_{\text{in}} u + b \quad (133)$$

حيث يشير  $W_{\text{rec}}$  إلى وزن المصفوفة المتكررة، ويشير  $W_{\text{in}}$  إلى أوزان الإدخال، و  $u$  هو الانحياز، و  $b$  يشير إلى دالة عنصرية.

يوضح الشكل 17 بنية شبكة عصبية متكررة موسعة. على الرغم من المزايا المذكورة أعلاه، فإن الشبكة العصبية المتكررة عرضة لمشكلتي تلاشي التدرج وإنفجار التدرج عندما ينتشر خطأ خوارزمية هبوط التدرج عكسياً عبر الشبكة. [181]



الشكل 17. الشبكات العصبية المتكررة القياسية للذاكرة طويلة المدى (LSTM) / GRU.

الذاكرة طويلة المدى (LSTM)

تم تقديم نموذج LSTM بواسطة إس. هوكريتر وج. شميدهوبير. [182] ومنذ ذلك الحين تم تطويره وتحسينه. من قبل العديد من علماء الأبحاث. [183]

تُعد LSTM، كأحد شبكات العصبية المتكررة (RNN)، التي تعالج هذه المشكلات من خلال الحفاظ على التبعية طولية المدى بطريقة أكثر فعالية مقارنة بشبكة RNN الأساسية. وتعد LSTM مفيدة بشكل خاص في التغلب على مشكلة تلاشي التدرج. [184] على الرغم من أن LSTM لها بنية شبيهة بالسلسلة مثل RNN، إلا أنها تستخدم بوابات متعددة لتنظيم كمية المعلومات المسموح بها في كل حالة عقدة بدقة. يوضح الشكل 18 الخلية الأساسية لنموذج LSTM، وفيما يلي شرح تفصيلي لخلية LSTM:

$$it = \sigma(W_i [xt, ht \odot 1] + b), \quad (136)$$

$$i_t = \tanh(W_c[xt, ht \odot 1] + bc), \quad (135)$$

$$ft = \sigma(W_f[xt, ht \odot 1] + bf), \quad (134)$$

$$ot = \sigma(W_o[xt, ht \odot 1] + bt), \quad (137)$$

$$(139)$$

$$(138)$$

حيث تمثل المعادلة (134) بوابة الإدخال، وتتمثل المعادلة (135) قيمة خلية الذاكرة الأصلية، وتحدد المعادلة (136) تنشيط بوابة النسيان، وتحسب المعادلة (137) قيمة خلية الذاكرة الجديدة، وتحدد المعادلتان (138) و (139) قيمة بوابة الإخراج النهائية. في الوصف أعلاه، كل  $b$

يمثل متجه الانحياز، ويمثل كل  $W$  مصفوفة وزن، ويمثل  $x_t$  مدخلات خلية الذاكرة في الوقت  $t$ .علاوة على ذلك، تشير المؤشرات أو  $c$  و  $f$  إلى بوابات الإدخال وذاكرة الخلية والنسيان والإخراج على التوالي.

يوضح الشكل 18 تمثيلاً لبنيته هذه البوابات.

قد تكون الشبكة العصبية المترکزة مت Hickie عندما تكون الكلمات اللاحقة أكثر تأثيراً من الكلمات السابقة. الشبكات العصبية الالتفافية

تم تقديم نماذج الشبكة العصبية (CNN) التي تمت مناقشتها في القسم 4.10.3 للتلقي على هذا التحيز من خلال نشر طبقة التجميع القصوى لتحديد العبارات المميزة في البيانات النصية. [6]

### وحدة التكرار البوابية (GRU)

تُعد GRUs آلية بوابات لـ RNN صاغها K. Cho et al. [101]. Chung et al. [185] و.

تُعد وحدات التكرار البوابية (GRUs) نسخة مبسطة من بنية LSTM. ولذلك، تختلف وحدة التكرار البوابية عن LSTM لأنها تحتوي على بوابتين، كما أنها لا تمتلك ذاكرة داخلية (أي  $C_t$ ) في الشكل 18.

علاوة على ذلك، لا يتم تطبيق أي دالة غير خطية ثانية (أي دالة الظل الزائد) في الشكل 18. وفيما يلي شرح مفصل لخلية GRU:

$$z_t = \sigma(W_{zxt} + U_{zht} h_{t-1} + b_z), \quad (140)$$

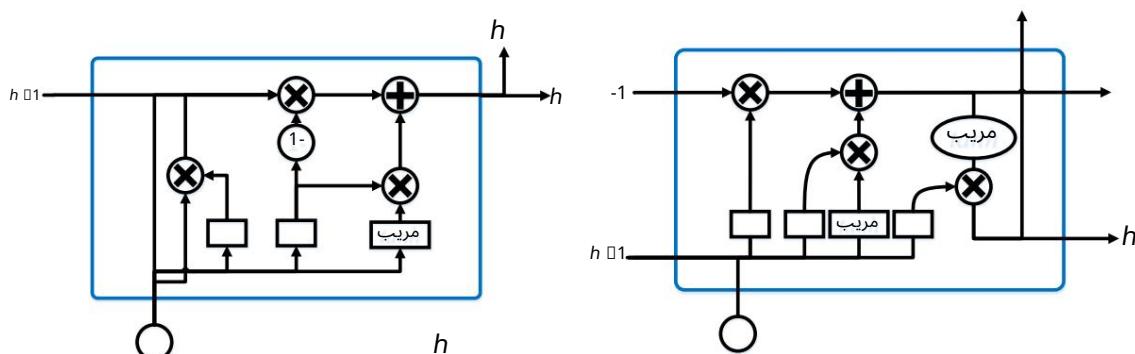
حيث يشير  $z_t$  إلى متجه بوابة التحديث لـ  $t$ ، و  $x_t$  إلى متجه الإدخال، و  $W$  و  $U$  تمثل مصفوفات/متجهات المعاملات. دالة التنشيط ( $\sigma$ ) هي إما دالة سigmoid أو دالة ReLU، ويمكن صياغتها على النحو التالي:

$$r_t = \sigma(W_{rxt} + U_{rht} h_{t-1} + b_r), \quad (141)$$

حيث يمثل  $r_t$  متجه بوابة إعادة الضبط لـ  $t$ ، و  $z_t$  هو متجه بوابة التحديث لـ  $t$ .

$$\begin{aligned} h_t &= \sigma(W_{oh} (Wh_{t-1} + Uh_t r_t) + bh) \\ &= z_t h_{t-1} + (1 - z_t) h_t \end{aligned} \quad (142)$$

حيث يمثل  $h_t$  متجه الإخراج لـ  $t$ ، و  $h_{t-1}$  يشير إلى دالة الظل الزائد.



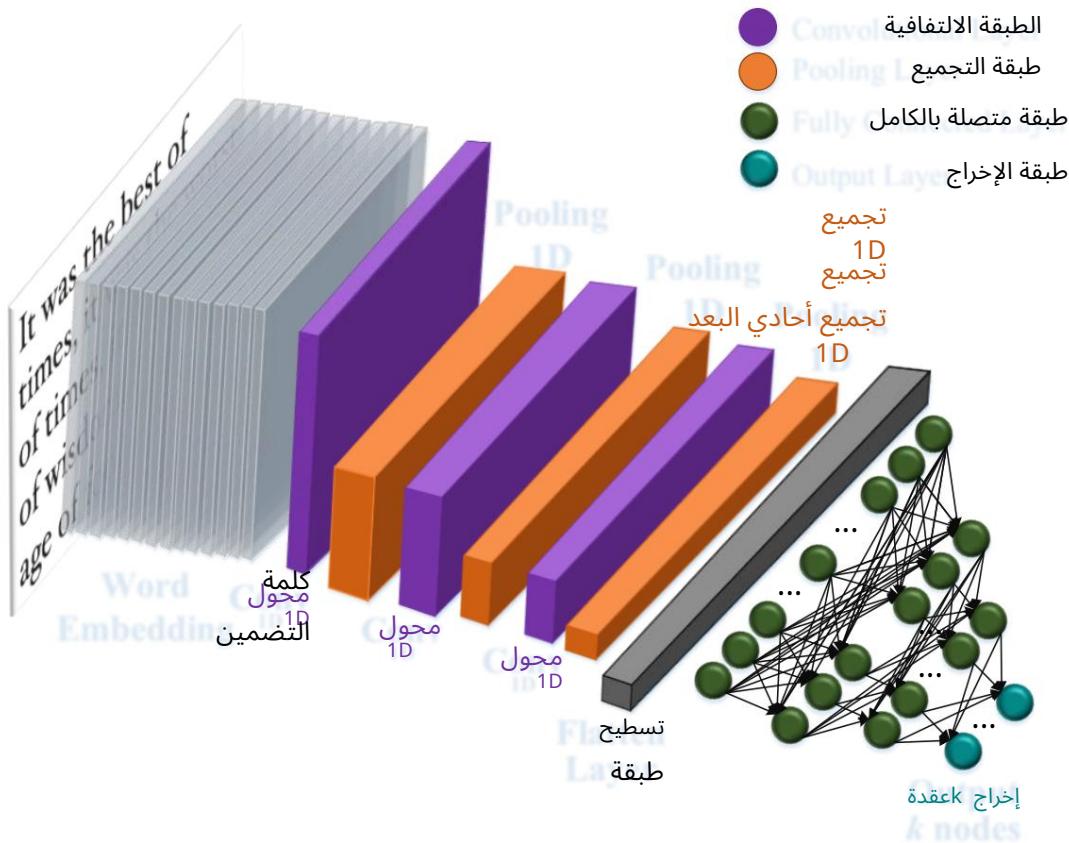
الشكل 18. الشكل الأيسر هو خلية LSTM بينما الشكل الأيمن هو خلية GRU.

### 4.10.3. الشبكات العصبية الالتفافية (CNN)

الشبكة العصبية الالتفافية (CNN) هي بنية تعلم عميق شائعة الاستخدام في تصنیف المستندات الهرمي. [6, 186] على الرغم من أنها صُممت في الأصل لمعالجة الصور، فقد استُخدمت الشبكات العصبية الالتفافية بكفاءة في تصنیف النصوص أيضًا. [27, 187] في شبكة عصبية التفافية أساسية لمعالجة الصور، يتم تطبيق عملية الالتفاف على موتر الصورة باستخدام مجموعة من النوى بحجم  $d \times d$ . تُسمى طبقات الالتفاف بهذه بخراط الميزات، ويمكن تكديسها ل توفير مرشحات متعددة على المدخلات. ولتقليل التعقيد الحسابي، تستخدم الشبكات العصبية الالتفافية التجميع لتقليل حجم المخرجات من طبقة إلى أخرى في الشبكة.

تُستخدم تقنيات تجميع مختلفة لتقليل المخرجات مع الحفاظ على الميزات المهمة. [188] أكثر طرق التجميع شيوعاً هي التجميع الأقصى، حيث يتم اختيار العنصر الأقصى في نافذة التجميع، وتغذية الطبقة التالية بالنتائج المُجمع من خرائط الميزات المكعبة، دمج الخرائط في عمود واحد. عادةً ما تكون الطبقات الأخيرة في الشبكة العصبية التلتفافية متصلة اتصالاً كاملاً.

بشكل عام، خلال خطوة الانتشار العكسي في الشبكة العصبية الالتفافية، يتم تعديل كل من الأوزان ومرشحات كاشف الميزات. تمثل إحدى المشكلات المحتملة عند استخدام الشبكات العصبية الالتفافية لتصنيف النصوص في عدد "القنوات". (حجم فضاء الميزات). بينما تحتوي تطبيقات تصنيف الصور عادةً على عدد قليل من القنوات (مثل 3قنوات RGB)، قد يكون [كبيرًا جدًا] (مثلاً 50 ألف قناة) في تطبيقات تصنيف النصوص [189]. مما يؤدي إلى أبعاد عالية جدًا. يوضح الشكل 19 بنية الشبكة العصبية الالتفافية لتصنيف النصوص، والتي تتضمن تصميم الكلمات كطبقة إدخال، وطبقات التفافية أحادية البعد، وطبقات تجميع أحادية البعد، وطبقات متصلة بالكامل، وأخيرًا طبقة الإخراج.



الشكل 19. بنية الشبكة العصبية الالتفافية (CNN) لتصنيف النصوص.

#### 4.10.4. شبكة الاعتقاد العميق (DBN).

شبكة الاعتقاد العميق (DBN) هي بنية تعلم عميق مُرگبة على آلات بولتزمان المقيدة. [1] آلة بولتزمان المقيدة هي شبكة عصبية اصطناعية توليدية قادرة على تعلم توزيع الاحتمالات على العينات. التباعد التباعي [190] (CD) هو أسلوب تدريب يُستخدم مع آلات بولتزمان المقيدة. [191, 192]

دالة الطاقة هي كما يلي:

$$E(v, h) = \sum_i a_i v_i \sum_j b_j h_j - \sum_{ij} v_i w_{ij} h_j \quad (143)$$

حيث تمثل  $a_i$  الوحدات المرئية، بينما تمثل  $b_j$  الوحدات المخفية في تدوين المصفوفات. ويمكن تبسيط هذا التعبير على النحو التالي:

$$E(v, h) = \frac{1}{2} v^T W h + b^T v - \frac{1}{2} \|v\|^2 - \frac{1}{2} \|h\|^2 \quad (144)$$

بالنظر إلى تكوين الوحدات المخفية، يتم تعريفها على النحو التالي:

$$\text{وحدة مخفية } i = \frac{\exp(a_i^T \cdot \mathbf{h}_j)}{1 + \sum_{k=1}^K \exp(a_k^T \cdot \mathbf{h}_j)} \quad (145)$$

بالنسبة لبيرنولي، يتم استبدال الدالة اللوجستية للوحدات المرئية على النحو التالي:

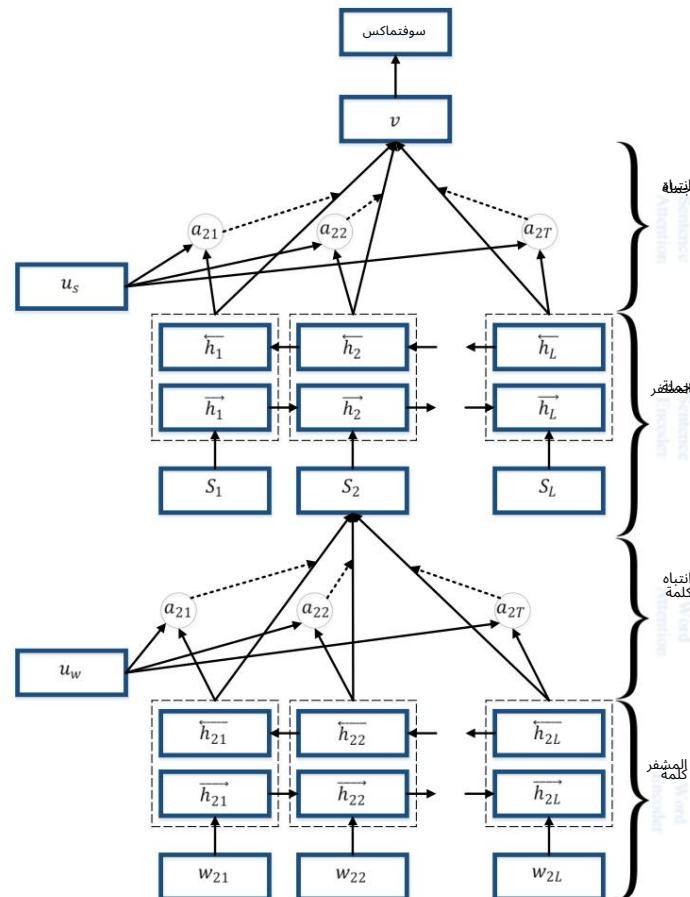
$$\text{وحدة مخفية } i = \frac{\exp(a_i^T \cdot \mathbf{h}_j) + \sum_j W_k h_j}{1 + \sum_k \exp(a_k^T \cdot \mathbf{h}_j) + \sum_j W_k h_j} \quad (146)$$

تكون دالة التحديث باستخدام خوارزمية التدرج الهبوطي كما يلي:

$$\frac{\partial \text{سجل}(ص(t))}{\partial w_{ij}(t)} = w_{ij}(t) + \eta \Delta w_{ij} \quad (147)$$

#### 4.10.5. شبكات الانتباه الهرمية (HAN).

تُعد شبكات الانتباه الهرمية (HAN) إحدى البنى العميقية الناجحة لتصنيف النصوص والوثائق. وقد ظهرت هذه التقنية بواسطة Z. Yang وآخرون [194] و PS HongSuck [193] الآخرين. يركز هيكل شبكة HAN على تصنيف المستندات على مستوى المستند، حيث يحتوي المستند على جملة وتحتوي كل جملة على  $T_i$  الكلمات، حيث يمثل  $w_{it}$  الكلمة رقم  $i$  في الجملة رقم  $t$ . تم توضيح بنية HAN في الشكل 20، حيث يحتوي المستوى السفلي على ترميز الكلمات وانتباه الكلمات ويحتوي المستوى العلوي على ترميز الجمل وانتباه الجمل.



الشكل 20. شبكات الانتباه الهرمية لتصنيف المستندات.

#### 4.10.6. تقييات الجمع

يلجأ العديد من الباحثين إلى دمج أو دمج بنى التعلم العميق القياسي لتطوير تقنيات جديدة ذات بنى أكثر قوّة ودقة لمهام التصنيف. في هذا القسم الفرعي، نصف بنى التعلم العميق الحديثة والشائعة وهياكلها.

#### التعلم العميق متعدد النماذج العشوائي (RMDL)

قدم ك. كوصاري وأخرون [5] تقنية التعلم العميق متعدد النماذج العشوائي (RMDL) كتقنية جديدة للتعلم العميق في مجال التصنيف. ويمكن استخدام RMDL مع أي نوع من مجموعات البيانات لأغراض التصنيف. يُظهر الشكل 21 نظرة عامة على هذه التقنية، موضحاً بنية الشبكة باستخدام الشبكات العصبية العميقه المتعددة، والشبكات العصبية الالتفافية العميقه، والشبكات العصبية المتكررة العميقه. يتم توليد عدد الطبقات والعقد لجميع نماذج التعلم العميق المتعددة هذه عشوائياً (على سبيل المثال، 9 نماذج عشوائية في RMDL مُنشأة من 3 شبكات عصبية التفافية، و 3 شبكات عصبية عميقه، وكلها فريدة من نوعها نظرًا لإنشائها العشوائي).

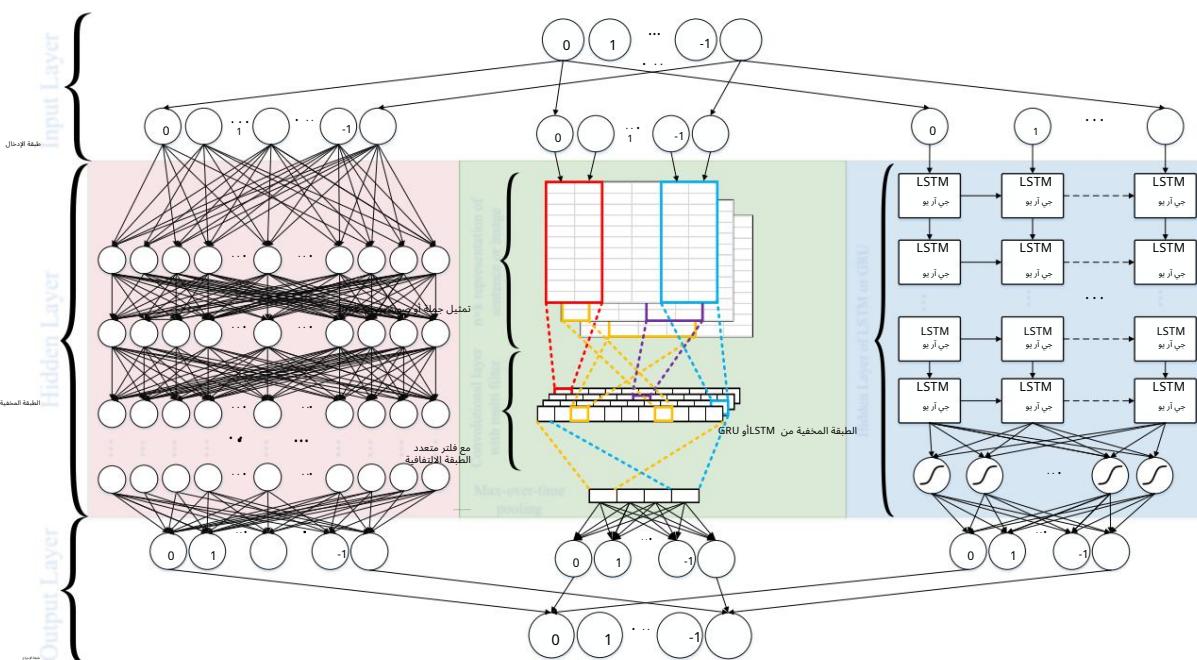
$$\text{ص}^{\wedge} \text{ي} = \text{ص}^{\wedge} 1 \text{ا}^{\wedge} \text{ي} 1 \text{.....} \text{ا}^{\wedge} \text{ي} \text{ن} = \frac{\sum_{j=1}^{n^2} y_{ij}}{n} \quad (148)$$

حيث  $n$  هو عدد النماذج العشوائية، و  $\hat{y}$  هو تنبؤ النموذج لنقطة البيانات أفي النموذج (ز) تُستخدم المعادلة (148) للتصنيف الثنائي، 0 أو 1). يستخدم فضاء المخرجات التصويت بالأغلبية لحساب القيمة النهائي  $\hat{z}$ . لذلك، تُعطى  $\hat{z}$  كما يلي:

$$\text{ص}^{\wedge} \text{ي} = \text{ص}^{\wedge} 1 \text{ا}^{\wedge} \text{ي} 1 \text{.....} \text{ا}^{\wedge} \text{ي} \text{ن} \quad (149)$$

حيث  $n$  هو رقم النموذج العشوائي، و  $\hat{y}_i$  يوضح التنبؤ بتصنيف نقطة البيانات (مثل المستند) لـ  $\{x_i\}$  للنموذج، ويتم تعريف  $\hat{z}$  على النحو التالي:

$$\hat{z} = \text{f}(\text{tm}^{\wedge} \text{ax}^{\wedge} \text{t} \text{m}^{\wedge} \text{ax}^{\wedge} \text{t}) \quad (150)$$



الشكل 21. بنية التعلم العميق متعدد النماذج العشوائي (RMDL) للتصنيف. تتضمن RMDL ثلاثة نماذج عشوائية: مصنف شبكة عصبية عميقه (DNN) (يسار)، ومصنف شبكة عصبية تلقيفية عميقه (CNN) (وسط)، ومصنف شبكة عصبية متكررة عميقه (RNN) (يمين). يمكن أن تكون كل وحدة من وحدات GRU أو LSTM.

بعد تدريب جميع نماذج التعلم العميق العشوائي،  $\text{RMDL}$  تُحسب النتائج النهائية باستخدام التصويت بالأغلبية على مخرجات هذه النماذج. تكمن الفكرة الرئيسية لاستخدام نماذج متعددة مع محسّنات مختلفة في أنه إذا لم يُوفّر محسّن واحد ملائمة جيدة لمجموعة بيانات محدّدة، فإن نموذج  $\text{RMDL}$  يحتوي على  $n$  نموذجاً عشوائياً (حيث قد يستخدم بعضها محسّنات مختلفة) يمكنه تجاهل  $n$  نموذجاً غير فعال إذا كان  $k < n$ . يُساعد استخدام تقنيات محسّنات متعددة (مثل  $\text{Adam}$  و  $\text{SGD}$  و  $\text{Adagrad}$  و  $\text{Adamax}$  و  $\text{RMSProp}$ ) على أن يكون أكثر ملاءمة لأي نوع منمجموعات البيانات. على الرغم من أننا استخدمنا محسّنين فقط ( $\text{Adam}$  و  $\text{RMSProp}$ ) للتقييم النموذج في هذا البحث، إلا أن نموذج  $\text{RMDL}$  يمكنه استخدام أي نوع من المحسّنات. في هذا الجزء، نصف تقنيات التحسين الشائعة المستخدمة في بناء التعلم العميق.

### محسّن التدرج العشوائي (SGD):

تظهر المعادلة الأساسية لأنحدار التدرج العشوائي [195] في المعادلة (151) في المعادلة (152) لتحديث المعلمات. يستخدم SGD برمداً على التدرج المعاوقيا والذى يظهر في المعادلة (152) لتحديث المعلمات.

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} J(\theta, x_i, y_i) \quad \theta \leftarrow \theta - \gamma \theta + \alpha \nabla_{\theta} J(\theta, \text{all}) \quad (151)$$

### RMSprop:

قدم T. Tieleman و G. Hinton [196] RMSprop كمحسن جديد يقسم معدل التعلم لوزن ما على المتوسط المتحرك لأحجام التدرجات الأخيرة لهذا الوزن.

معادلة طريقة الزخم لـ RMSprop هي كما يلي:

$$v(t) = \alpha v(t-1) + \frac{\nabla E}{\|w\|} \quad (152)$$

$$\begin{aligned} \nabla w(t) &= v(t) \\ &= \alpha v(t-1) + \frac{\nabla E}{\|w\|_E} \\ &= \alpha \nabla v(t-1) + \frac{\nabla E}{\|w\|_E} \end{aligned} \quad (153)$$

لا يقوم RMSProp بتصحيح الانحراف، مما يسبب مشكلات كبيرة عند التعامل مع تدرج متفرق.

### محسن آدم

آدم هو محسن تدرج عشوائي آخر يستخدم فقط أول عزمين للتدرج  $v$  و  $m$ . الموضحين في المعادلات (155)-(158) ويحسب المتوسط عليهم. يمكنه التعامل مع عدم استقرار دالة الهدف كما في RMSProp مع التغلب على مشكلة التدرج المتفرق التي تعاني منها. [197]

$$\theta \leftarrow \theta - \frac{\nabla}{\sqrt{m^2 + (\beta_1 m)^2}} \quad (155)$$

$$+ (1 - \beta_1) g_i t \quad mt = \beta_2 v^2 t \quad (156)$$

$$\nabla \theta J(\theta, x_i, y_i) \quad mt = \beta_1 m t \quad (157)$$

$$\nabla v^2 + g_i t \quad (158)$$

$$\begin{aligned} m^2 t &= \text{مقدار العزم الأول، ويشير } v^2 t \text{ إلى العزم الثاني، وكلاهما يُقدر.} \\ \beta_1 &= \frac{1}{1 - \beta_1} \\ v^2 t &= 1 - \beta_1 \end{aligned} \quad (159)$$

يضيف:

تمت معالجة Adagrad في [198] باعتبارها عائلة جديدة من طرق التدرج الفرعي التي تمت بشكل ديناميكي معرفة هندسة البيانات لإجراء تعلم قائم على التدرج أكثر إفادة.

هو امتداد لخوارزمية SGD في التكرار، يتم تعريف التدرج على النحو التالي:

$$\text{G}(k) = \frac{\text{نصف}}{\sum_{i=1}^k j^{(i)(g(i))}} \quad (159)$$

المصفوفة القطرية:

$$\text{G}(k) = \frac{\text{نصف}}{\sum_{i=1}^k (g_i)^2} \quad (160)$$

تحديث القاعدة:

$$\begin{aligned} x &= \arg \min_{x \in X} (k+1) \|f(x(k))\|_2^2 \\ &= \frac{1}{(k+1) \|f(x(k))\|_2^2} \cdot x + \frac{2ak(k)}{(k+1) \|f(x(k))\|_2^2} \cdot g(k) \\ &= R_n(\text{إذا كان}) \end{aligned} \quad (161)$$

من آدا:

تستخدم خوارزمية AdaDelta، التي قدمها إم دي زيلر [199] المتوسط المتناظر أسيًا لـ  $\text{gt}$  ويزعم ثان للتدرج. تُعد هذه الطريقة نسخة مُحدثة من AdaDelta تعتمد فقط على معلومات الدرجة الأولى. قاعدة التحديث لـ AdaGrad هي:

$$g_{t+1} = \gamma g_t + (1 - \gamma) L(\theta)^2 \quad (162)$$

$$x_{t+1} = x_t + (1 - \gamma) v \quad (163)$$

$$v_{t+1} = \frac{g_{t+1}}{\|g_{t+1}\|} + \gamma v_t \quad (164)$$

### التعلم العميق الهرمي للنصوص (HDLTex)

تمثل المساهمة الرئيسية لبنية التعلم العميق الهرمي للنصوص (HDLTex) في التصنيف الهرمي للوثائق [2]. قد تُجدي تقنية التصنيف التقليدية متعددة الفئات نفعاً مع عدد محدود من الفئات، لكن الأداء يتراجع مع ازدياد عدد الفئات، كما هو الحال في الوثائق المنظمة هرمياً. في نموذج التعلم العميق الهرمي هذا، تم حل هذه المشكلة من خلال إنشاء بنية تخصّص مناهج التعلم العميق وفقاً لمستوى التسلسل الهرمي للوثيقة (انظر، على سبيل المثال، الشكل). فيما يلي بنية HDLTex كل نموذج من نماذج التعلم العميق :

الشبكة العصبية العميقية: 8 طبقات مخفية مع 1024 خلية في كل طبقة مخفية.

يتم استخدام LSTM و GRU في هذا التطبيق، 100 خلية مع GRU مع طبقتين مخفيتين. أحجام المرشحات {3, 4, 5, 6, 7} أو CNN: max-pooling بقيمة 5، وأحجام الطبقات {128, 128, 128} مع CNN: وتحتوي على 8 طبقات مخفية.

تم بناء جميع النماذج باستخدام المعايير التالية: حجم الدفعه 128 = معلمات التعلم ، معدل التضاؤل ، 0.0 = معدل التسرّب .

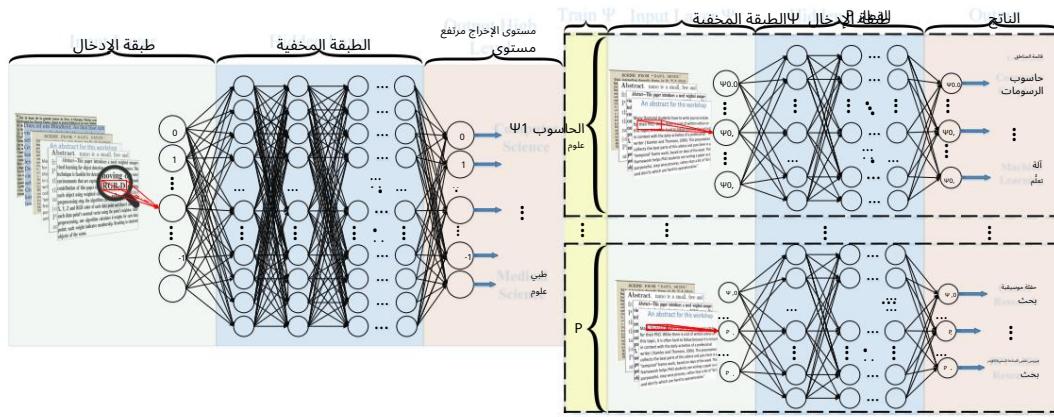
$\text{loss} = \text{loss}_{\text{softmax}}(\text{output}, \text{target})$  .

يستخدم HDLTex دالة التكلفة التالية لتقدير نماذج التعلم العميق:

$$\text{Acc}(X\Psi) = \frac{\sum_{k=1}^K \text{Acc}(X\Psi_k)}{\sum_{k=1}^K \Psi_k}$$

(165)

حيث يمثل عدد المستويات، و  $\Psi_k$  يشير إلى عدد الفئات في كل مستوى، و  $\Psi$  يشير إلى عدد الفئات في مستوى الطفل في النموذج الهرمي.



الشكل 22. التعلم العميق الهرمي لتصنيف النصوص. نهج الشبكة العصبية العميقه لتصنيف النصوص. يوضح الشكل العلوي المستوى الأبوى لمودجنا، بينما يوضح الشكل السفلى نماذج المستوى الفرعى ( $\Psi_k$ ) كمستندات إدخال في المستوى الأبوى.

#### تقنيات أخرى

في هذا القسم، نناقش تقنيات أخرى لتصنيف النصوص ناتجة عن دمج بين التعلم العميق. تُستخدم الشبكات العصبية الالتفافية المتكررة (RCNN) لتصنيف النصوص [6, 200]. تستطيع هذه الشبكات استخلاص المعلومات السياقية من خلال بنيتها المتكررة، وبناء تمثيل النص باستخدام شبكة عصبية التفافية [6] (CNN) مزيجاً من الشبكات العصبية المتكررة (RNN) والشبكات العصبية الالتفافية (CNN)، حيث تستفيد من مزايا كلتا التقنيتين في نموذج واحد.

تُعد C-LSTM-CNN تقنية أخرى لتصنيف النصوص والوثائق، وقد طورها سي. تشو وآخرون [201]. تجمع C-LSTM-CNN وشبكات الذاكرة طويلة المدى (LSTM) لتعلم خصائص العبارات باستخدام طبقات التفافية. يُغذّي هذا التصميم سلسلات من تمثيلات عالية المستوى إلى شبكة LSTM لتعلم الارتباطات طويلة المدى.

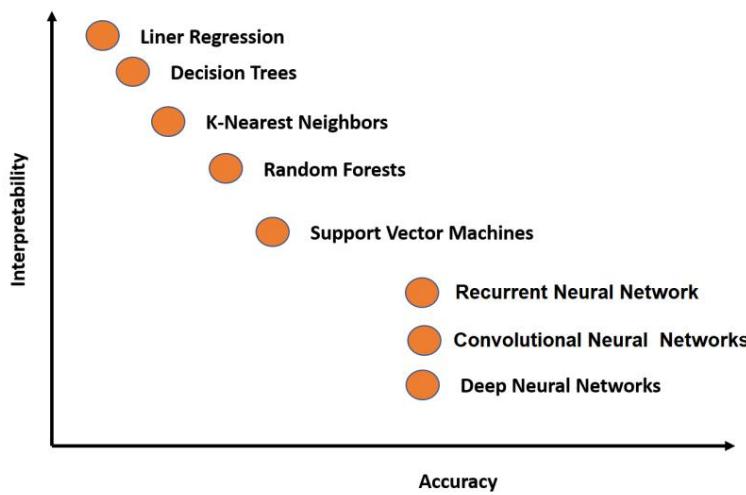
#### 4.10.7. قيود التعلم العميق

لطالما شكلت قابلية تفسير نماذج التعلم العميق، وخاصة الشبكات العصبية العميقه، عائقاً أمام استخدامها في الحالات التي تتطلب شرحاً للخصائص المستخدمة في النموذج، كما هو الحال في العديد من مشاكل الرعاية الصحية. ويعود هذا إلى تفضيل العلماء استخدام التقنيات التقليدية ، مثل النماذج الخطية، والنماذج البيانية، وألات المنتجات الداعمة، وأشجار القرار، وغيرها. تُعد الأوزان في الشبكة العصبية مقاييسًا لقوه كل اتصال بين كل عصيون، وذلك لتحديد فضاء الخصائص المهمة. وكما هو موضح في الشكل 23، كلما زادت دقة النموذج، قلت قابلية للتفسير، مما يعني صعوبة فهم الخوارزميات المعقدة، مثل التعلم العميق.

يُعد التعلم العميق (DL) أحد أقوى تقنيات الذكاء الاصطناعي (AI) ويركز العديد من الباحثين والعلماء على بنى التعلم العميق لتحسين مثانة هذه الأداة وقدرتها الحسابية. مع ذلك، تُعاني بنى التعلم العميق من بعض العيوب والقيود عند تطبيقها على مهام التصنيف. إحدى المشكلات الرئيسية لهذا النموذج هي أن التعلم العميق لا يُسهل فهماً نظرياً شاملًا للتعلم [202].

من عيوب أساليب التعلم العميق طبيعتها "المبهمة". [203, 204] أي أن الطريقة التي تُنْتَجُ بها هذه الأساليب المخرجات المختلفة غير مفهومة بسهولة. ومن القيود الأخرى للتعلم العميق أنه يتطلب عادةً بيانات أكثر بكثير من خوارزميات التعلم الآلي التقليدية، مما يعني أن هذه التقنية لا يمكن تطبيقها على مهام التصنيف التي تعتمد على مجموعات بيانات صغيرة. [205, 206]

بالإضافة إلى ذلك، فإن الكمية الهائلة من البيانات اللازمة لخوارزميات تصنيف التعلم العميق تزيد من تعقيد الحساب أثناء خطوة التدريب. [207]



الشكل .23. مقارنة قابلية تفسير النموذج بين تقنيات التعلم التقليدية والعميقة.

#### 4.11. التعلم شبه الموجة لتصنيف النصوص

طور العديد من الباحثين العديد من المصنفات الفعالة للوثائق المصنفة وغير المصنفة. يُعد التعلم شبه الموجة نوعاً من أنواع التعلم الموجة، حيث يستخدم بيانات غير مصنفة لتدريب نموذج، عادةً ما يُفضل الباحثون والعلماء استخدام تقنيات التعلم شبه الموجة عندما يحتوي جزء صغير من مجموعة البيانات على نقاط بيانات مصنفة، بينما لا يحتوي الجزء الأكبر منها على تصنيفات. [208] تستخدم معظم خوارزميات التعلم شبه الموجة لمهام التصنيف تقنية التجميع (المستخدمة عادةً في التعلم غير الموجة) [209] كما يلي: في البداية، تُطبق تقنية التجميع على شجرة القرار (DT) التي تحتوي على  $K = K$  (عدد الفئات)، نظرًا لاحتواء شجرة القرار على عينات مصنفة من جميع الفئات. [208] إذا أحتوى قسم  $P_i$  على عينات مصنفة، فإن جميع نقاط البيانات في تلك المجموعة تتبع إلى

هذا التصنيف.

الهدف البحثي لتقنيات التجميع هو تحديد ما إذا كان لدينا أكثر من فئة واحدة مصنفة في مجموعة واحدة، وماذا يحدث إذا لم يكن لدينا أي نقطة بيانات مصنفة في مجموعة واحدة. [210] في هذا الجزء، نصف بإيجاز أكثر تقنيات تصنيف النصوص والوثائق شبه الخاضعة للإشراف شيوعًا. عمل كل من أ. شابيل وأ. زين [211] على التصنيف شبه الخاضع للإشراف عبر فصل الكثافة المنخفضة، والذي يجمع بين حساب مسافة الرسم البياني وتدرير آلة المتغيرات الداعمة الاستقرائية (TSVM) (طور ك. نيفام وآخرون [212] تقنية لتصنيف النصوص باستخدام خوارزمية تنظيم التوقع (EM) والنماذج التوليدية للتعلم شبه الخاضع للإشراف مع البيانات المصنفة وغير المصنفة في مجال تصنيف النصوص. قدم ل. شي وآخرون [213] طريقة لنقل معرفة التصنيف بين اللغات عبر الميزات المترجمة. تستخدم هذه التقنية خوارزمية EM التي تأخذ في الاعتبار بشكل طبيعي الغموض المرتبط بترجمة الكلمة. قدم ج. سو وآخرون [213] "تقدير التردد شبه الخاضع للإشراف" (SFE)، وهي طريقة MNBC لتصنيف النصوص على نطاق واسع. [214] ابتكرروا طريقة جديدة للتعلم العميق تستند شبكة دي بي إن الضبابية لتصنيف المشاعر شبه الخاضع للإشراف. تستخدم هذه الطريقة دالة عضوية ضبابية لكل فئة من فئات المراجعات بناءً على البنية المعلمة.

## 5. التقييم.

في الأوساط البحثية، يُفضل وجود معايير أداء مشتركة وقابلة للمقارنة لتقدير الخوارزميات. مع ذلك، في الواقع، قد لا تتوفر هذه المعايير إلا لعدد قليل من الطرق.

تكمن المشكلة الرئيسية عند تقييم أساليب تصنيف النصوص في غياب بروتوكولات موحدة لجمع البيانات. حتى في حال وجود طريقة جمع بيانات مشتركة (مثل مجموعة بيانات روبيز الإخبارية)، فإن مجرد اختيار مجموعات تدريب وختبار مختلفة قد يؤدي إلى تباينات في أداء التمودج [215].

يتمثل أحد التحديات الأخرى المتعلقة بتقدير الأساليب في القدرة على مقارنة مقاييس الأداء المختلفة المستخدمة في تجارب منفصلة. **تقدير مقاييس الأداء** عموماً جواب محددة من أداء مهمة التصنيف، وبالتالي لا تقدم دليلاً معلومات متطابقة. في هذا القسم، نناقش مقاييس التقييم ومقاييس الأداء ونسلط الضوء على طرق مقارنة أداء المصنفات. نظراً لاختلاف الآليات الكامنة وراء مقاييس التقييم المختلفة، فإن فهم ما يمثله كل مقاييس منها تحدياً، ونوع المعلومات التي يحاول نقلها، أمر بالغ الأهمية للمقارنة. تتضمن بعض الأمثلة على هذه المقاييس: الاستدعاء، والدقة، والصحة، ومقاييس F، والمتوسط الجرئي، والمتوسط الكلي. تستند هذه المقاييس إلى "مصفوفة الارتباك" (الموضحة في الشكل 24) التي تتضمن الإيجابيات الحقيقية (TP) والإيجابيات الخاطئة (FP) والسلبيات الخاطئة (FN).

قد تختلف أهمية هذه العناصر الأربع بناءً على تطبيق التصنيف. يُطلق على نسبة التنبؤات الصحيحة من إجمالي التنبؤات اسم الصحة (المعادلة 166). تُسمى نسبة النتائج الإيجابية المعروفة التي تم التنبؤ بها بشكل صحيح بالحساسية، أي معدل الإيجابية الحقيقية أو الاستدعاء (المعادلة 167). تُسمى نسبة النتائج السلبية التي تم التنبؤ بها بشكل صحيح بالنوعية (المعادلة 168). أما نسبة النتائج الإيجابية التي تم التنبؤ بها بشكل صحيح إلى إجمالي النتائج الإيجابية فتُسمى الدقة، أي القيمة التنبؤية الإيجابية (المعادلة 169).

$$\text{دقة} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (166)$$

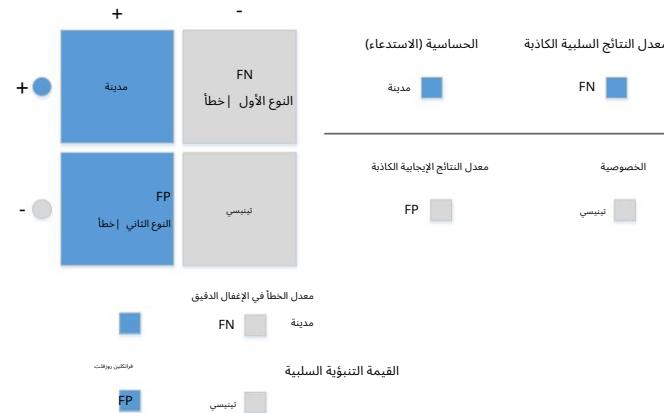
$$\text{الحساسية} = \frac{\text{مدينة}}{(TP + FN)} \quad (167)$$

$$\text{الخصوصية} = \frac{\text{تبنيسي}}{(TN + FP)} \quad (168)$$

$$\text{الدقة} = \frac{\Box = 1 TPI}{I = 1 TPI + FPI} \quad (169)$$

$$\text{الاستدعاء} = \frac{\Box}{I = 1 TPI + FNI} \quad (170)$$

$$\text{النسبة المئوية} = \frac{2 TPI + FPI - FNI}{FPI + FNI} \quad (171)$$



الشكل 24. مصفوفة الارتباك.

## 5.1. المتوسط الكلي والمتوسط الجزئي.

يلزم استخدام مقياس إجمالي واحد عند استخدام عدة مصنفات ثنائية الفئات لمعالجة مجموعة بيانات. يعطي المتوسط الكلي متوسطاً بسيطًا عبر الفئات، بينما يجمع المتوسط الجزئي القرارات الخاصة بكل مستند عبر الفئات، ثم يخرج مقياساً فعالاً على جدول التوافق المجمع [217]. يمكن حساب نتائج المتوسط الكلي كما يلي:

$$\text{B} = \frac{1}{\lambda=1} \frac{q}{q} \frac{B(TP\lambda + FP\lambda + TN\lambda + FN\lambda)}{B(\lambda=1)}$$

حيث  $B$  هو مقياس تقييم ثنائي يتم حسابه بناءً على الإيجابيات الحقيقة ( $TP$ ) والإيجابيات الخاطئة ( $FP$ ) والسلبيات الخاطئة ( $FN$ ) والسلبيات الحقيقة،  $\lambda = j : j = 1 \dots q$  هي مجموعة جميع التصنيفات. يمكن حساب النتائج المتوسطة الجزئية [156, 218] على النحو التالي:

$$B_{macro} = \frac{q}{\lambda=1} \frac{q}{\lambda=1} \frac{q}{\lambda=1} \frac{q}{\lambda=1} \frac{TP\lambda}{FP\lambda} \frac{TN\lambda}{FN\lambda} \quad (173)$$

نتيجةً لذلك، يعطي متوسط التقييم الجزئي وزناً متساوياً لكل وثيقة، ويُعتبر متوسطاً لكل وثيقة على حدة. أما متوسط التقييم الكلي، فيعطي وزناً متساوياً لكل فئة دون مراعاة التكرار، وبالتالي فهو متوسط لكل فئة على حدة.

## 5.2. درجة $F\beta$ .

$F\beta$  هو أحد أكثر مقاييس التقييم المجمعة شيوعاً لتقييم المصنفات [216]. تُستخدم المعلمة  $\beta$  لتحقيق التوازن بين الاستدعاء والدقة، ويتم تعريفها على النحو التالي:

$$F\beta = \frac{\frac{(1+\beta)^2}{(1+\beta) \times \text{الدقة} \times \text{الاستدعاء}}}{\frac{\text{الدقة} \times \text{الاستدعاء}}{\text{الدقة} + \text{الاستدعاء}}} \quad (174)$$

بالنسبة لقيمة  $\beta = 1$  الشائعة الاستخدام، أي  $F1$ ، يتم إعطاء الاستدعاء والدقة أوزاناً متساوية، ويمكن تبسيط المعادلة (174) إلى:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (175)$$

بما أن  $F\beta$  يعتمد على الاستدعاء والدقة، فإنه لا يمثل مصفوفة الارتباط بشكل كامل.

## 5.3. معامل ارتباط ماثيوز (MCC).

يقيس معامل ارتباط ماثيوز [30] (MCC) جودة أساليب التصنيف الثنائي، إذ يجمع جميع البيانات في مصفوفة الارتباط. يمكن استخدام MCC في المشكلات ذات أحجام الفئات غير المتساوية، ويعتبر مقياساً متوازناً. تراوح قيمة MCC بين -1 و 1 (أي أن التصنيف خاطئ دائمًا وصحيح دائمًا، على التوالي). يمكن حساب MCC كما يلي:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (176)$$

عند مقارنة مصنفين، قد يكون لأحدهما درجة أعلى باستخدام MCC والآخر درجة أعلى باستخدام  $F1$ . ونتيجة لذلك لا يمكن لمقياس محدد واحد أن يلتقط جميع نقاط القوة والضعف للمصنف [216].

## 5.4. خصائص تشغيل جهاز الاستقبال (ROC).

تُعد منحنيات خصائص تشغيل المستقبل [219] (ROC) أدوات بيانية قيمة لتقييم المصنفات. ومع ذلك، فإن عدم توازن الفئات (أي الاختلافات في احتمالات الفئات المسبقة) [220] يمكن أن يُسبب تبايناً في منحنيات ROC.

لا تُظهر المنحنيات أداء المصنف بدقة. يرسم منحنى ROC مُعدل الإيجابية الحقيقية (TPR) ومُعدل الإيجابية الكاذبة (FPR):

$$\frac{\text{مُعدل}}{\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}} \quad (177)$$

$$\frac{\text{FP}}{\text{RPF} = \frac{\text{FP}}{\text{FP} + \text{TN}}} \quad (178)$$

## 5.5 المساحة تحت منحنى ROC (AUC).

تقيس المساحة تحت منحنى ROC (AUC) [31,32] المساحة الكاملة أسفل منحنى ROC. يستفيد AUC من خصائص مفيدة مثل زيادة الحساسية في اختبارات تحليل التباين ، و الاستقلال عن عتبة القرار، وعدم الثبات بالنسبة لاحتمالات الفئة المسبقة، والإشارة إلى مدى جودة الفئات السلبية والإيجابية فيما يتعلق بمؤشر القرار [221].

بالنسبة لمهام التصنيف الثنائي، يمكن صياغة AUC على النحو التالي:

$$\begin{aligned} \text{AUC} &= \int_{-\infty}^{\infty} \text{TPR}(T) \text{FPR}(T) dT \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(T > T') f_1(T) f_0(T) dT dT' \\ &= P(X_1 > X_0) \end{aligned} \quad (179)$$

بالنسبة لـ AUC متعدد الفئات، يمكن تعريف متوسط [222] AUC على النحو التالي:

$$\frac{\text{AUC}}{|C|(|C| - 1)} = \frac{1}{|C|} \sum_{i=1}^{|C|} \text{AUC}_i \quad (180)$$

حيث  $C$  هو عدد الفصوص.

قام يانغ [215] بتقييم الأساليب الإحصائية لتصنيف النصوص، وأفاد بما يلي:  
العوامل المهمة التي ينبغي مراعاتها عند مقارنة خوارزميات التصنيف:

- \* التقييم المقارن بين الطرق والتجارب، والذي يتيح فهّماً أعمق للعوامل الكامنة وراء اختلافات الأداء، مما يسهم في تطوير منهجية تقييم أفضل في المستقبل؛
- \* تأثير تباين البيانات، مثل تضمين مستندات غير مصنفة في مجموعة التدريب أو الاختبار.

إن التعامل معها كحالات سلبية قد يشكل مشكلة خطيرة؛

يُظهر تقييم تصنيف الفئات وتقييم التصنيف الثنائي فائدته المصنفة في التطبيقات التفاعلية، وبؤدان على استخدامها في وضع المعالجة الدفعية على التوالي. إن وجود كلا النوعين من مقاييس الأداء لترتيب المصنفات يساعد في الكشف عن تأثيرات استراتيجيات تحديد العتبة.

\* يُعد تقييم قابلية التوسيع للمصنفات في مساحات الفئات الكبيرة مجالاً نادراً ما يتم بحثه.

## 6. المناقشة.

هدفت هذه المقالة إلى تقديم لمحة موجزة عن تقنيات تصنيف النصوص، إلى جانب مناقشة خوطات المعالجة المسبقة وأساليب التقييم ذات الصلة. في هذا القسم، نقارن بين كل تقنية وخوارزمية من هذه التقنيات ونوضح أوجه الاختلاف بينها.علاوة على ذلك، نناقش قيود تقنيات التصنيف وأساليب التقييم الحالية. يمكن التحدى الرئيسي في اختبار نظام تصفيف فعال في فهم أوجه التشابه والاختلاف بين التقنيات المتاحة في مختلف مراحل معالجة البيانات.

### 6.1. استخراج خصائص النصوص والمستندات.

لقد حددنا نهجين رئيسيين لاستخلاص الميزات: الكلمات الموزونة (حقيقة الكلمات) وتضمين الكلمات. تتعلم تقنيات تضمين الكلمات من تسلسلات الكلمات عن طريق

مع الأخذ في الاعتبار معلومات التكرار والتواجد المشترك، كما أن هذه الطرق نماذج غير خاضعة للإشراف لتوليد متجهات الكلمات. في المقابل، تعتمد ميزات الكلمات الموزونة على عد الكلمات في المستندات، ويمكن استخدامها كآلية تقييم بسيطة لتمثيل الكلمات.

#### لكل تقنية قيودها الخاصة.

تحسب الكلمات الموزونة تشابه المستندات مباشرةً من فضاء عدد الكلمات، مما يزيد من وقت الحساب للمفردات الكبيرة. [223] ورغم أن عدد الكلمات الفريدة يوفر دليلاً مستقلاً على التشابه، إلا أنه لا يأخذ في الحسبان التشابهات الدلالية بين الكلمات (مثل "Hello" و "Hi"). تعالج طرق تضمين الكلمات هذه المشكلة، لكنها محدودة بسبب الحاجة إلى مجموعة ضخمة من بيانات النصوص للتدريب. [224] ونتيجةً لذلك، يفضل العلماء استخدام متجهات تضمين الكلمات المدرية مسبقاً. [224] مع ذلك، لا يمكن لهذا النهج أن ينجح مع الكلمات المفقودة منمجموعات بيانات النصوص هذه.

على سبيل المثال، في بعضمجموعات بيانات خدمة الرسائل القصيرة (SMS) يستخدم الناس كلمات ذات معانٍ متعددة، مثل العافية أو الاختصارات، والتي لا تتشابه دليلاً.علاوة على ذلك، لا تدرج الاختصارات في متجهات تضمين الكلمات المدرية مسبقاً. حل هذه المشكلة، يعمل العديد من الباحثين على تنظيف النصوص، كما ناقشنا في القسم 2. تدرب تقنيات تضمين الكلمات ، مثل VecFastText و Word2Vec و GloVe و FastText ببناء على الكلمة وأقرب جار لها ، وهذا ينطوي على قيد بالغ الأهمية (قد يختلف معنى الكلمة في جملتين مختلفتين). حل هذه المشكلة، ابتكر العلماء أساليب جديدة تُسمى تمثيلات الكلمات السياقية، والتي تدرب بناءً على سياق الكلمة في المستند.

كما هو موضح في الجدول 1، نقوم بمقارنة وتقييم كل تقنية بما في ذلك الكلمات الموزونة، وFastText، وIDF، وFT، وGlove، وWord2Vec، وGloVe.

الجدول 1. مقارنة استخلاص الميزات.

القيود	المزايا	نموذج
<b>لا يهدى المطبع بالمعنى سهل التعلم</b> بين مستندين باستخدامه		
لا يستوعب المعنى في النص (الدلالة). • يؤثر الكلمات الشائعة على النتائج	يُعمل مع الكلمات غير المعرفة • (مثال: كلمات جديدة في اللغات)	موزون كلمات
لا يحدد موقع النص (من الناحية النحوية).	سهل الحساب • سهل حساب التشابه بين مستندين باستخدامه	TF-IDF
لا يستوعب المعنى في النص (الدلالة)	مقاييس أساسية لاستخراج أكبر قدر من المعلومات المصطلحات الوصفية في المستند • الكلمات الشائعة لا يؤثر على النتائج الناتجة عن IDF (مثل "am" و "is" وما إلى ذلك)	Word2Vec
لا يمكنها استيعاب معنى الكلمة من النص (تفشل في استيعاب تعدد المعاني)	إنها تحدد موقع الكلمات في النص (التحوي)	فاز (مدرب مسبقاً)
لا يمكنه استخراج الكلمات غير الموجودة في القاموس من المدونة	إنها تلتقط المعنى في الكلمات (علم الدلالة)	
لا يمكنه استخلاص معنى الكلمة من النص (يفشل في استخلاص تعدد المعاني) • استهلاك الذاكرة للت تخزين	إنها تحدد موقع الكلمات في النص (التحوي)	
لا يمكنه استخراج الكلمات غير الموجودة في القاموس من المدونة	إنها تلتقط المعنى في الكلمات (علم الدلالة) تم تدريبه على مجموعة بيانات ضخمة	

## الجدول .1تابع

نموذج	المزايا: يتميز هذا النموذج بـ <b>القيود</b> استخدامه على سبل المثال، لفرض قدرة مجاهات الكلمات على استيعاب العلاقات شبه الخطية في فضاء المتجهات (يتحقق أداوه على Word2vec). كما أنه يقلل من وزن أزواج الكلمات المترددة بكثرة، مثل الكلمات الشائعة كـ "is" و "am" مما يمنعها من التأثير بشكل <b>لا يستهلك الذاكرة للتخزين الحاجة إلى مجموعة بيانات</b>
قفاز (متدرّب)	• ضخمة للتعلم • لا يمكنه استخراج الكلمات غير الموجودة في القاموس من المدونة • لا يمكنها استيعاب معنى الكلمة من النص (تفشل في استيعاب تعدد المعاني)
نص سريع	• لا يمكنه الاستخلاص <b>معنى</b> الكلمة من النص (يفشل في التمييز بين الكلمات مع كلمات أخرى) استيعاب تعدد المعاني) • استهلاك الذاكرة للتخزين أكثر <b>كلفة حسابياً مقارنة بـ Word2Vec و GloVe</b> حل الكلمات الخارجة عن المفردات باستخدام n-gram على مستوى الأحرف
سياسيًّا	• استهلاك الذاكرة للتخزين • تحسين الأداء بشكل ملحوظ على المهام اللاحقة. من الناحية الحسابية، هي أكثر تكلفة مقارنة <b>باتجاهين يعني الكلمة من النص كمية أكبر للحبيبية، ويعامل مع تعدد المعاني</b>
كلمة التمثيلات	• طبقات LSTM والتجذبة الأمامية لا يمكنه استخراج الكلمات غير الموجودة في القاموس من المدونة • يعمل فقط على مستوى الجملة والوثيقة (لا يمكنه العمل على مستوى الكلمة الفردية)

## 6.2. تقليل الأبعاد.

في القسم الثالث، استعرضنا العديد من تقنيات تقليل الأبعاد. في هذا القسم، نناقش مدى فعالية هذه الخطوة فيما يتعلق بوقت الحساب وقوة نظام تصنيف النصوص . يُستخدم تقليل الأبعاد في الغالب لتحسين وقت الحساب وتقليل تعقيد الذاكرة.

تسعى تقنية تحليل المكونات الرئيسية (PCA) إلى إيجاد إسقاطات متعمدة لمجموعة البيانات تحتوي على أعلى تباين ممكن، وذلك لاستخراج الارتباطات الخطية بين متغيرات مجموعة البيانات. يمثل القيد الرئيسي لتقنية PCA في تعقيدها الحسابي عند تقليل الأبعاد . [225] ولحل هذه المشكلة، قدم العلماء تقنية الإسقاط العشوائي (القسم . 3).

تُعد تقنية تحليل التمييز الخططي (LDA) أسلوبًا مُشرقاً لتقليل الأبعاد، يُمكنه تحسين الأداء التنبؤي للميزات المستخرجة. مع ذلك، تتطلب هذه التقنية من الباحثين إدخال عدد المكونات يدوياً، وتتطلب بيانات مُصنفة، وتنتج ميزات يصعب تفسيرها . [226]

يُعد الإسقاط العشوائي أسرع بكثير من الناحية الحسابية من تحليل المكونات الرئيسية. ومع ذلك، فإن هذه الطريقة لا تعمل بشكل جيد مع مجموعات البيانات الصغيرة . [227]  
تتطلب المشفرات التلقائية بيانات أكثر للتدريب مقارنة بطرق تقليل الأبعاد الأخرى، وبالتالي لا يمكن استخدامها كخوارزمية عامة لتقليل الأبعاد بدون بيانات كافية.  
تُستخدم تقنية T-SNE في الغالب لتصور البيانات في مجموعات بيانات النصوص والمستندات.

## 6.3. تقنيات التصنيف الحالية.

في هذا القسم، نناقش القيود والمزايا للنصوص والوثائق الموجودة خوارزميات التصنيف. ثم نقارن أحدث التقنيات في جدولين.

### القيود والمزايا 6.3.1.

كما هو موضح في الجدولين [2](#) و [3](#), فإن خوارزمية روكيو محدودة بقدرها على استرجاع عدد قليل فقط من المستندات ذات الصلة باستخدام هذا النموذج [\[108\]](#). علاوة على ذلك, تُظهر نتائج الخوارزميات العديد من القيود في تصنيف النصوص, والتي يمكن معالجتها من خلال مراعاة الدلالات [\[109\]](#). كما أن لأساليب التعزيز والتجميع العديد من القيود والعيوب, مثل التعقيد الحسابي وفقدان قابلية التفسير [\[117\]](#)[12]. يعمل الانحدار اللوجستي بشكل جيد في التنبؤ بالنتائج الفئوية. ومع ذلك, يتطلب هذا التنبؤ أن تكون كل نقطة بيانات مستقلة [\[124\]](#)[12]. وهو ما يحاول التنبؤ بالنتائج بناءً على مجموعة من المتغيرات المستقلة [\[125\]](#)[12]. كما أن لخوارزمية بايز الساذجة العديد من القيود. تفترض هذه الخوارزمية افتراضًا قويًا حول شكل توزيع البيانات [\[134\]](#), [\[135\]](#)[13]. كما أنها محدودة بقدرة البيانات, حيث يجب تقدير قيمة الاحتمالية لأي قيمة محتملة في فضاء الميزات بواسطة إحصائية تكرارية [\[136\]](#)[13]. أما خوارزمية أقرب جار (KNN) فهي طريقة تصنيف سهلة التنفيذ وتتكيف مع أي نوع من فضاء الميزات. يتعامل هذا النموذج بشكل طبيعي مع حالات التصنيف المتعدد [\[140\]](#), [\[141\]](#)[14]. مع ذلك, يعني نموذج أقرب الجيران (KNN) من قيود تخزين البيانات في مسائل البحث الكبيرة للعثور على أقرب الجيران. إضافًةً إلى ذلك, يعتمد أداء KNN على إيجاد دالة مسافة ذات دالة, مما يجعل هذه التقنية خوارزمية تعتمد بشكل كبير على البيانات [\[142\]](#), [\[143\]](#)[15]. يُعد نموذج آلة المتوجهات الداعمة (SVM) من أكثر خوارزميات التعلم الآلي كفاءةً منذ ظهوره في التسعينيات [\[159\]](#).

إلا أن هذه النتائج محدودة بسبب نقص الشفافية فيها نتيجة العدد الكبير من الأبعاد. ونتيجة لذلك, لا يمكن عرض تقييم الشركة دالة بaramترية تعتمد على النسب المالية أو أي شكل وظيفي آخر [\[159\]](#)[15]. ومن القيود الأخرى معدل النسب المالية المتغير [\[160\]](#).

تُعدّ شجرة القرار خوارزمية سريعة جدًا للتعلم والتنبؤ, لكنها شديدة الحساسية للتغيرات الطفيفة في البيانات [\[166\]](#)[16]. يمكن التغلب على هذه الآثار باستخدام أساليب التحقق والتقليم, لكن هذا الأمر غير واضح تماماً [\[166\]](#)[16]. كما يعني هذا النموذج من مشاكل في التنبؤ خارج العينة [\[168\]](#)[16]. تتميز الغابات العشوائية (أي مجموعات أشجار القرار) بسرعة تدريبها مقارنةً بالتقنيات الأخرى, لكنها بطبيعة الحال مبطأة في خطوة التنبؤ. أما بالنسبة لنموذج الحقول العشوائية الشرطية (CRF), فإن أبرز عيوبه هو التعقيد الحسابي العالي لخطوة التدريب [\[176\]](#)[17]. كما أن هذه الخوارزمية لا تعمل مع الكلمات التي لم تكن موجودة في عينة بيانات التدريب [\[177\]](#)[17]. يُعدّ التعلم العميق من أقوى تقنيات الذكاء الاصطناعي, ويرجع العديد من الباحثين والعلماء على بني التعلم العميق لتحسين مثانته وقدرته الحاسوبية. مع ذلك, يعني بني التعلم العميق من بعض العيوب والقيود عند تطبيقها على مهام التصنيف. من أبرز مشاكل هذا النموذج عدم قدرة التعلم العميق على توفير فهم نظري شامل لعملية التعلم [\[202\]](#)[20]. ومن عيوب أساليب التعلم العميق المعروفة طبيعتها "المبهمة" [\[203\]](#), [\[204\]](#)[20]. أي أن الطريقة التي تُنتج بها أساليب التعلم العميق المخرجات المعقّدة غير مفهومة بسهولة.

من عيوب التعلم العميق أنه يتطلب عادةً بيانات أكثر بكثير من خوارزميات التعلم الآلي التقليدية, مما يعني عدم إمكانية تطبيق هذه التقنية على مهام التصنيف التي تعتمد علىمجموعات بيانات صغيرة [\[205\]](#), [\[206\]](#)[20]. إضافًةً إلى ذلك, فإن الكم الهائل من البيانات الازمة لخوارزميات تصنيف التعلم العميق يزيد من التعقيد الحسابي خلال مرحلة التدريب [\[207\]](#).

الجدول 2. مقارنة تصنيف النصوص (خوارزمية روكيو، والتعزيز، والتجميع، والانحدار اللوجستي، ومصنف بايز الساذج، وأقرب جار، آلة المتجهات الداعمة).

نموذج	المزايا	الضعف
روكيو الخوارزمية	<ul style="list-style-type: none"> <li>سهل التنفيذ • منخفض التكلفة الحسابية • آلية التغذية الراجعة للملاءمة</li> <li>فوائد تصنيف المستندات على أنها غير ذات صلة</li> </ul>	<ul style="list-style-type: none"> <li>لا يستطيع المستخدم استرداد سوى عدد قليل من المستندات ذات الصلة غالباً ما يصنف روكيو النوع بشكل خاطئ على أنه فئة متعددة الوسائط</li> <li>هذا التقنية ليست قوية للغاية لأنّ الجمع الخطّي في هذه الخوارزمية مناسبة لمجموعات البيانات متعددة الفئات.</li> </ul>
تعزيز و التعبئة والتغليف	<ul style="list-style-type: none"> <li>يحسن الاستقرار و الدقة (تستفيد من التعلم الجامعي حيث يتفوق أداء عدة متعلمين ضعفاء على أداء متعلم قوي واحد) • تقليل التباين مما يساعد على تجنب مشاكل التخصيص الزائد</li> </ul>	<ul style="list-style-type: none"> <li>فقدان قابلية التفسير (إذا كان عدد النماذج كبيراً، يصبح فهم النموذج صعباً للغاية)</li> <li>يتطلب ذلك ضبطاً دقيقاً لمختلف المعلمات الفائقية</li> </ul>
الانحدار اللوجستي	<ul style="list-style-type: none"> <li>سهل التنفيذ • لا يتطلب الكثير من الموارد الحاسوبية</li> <li>لا يتطلب الأمر تغيير حجم ميزات الإدخال (المعالجة المسبيقة).</li> <li>لا يتطلب أي ضبط</li> </ul>	<ul style="list-style-type: none"> <li>لا يمكنها حل المشكلات غير الخطية</li> <li>يتطلب التنبؤ أن تكون كل نقطة بيانات مستقلة • محاولة التنبؤ بالنتائج بناءً على مجموعة من المتغيرات المستقلة</li> </ul>
بايز الساذج المصنف	<ul style="list-style-type: none"> <li>يعمل بشكل ممتاز مع البيانات التنصية • سهل التطبيق</li> <li>سرع مقارنة بالخوارزميات الأخرى</li> </ul>	<ul style="list-style-type: none"> <li>افتراض قوي حول شكل توزيع البيانات • محدودية البيانات بسبب ندرة البيانات، حيث يجب تقدير قيمة الاحتمالية لأي قيمة محتملة في فضاء الميزات بواسطة إحصائية التكرار.</li> </ul>
أقرب رقم جار	<ul style="list-style-type: none"> <li>فعال لمجموعات بيانات النصوص • غير معلم • يتم مراعاة المزيد من الخصائص المحلية للنص أو المستند من الصعب إيجاد القيمة المثلث <math>k</math></li> <li>قىود على مسائل البحث الكبيرة لإيجاد أقرب الجيران • بعد إيجاد دالة مسافة ذات معنى أمراً صعباً بالنسبة لبيانات النصوص يتعامل بشكل طبيعي مع مجموعات البيانات متعددة الفئات</li> </ul>	<ul style="list-style-type: none"> <li>حساب هذا النموذج مكلف للغاية</li> </ul>
		مجموعات
آلة الدعم (SVM)	<ul style="list-style-type: none"> <li>يمكن لآلية المتجهات الداعمة (SVM) تمثيل النماذج غير الخطية حدود القرار</li> <li>يؤدي وظائف مشابهة للخدمات اللوجستية</li> <li>الانحدار عند الفصل الخطّي • مقاومة لمشاكل التوفيق الزائد (خاصة بالنسبة لمجموعات بيانات النصوص بسبب الفضاء عالي الأبعاد)</li> <li>تعُد هذه الوظيفة صعبة (عرضة لمشاكل التجاوز / التدريب اعتماداً على النواة) • تعقيد الذاكرة</li> </ul>	<ul style="list-style-type: none"> <li>عدم شفافية النتائج بسبب كثرة الأبعاد (خاصة بالنسبة لبيانات النصوص). • اختيار نواة فعالة</li> </ul>

الجدول 3. مقارنة تصنيف النصوص (شجرة القرار، المدخل الشعواني الشرطي، CRF) (الغابة العشوائية، والتعلم العميق).

نموذج	المزايا	مشكلات القراءة
شجرة القرار	<ul style="list-style-type: none"> <li>يمكنه التعامل بسهولة مع السمات النوعية (الفئوية) • يعمل بشكل جيد مع حدود القرار</li> <li>مواءٌ لمحور الميزة</li> <li>تُعد شجرة القرار خوارزمية سريعة للغاية لكل من التعلم والتنبؤ</li> </ul>	<ul style="list-style-type: none"> <li>مشكلات حدود القراءة القراءة • سهولة الإفراط في التخصيص • حساسية شديدة لاضطرابات الصغيرة في البيانات • مشكلات في التنبؤ خارج العينة</li> </ul>
(CRF)	<ul style="list-style-type: none"> <li>يتميز تصميمه بالمرنة</li> <li>ما أن خوارزمية المدخل الشعواني الشرطي (CRF) تحسن الاحتمالية الشرطية لعقد الإخراج المثلث عالميا ، فإنها تتغلب على عيوب تحيز الفطري.</li> <li>عقل عشوائي</li> <li>يعين مزايا التصنيف والنمذجة الرسمية، مما يتيح نمذجة البيانات متعددة المتغيرات بشكل مختصر</li> </ul>	<ul style="list-style-type: none"> <li>التعقيد الحسابي العالي لخطوة التدريب</li> <li>لا تعمل خوارزمية Rhis مع الكلمات غير المعروفة</li> <li>مشكلة تتعلق بالتعلم عبر الإنترنت</li> <li>( يجعل ذلك من الصعب للغاية إعادة تدريب النموذج عند توفر بيانات أحدث )</li> </ul>
الغابة العشوائية	<ul style="list-style-type: none"> <li>تعتمد على التعلم العشوائى</li> <li>إذاً أو معالجة مسبقةً لبيانات الدخال.</li> <li>زيادة عدد الأشجار في الغابة يزيد من ال وقت</li> <li>تعقيد في خطوة التنبؤ ليس من السهل تفسيرها بصرياً</li> <li>قد يحدث التخصيص الزائد بسهولة يجب اختيار عدد أشجار في الغابة</li> </ul>	<ul style="list-style-type: none"> <li>• تعيق التعلم العشوائى في الغابة بـ التعلم العشوائى تدريبياً مقارنةً بالتقنيات الأخرى. • انخفاض التباين (مقارنةً بالأشجار العادمة). • لا تتطلب</li> </ul>
التعلم العميق	<ul style="list-style-type: none"> <li>يتميز بالمرونة في تصميم الميزات (يقلل الحاجة إلى هندسة الميزات، وهي واحدة من أكثر أجزاء ممارسة التعلم الآلي التقطيع كثافةً) كثافةً من البيانات (إذا كان لديك فقط بيانات نسبة صغيرة، فمن غير المرجح أن يتتفوق التعلم العميق على الأساليب الأخرى).</li> <li>بنية قابلة للتكييف مع المشكلات الجديدة • قادرة على التعامل مع عمليات ربط المدخلات والمخرجات المعقدة • قادر على التعلم على التعلم العميق (أذ يغير التعلم العميق في أغلب الأحيان مشكلة في التعلم العميق (أذ يغير التعلم العميق في أغلب الأحيان صندوقاً أسود). لا يزال إيجاد بنية وهيكلاً فعالة يمثل التحدي الرئيسي لهذه التقنية.</li> </ul> <p>(يسهل ذلك إعادة تدريب النموذج عند ظهور بيانات أحدث)</p>	<ul style="list-style-type: none"> <li>القدرة على المعالجة المتوازية (يمكنها تنفيذ أكثر من مهمة في نفس الوقت)</li> </ul>

### 6.3.2. مقارنة أحدث التقنيات

فيما يتعلق بالجدولين 4 و 5، تتم مقارنة تقنيات تصنيف النصوص وفقاً للمعايير التالية: البنية، المؤلف (المؤلفون)، والنماذج، والحداثة، واستخلاص الميزات، والتفاصيل، ومجموعة النصوص، وقياس التحقق، وقيود كل تقنية. تحتوي كل تقنية (نظام) لتصنيف النصوص على نموذج يمثل خوارزمية التصنيف، كما تتطلب تقنية لاستخلاص الميزات، أي تحويل مجموعة بيانات النصوص أو المستندين إلى بيانات رقمية (كما هو موضح في القسم 2). ويُعد مقياس التحقق عنصراً هاماً آخر في مقارنتنا، حيث يُستخدم لتقدير النظام.

## الجدول 4. مقارنة تقنيات تصنيف النصوص.

القيود	تصديق	مجموعه النصوص	تفاصيل	مثرة استخلاص	بدعة	بيان	المؤلف(ون)	نموذج
يعمل فقط على مجموعات البيانات الهرمية ويسترجع بعض المستندات ذات الصلة.	ماקרו F1	ويكيبيديا	استخدم CUDA على وحدة معالجة الرسومات لحساب المسافات ومقاربتها.	TF-IDF	للمضييف على البيانات الهرمية روكيو	B.J. Sowmya et al. [106]	روكيو الخوارزمية	
ال فقدان قابلية Macro F1-Micro التقسيم	روبرز-87512	خوارزمية التعلم الجماعي	قوس	مع الميزات الدلالية AdaBoost	S. Bloehdorn et al. [114]	تعزيز		
تعتمد نتائج التبؤ على مجموعة من المتغيرات المستقلة	ماקרו F1	RCV1-v2	وهو يعتمد على التوزيع الفاوسي شركة بريورز آند ريدج للخدمات الهرمية الانحدار	TF-IDF	تبليط الانحدار اللوجستي للبيانات هاليفا-البعادنة الانحدار	أ. جينكين وآخرون. [120]	الخدمات الهرمية الانحدار	
تعتمد هذه الطريقة على افتراض قوي بشأن شكل توزيع البيانات	روبرز-87512 - ماקרו	لتقدير توزيع بواسون	توحيد تردد المصطلحات لكل مستند كلمات الأوزان	وتفوّج بواسون متعدد المتغيرات لل الموضوع ظبيقة	كيم، إس. بي. وآخرون [131]	ساذج بايز		
يعجز عن استيعاب تعدد المعاني، كما أن الجوانب الدلالية وال نحوية لا تزال غير محلولة.	ماקרו F1	مجموعات الأخبار و 20 بدقة	يتضمن نموذجاً إحصائياً لقياس الفئة	TF-IDF	تم تقديم TFIGM (تردد الحد وزعم الحادبية العكسي) بي إف آي جي إم	معكوس جاذبة لحظة	ك. تشين وآخرون [148]	SVM و KNN
انعدام الشفافية في النتائج	روبرز-87512 - ماקרו	TF-IDF	النواة هي حاصل ضرب داخلي في فضاء التتشابه باستخدام المبريات المألأ عن جميع المتambilيات الفرعية	خيط تسلسل النواة	استخدام نواة خاصة	H. Lodhi et al. [151]	يدعم متوجه الآلات	
تتميز هذه الخوارزمية بتعقيد حسابي عالي، كما أنها لا تعمل مع الكلمات غير المزنة.	دقة	آراء العملاء	تحسين تحليل المشاعر على مستوى الجملة من خلال تصنيف أنواع الجمل	تضمين الكلمات	قم بتطبيق نموذج تسلسل قائم على الشبكة العصبية لتصنيف الجمل التي تعبر عن الرأي إلى ثلاثة أنواع وفقاً لعدد الأهداف التي تظهر في الجملة	BiLSTM-CRF	ت. تشين وآخرون [175]	شرط عشوائي مجال (CRF)

الجدول 5. مقارنة تقنيات تصنيف النصوص (تابع).

نوع	المؤلف(ون)	بيان	بدعة	ميزة إسخلاص	تفاصيل	مجموع النصوص	تصديق رقمي	القيود
عميق تعلم	Z. Yang et al. [193]	هرمي انتيه الشبكات	له هيكل هرمي بناء	مستويات من الانتهاء توسيع الكلمة على مستوى الكلمة والجملة	تقديرات IMDb و Yelp و Amazon مراجعة	دققة	يعمل فقط على مستوى المستند	
عميق تعلم	ج. تشين وآخرون [228]	الأعصاب العميقه الشبكات	الشبكات العصبية الالتفافية (CNN) باستخدام ثانوي الأبعاد TF-IDF ميزات	حل جديد ل مهمة الكشف عن العوائق اللقطي	تقرير تعليقات	F1-Macro و F1-Micro	تعتمد البيانات على تصميم بنية النموذج	
عميق تعلم	م. جيانغ وآخرون [1]	إيمان عميق شبكة	نموذج تصنيف النصوص المجهين يعتمد على شبكة الاعتقاد العميق وانحدار سوفتماكس.	تحتل شبكة DBN عملية الميزات لحل مشكلة الأبعاد العالية والمصفوفة المفرغة . و يتم استخدام انحدار softmax لتصنيف النصوص.	روبيترز- 87512 و معدل الخطأ	معدل الخطأ	تعد هذه العملية مكلفة حسابياً، ولا تزال قابلية تفسير النموذج تمثل مشكلة في هذا النموذج.	
عميق تعلم	X. Zhang et al. [229]	سي إن إن	الشبكات العصبية الالتفافية على مستوى الأحرف (ConvNets) لتصنيف النصوص	تحتوي الشبكة العصبية الالتفافية على مستوى التصفيق على 6 طبقات التفافية للشخصية والبيانات متصلة بالكامل	بلب، ومراجعات Amazon، وباهوا! بيانات الإجابات تعدين	نسبي أخطاء	تم تصميم هذا النموذج فقط لاكتشاف السمات الثابتة للموقع الخاصة بها	
عميق تعلم	ك. كوصاري [4]	العميق	خوارزمية التعلم العميق الجماعي يجل منشأة إجاد أفضل بنية وهيكل للتعلم العميق (الشبكات العصبية الالتفافية، والشبكات العصبية المدمجة، والشبكات العصبية المتكررة)	مراجعة موقع IMDB، نماذج متعددة عشوائية (RDMIL) التعلم العميق (WOS و 20NewsGroup،	دققة	الحساب مكلف		
عميق تعلم	ك. كوصاري [2]	هرمي بناء	يستخدم مجموعات من بنى التعلم العميق ل توفير فهم متخصص على كل مستوى من مستويات التسلسل الهرمي للواثائق	هرمي عميق التعلم من أجل النصوص (HDLtex)	مجموعة بيانات Web of Science	دققة	يعمل فقط مع مجموعات البيانات الهرمية	

#### 6.4. التقييم

يقيس التقييم التجاري لمصنفات النصوص فعاليتها (أي قدرتها على اتخاذ قرار التصنيف أو التنبؤ الصحيح)، ويستخدم كل من الدقة والاستدامة على نطاق واسع لقياس فعالية مصنفات النصوص. أما الدقة والخطأ (النسبة  $= 1 - \text{الدقة}$ )، فلا يُستخدمان على نطاق واسع في تطبيقات تصنيف النصوص لأنهما غير حساسين للتغيرات في عدد القرارات الصحيحة نظرًا لكبر قيمة المقام. [215]

$$\frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

ويوضح الجدول 6 عيوب كل مقياس من المقاييس المذكورة أعلاه.

#### الجدول 6. مآذق المقاييس.

القيود	
دقة	لا يقدم لنا أي معلومات عن النتائج السلبية الكاذبة (FN) والناتج الإيجابية الكاذبة (FP).
حساسية	لا يقوم بتقييم النتائج السلبية الحقيقية (TN) والناتج الإيجابية الكاذبة (FP) وأي مصنف يتبعها بأن نقاط البيانات إيجابية، والتي تعتبر ذات حساسية عالية.
الخصوصية	يشبه ذلك الحساسية ولا يأخذ في الاعتبار النتائج السلبية الكاذبة والناتج الإيجابية الحقيقية
دقة	لا يُقيّم النتائج السلبية الحقيقة والناتج السلبية الكاذبة، ويُعتبر متحفظاً للغاية، ويتجه نحو الحالة الأكثر ترجيحاً أن تكون إيجابية.

#### 7. استخدام تصنيف النصوص.

في بدايات تاريخ التعلم الآلي والذكاء الاصطناعي، استُخدمت تقنيات تصنيف النصوص في الغالب لأنظمة استرجاع المعلومات. ومع ذلك، ومع تطور التكنولوجيا بمرور الوقت، أصبح تصنيف النصوص وتصنيف الوثائق يستخدم عالمياً في العديد من المجالات، مثل الطب، والعلوم الاجتماعية، والرعاية الصحية، وعلم النفس، والقانون، والهندسة، وغيرها. في هذا القسم، نسلط الضوء على بعض المجالات التي تستخدم تقنيات تصنيف النصوص.

#### 7.1. تطبيقات تصنيف النصوص

##### 7.1.1. استرجاع المعلومات

يُعرف استرجاع المعلومات بأنه إيجاد وثائق من بيانات غير مهيكلة تُلبي حاجة معلوماتية ضمن مجموعات كبيرة من الوثائق. [230] ومع التمو السريع للمعلومات المتوفرة عبر الإنترنت، ولا سيما في شكل نصوص، أصبح تصنيف النصوص أسلوباً بالغ الأهمية لإدارة هذا النوع من البيانات. [231] ومن أهم الأساليب المستخدمة في هذا المجال: Naïve Bayes و MVS و KNN و IBK و 84. و شجرة القرار، [232] و يُعد تطبيق أساليب تصنيف الوثائق لاسترجاع المعلومات من أكثر التطبيقات تحدّياً في معالجة مجموعات بيانات الوثائق والنصوص. [34, 233]

##### 7.1.2. تصفية المعلومات

يشير ترشيح المعلومات إلى اختبار المعلومات ذات الصلة أو رفض المعلومات غير ذات الصلة من تدفق البيانات الواردة. تُستخدم أنظمة ترشيح المعلومات عادةً لقياس وتوقع اهتمامات المستخدمين على المدى الطويل. [234] تُستخدم النماذج الاحتمالية، مثل شبكة الاستدلال البayesian ، بشكل شائع في أنظمة ترشيح المعلومات. تستخدم شبكات الاستدلال bayesian الاستدلال التكراري لنشر القيم عبر شبكة الاستدلال وإعادة المستندات ذات التصنيف الأعلى. [34] تُستخدم باكلي، سي. [235] نموذج فضاء المتجهات مع التحسين التكراري لمهمة الترشيح.

##### 7.1.3. تحليل المشاعر

تحليل المشاعر هو منهج حاسوبي لتحديد الرأي والمشاعر الذاتية في النصوص. [236] تصنف أساليب تصنيف المشاعر المستند المرتبط برأي ما إلى إيجابي أو سلبي. ويفترض أن المستند (د) يعبر عن رأي حول كيان واحد (هـ)، وأن الآراء تتشكل من خلال صاحب رأي واحد (ح). [237] بايزى السادس

يُعد التصنيف وخوارزمية آلة المتجهات الداعمة (SVM) من أكثر أساليب التعلم الخاضع للإشراف شيوعاً والتي استُخدمت لتصنيف المشاعر [238] وقد استُخدمت في تقنيات تصنيف المشاعر ميزات مثل المصطلحات وتكرارها، وأجزاء الكلام، وكلمات وعبارات الرأي، والنفي، والتبعية النحوية.

#### 7.1.4. أنظمة التوصية

تقترن أنظمة التوصية القائمة على المحتوى عناصر المستخدمين بناءً على وصف العنصر وملف تعريف اهتمامات المستخدم [239].

يمكن استخلاص معلومات عن ملف تعريف المستخدم من خلال ملاحظاته (سجل استعلامات البحث أو تقاريره الذاتية) حول المنتجات، بالإضافة إلى السمات المُفتشة ذاتياً (مثل الفلاتر أو الشروط المفروضة على الاستعلامات) في ملفه الشخصي. وبهذه الطريقة، يمكن أن تكون مدخلات أنظمة التوصية شبه مُهيكلة، بحيث تُستخرج بعض السمات من حقل نصي حر، بينما تُحدّد سمات أخرى مباشرةً [240]. وقد استُخدمت أنواع عديدة من أساليب تصنيف النصوص، مثل أشجار القراء، وأساليب أقرب جار، وخوارزمية روكيو، والمصنفات الخطية، والأساليب الاحتمالية، وخوارزمية بايز البيسيطة، لنموذج تفضيلات المستخدم.

#### 7.1.5. إدارة المعرفة

تُعد قواعد البيانات النصية مصادر مهمة للمعلومات والمعرفة، إذ توجد نسبة كبيرة من معلومات الشركات (حوالى 80%) في صيغ بيانات نصية (غير مُهيكلة). في عملية استخلاص المعرفة، تُستخرج الأنماط أو المعرفة من أشكال مباشرة قد تكون شبه مُهيكلة (مثل تمثيل بياني مفاهيمي) أو مُهيكلة/علائقية (مثل تمثيل البيانات). يمكن أن يكون الشكل الوسيط قائماً على المستندات، بحيث يُمثل كل كيان موضوعاً أو مفهوماً ذو أهمية في مجال محدد. يُعدّ تصنيف المستندات أحد أكثر الطرق شيوعاً لاستخراج المعلومات من الأشكال الوسيطة القائمة على المستندات [241] في دراسات أخرى، استُخدم تصنيف النصوص لإيجاد العلاقة بين أسباب حوادث السكك الحديدية والأوصاف المقابلة لها في التقارير [242].

#### 7.1.6. تلخيص الوثائق

يُستخدم تصنيف النصوص في تلخيص الوثائق، حيث قد يتضمن الملخص كلمات أو عبارات غير موجودة في الوثيقة الأصلية [243] كما أن تلخيص الوثائق المتعددة ضروريٌّ نظرًا للزيادة السريعة في المعلومات المتاحة عبر الإنترنت [244]. لذا، يركز العديد من الباحثين على هذه المهمة باستخدام تصنيف النصوص لاستخلاص السمات المهمة من الوثائق.

#### 7.2. دعم تصنيف النصوص

##### 7.2.1. الصحة

تُقدم معظم المعلومات النصية في المجال الطبي بشكل غير منظم أو سريدي، مع استخدام مصطلحات غامضة وأخطاء مطبعية. يجب أن تكون هذه المعلومات متاحة فوراً خلال جمع مراحل التشخيص والعلاج التي يلتقي فيها المريض بالطبيب [245].

يُعد الترميز الطبي، الذي يتضمن تصنيف التشخيصات الطبية إلى فئات محددة مستمدّة من مجموعة كبيرة من التصنيفات، مجالاً بالغ الأهمية في تطبيقات الرعاية الصحية، حيث تُعتبر تقنيات تصنيف النصوص ذات قيمة عالية. في بحث آخر، قدم ج. تشانغ وزملاؤه تقنية Patient2Vec لتعلم تمثيل عميق قابل للتفسير لبيانات السجلات الصحية الإلكترونية الطويلة، وهو تمثيل مُخصص لكل مريض [246]. تُعتبر Patient2Vec تقنية مبتكرة لتضمين ميزات مجموعات البيانات النصية، حيث يمكنها تعلم تمثيل عميق مُخصص وقابل للتفسير لبيانات السجلات الصحية الإلكترونية بالاعتماد على الشبكات العصبية المترکزة وآلية الانتباه. كما ظهرت تقنيات النصوص في تطوير رؤوس الموضوعات الطبية (MeSH) [247] وعلم الجينات (GO).

### 7.2.2. العلوم الاجتماعية

ازداد استخدام تصنيف النصوص وتصنيف الوثائق في فهم السلوك البشري خلال العقود الماضية. [38, 248] وقد ركزت الجهود الحديثة القائمة على البيانات في أبحاث السلوك البشري على استخراج اللغة من الملاحظات غير الرسمية ومجموعات البيانات النصية، بما في ذلك الرسائل النصية القصيرة (SMS) والملاحظات السريرية ووسائل التواصل الاجتماعي، وغيرها. [38] وركزت هذه الدراسات في الغالب على استخدام مناجع تعتمد على تكرار الكلمات (أي عدد مرات ظهور الكلمة في الوثيقة) أو على خصائص تستند إلى عد الكلمات في الاستقصاء اللغوي، [249] وهو معجم مدقق جيداً لفئات الكلمات ذات الصلة النفسية. [250]

### 7.2.3. الأعمال والتسويق

تستخدم الشركات والمؤسسات الرابحة وسائل التواصل الاجتماعي بشكل متزايد لأغراض التسويق. [251] ويُعد استخراج البيانات من منصات التواصل الاجتماعي مثل فيسبوك وتويتر وغيرها هدفاً رئيسياً للشركات لزيادة أرباحها بسرعة. [252] كما يُعد تصنيف النصوص والوثائق أدلة فعالة للشركات للوصول إلى عملائها بسهولة أكبر.

### 7.2.4. القانون

أنتجت المؤسسات الحكومية كميات هائلة من المعلومات والوثائق القانونية. ولا يقتصر دور استرجاع هذه المعلومات وتصنيفها تلقائياً على مساعدة المحامين فحسب، بل يشمل أيضاً موكليهم. [253] في الولايات المتحدة، يستمد القانون من خمسة مصادر: القانون الدستوري، والقانون التشريعي، والمعاهدات، واللوائح الإدارية، والقانون العام. [254] ويتم إصدار العديد من الوثائق القانونية الجديدة سنوياً، ويعُد تصنيف هذه الوثائق التحدي الأكبر الذي يواجه مجتمع المحامين.

## 8. الاستنتاجات

تُعد مهمة التصنيف من أهم المشكلات في مجال التعلم الآلي. ومع تزايد مجموعات بيانات النصوص والوثائق، أصبح تطوير وتوسيع خوارزميات التعلم الآلي الخاضعة للإشراف ضرورةً ملحةً. لا سيما في تصنيف النصوص. ويتباطّب وجود نظام تصنيف وثائق أفضل لهذه المعلومات فهو دقيقاً لهذه الخوارزميات. ومع ذلك، تعمل خوارزميات تصنيف النصوص بشكل رئيسي على النحو التالي: (أ) أساليب استخلاص الميزات وكيفية تقديرها بشكل صحيح. حالياً، يمكن تصنيف خوارزميات تصفييف النصوص بشكل رئيسي على النحو التالي: (أ) أساليب استخلاص الميزات، مثل تردد المصطلح - تردد الوثيقة العكسي، (TF-IDF) وتردد المصطلح، (TF) وتضمين الكلمات (مثل، Word2Vec وتمثيل الكلمات السياقية، والمتوجهات العالمية لممثل الكلمات، (FastText)، (GloVe) وـ (FastText)، (GloVe) تُستخدم على نطاق واسع في التطبيقات الأكادémية والتجارية. وقد تناولنا هذه التقنيات في هذه الورقة. ومع ذلك، يمكن أن يُساهم تنظيف النصوص والوثائق في تحسين دقة التطبيق وقوته. وقد وصفنا الأساليب الأساسية لخطوة المعالجة المسبقة للنصوص. (ب)

تُعدّ أساليب تقليل الأبعاد، مثل تحليل المكونات الرئيسية، (PCA) وتحليل المصفوفات غير السالبة، (NMF) والإسقاط العشوائي، والمشفر الثنائي، وتضمين الجوار العشوائي الموزع، (t-SNE) مفيدةً في تقليل تعقيد الوقت والذاكرة لخوارزميات تصنيف النصوص الحالية. وقد عُرضت في قسم منفصل أكثر أساليب تقليل الأبعاد شيئاً، (ثالثاً) تُركّز هذه الورقة البحثية بشكل أساسى على خوارزميات التصنيف الحالية، مثل خوارزمية روكيو، والتجميع والتعمير، والانحدار اللوجستي، (LR) ومصنف بايز الساذج، (NBC) وخوارزمية أقرب جار، (KNN) وآلة المنتجات الداعمة، (SVM) ومصنف شجرة القرار، (DTC) والغاية العشوائية، والحقول العشوائي الشرطي، (CRF) والتعلم العميق. (رابعاً) تم شرح أساليب التقييم، مثل الدقة، و،  $F\beta$  ومعامل ارتباط ماثيوز، (MCC) وخصائص تشغيل المستقبل، (ROC) والمساحة تحت المنحنى، (AUC).

باستخدام هذه المقاييس، يمكن تقييم خوارزمية تصنيف النصوص. (خامساً) تم تناول القيود الحرجة لكل مكون من مكونات مسار تصنيف النصوص (أي استخراج الميزات، وتقليل الأبعاد، وخوارزميات التصنيف الحالية، والتقييم) من أجل كل تقنية. وأخيراً،

قارن بين أكثر خوارزميات تصنيف النصوص شيئاً في هذا القسم. (خامسًا) أخيرًا، يُقطع استخدام تصنيف النصوص كتطبيق وأو دعم لتخصصات أخرى مثل التخصصات العامة والطبية، وما إلى ذلك، في قسم منفصل.

في هذا الاستطلاع، تمت مناقشة التقنيات الحديثة واجهات خوارزميات تصنيف النصوص.

مساهمات المؤلفين: عمل كل من KJK و GitHub على فكرة وتصميم المنصة، كما عمل على نماذج التعليمات البرمجية على منصة GitHub لجميع هذه النماذج. قام كل من MS و MH بتنظيم الورقة ومراجعتها. هذا العمل تحت إشراف LB و DB.

التمويل: تم دعم هذا العمل من قبل مختبر أبحاث الجيش الأمريكي بموجب المنحة W911NF-17-2-0110.

شكراً وتقدير: يود المؤلفون أن يشكرؤ مايكل إس. جيربر على ملاحظاته وتعليقاته.

تضارب المصالح: لا يعلن المؤلفون عن أي تضارب في المصالح. لم يكن للجهات الممولة أي دور في تصميم الدراسة، أو في جمع البيانات أو تحليلها أو تفسيرها، أو في كتابة المخطوطة، أو في قرار نشر النتائج.

## مراجع

1. جانب، م؛ ليانغ، ي؛ فينغ، ش؛ فان، ش؛ باي، ز؛ شيووه، ي؛ غوان، ر. تصنيف النصوص بناءً على الاعتقاد العميق الشبكة وإنحدار سوفتماس. *تطبيقات الحوسبة العصبية* [CrossRef]. 2018, 29, 61-70.
2. كوصاري، ك؛ براون، دي إي؛ حيدري صفا، م؛ جعفرى ميماندى، ك. HDLTex: التعلم العميق الهرمي لتصنيف النصوص. التعلم الآلي والتطبيقات (ICMLA). 2017، 26-27 يونيو، 1998، المجلد 752، 41-48.
3. ماكالوم، أ.؛ نيفام، ك. مقارنة نماذج الأحداث لتصنيف النصوص باستخدام خوارزمية بايز البسيطة. في وقائع ورشة عمل AAAI-98 حول التعلم لتصنيف النصوص، ماديسون، ويسكونسن، الولايات المتحدة الأمريكية، 27-26 يونيو، 1998، المجلد 11، 3206098. doi:10.1145/3206098.3206111.
4. كوصاري، ك؛ حيدري صفا، م؛ براون، دي؛ جعفرى ميماندى، ك؛ بارنز، ل. RMDL: التعلم العميق متعدد النماذج العشوائي للتصنيف. في وقائع المؤتمر الدولي لعام 2018 حول نظم المعلومات واستخراج البيانات، ليكلاند، فلوريدا، الولايات المتحدة الأمريكية، 11-12 أبريل، 2018: doi:10.1145/3206098.3206111.
5. حيدري صفا، م؛ كوثري، ك.؛ براون، دي؛ جعفرى ميماندى، ك.؛ بارنز، ل. تحسين تصنيف البيانات باستخدام التعلم العميق متعدد النماذج العشوائي (RMDL). *المجلة الدولية للتعلم الآلي والتعلم الآلي* [CrossRef]. 2018, 8, 298-310.
6. لاي، إس؛ شو، إل؛ ليو، ك.؛ تشاو، ج. الشبكات العصبية الالتفافية المتكررة لتصنيف النصوص. في وقائع المؤتمر التاسع والعشرين لجمعية الذكاء الاصطناعي الأمريكية (AAAI)، أوستن، تكساس، الولايات المتحدة الأمريكية، 30-25 يناير، 2015، المجلد 333، الصفحات 2267-2273.
7. أغراوال، س. سي؛ تشاي، س. مسح لخوارزميات تصنيف النصوص. في كتاب استخراج البيانات التصبية؛ سيرينغر: برلين/هابيلر، ألمانيا، 2012، 163-222.
8. أغراوال، س. سي؛ تشاي، س. إكس، اشتراك البيانات التصبية؛ سيرينغر: برلين/هابيلر، ألمانيا، 2012، 163-222.
9. سالتون، جي؛ باكتي، س. أساليب ترجيح المصطلحات في استرجاع النصوص الآلي. معالجة المعلومات وإدارتها. 1988، 24، 513-523. [CrossRef]
10. عولادي، ر.؛ إيفي، أو. شرح Word2vec: اشتقاق تضمين الكلمات باستخدام أخذ العينات السلبية لميكروفواف آخر. طريقة. أرخايف. 2014، 2041-2273.
11. برينجتون، ج؛ سوشر، ر؛ مانينج، س. د. غلوف: المتجهات العالمية لتمثيل الكلمات. في وقائع مؤتمر 2014 حول الأساليب التجريبية في معالجة اللغة الطبيعية (EMNLP). قطر، 29-25 أكتوبر، 2014، المجلد 14، 1532-1543، الصفحات 1532-1543.
12. آماميتسوكا، ن. ه. استراتيجيات تعلم الاستعلام باستخدام التعزيز والتجميع. في: تعلم الآلة: وقائع المؤتمر الدولي الخامس عشر (ICML'98)؛ دار مورغان كوفمان للنشر؛ برينجتون، ماساتشوستس، الولايات المتحدة الأمريكية، 1998، المجلد 1.
13. كيم، واي إتش؛ هان، إس واي؛ تشانغ، بي تي. تصفية النصوص باستخدام مصنفات بايز الساذحة المعززة. في وقائع المؤتمر الدولي السنوي الثالث والعشرين لجمعية ACM SIGIR حول البحث والتطوير في استرجاع المعلومات، ألبينا، اليونان، 24-28 يونيو، 2000، 168-175.
14. شابير، ر. إي؛ سينغر، ي. بوس تكستر: نظام قائم على التعزيز لتصنيف النصوص. تعلم الآلة. 2000، 39، 135-168. [CrossRef]
15. هاريل، إف إي. الانحدار اللوجستي التربيني. في استراتيجيات نمذجة الانحدار؛ سيرينغر: برلين/هابيلر، ألمانيا، 2001، 331-343.

16. هوسمير، د. و. الابن؛ ليميشو، س.؛ ستوريديفانت، ر. إكس. الانحدار اللوجستي التطبيقي؛ جون وايلي وأولاده: هووكين، نيوجيرسي، الولايات المتحدة الأمريكية، 2013، المجلد 398.

17. دو، ج؛ ياماغيشي، ه.؛ تشو، ز.؛ يونس، أ.ب.؛ تشين، س.و. 1.081-6.1 دراسة مقارنة بين نموذجي الانحدار اللوجستي الثنائي (BLR) والشبكة العصبية الاصطناعية (ANN) للتنبؤ المكانى بالانهيارات الأرضية على المستوى الإقليمي باستخدام نظم المعلومات الجغرافية. في: ديناميكيات الانهيارات الأرضية: أدوات تعليمية تفاعلية للأنهيارات الأرضية من ICL-ISB-Sympo: برلين/هايدلبرغ، ألمانيا، 2018، ص. 139-151.

18. تشن، و.؛ شي، ش.؛ وانغ، ج.؛ بربادن، ب.؛ هوونغ، ه.؛ بوي، د.ت.؛ دون، ز.؛ ما، ج. دراسة مقارنة لنماذج شجرة الانحدار اللوجستي، والغاية العشوائية، وشجرة التصنيف والانحدار للتنبؤ المكانى بقابلية حدوث الانهيارات الأرضية. كانتا [CrossRef] 2017, 151, 147-160.

19. لارسون، آر. آر. مقدمة في استرجاع المعلومات. مجلة الجمعية الأمريكية لعلوم وتكنولوجيا المعلومات. 2010, 61, 852-853. [CrossRef]

20. لـ، لـ؛ وابيرغ، سـ آرـ؛ دارـنـ، تـ إـيـهـ؛ بـيـدرـسـنـ، إـلـ جـيـ. اختيار الجينات لتصنيف العينات بناءً على بيانات التعبير الجيني: دراسة حساسية اختيار معلمات طريقة GA/KNN. 2001, 17, 1131-1142. [CrossRef]

21. مانيفيتز، إـلـ؛ يـوسـفـ، إـمـ. آلات المـتجـهـاتـ الدـاعـمـةـ أحـادـيـةـ الفـتـنـةـ لـتـصـنـيـفـ الـمـسـتـنـدـاتـ. مجلـةـ أـبـجـاثـ تـعـلـمـ الـأـلـةـ. 2001, 2, 139-154.

22. هـانـ، إـيـ إـشـ إـسـ؛ كـارـيـسـ، جـيـ. تـصـنـيـفـ الـمـسـتـنـدـاتـ القـائـمـ عـلـىـ الـمـرـكـزـ: الـتـحلـيلـ وـالـنـتـائـجـ الـتـجـرـيـبـيـةـ. في المؤتمر الأوروبي حول مبادئ استخراج البيانات واكتشاف المعرفة: سـيرـينـغـ: برـلـينـ/ـهاـيدـلـبـرـغـ، أـلمـانـيـاـ، 2000ـ الصـفـحـاتـ 424ـ431ـ.

23. Xu, B.; Guo, X.; Ye, Y.; Cheng, J. مصنف الغابات العشوائية المحسن لتصنيف النصوص. JCP 2012, 7, 2913-2920. [CrossRef]

24. لـافـيرـتـيـ، جـ. مـاكـالـومـ، أـ.؛ بـيـرـيرـاـ، فـ.ـسـ. الحقـولـ الـعـشـوـائـيـةـ الشـرـطـيـةـ: نـمـاذـجـ اـحـتمـالـيـةـ لـتـقـسـيمـ وـتـصـنـيـفـ بـيـانـاتـ التـسـلـسـلـ. في وـاقـعـ الـمـؤـتـمـرـ الـدـولـيـ الـثـامـنـ عـشـرـ لـلـتـعـلـمـ الـآـلـيـ، 2001ـ ICMLـ وـبـلـيـمـراـتونـ، مـاسـاـشـوـسـيـسـ، الـولـاـيـاتـ الـمـتـحـدةـ الـأـمـرـيـكـيـةـ، 28ـ يـونـيوـ 1ـ يولـيوـ 2001ـ.

25. تـشـينـ، دـ.ـ؛ صـنـ، جـيـهـ تـيـ؛ لـيـ، إـتـشـ؛ يـانـغـ، كـيـوـ؛ تـشـينـ، زـدـ. تـلـخـيـصـ الـمـسـتـنـدـاتـ باـسـتـخـارـ الـحـقـولـ الـعـشـوـائـيـةـ الشـرـطـيـةـ. IJCAI 2007, 7, 2862-2867.

26. تـشـانـغـ، سـيـ. استـخـارـ الـكـلـمـاتـ الـمـفـاتـحـيـةـ تـلـقـائـيـاـ مـنـ الـمـسـتـنـدـاتـ باـسـتـخـارـ الـحـقـولـ الـعـشـوـائـيـةـ الشـرـطـيـةـ. مجلـةـ الـجـوسـيـةـ. Inf. Syst. 2008, 4, 1169-1180.

27. لـوـكـونـ، واـيـ؛ بـيـنـجـيوـ، واـيـ؛ هـيـنـتوـنـ، جـيـ. التـعـلـمـ الـعـمـيقـ. نـيـتـشـرـ 2015, 521, 436-444. [CrossRef] [PubMed]

28. هـوـانـغـ، جـيـهـ؛ لـيـنـغـ، سـيـ؛ إـكـسـ. اـسـتـخـارـ الـمـسـاحـةـ تـحـتـ الـمـنـحـنـيـ وـالـدـقـقـةـ فـيـ تـقـيـيـمـ خـواـزـمـيـاتـ الـتـعـلـمـ. معـالـمـاتـ IEEEـ الـلـمـعـرـفـةـ. 2005, 17, 299-310. [CrossRef]

29. لـوكـ، جـيـ. نقـصـ الـتـرـوـيـةـ الـمـسـارـيـقـيـةـ الـحـادـ: التـصـنـيـفـ وـالـتـقيـيـمـ وـالـعـلاـجـ. مجلـةـ الـجـيـجـيـكاـ. Acta Gastro-Enterol. 2002, 29, 220-225.

30. مـاـيـوـزـ، بـيـ دـبـلـيـوـ مـقـارـنـةـ الـبـيـنـةـ الـثـانـوـيـةـ الـمـتـوـقـعـةـ وـالـمـلـاـحظـةـ لـإـنـزـيمـ الـلـبـزـوـزـيـمـ لـفـيـروـسـ T4ـ. Biochim. Biophys. Acta (BBA)-Protein Struct. 1975, 405, 442-451. [CrossRef]

31. هـانـلـيـ، جـيـهـ إـيـهـ؛ مـاـكـيـلـ، بـيـ. معـنـىـ وـاسـتـخـارـ الـمـسـاحـةـ تـحـتـ منـحـنـيـ خـصـائـصـ تـشـغـيلـ الـمـسـتـقـبـلـ. ROCـ عـلـمـ الـأشـعـةـ 1982, 143, 29-36. [CrossRef]

32. بـيـنـسـيـنـاـ، إـمـ؛ دـاغـوـسـتـيـنـوـ، آرـ بـيـ؛ فـاسـانـ، آرـ إـسـ. تـقـيـيـمـ الـقـدـرـةـ الـتـنـبـؤـيـةـ الـإـضـافـيـةـ لـعـلـمـةـ جـديـدـةـ منـ الـمـسـاحـةـ تـحـتـ منـحـنـيـ ROCـ إـلـىـ إـعادـةـ التـصـنـيـفـ وـمـاـ بـعـدـهـ. مجلـةـ الـإـحـصـاءـ الـطـبـيـ. 2008, 27, 157-172. [CrossRef]

33. جـاكـوبـسـ، بـيـ إـسـ الـأـنـظـمـةـ الـذـكـيـةـ الـقـائـمـةـ عـلـىـ النـصـوـصـ: الـبـحـوـثـ وـالـمـارـسـاتـ الـحـالـيـةـ فـيـ اـسـتـخـارـ الـمـعـلـومـاتـ وـاسـتـرـجـاعـهـاـ. 2008.

34. كـوـرـفـتـ، دـبـلـيـوـ بـيـ؛ مـيـتـزـلـرـ، دـيـ؛ سـتـرـوـمـانـ، تـيـ. مـحـركـاتـ الـبـحـثـ: اـسـتـرـجـاعـ الـمـعـلـومـاتـ فـيـ الـمـارـسـةـ الـعـمـلـيـةـ: أـدـيـسـونـ وـبـيـسـلـيـ وـرـيـدـينـغـ: بـوـسـطـنـ، مـاسـاـشـوـسـيـسـ، الـولـاـيـاتـ الـمـتـحـدةـ الـأـمـرـيـكـيـةـ، 2010ـ المـجلـدـ 283ـ.

35. يـمـاـحـيـ، مـ؛ كـوـصـارـيـ، لـكـ؛ شـيـنـ، سـ؛ بـيـرـوكـفـيـشـ، سـ. تقـنـيـةـ فـعـالـةـ لـلـبـحـثـ فـيـ الـمـلـفـاتـ الـكـبـيـرـةـ جـدـاـ بـعـدـ ضـبـابـيـةـ باـسـتـخـارـ مـبـدـأـ خـانـةـ الـحـامـمـ. في وـاقـعـ الـمـؤـتـمـرـ الـدـولـيـ الـخـامـسـ للـجـوسـيـةـ منـ أـجـلـ الـبـحـوـثـ وـالـتـطـبـيقـاتـ الـجـغـرافـيـةـ الـمـكـانـيـةـ لـعـامـ 2014ـ واـشـنـطـنـ الـعـاصـمـةـ، الـولـاـيـاتـ الـمـتـحـدةـ الـأـمـرـيـكـيـةـ، 6ـ أـغـسـطـسـ 2014ـ الصـفـحـاتـ 82-86ـ.

36. تـشـ، زـ؛ جـانـفـيكـيـوـ، سـ؛ وـانـغـ، هـ؛ جـاجـودـيـاـ، سـ. مـنـ يـغـرـدـ عـلـىـ توـيـتـرـ: إـنـسـانـ، روـبـوتـ، أـمـ سـاـيـبورـغـ؟ في وـاقـعـ الـمـؤـتـمـرـ السـنـوـيـ الـسـادـسـ وـالـعـشـرـينـ لـتـطـبـيقـاتـ أـمـنـ الـحـاسـوبـ، أـوـسـتنـ، تـكـسـاسـ، الـولـاـيـاتـ الـمـتـحـدةـ الـأـمـرـيـكـيـةـ، 2010ـ دـيـسـمـبـرـ 21ـ30ـ.

37. غـورـدونـ، آرـ إـسـ، الـابـنـ. تـصـنـيـفـ عـلـىـ لـلـوـقـاـيـةـ مـنـ الـأـمـراضـ. تـقارـيرـ الصـحةـ الـعـامـةـ 1983, 98, 107ـ.

[PubMed]

38. تـولـزـ، أـلـ؛ غـيلـنـ، جـ.ـ؛ كـوـصـارـيـ، لـكـ؛ بـارـزـ، لـ.ـ؛ تـيـتـشـمـانـ، بـأـ.ـ؛ تـدـدـيـدـ خـطـرـ الـانـتـخـارـ الـوـشـكـ بـيـنـ الشـيـابـ باـسـتـخـارـ الرـسـائـلـ النـصـيـةـ. في وـاقـعـ مؤـتـمـرـ CHIـ لـعـامـ 2018ـ حولـ الـعـوـامـلـ الـبـشـرـيـةـ فـيـ الـأـنـظـمـةـ الـجـوسـيـةـ، مـونـتـرـالـ، كـيـبـكـ، كـنـداـ، 26ـ27ـأـبـرـيلـ 2018ـ صـ3ـ.

39. غوبتا، جي؛ مالهوترا، إس. تجزئة المستندات النصية لحساب تكرار الكلمات باستخدام برنامج رايد ماينر (أخذ السيرة الذاتية كمثال). *المجلة الدولية لتطبيقات الحاسوب*. 2015, 975, 8887.
40. فيرما، ت؛ رينو، ر؛ غور، د. عملية التجزئة والتصفيحة في رايد ماينر. *المجلة الدولية لنظم المعلومات التطبيقية*. 2014, 7, 16-18. [CrossRef]
41. أغراوال، التعلم الآلي للنصوص؛ سيرينغر؛ برلين/هايدلبرغ، ألمانيا. 2018.
42. سيف، ح؛ فرنانديز، م؛ هي، ي؛ علانى، ح. حول الكلمات الموقوفة، والترشيح، وندرة البيانات لتحليل المشاعر على تويتر. في وقائع المؤتمر الدولي التاسع حول موارد اللغة وتقديرها. (LREC 2014) ريكافيوك، أيسلندا، 26-31 مايو 2014.
43. غوبتا، ف؛ ليهال، ج-س. مسح لتقنيات وتطبيقات استخراج النصوص. *مجلة التقنيات الناشئة وذكاء الوب*. 2009.
44. دلال، م.ك؛ زافيري، م.أ. التصنيف الآلي للنصوص: مراجعة فنية. *المجلة الدولية لتطبيقات الحاسوب*. 2011, 28, 37-40. [CrossRef]
45. ويتنى، دي إل؛ إيفانز، بي ديليو. اختصارات أسماء المعادن المكونة للصخور. *المجلة الأمريكية لعلم المعادن*. 2010, 95, 185-187.
46. هيلم، أ. التعافي والاستعادة: رحلة في فهم من نحن وما هيتنا. في التمريض النفسي والصحة العقلية: فن الرعاية؛ روتلidge: لندن، المملكة المتحدة، 2003 ص. 50-55.
47. Dhuliawala, S.; Kanojia, D.; Bhattacharyya, P. *SlangNet: مورد يشبه WordNet للغة العامية الإنجليزية*. في وقائع مؤتمر LREC 2016، 23-28 مايو 2016.
48. باهوا، ب؛ تارونا، س؛ كاسليوال، ن. تحليل المشاعر -استراتيجية للمعالجة المسبقة للنصوص. *المجلة الدولية لتطبيقات الحاسوب*. 2018, 180, 15-18. [CrossRef]
49. ماواردي، في سي؛ سوسانتو، ناغا، دي إس. تصحيح الأخطاء الإملائية في المستندات النصية باللغة الإندونيسية باستخدام الأوتومات ذات الحالة المحدودة وطريقة مسافة ليفينشتاين. *مجلة علوم معالجة البيانات الإلكترونية*. 2018, 164. [CrossRef]
50. زيدايك، ج؛ هنريكسون، أ؛ دونيلد، م. تحسين ربط المصطلحات في النصوص السريرية باستخدام تصحيح إملائي حساس للسياق. في المعلوماتية الصحية: الصحة العامة المتصلة بقيادة المواطنين؛ دار نشر IOS؛ أمستردام، هولندا، 2017؛ المجلد 235، الصفحات 241-245.
51. ماواردي، في سي؛ روبي، آر؛ ناجا، دي إس تصحيح سريع ودقيق للتجهزة باستخدام شجرة البحث والثانيات. *تيلكومنيكا (الاتصالات والحوسبة والإلكترونيات والتحكم)*. 2018, 16, 827-833. [CrossRef]
52. سيبروف斯基، ك؛ ستيفانوسكا، إ؛ كولاكوف، أ؛ بوبيسكا، ز؛ فيلينوف، ج. مقارنة أداء نماذج مختلفة في مهمة تصنيف المستندات. في وقائع المؤتمر الدولي الثامن حول ذكاء الوب والتقييم والدلائل، نوفي ساد، صربيا، 25-27 يونيو 2018؛ ص. 10.
53. سينغ، ج؛ غوبتا، ف. تجذير النصوص: المناهج والتطبيقات والتحديات. *مجلة ACM للحوسبة (CSUR)*. 2016, 49, 45. [CrossRef]
54. سامبسون، ج. جدول "غيرة اللغة": طبعة منقحة؛ A&C Black: لندن، المملكة المتحدة، 2005.
55. بليسون، ج؛ لافراك، ن؛ ملدينبيتش، د. نهج قائم على القواعد لتحليل الكلمات إلى جذورها. في وقائع المؤتمر الدولي السابع متعدد التخصصات لجمعية المعلومات IS لليوبليانا، سلوفينيا، 14-13 أكتوبر 2004.
56. كورنيوس، ت؛ لوريكا، ج؛ يارفيلين، ك؛ جوهولا، م. التجريد والتحليل الصرفي في تجميع وتألق النصوص الفنلندية. في وقائع المؤتمر الدولي الثالث عشر لجمعية الحوسبة الآلية حول إدارة المعلومات والمعرفة، واشنطن العاصمة، الولايات المتحدة الأمريكية، 13-18 نوفمبر 2004؛ ص. 625-633.
57. كاربوريسو، إم؛ ماتوين، إس. ما وراء حقيقة الكلمات: تمثيل نصي لاختيار الجملة. في مؤتمر الجمعية الكندية للدراسات الحاسوبية للذكاء؛ سيرينغر؛ برلين/هايدلبرغ، ألمانيا، 2006؛ ص. 324-335.
58. سيدوروفر، ج؛ فيلاسكيز، ف؛ ساتاماتاتوس، إ؛ جيلوخ، أ؛ تشانونا-هيرنانديز، ل. النماذج النحوية القائمة على التباعية التركيبية كميزات تصنيف. في المؤتمر المكسيكي الدولي للذكاء الاصطناعي؛ سيرينغر؛ برلين/هايدلبرغ، ألمانيا، 2012؛ ص. 11-12.
59. سبارك جونز، ك. تفسير إحصائي لخصوصية المصطلح وتطبيقه في الاسترجاع. *مجلة التوثيق*. 1972, 28, 11-21. [CrossRef]
60. توكوناغا، ت؛ ماكونو، إ. تصنيف النصوص بناءً على التردد العكسي المرجح للمستندات. *معالجة المعلومات*. Soc. Jpn. SIGNL 1994, 94, 33-40.
61. ميكولوف، ت؛ تشين، ك؛ كورادو، ج؛ دين، ج. التقدير الفعال لمتمثيلات الكلمات في الفضاء المتجهي. arXiv 2013, arXiv:1301.3781.
62. ميكولوف، ت؛ سوتسيكير، إ؛ تشين، ك؛ كورادو، ج. التمثيلات الموزعة للكلامات والعبارات وتركيبها. Adv. Neural Inf. Process. Syst. 2013, 26, 3111-3119.
63. word2vec. arXiv 2014, arXiv:1411.2738.

64. ماتن، إل في دي؛ هينتون، جي. تصوير البيانات باستخدام SNE-مجلة أبحاث التعلم الآلي. 2008, 9, 2579-2605.
65. بوجانوفسكي، ب.; جريف، إ؛ جولين، أ؛ ميكولوف، ت. إثراء متجهات الكلمات بمعلومات الكلمات الفرعية. arXiv 2016. arXiv:1607.04606.
66. ميلامود، أ؛ غولديبرغر، ج؛ داغان، إ. تعلم تصميم السياق العام باستخدام LSTM ذاتي الاتجاه. في وقائع المؤتمر العشرين لـ SIGNLL حول التعلم الحاسوبي للغة الطبيعية، برلين، ألمانيا، 11-12 أغسطس 2016! الصفحات 51-61.
67. بيترز، إم إيه؛ نيومان، إم؛ إير، إم؛ غاردنر، إم؛ كلارك، س؛ لي، ك؛ زيتلوبير، إل. تمثيلات الكلمات السياقية العميقه. arXiv 2018. arXiv:1802.05365.
68. عبدي، ح؛ ويليامز، ل. ج. تحليل المكونات الرئيسية. مجلة وايلي متعددة التخصصات لمراجعات الإحصاء الحاسوبي. 2010, 2, 433-459.
69. جوليف، آي تي؛ كاديماء، جие. تحليل المكونات الرئيسية: مراجعة وتطورات حديثة. معاملات الجمعية الملكية الفلسفية. Soc. A 2016, 374, 20150202. [CrossRef]
70. نغ، أ. تحليل المكونات الرئيسية. الخوارزميات التوليدية، والتنظيم، و اختيار النموذج، CS 2015 229, 71.
71. كاو، ل؛ تشوا، ك. س؛ تشونغ، و؛ لي، ه؛ غو، ك. مقارنة بين تحليل المكونات الرئيسية (PCA) وتحليل المكونات المستقلة (ICA) من حيث الأبعاد انخفاض في آلة المتجهات الداعمة. الحوسية العصبية [CrossRef] 2003, 55, 321-336.
72. هيرولت، ج. الشبكات العصبية ذات المشابك القابلة للتتعديل: فك تشفير الرسائل الحسية المركبة عن طريق التعلم غير الخاضع للإشراف والمستمر. CR Acad. Sci. Paris 1984, 299 525-528.
73. جوت، س.؛ هيرولت، ج. الفصل الأعمى للمصادر، الجزء الأول: خوارزمية تكيفية قائمة على المحاكاة العصبية الهندسة المعمارية. معالجة الإشارات. 1991, 24, 1-10. [CrossRef]
74. هيغارين، أ.؛ هوير، ب. و؛ إنكي، م. تحليل المكونات المستقلة الطبوغرافية. الحوسية العصبية، 2001, 13, 1527-1558. [CrossRef] [PubMed]
75. هيغارين، أ.؛ أوجا، إ. تحليل المكونات المستقلة: الخوارزميات والتطبيقات. الشبكات العصبية. 2000, 13, 411-430. [CrossRef]
76. سوجياما، م. تقليل أبعاد البيانات المصنفة متعددة الوسائل بواسطة تحليل التمييز المحلي لفيشر. J. Mach. Learn. Res. 2007, 8, 1027-1061.
77. بالاكريشنا، س.؛ غالابايراجو، أ. تحليل التمييز الخطى - دليل موجز. معهد معالجة الإشارات والمعلومات. 1998, 18, 1-8.
78. سوجياما، م. تحليل التمييز المحلي لفيشر لتقليل الأبعاد الخاضع للإشراف. في وقائع المؤتمر الدولي الثالث والعشرين للتعلم الآلي، بيتسبurg، بنسلفانيا، الولايات المتحدة الأمريكية، 29-25 يونيو 2006 ص 905-912.
79. باوكا، نائب الرئيس؛ شاهناز، ف؛ بيري، م؛ بلومونز، رج. استخراج النصوص باستخدام تحليل المصروفات غير السالبة. في وقائع المؤتمر الدولي لجمعية الرياضيات التطبيقية والصناعية لعام 2004 حول استخراج البيانات، ليك بوينا فيستا، فلوريدا، الولايات المتحدة الأمريكية، 22-24 أبريل 2004! الصفحات 452-456.
80. سوجي، س.؛ شيشيبوري، م.؛ كوروبوا، س.؛ كيتا، ك. تقليل الأبعاد باستخدام تحليل المصروفات غير السالبة لاسترجاع المعلومات. في وقائع المؤتمر الدولي لعام 2001 التابع لمعهد مهندسي الكهرباء والإلكترونيات حول الأنظمة والإنسان وعلم التحكم الآلي، توسمون، أريزونا، الولايات المتحدة الأمريكية، 10-12 أكتوبر 2001! المجلد 2، الصفحات 960-965.
81. كوليak، س.؛ ليبلر، ر.أ. حول المعلومات والكافية. حوليات الإحصاء الرياضي [CrossRef] 1901, 12, 79-86.
82. جونسون، د.؛ سينانيفينش، س. تماثل مسافة كوليak-ليبلر. معاملات IEEE على النظرية المعلومات. 2001. متاح عبر الإنترنت: <https://scholarship.rice.edu/bitstream/handle/1911/19969/joh2001Mar1Symmetrizi.pdf?sequence=1> (تم الاطلاع عليه في 23 أبريل 2019).
83. بینغهام، ای؛ مانیلا، اتش. الإسقاط العشوائي في تقليل الأبعاد: تطبيقات على بيانات الصور والنصوص . في وقائع المؤتمر الدولي السابع لجمعية ACM SIGKDD حول اكتشاف المعرفة واستخراج البيانات، سان فرانسيسكو، كاليفورنيا، الولايات المتحدة الأمريكية، 29-26 أغسطس 2001! الصفحات 245-250.
84. نشاكرابارتي، إس؛ رو، إس؛ ساوندالجيكار، إم، في. تصنیف سریع ودقیق للنصوص عبر خطوط متعددة الإسقاطات التميیزیة. مجلة VLDB. 2003, 12, 170-185. [CrossRef]
85. رحیمی، ا.؛ ریخت، ب. الماجیع المرجحة لأحواض المطبخ العشوائی: استبدال التقلیل بالعشوائیة. في التعلم. Adv. Neural Inf. Process. Syst. 2009, 21, 1313-1320.
86. موروکوف، دبلیو. جیه؛ کافلیش، آر.ای. التکامل شبه مونت کارلو. مجلة الفیزیاء الحاسوبیة [CrossRef] 1995, 122, 218-230.
87. جونسون، دبلیو بی؛ لیندنشتراوس، جی؛ شیختمان، جی. امتدادات خرائط لیپشیتز إلى فضاءات باناخ. المجلة الإسرائلية. الرياضيات. 1986, 54, 129-138. [CrossRef]
88. داسغوبتا، س.؛ غوبتا، أ. برہان أولی لنظرية جونسون ولیندنشتراوس. الهیاکل العشوائیة. ACM SIGKDD، 2003، 22، 60-65. [CrossRef]
89. فیمبالا، إس طریقة الإسقاط العشوائی؛ الجمیعیة الیاضیة الامیرکیة: بروفیدنس، رود آیلاند، الولايات المتحدة الامیرکیة، 2005.

90. ماو، إكس؛ يوان، سي. المعادلات التفاضلية العشوائية مع التبديل الماركوفي؛ وورلد ساينتيفيك: سنغافورة. 2016.
91. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. Deep Learning; MIT Press: Cambridge, MA, USA, 2016;
- المجلد 1.
92. وانغ، و.؛ هوانغ، ي.؛ وانغ، ي.؛ وانغ، ل. المشفر التقائي المعمم: إطار عمل للشبكة العصبية لتقليل الأبعاد. في وقائع مؤتمر IEEE حول ورش عمل رؤية الحاسوب والتعرف على الأنماط، كولومبوس، أوهايو، الولايات المتحدة الأمريكية، 28-23 يونيو 2014: الصفحات 490-497.
93. روميلهارت، دي إي؛ هينتون، جي إي؛ ويليامز، آر جي. تعلم التمثيلات الداخلية عن طريق انتشار الخطأ؛ تبني تقرير، جامعة كاليفورنيا سان دييغو، معهد العلوم المعرفية: لا جولا، كاليفورنيا، الولايات المتحدة الأمريكية. 1985.
94. ليانغ، ه.؛ صن، واي.؛ غاو، واي. استخلاص ميزات النص باستخدام التعلم العميق: مراجعة. مجلة EURASIP Wirel. Commun. Netw. 2017, 2017, 211. [CrossRef]
95. بالدي، ب. المشفرات التقائية، والتعلم غير الخاضع للإشراف، والبني العميق. في وقائع ورشة عمل ICML حول التعلم غير الخاضع للإشراف والتعلم الانتقالي، بيلفيو، واشنطن، الولايات المتحدة الأمريكية، 20 يونيو 2011: ص 49-37.
96. AP, SC; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V.C; Saha, A. نهج التشفير التقائي لتعلم تمثيلات الكلمات ثنائية اللغة. 27, 1853-1861. Adv. Neural Inf. Process. Syst. 2014.
97. ماسكي، ج.؛ مابر، يو.؛ سيريسان، د.؛ شميدهور، ج. مشفرات تقائية تلقيحية مكذبة لاستخلاص الميزات الهرمية. في المؤتمر الدولي للشبكات العصبية الاصطناعية؛ سبرينغر: برلين/هابيلرغ، ألمانيا، 2011: ص 59-52.
98. تشين، ك.؛ سوريا، م.؛ ليوكي، م.؛ هينبيرت، ج.؛ إنجلولد، ر. تجزئة صفحات صور الوثائق التاريخية باستخدام المشفرات التقائية الالتفافية. في وقائع المؤتمر الدولي الثالث عشر لتحليل الوثائق والتعرف عليها (ICDAR)، تونس، 23-26 أغسطس 2015: الصفحات 1011-1015.
99. جينغ، ج.؛ فان، ج.؛ وانغ، ه.؛ ما، ش.؛ لي، ب.؛ تشين، ف. تصنیف صور الرادار ذي الفتحة التركيبية عالية الدقة باستخدام مشفرات تقائية تلقيحية عميق. رسائل IEEE علوم الأرض والاستشعار عن بعد. 2010, 12, 2351-2350. [CrossRef]
100. سوتسيكفي، آي.؛ فينيالز، أو.؛ لي، كيو. في. تعلم التسلسل إلى التسلسل باستخدام الشبكات العصبية. التقدم في المعلومات العصبية. Process. Syst. 2014, 27, 3104-3112.
101. تشون، ك.؛ فان ميرينبور، ب.؛ جوليبيه، آ.؛ ليهداوا، د.؛ بوجارس، ف.؛ شوينك، H.؛ بنجيوي، Y. تعلم تمثيلات العبارات باستخدام نموذج التشفير-فك التشفير RNN للترجمة الآلية الإحصائية. arXiv 2014, arXiv:1406.1078.
102. هينتون، جي إي؛ رويس، إس تي. تضمين الجوار العشوائي. أنظمة معالجة المعلومات العصبية المتقدمة. 2002, 15, 857-864.
103. جويس، جيه إم، تباعد كولباك-لابير. في الموسوعة الدولية للعلوم الإحصائية؛ سبرينغر: برلين/هابيلرغ، ألمانيا، 2011: ص 720-722.
104. روكيو، جيه جيه. التغذية الراجعة المتعلقة بالملاعبة في استرجاع المعلومات. في نظام الاسترجاع الذكي: تجارب في المعالجة الآلية للوثائق؛ إنجلوود كليفز: برتبس هول، نوجرس، الولايات المتحدة الأمريكية، 1971: الصفحات 313-323.
105. بارتلس، آي.؛ كوزموبولوس، أ.؛ باسيوتيس، ن.؛ أربير، آ.؛ باليوراس، ج.؛ جاوسيبيه، إي.؛ أندروتسوبولوس، أنا؛ أميني، السيد. Galinari, P. LSHTC: معيار لتصنيف النصوص على نطاق واسع. أرشيف. 2015, 4, 18580.3051.
106. سويميا، ب.؛ سرينيفاسا، ك. تصنیف النصوص متعدد التصنيفات على نطاق واسع لمجموعة بيانات هرمية باستخدام خوارزمية روكيو. في وقائع المؤتمر الدولي لعام 2016 حول أنظمة الحوسبة وتقنيات المعلومات من أجل حلول مستدامة (CSITSS)، بانغالور، الهند، 8-6 أكتوبر 2016: الصفحات 291-296.
107. كوردي، ف.؛ ماهندر، سبي إن. تصنیف النصوص والمصنفات: دراسة استقصائية. المجلة الدولية لتطبيقات الذكاء الاصطناعي. 2012, 3, 85.
108. سيلفي، إس تي؛ كاريكيان، بي.؛ فينسنت، إيه؛ ديكا، جي.؛ تيراجا، جي.؛ ديبكا، آر. تصنیف النصوص باستخدام خوارزمية روكيو وخوارزمية الغابة العشوائية. في وقائع المؤتمر الدولي الثامن للحوسبة المتقدمة، تشيناي، الهند، 21-19 يناير 2017: الصفحات 7-12.
109. ألبكار، س.؛ إسبينس، ب.؛ فورنييه، س. نحو تصنیف دللي مُشرف قائم على خوارزمية روكيو للويب الصفحات. في وقائع مؤتمر KES، إسبانيا، 10-12 سبتمبر 2012: 460-469.
110. فرزى، ر.؛ بولندي، ف. تقدیر السخنات العضوية باستخدام أساليب التجميع مقارنةً بالأساليب الذكية التقليدية: دراسة حالة لحقن غاز جنوب فارس، الخليج العربي، إيران. نمذجة نظام الأرض. Environ. 2016, 2, 105. [CrossRef]
111. باور، إى.؛ كوهفي، ر. مقارنة تجريبية لخوارزميات تصنیف التصویت: التجمیع، والتعزیز، والمتغيرات. تعلم الآلة. 1999, 36, 105-139. [CrossRef]
112. شابير، ر. إى. قوة خصف قابلية التعلم. تعلم الآلة. 1990, 5, 197-227. [CrossRef]
113. فرونوند، واي. خوارزمية تعزیز محشنة وآثارها على تعقید التعلم. في وقائع ورشة العمل السنوية الخامسة حول نظرية التعلم الحسابي، بیتسبرغ، بنسلفانيا، الولايات المتحدة الأمريكية، 29-27 يولیو 1992: الصفحات 391-398.

114. بلوهورن، س.؛ هونو، أ. تعزيز تصنيف النصوص باستخدام السمات الدلالية. في ورشة العمل الدولية حول اكتشاف المعرفة على الويب؛ سيرينغر؛ برلين/هایدلبرگ، ألمانيا، 2004 ص 149-166.
115. فوند، ي.؛ كيرنز، م.؛ منصور، ي.؛ رون، د.؛ روينفيلد، ر.؛ شابر، ر. إ. خوارزميات فعالة لتعلم لعب ألعاب متكررة ضد خصوم محدودين حسابياً. في وقائع الندوة السنوية السادسة والثلاثين حول أساس علوم الحاسوب، ميلووكى، ويسكونسن، الولايات المتحدة الأمريكية، 25-23 أكتوبر 1995: 341-332.
116. بريمان، ل. تجميع المتنبئات. تعلم الآلة. 1996, 24, 123-140. [CrossRef]
117. جورتس، ب. بعض التحسينات على تجميع شجرة القراء. في المؤتمر الأوروبي حول مبادئ استخراج البيانات واكتشاف المعرفة؛ سيرينغر؛ برلين/هایدلبرگ، ألمانيا، 2000 ص 136-147.
118. كوكس، تحليل البيانات الثنائية؛ روتليج؛ لندن، المملكة المتحدة، 2018.
119. قان، ر.؛ تشاغن، ل.؛ هسيه، س.؛ وانغ، إ.؛ إكس-آر، لين، س.؛ ج. TBLINEAR: مكتبة للتصنيف الخطى الكبير. *J. Mach. Learn. Res.* 2008, 9, 1871-1874.
120. جينكين، أ.؛ لويس، د.د.؛ ماديغان، د. الانحدار اللوجستي البايزى واسع النطاق لتصنيف النصوص. *Technometrics* 2007, 49, 291-304. [CrossRef]
121. خوان، أ.؛ فيداي، إ. حول استخدام نماذج خليط برنولي لتصنيف النصوص. التعرف على الأنماط 2002, 35, 2705-2710. [CrossRef]
122. تشنج، و.؛ هولماربر، إ. دمج التعلم القائم على الحالات والانحدار اللوجستي للتصنفيات المتعددة. *Tعلم الآلة*. 2009, 76, 211-225. [CrossRef]
123. كريشناورام، ب.؛ كارين، ل.؛ فيغيريدو، م.؛ هارتمينك، أ.ج. الانحدار اللوجستي متعدد الحدود المتاثر: خوارزميات سريعة وحدود التعميم. معاملات IEEE للتحليل الأنماط والذكاء الاصطناعي، 200. ٢٠٥، ٢٧، ٩٥٧-٩٦٨. [CrossRef]
124. هوانغ، ك. الاستشعار غير المقيد للهواتف الذكية دراسة تجريبية لمراقبة اليوم والإدارة الذاتية. أطروحة دكتوراه، جامعة ماساتشوستس لويل، لويل، ماساتشوستس، الولايات المتحدة الأمريكية، 2015.
125. غورين، أ. استخدام المتغيرات الديموغرافية وخصائص الطالب داخل الكلية للتبؤ بمعدل استبقاء الطالب على مستوى المقرر الدراسي في المجتمع طلاب اللغة الإسانية في الجامعة؛ جامعة نورث سترال: سكوتسليل، أريزونا، الولايات المتحدة الأمريكية، 2016.
126. كوفمان، إس. كوفمان، إس. التفاعل الاجتماعي الاصطناعي وتحليل سلوك المستخدم في خلاصات الويب. أطروحة دكتوراه، جامعة هامبورغ، هامبورغ، ألمانيا، 1969.
127. بورت، إم. إف. خوارزمية لإزالة اللواحق. برنامج 1980, 14, 130-137. [CrossRef]
128. بيرسون، إي. إس. نظرية بايز، تم فحصها في ضوءأخذ العينات التجريبية. بومتيكا 1925, 17, 388-442. [CrossRef]
129. هيل، بي. إم. التوزيع الخلفي للنسبة المئوية: نظرية بايز لأخذ العينات من مجتمع إحصائي. مجلة الجمعية الأمريكية للإحصاء. *Stat. Assoc.* 1968, 63, 677-691.
130. أاكو، ز.؛ سونغ، ش.؛ تشنج، س.؛ وانغ، ش.؛ سونغ، ش.؛ لي، ز. طريقة بايز المحسنة القائمة على ميزة TF-IDF وميزة عامل الدرجة لتصنيف المعلومات الصбинية. في وقائع المؤتمر الدولي لعام ٢٠١٨ حول البيانات الضخمة والحوسبة الذكية (BigComp) شنغهاي، الصين، ١٧-١٨يناير ٢٠١٨: ٦٧٧-٦٨٠. [CrossRef]
131. كيم، إس. بي؛ هان، كيه إس؛ ريم، إتش سي؛ ميانغ، إس. إتش بعض التقنيات الفعالة لتصنيف النصوص باستخدام بايز الساذج. *IEEE Trans. Knowl. Data Eng.* 2006, 18, 1457-1466.
132. فرانك، إي.؛ بوكرت، آر. آر. خوارزمية بايز البسيطة لتصنيف النصوص ذات الفئات غير المتوازنة. في المؤتمر الأوروبي حول مبادئ استخراج البيانات واكتشاف المعرفة؛ سيرينغر؛ برلين/هایدلبرگ، ألمانيا، 2006 ص 503-510.
133. ليو، واي.؛ لوه، إتش تي؛ صن، إيه. تصنیف النصوص غير المتوازنة: نهج ترجيح المصطلحات. أنظمة الخبراء وتطبيقاتها. 2009, 36, 690-701. [CrossRef]
134. سمهيل-خا، س.؛ مارتون، ب.ف.؛ بيشيه، ن. كشف الاختراقات في أنظمة التيقن الهجينية الخاصة للإشراف وغير الخاضعة للإشراف - دراسة حالة مفصلة على مجموعة بيانات ISCX للمعايير. HAL 2017. [CrossRef]
135. وانغ، واي.؛ خاردون، آر.؛ بروتوباباس، بي. التقدير البايزى غير البارامترى لمنحنيات الضوء الدورية. *Astrophys. J.* 2012, 756, 67. [CrossRef]
136. راجان، إم إن؛ غورياد، واي آر؛ كانثال، جي آر؛ غورياد، إيه آر؛ دوبي، إيه إس. تصنیف المستندات باستخدام الشبكة العصبية LSTM. مجلة إدارة البيانات والتقييم. 2017, 2، 2017 متاح على الإنترت: <http://matjournals.in/index.php/joDM&M/article/view/1534> (تم الاطلاع عليه في 20أبريل 2019).
137. جيانغ، إس؛ بانج، ج.؛ وو، م. K-nearest-neighbor خوارزمية Kuang، L.؛ المحسنة لتصنيف النص. [CrossRef] [PubMed]
138. Expert Syst. Appl. 2012, 39, 1503-1509. [CrossRef]
139. هان، إي إتش إس؛ كاريسب، جي؛ كومار، في. تصنیف النصوص باستخدام تصنیف أقرب جار  $k$  المعبد بالوزن. في مؤتمر آسيا والمحيط الهادئ لاكتشاف المعرفة واستخراج البيانات؛ سيرينغر؛ برلين/هایدلبرگ، ألمانيا، 2001 ص 53-65.

139. سالتون، ج. المعالجة الآلية للنصوص: تحويل وتحليل واسترجاع النصوص؛ أديسون-ويسلي: ريدينغ، المملكة المتحدة، 1989.
140. ساهفال، د؛ راميش، أ. حول كشف المركبات على الطرق باستخدام ميزات موجات غابور مع تقنيات تصنيف متعددة. في وقائع المؤتمر الدولي الرابع عشر لمعالجة الإشارات الرقمية، 2002 (رقم الكatalog 02TH8628)، سانتوريوني، اليونان، 3-1 يوليو 2002. doi:10.1109/ICDSP.2002.1028263.
141. بايتل، د؛ سريفاستافا، ت. نموذج تحسين مستعمرة النمل لمشاكل التصوير المقطعي المنفصل. في وقائع المؤتمر الدولي الثالث حول الحوسبة المزنة حل المشكلات، سبرينغر؛ برلين/هادلبرغ، ألمانيا، 2014 ص. 785-792.
142. ساهفال، د؛ باريدا، م. التعرف على الأشياء باستخدام ميزات موجات غابور مع تقنيات تصنيف متعددة. في وقائع المؤتمر الدولي الثالث حول الحوسبة المزنة حل المشكلات، سبرينغر؛ برلين/هادلبرغ، ألمانيا، 2014 ص. 793-804.
143. سانجاي، جي بي؛ ناجوري، في؛ سانجاي، جي بي؛ ناجوري، في. مقارنة الطرق الحالية للتنبؤ باحتمالات الكشف عن سرطان الدم من خلال تحليل البيانات الصحية. المجلة الدولية للبحوث المبنية في العلوم والتكنولوجيا، 2018، 4، 10-14.
144. فابنيك، ف؛ Chervonenkis، A. Y. فئة من الخوارزميات لتعلم التعرف على الأنماط. تلقائي. تيليميك 1964، 25، 937-945.
145. بوسن، بي؛ غابون، آي إم؛ فابنيك، في إن. خوارزمية تدريب لمصنفات الهاشم الأمثل. في وقائع ورشة العمل السنوية الخامسة حول نظرية التعلم الحاسبي، بيتسبيرغ، بنسلفانيا، الولايات المتحدة الأمريكية، 27-29 يوليو 1992؛ الصفحات 144-152.
146. بوسن، جي؛ شيان وو، إتش. تصنيف SVM متعدد الفئات. مجلة معالجة واكتساب البيانات، 2006، 3، 017.
147. مهري، م؛ رستمي زاده، أ؛ تالواكار، أ. أسس التعليم الآلي؛ مطبعة معهد ماساتشوستس للتكنولوجيا: كامبريدج، ماساتشوستس، الولايات المتحدة، 2012.
148. تشين، ك؛ تشناغ، ز؛ لونغ، ج؛ تشناغ، ه. التحول من TF-IDF إلى TF-IGM لترجيح المصطلحات في النص التصنيف. أنظمة الخبراء وتطبيقاتها [CrossRef] 2016، 66، 245-260.
149. ويستون، ج؛ واتكينز، سي. آلات المتجهات الداعمة متعددة الفئات، التقرير الفني 04-98-TR-CSD، قسم علوم الحاسوب، رووال هولوي، جامعة لندن: إيقام، المملكة المتحدة، 1998.
150. Zhang، W.; Yoshida، T.; Tang، X. أنظمة قائمة على المعرفة [CrossRef] 2008، 21، 879-886.
151. لودهي، ه؛ سوندرز، س؛ شاو-تايلور، ج؛ كريستيانيني، ن؛ واتكينز، س. تصنيف النصوص باستخدام نوى السلاسل. J. Mach. Learn. Res. 2002، 2، 419-444.
152. ليزلي، سي إس؛ إسكيين، إي؛ نوبل، ديليو إس نواة الطيف: نواة سلسلة لتصنيف بروتين SVM. الحوسبة الجوية، 2002، 7، 566-575.
153. Eskin، E.; Weston، J.; Noble، WS; Leslie، CS SVM نوى سلسلة عدم التطابق لتصنيف بروتين. Adv. Neural Inf. Process. Syst. 2002، 15، 1417-1424.
154. كوساري، ك؛ لانشانين، ج؛ وانغ، ب؛ تشي، ي. خوارزمية سريعة وقابلة للتتوسيع لحساب نواة سلسلة k-mers. bioRxiv 2017. [CrossRef]
155. صن، أ؛ لم، إي بي. التصنيف والتقييم الهرمي للنصوص. في وقائع المؤتمر الدولي لهندسة البيانات، (ICDM 2001) (سان خوسيه، كاليفورنيا، الولايات المتحدة الأمريكية، 29 نوفمبر 2001)؛ الصفحات 521-528.
156. سيباستيانى، ف. التعلم الآلي في تصنيف النصوص الآلي. مجلة ACM للحوسبة، (CSUR) 2002، 34، 1-47.
157. مارون، أو؛ لوزانو-بيريز، ت. إطار عمل للتعلم متعدد الحالات. مجلة أنظمة معالجة المعلومات العصبية المتقدمة، 1998، 10، 570-576.
158. أندرزون، إس؛ تسوتشاريديس، آي؛ هوفرمان، تي. آلات المتجهات الداعمة للتعلم متعدد الحالات. Adv. Neural Inf. Process. Syst. 2002، 15، 577-584.
159. كاراميزاده، س؛ عبد الله، س. م؛ حليمي، م؛ شابان، ج؛ جواد رجبى، م؛ مزايا وعيوب وظائف آلة المتجهات الداعمة. في وقائع المؤتمر الدولي لعام 2014 حول تكنولوجيا الحاسوب والاتصالات والتحكم (I4CT)، (لانكاوى، ماليزيا، 4-6 سبتمبر 2014)؛ الصفحات 63-65.
160. غو، جي. القياسات الحيوية الناعمة من صور الوجه باستخدام آلات المتجهات الداعمة. في آلات المتجهات الداعمة التطبيقات؛ سبرينغر؛ برلين/هادلبرغ، ألمانيا، 2014؛ الصفحات 269-302.
161. مورغان، جي إن؛ سونكوبست، جي إيه. مشاكل في تطبيق بيانات المسح، واقتراح. مجلة الجمعية الإحصائية الأمريكية، 1963، 58، 415-434. [CrossRef]
162. سافافيان، إس آر؛ لاندغريب، دي. دراسة استقصائية لمنهجية مصنف شجرة القرارات، معاملات IEEE في أنظمة الإنسان والسيبرانية، 1991، 21، 660-674. [CrossRef]

163. ماجرمان، د.م. نماذج شجرة القرار الإحصائية للتحليل النحوي. في وقائع الاجتماع السنوي الثالث والثلاثين لجمعية اللغويات الحاسوبية، كامبريدج، ماساتشوستس، الولايات المتحدة الأمريكية، 30-26 يونيو 1995؛ جمعية اللغويات الحاسوبية: ستورسيبرغ، بنسلفانيا، الولايات المتحدة الأمريكية، 276-283. الصفحات.
164. كوبيلان، جيه آر. استقراء أشجار القرار، تعلم الآلة. [CrossRef] 1986, 1, 81-106.
165. دن مانتراس، آر إل. مقياس اختبار السمات قائم على المسافة لاستقراء شجرة القراء، تعلم الآلة. 1991, 6, 81-92. [CrossRef]
166. جيفانيلي، س.؛ ليو، إكس؛ سيرلا، إس.؛ فياتكين، في؛ إيتسيزي، آر. نحو مجتمع يستغل البيانات الضخمة للمزايدة على سوق احتياطي احتواء التردد. في وقائع المؤتمر السنوي الثالث والأربعين لجمعية الإلكترونيات الصناعية التابعة لمعهد مهندسي الكهرباء والإلكترونيات (IECON 2017) بكين، الصين، 29 أكتوبر - 1 نوفمبر 2017؛ الصفحات 7514-7519.
167. كوبيلان، جيه آر. تبسيط أشجار القرار، المجلة الدولية لدراسات الإنسان والآلة. 1987, 27, 221-234. [CrossRef]
168. جاسم، main-steps-for-doing-data-mining-project-using-weka.pdf. متاح على الإنترنت: [https://www.researchgate.net/profile/Dalia\\_Jasim/publication/293464737\\_main\\_steps\\_for\\_doing\\_data\\_mining\\_project\\_using\\_weka/links/56b8782008ae44bb330d2583/](https://www.researchgate.net/profile/Dalia_Jasim/publication/293464737_main_steps_for_doing_data_mining_project_using_weka/links/56b8782008ae44bb330d2583/) (تم الاطلاع عليه في 23 أبريل 2019). (<https://www.researchgate.net/>)
169. هو، تي كيه. غابات القرار العشوائية. في وقائع المؤتمر الدولي الثالث لتحليل المستندات والتعرف عليها، مونتريال، كيبك، كندا، 16-14 أغسطس 1995؛ المجلد 1، الصفحات 278-282. [CrossRef]
170. بريمان، ل. الغابات العشوائية؛ يو سى بيركلي: جامعة كاليفورنيا: بيركلي، كاليفورنيا، الولايات المتحدة الأمريكية، 1999.
171. وو، تي إف؛ لين، سى جيه؛ وينغ، آر سى. تقديرات الاحتمالية للتصنification متعدد الفئات عن طريق الاقتران الروجي. J. Mach. Learn. Res. 2004, 5, 975-1005.
172. بانسال، ه.؛ شريفاستافا، ج؛ فهو، ن.؛ ستانسيو، ل. تحليلات الشبكات الاجتماعية لمنظمات الأعمال المعاصرة؛ آي جي آي جلوبال: هيرشي، بنسلفانيا، الولايات المتحدة الأمريكية، 2018.
173. ساتون، سى.؛ ماكلوم، أ. مقدمة في الحقول العشوائية الشرطية. أنسس. تريندز® للتعلم الآلي، 2012.
174. فايل، دي إل؛ فيلوسو، إم؛ لافيرتي، جي دي. الحقول العشوائية الشرطية للتعرف على الأنشطة. في وقائع المؤتمر الدولي المشترك السادس حول الوكاء المستقلين وأنظمة الوكاء المتعددين، هونولولو، هاواي، الولايات المتحدة الأمريكية، 18-14 مايو 2007؛ من 235.
175. تشين، ت.؛ شو، ر.؛ هي، ي.؛ وانغ، إكس. تحسين تحليل المشاعر من خلال تصنification أنواع الجمل باستخدام CNN. Expert Syst. Appl. 2017, 72, 221-230. [CrossRef]
176. ساتون، سى.؛ ماكلوم، أ. مقدمة في الحقول العشوائية الشرطية للتعلم العائقي. في المقدمة إلى التعلم الإحصائي العائقي؛ مطبعة معهد ماساتشوستس للتكنولوجيا: كامبريدج، ماساتشوستس، الولايات المتحدة الأمريكية، 2006؛ المجلد 2.
177. تسينغ، ه.؛ تسانغ، ب.؛ أندرو، ج.؛ جروفاسكي، د.؛ مانينغ، س. مجزى كلمات حقل عشوائي شرطي لمسابقة سيفان بيك أوف 2005 في وقائع ورشة عمل سيفان الرابعة حول معالجة اللغة الصينية، جزيرة جيجو، كوريا، 15-14 أكتوبر 2005.
178. ناير، ف.؛ هينتون، ج. إي. وحدات خطية معدلة تحسن آلات بولتزمان المقيدة. في وقائع المؤتمر الدولي السابع والعشرين للتعلم الآلي (ICML-10)، حيفا، إسرائيل، 24-21 يونيو 2010؛ الصفحات 807-814.
179. سوتسيفير، آي. مارتنز، ج.؛ هينتون، جي إي. توليد النصوص باستخدام الشبكات العصبية المترکزة. في وقائع المؤتمر الدولي الثامن والعشرين للتعلم الآلي (ICML-11)، بيلفيو، واشنطن، الولايات المتحدة الأمريكية، 28 يونيو 2011؛ بيلفيو 10-10124. [CrossRef]
180. مانديتش، د.ب.؛ تشارميرز، ج.أ. الشبكات العصبية المترکزة للتنبؤ: خوارزميات التعلم، والهايكل، والاستقرار؛ مكتبة وايلي الإلكترونية: هوبيكين، نيوجيرسي، الولايات المتحدة الأمريكية، 2001.
181. بینجیو، واي؛ سیمارد، بی؛ فراسکونی، بی. تعلم التبعيات طويلة المدى باستخدام انحدار التدرج أمر صعب. IEEE Trans. Neural Netw. 1994, 5, 157-166. [CrossRef]
182. هوكریت، س.؛ شمیدهور، ج. الذكرة طويلة المدى قصيرة المدى. الحوسنة العصبية. 1997, 9, 1735-1780. [CrossRef]
183. غريفز، أ.؛ شمیدهور، ج. تصنification الصوتيات على مستوى الإطار باستخدام LSTM الثنائي الاتجاه وشبكات عصبية أخرى بنى الشبكات، الشبكات العصبية. 2005, 18, 602-610. [CrossRef] [PubMed]
184. باسكانو، ر.؛ میکولوف، ت.؛ بینجیو، بی. حول صعوبة تدريب الشبكات العصبية المترکزة. المؤتمر الدولي للتعلم الآلي، 2013، 28، 1310-1318.
185. تشونغ، ج.؛ عولجيهري، س.؛ تشو، ل.؛ بینجیو، بی. التقييم التجاري للشبكات العصبية المترکزة ذات البوابات على نمذجة التسلسل. arXiv 2014, arXiv:1412.3555.
186. جادربیرغ، م.؛ سیمونیان، ل.؛ زیدالدی، أ.؛ زیرمان، أ. قراءة النصوص في بيانات واقعية باستخدام الشبكات العصبية الالتفافية للشبكات، المجلة الدولية لرؤية الحاسوب 2016, 116, 1-20. [CrossRef]

187. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. التعلم القائم على التدرج المطبق على التعرف على المستندات. *1998*, 86, 2278-2324. [CrossRef]
- ICANN 2010، شيرر، د؛ مولر، أ؛ بينكه، س. تقييم عمليات التجميع في البن الافتافية للتعرف على الأشياء. في وقائع مؤتمر الشبكات العصبية الاصطناعية، 188. ثيسالونيكي، اليونان، 18-15 سبتمبر 2010 الصفحات 92-101.
189. جونسون، ر؛ تشانغ، ت. الاستخدام الفعال لترتيب الكلمات لتصنيف النصوص باستخدام الشبكات العصبية الافتافية. *arXiv* 2014, arXiv:1412.1058.
190. هينتون، جي اي. تدريب منتجات الخبراء عن طريق تقليل التباعد التباعي. *الحوسبة العصبية*. 2002, 14, 1771-1800. [CrossRef]
191. هينتون، جي اي؛ أوسيندريو، إس؛ تيه، واي ديليو. خوارزمية تعلم سريعة لشبكات الاعتقاد العميق. *الحوسبة العصبية*. 2006, 18, 1527-1554. [CrossRef]
192. محمد، أ. ر؛ دال، ج. إ؛ هينتون، ج. النمذجة الصوتية باستخدام شبكات الاعتقاد العميق. معاملات IEEE في معالجة الصوت والكلام واللغة. *HLT-NAACL* سان دييغو، كاليفورنيا، الولايات المتحدة الأمريكية، 17-12 يونيو 2016 الصفحات 1480-1489.
193. يانغ، ز؛ داير، س؛ هي، إكس؛ سمول، أ. ج؛ هو، اي. ه. شبكات الانتباه الهرمية لتصنيف المستندات . في وقائع مؤتمر COMPSTAT'2010، ل. التعلم الآلي واسع النطاق باستخدام التدرج العشوائي. في وقائع مؤتمر 2016، سبيو، بي إتش؛ لين، زد؛ كوهين، إس؛ شين، إكس؛ هان، بي. شبكات الانتباه الهرمية. *arXiv* 2016, arXiv:1606.02393.
194. دوتشي، ج؛ هازان، إ؛ سينجر، ي. طرق التدرج الفرعى التكيفى للتعلم عبر الإنترنٌت والتحسين العشوائى. *J. Mach. Learn. Res.* 2011, 12, 2121-2159.
195. زيلر، دكتور في الطب، آدم: طريقةADADELTA معدل التعلم التكيفي. *arXiv* 2012, arXiv:1212.5701.
196. غاري، أ. ماكدونيل، س. بدائل لنماذج الانحدار لتقدير مشاريع البرمجيات. متاح على الإنترنت: [2747623\\_Alternatives\\_to\\_Regression\\_Models\\_for\\_Technical\\_Programs](https://www.researchgate.net/publication/2747623_Alternatives_to_Regression_Models_for_Technical_Programs) (تم الاطلاع عليه في 23 أبريل 2019).
197. شوارتز-زيف، ر؛ تيشين، ن. فتح الصندوق الأسود للشبكات العصبية العميقه عبر المعلومات. *arXiv* 2017, arXiv:1703.00810.
198. عربى، أ؛ كوندومار، أ؛ جرينسايد، ب؛ كونداجي، أ. تعلم السمات المهمة من خلال نشر التشبيث الأخلاقات. *arXiv* 2017, arXiv:1704.02685.
199. غارى، أ. ماكدونيل، س. بدائل لنماذج الانحدار لتقدير مشاريع البرمجيات. متاح على الإنترنت: [2747623\\_Alternatives\\_to\\_Regression\\_Models\\_for\\_Technical\\_Programs](https://www.researchgate.net/publication/2747623_Alternatives_to_Regression_Models_for_Technical_Programs) (تم الاطلاع عليه في 23 أبريل 2019).
200. شريكومار، أ؛ جرينسايد، ب؛ كونداجي، أ. تعلم السمات المهمة من خلال نشر التشبيث الأخلاقات. *arXiv* 2017, arXiv:1704.02685.
201. تشن، سي؛ صن، سي؛ ليو، زد؛ لاو، إف. شبكة عصبية LSTM-C لتصنيف النصوص. *arXiv* 2015, arXiv:1511.08630.
202. شوارتز-زيف، ر؛ تيشين، ن. فتح الصندوق الأسود للشبكات العصبية العميقه عبر المعلومات. *arXiv* 2017, arXiv:1703.00810.
203. لامبيين، أ.ك؛ ماكيلاند، ج.ل. تعلم تضمينات الكلمات باستخدام لقطة واحدة وعدد قليل من اللقطات. *arXiv* 2017, arXiv:1710.10280.
204. Severyn, A.; Moschitti, A. التعلم لترتيب أزواج النصوص القصيرة باستخدام الشبكات العصبية الافتافية. *ACM SIGIR* 2013, 56, 13-15. [CrossRef]
205. شريكومار، أ؛ جرينسايد، ب؛ كونداجي، أ. تعلم السمات المهمة من خلال نشر التشبيث الأخلاقات. *arXiv* 2017, arXiv:1704.02685.
206. غودا، إتش إس؛ سهيل، إم؛ غورو، دي؛ راجو، إل إن. تصنیف النصوص شبه الخاضع للإشراف باستخدام التجميع المتكرر لخوارزمية K-means. في المؤتمر الدولي لـ SIGIR حول البحث والتطوير في استرجاع المعلومات، سانتياغو، تشيلي، 13-19 أغسطس 2015، الصفحات 373-382.
207. غودا، إتش إس؛ سهيل، إم؛ غورو، دي؛ راجو، إل إن. تصنیف النصوص شبه الخاضع للإشراف باستخدام التجميع المتكرر لخوارزمية K-means. في المؤتمر الدولي ACM SIGIR حول البحث والتطوير في استرجاع المعلومات، سانتياغو، تشيلي، 13-19 أغسطس 2015، الصفحات 217-227.
208. كوصاري، ك. دراسة البحث الضبابي باستخدام تحويلات كود غولي. أطروحة دكتوراه، قسم علوم الحاسوب، جامعة جورج واشنطن، واشنطن العاصمة، الولايات المتحدة الأمريكية، 2014.
209. كوصاري، ك؛ يماحي، م؛ باري، ن؛ فيشر، ر؛ السابي، ف؛ بيركوفيتش، س.ي. بناء قاموس البحث الضبابي باستخدام تحويل ترميز غولي لتطبيقات البحث. *arXiv* 2015, arXiv:1503.06483.
210. شabil، أو؛ زين، أ. التصنیف شبه الخاضع للإشراف عن طريق فصل الكثافة المنخفضة. في وقائع المؤتمر AISTATS، فندق سافانا، باربادوس، 8-6يناير 2005، الصفحات 57-64.
211. شabil، أو؛ زين، أ. التصنیف شبه الخاضع للإشراف عن طريق فصل الكثافة المنخفضة. في وقائع المؤتمر AISTATS، فندق سافانا، باربادوس، 8-6يناير 2005، الصفحات 57-64.
212. شفقام، ك؛ ماكالوم، أ؛ ميتشل، ت. تصنیف النصوص شبه الخاضع للإشراف باستخدام خوارزمية EM. في: التصنیف شبه الخاضع للإشراف

231. شي، ل.; ميهالسا، ر.; تيان، م. تصنیف النصوص عبر اللغات باستخدام الترجمة النموذجية والتعلم شبه الموجه . في وقائع مؤتمر عام 2010 حول الأساليب التجريبية في معالجة اللغة الطبيعية، كامبريدج، ماساتشوستس، الولايات المتحدة الأمريكية، 11-19 أكتوبر؛ 2010؛ رابطة اللغويات الحاسوبية: ستورديبورغ، بنسلفانيا، الولايات المتحدة الأمريكية، 2010؛ الصفحات 1057-1067.

214. Zhou, S.; Chen, Q.; Wang, X. Fuzzy deep belief networks for semi supervised sentiment classification.

الحوسبة العصبية [CrossRef] 2014, 131, 312-322.

215. بانغ، واي. تقييم الأساليب الإحصائية لتصنيف النصوص. استرجاع المعلومات [CrossRef] 1999, 1, 69-90.

216. ليفر، ج.; كريزونسكي، م.; ألمان، ن. نقاط الأهمية: تقييم التصنيف. نات. ميثودز، 13, 2016, 1067-1074.

603-604. [CrossRef]

217. مانيغ، سي دي؛ راغافان، بي؛ شوتزه، إتش. تحليلات المصفوفة والفهرسة الدلالية الكامنة.

في مقدمة استرجاع المعلومات: مطبعة جامعة كامبريدج: كامبريدج، المملكة المتحدة، 2008؛ الصفحات 403-417.

218. توسماكاس، جي.; كاتاكيش، آي.; فلاهاواس، آي. استخراج البيانات متعددة التصنيفات. في كتاب "دليل استخراج البيانات واكتشاف المعرفة"; سيرينغر: برلين/هابيلرغ، ألمانيا، 2009؛ الصفحات 667-685.

219. بونيليانس، أ.ب.; باركس، س.م. خصائص تشغيل المستقبل (ROCs) في ذاكرة التعرّف: مراجعة.

مجلة علم النفس [CrossRef] 2007, 133, 800.

220. جاكوفيتش، ن.; ستيفن، س. مشكلة عدم توازن الفئات: دراسة منهجية. تحليل البيانات الذكية [CrossRef] 2002, 6, 429-449.

221. ببرادي، أ.ب. استخدام المساحة تحت منحنى ROC في تقييم خوارزميات التعلم الآلي.

التعرف على الأنماط [CrossRef] 1997, 30, 1145-1159.

222. هاند، دي جي؛ تيل، آر جي. تعميم بسيط للمساحة تحت منحنى ROC لتصنيف الفئات المتعددة

مشاكل. تعلم الآلة. [CrossRef] 2001, 45, 171-186.

223. وو، إتش س؛ لوك، آر دبليو بي؛ وونغ، كيه إف؛ كوك، كيه إل تفسير أوزان مصطلحات tf-idf على أنها اتخاذ قرارات تتعلق بالماءمة.

ACM Trans. Inf. Syst. (TOIS) 2008, 26, 13. [CrossRef]

224. رضائيينا، إس إم؛ غودسي، أ؛ رحماني، ر. تحسين دقة تضمينات الكلمات المدرية مسبقاً لتحليل المشاعر. arXiv 2017, arXiv:1711.08609.

225. شارما، أ.; باليوال، ل.ك. تحليل المكونات الرئيسية السريع باستخدام خوارزمية النقطة الثابتة. التعرّف على الأنماط.

Lett. 2007, 28, 1151-1155. [CrossRef]

226. Putthividhya, DP; Hu, J. التعرّف على البيانات المسممة باستخدام Bootstrapped لاستخراج سمات المنتج.

في وقائع المؤتمر حول الأساليب التجريبية في معالجة اللغة الطبيعية، إنديرا، المملكة المتحدة، 31-37 بوليو؛ 2011؛ رابطة اللغويات الحاسوبية:

ستورديبورغ، بنسلفانيا، الولايات المتحدة الأمريكية، 2011؛ الصفحات 1557-1567.

227. بانيرجي، م. إطار عمل واع بالفائدة يحافظ على الخصوصية لاستخراج البيانات الموزعة مع أسوأ حالة للخصوصية

الضمان؛ جامعة ماريلاند: مقاطعة بالتيمور، ماريلاند، الولايات المتحدة الأمريكية، 2011.

228. تشن، جي؛ يان، إس؛ وونغ، كيه س. الكشف عن العدوان اللغوي في تعليقات تويتر: شبكة عصبية تلقائية لتحليل المشاعر. تطبيقات الحوسبة العصبية

2018, 1-10. [CrossRef]

229. تشانغ، إكس؛ تشاو، جي. ليكان، واي. الشبكات الالتفافية على مستوى الأحرف لتصنيف النصوص. مجلة المعلومات العصبية المتقدمة.

Process. Syst. 2015, 28, 649-657.

230. شوتزه، ه؛ مانيغ، سي دي؛ راغافان، ب. مقدمة في استرجاع المعلومات: مطبعة جامعة كامبريدج:

كامبريدج، المملكة المتحدة، 2008؛ المجلد 39.

231. هوجيفين، د؛ وانغ، ل؛ بالدوين، KM. استرجاع منتدى الويب وتحليلات النص: دراسة استقصائية.

Found. Trends® Inf. Retr. 2018, 12, 1-163. [CrossRef]

232. دويفيدي، إس كيه؛ آريا، س. التصنيف التلقائي للنصوص في استرجاع المعلومات: دراسة استقصائية. في وقائع المؤتمر الدولي الثاني حول تكنولوجيا المعلومات والاتصالات

للأساليجيات التنافسية، أودايبور، الهند، 4-5 مارس؛ 2016؛ ص 131.

233. جونز، ك.س. التصنيف الآلي للكلمات المفتاحية لاسترجاع المعلومات. مجلة المكتبات الفصلية [CrossRef] 1971, 41, 338-340.

234. أوربوردان، س.؛ سورنسن، إتش. تصفية المعلومات واسترجاعها: نظرة عامة. في وقائع المؤتمر الدولي السنوي السادس عشر لمعهد مهندسي الكهرباء والإلكترونيات، أتلانتا، جورجيا، الولايات المتحدة الأمريكية، 31-32 أكتوبر؛ 1997؛ ص 42.

235. باكري، س. تنفيذ نظام استرجاع المعلومات الذكي؛ تقرير فني؛ جامعة كورنيل: إيشاكا، نيويورك، الولايات المتحدة الأمريكية، 1985.

236. بانغ، ب؛ لي، ل. استخراج الآراء وتحليل المشاعر. أنسس واتجاهات استرجاع المعلومات. [CrossRef] 2008, 2, 1-135.

237. ليو، ب؛ تشانغ، ل. دراسة استقصائية حول استخراج الآراء وتحليل المشاعر. في كتاب استخراج البيانات النصية؛ سيرينغر:

برلين/هابيلرغ، ألمانيا، 2012؛ الصفحات 415-463.

238. بانغ، ب.; لي، ل.; فاينانثان، س. هل هذا جيد؟؛ تصنيف المشاعر باستخدام تقنيات التعلم الآلي. في مؤتمر ACL-02 حول الأساليب التجريبية في معالجة اللغة الطبيعية؛ رابطة اللغويات الحاسوبية: ستوردسبرغ، بنسلفانيا، الولايات المتحدة الأمريكية، 2002؛ المجلد 10، الصفحات 79-86.
239. أغاروال، سي. سي. أنظمة التوصية القائمة على المحتوى. في أنظمة التوصية؛ سبرينغر: برلين/هایدلبرگ، ألمانيا، 2016؛ ص. 139-166.
240. بازانى، إم. جيه؛ بيلوس، د.ي. أنظمة التوصية القائمة على المحتوى. في الويب التكيفي؛ سبرينغر: برلين/هایدلبرگ، ألمانيا، 2007؛ ص. 325-341.
241. سومانى، ك.؛ تشيدامارام، م. التقييب في النصوص: المفاهيم والتطبيقات والأدوات والقضايا -نظرة عامة. المجلة الدولية. Comput. Appl. 2013, 80, 29-32.
242. حيدري صفا، م.; كونرى، ك.; بازنر، ل.؛ براون، د.؛ تحليل روايات حوادث السكك الحديدية باستخدام التعلم العميق. في وقائع المؤتمر الدولي السابع عشر لمعهد مهندسي الكهرباء والإلكترونيات حول التعلم الآلي والتطبيقات (ICMLA)؛ أورلاندو، فلوريدا، الولايات المتحدة الأمريكية، 20-22 ديسمبر 2018.
243. ماني، آي. التطورات في التلخيص الثنائي للنصوص؛ مطبعة معهد ماساتشوستس للتكنولوجيا: كامبريدج، ماساتشوستس، الولايات المتحدة الأمريكية، 1999.
244. Cao, Z.; Li, W.; Li, S.; Wei, F. تحسين تلخيص المستندات المتعددة من خلال تصنيف النصوص. وفي وقائع AAAI، سان فرانسيسكو، كاليفورنيا، الولايات المتحدة الأمريكية، 9-14 فبراير 2017؛ ص. 3053-3059. [CrossRef]
245. لوريا، إي. جيه؛ مارش، إيه. دى. دمج تصنيف النصوص الباياني والتقلص لأنتمة تميز الرعاية الصحية: تحليل جودة البيانات. مجلة جودة البيانات والمعلومات (JDIQ) 2011, 2, 13. [CrossRef]
246. Zhang, J.; Kowsari, K.; Harrison, JH; Lobo, JM; Barnes, LE. Patient2Vec: تمثيل عميق قابل للتفسير ومحضن للسجل الصحي الإلكتروني الطولي. IEEE Access 2018, 6, 65333-65346. [CrossRef]
247. ترينشيج، د.؛ بيزيك، ص. لي، د.؛ يونج، ف.؛ كريج، ديليو، D. MeSH Up: Reholz-Schuhmann. MeSH Up: MeSH Up: الفعال لتحسين استرجاع المستندات. المعلوماتية الحيوية. 2009, 25, 1412-1418. [PubMed] [CrossRef]
248. أوفوجي، ب.؛ فيرسبور، ك. تصنيف المشاعر النصية: دراسة قابلية التشغيل البيئي على مجموعات بيانات متعددة الأنواع. في المؤتمر الأسترالي الآسيوي المشترك حول الذكاء الاصطناعي؛ سبرينغر: برلين/هایدلبرگ، ألمانيا، 2017؛ ص. 262-273.
249. بينباقر، ج.؛ بوث، ر.؛ بويد، ر.؛ فرانسيس، م. البحث اللغوي وإحصاء الكلمات: LIWC2015 مجموعة بينباقر؛ أوستن، تكساس، الولايات المتحدة الأمريكية، 2015؛ متاح على الإنترنت: [www.LIWC.net](http://www.LIWC.net) (تم الاطلاع عليه في 10 يناير 2019).
250. بول، إم. جيه؛ دريدز، إم. الرصد الاجتماعي للصحة العامة. محاضرات تركيبية في مفاهيم المعلومات وخدمات الاسترجاع. 2017, 9, 1-183. doi:10.2200/S00791ED1V01Y201707ICR060. [CrossRef]
251. بول، كوك، ل. تصنيف رسائل التسويق التجاري على فيسبوك. في وقائع مجموعة الاهتمامات الخاصة باسترجاع المعلومات التابعة لجمعية آلات الحوسبة، بكين، الصين، 2011-24 يوليو.
252. كانغ، م.؛ آهن، ج.؛ لي، ك. استخراج الآراء باستخدام نماذج ماركوف المخفية للنصوص المجمعة لتصنيف النصوص. Expert Syst. Appl. 2018, 94, 218-227. [CrossRef]
253. تيرتل، ه. استرجاع النصوص في العالم القانوني. الذكاء الاصطناعي والقانون 1995, 3, 5-54. [CrossRef]
254. بيرغان، ب.؛ بيرمان، إس. جيه. تمثيل نفسك في المحكمة: كيفية التحضير لقضية رابحة ومحاكمتها؛ نولو: بيركل، كاليفورنيا، الولايات المتحدة، 2016.



© من تأليف المؤلفين. الناشر: MDPI، بازل، سويسرا. هذه المقالة متاحة للجميع بموجب شروط وأحكام رخصة المشاع الإبداعي.

(CC BY) (<http://creativecommons.org/licenses/by/4.0/>). رخصة.