استخدام خوارزميات تعلم الآلة لتصنيف همزتي الوصل والقطع

د. طلال أحمد شداد الثقفي
الأستاذ المساعد بقسم اللغة العربية،
الكلية الجامعية بتربة، جامعة الطائف

د. ياسر نصر الدين السيد علي
الأستاذ المساعد بقسم الرياضيات، برنامج علوم الحاسب بالكلية الجامعية بتربة، جامعة الطائف
البريد الجامعي/ ynali@tu.edu.sa

د. محمد فتحي عبد الفتاح عبدالمعطي
الأستاذ المشارك بقسم اللغة العربية،
الكلية الجامعية بتربة، جامعة الطائف
البريد الجامعي/ m.fathe@tu.edu.sa

د. طلال الطاهر قطبي بشير
الأستاذ المشارك بقسم اللغة العربية،
الكلية الجامعية بتربة، جامعة الطائف

تم تمويل هذا البحث برعاية ودعم عمادة البحث العلمي، بجامعة الطائف - المملكة العربية السعودية [مجموعة بحثية رقم 101-1441-1]

الملخص:

تعتبر اللغة العربية كائن حي ينمو ويتطور بالممارسة والتطبيق الصحيح لجميع آدابها وفروعها النحوية والصرفية والدلالية والمعجمية، وفي هذه الدراسة البحثية نستعرض مساهمة التقنية في تطور اللغة العربية لاسيما الرسم الصحيح لهمزتي الوصل والقطع.

يهدف هذا البحث بشكل خاص لبناء نموذج مصنف ذكي يعمل على تصنيف الكلمات العربية المبتدئة بحرف الألف، وتصنيف همزتها إلى همزة وصل أو همزة قطع من خلال استخدام تقنيات الذكاء الاصطناعي بصورة عامة، وخوارزميات تعلم الآلة بصورة خاصة من أجل تأسيس معايير دقيقة وصحيحة في رسم همزتي الوصل والقطع بالصورة الصحيحة، وبذلك نكون قد طوعنا التقنية لتسهم في خدمة اللغة العربية.

اعتمد هذا البحث على تجميع الكلمات العربية المبتدئة بالهمزة وذلك عن طريق تصميم استبانة رقمية مهمتها تجميع أكبر قدر من الكلمات المبتدئة بالهمزة وتصنيفها لهمزتي وصل أو قطع وفقاً للقواعد النحوية المتبعة في ذلك، تم نشر الاستبانة على الويب وتمت تعبئتها بواسطة 50 متخصص في النحو بدرجات علمية متفاوتة، حيث بلغ عدد الكلمات الكلية المصنفة 400 كلمة،

وبعد معالجة واستبعاد الكلمات المكررة وعددها 101 كلمة، حصلنا على عدد 299 كلمة صالحة لتطبيقها على نموذج المصنف، وبناءً على حجم ونوع البيانات المجمعة وآلية التصنيف المتبعة تم تطبيق خوارزميات تصنيف تتناسب مع العينة المجمعة مثل خوارزمية آلة دعم المتجه (SVM) وخوارزمية نايف بيز(NB) وخوارزمية الجار الأقرب (KNN) وذلك من خلال استخدام لغة Python ومكتبة (sk-learn).

بعد تدريب نماذج المصنفات المستخدمة وقياس دقة الخوارزميات تبين أن خوارزمية آلة دعم المتجه (SVM) قد حصلت على أعلى دقة للنموذج بنسبة 92% وهي نسبة مرتفعة وكافية لحل مشكلة البحث.

الكلمات المفتاحية:

الذكاء الاصطناعي، خوارزميات تعلم الآلة، همزة الوصل، همزة القطع، خوارزميات التصنيف، اللغة العربية.

Abstract

The Arabic language is considered a living organism that grows and develops through the correct practice and application of all its literature and its syntactic, morphological, semantic, and lexical branches. In this research study, we review the contribution of technology to the development of Arabic language, especially the correct writing of Hamzat Alwasl and Hamzat Alqatae.

This research aims to build a smart Classifier model to classify Arabic words beginning with the letter alif and their hamza into conjunctive and disjunctive through the using of artificial intelligence techniques in general and machine learning algorithms to establish accurate and correct criteria in writing the conjunctive and disjunctive hamza correctly.

Consequently, technology would be adapted to contribute to the service of Arabic language.

This research relied on the compilation of Arabic words beginning with hamza by designing a digital questionnaire. The task of this questionnaire is to collect the largest number of words beginning with hamza and to classify them as Alwasl and Hamzat Alqatae according to the grammatical rules followed in this process. The questionnaire was circulated in the internet web and was filled out by fifty specialists in syntax with different academic ranks. The total number of classified words reached 400 words, and after processing and excluding the repeated words, 101 words, we obtained 299 valid words to be applied to the Classifier model, and based on the size and type of the collected data and the classification mechanism followed, classification algorithms were applied that fit the collected sample such as the vector support machine algorithm , Naif Biz algorithm, and Nearest neighbor algorithm by using Python language and sk-learn library.

After training the used classificatory models and measuring the accuracy of the algorithms, it was quite apparent that the vector support machine (SVM) algorithm had obtained the highest accuracy of the model (92%), a high and enough percent to solve the research problem.

key words:

Artificial Intelligent, Machine Learning Algorithms, Hamza, Wasl, Gtaa, Classification Algorithms, Arabic Language.

تمهيد:

يمكننا تعريف التعلم الآلي على أنه يتبع لعلم الذكاء الاصطناعي الذي يهتم بتصميم خوارزميات وتقنيات تسمح لأجهزة الحاسب الآلي بعملية التعلم الذاتي مع إمكانية تطوير هذه الخوارزميات.

يمكننا استعراض نوعين من أنواع التعلم الآلي هما: التعلم الاستقرائي والتعلم الاستنتاجي، يعرف التعلم الاستقرائي أيضاً بالتعلم الاستكشافي ويعمل على مبدأ استنتاج قواعد عامة من البيانات، وهذا يختلف عن التعلم الاستنتاجي حيث يتم إعطاء المتعلم القواعد التي يحتاج لتطبيقها.

يعتمد التعلم الآلي على مبدأ تعلم الأنظمة من البيانات المتاحة وتحديد الأنماط المناسبة واتخاذ القرارات بدون تدخل بشري، وحيث أن المهمة الأساسية للتعلم الآلي هي استخلاص المعلومات من البيانات، فإن التعلم الآلي هو طريقة من طرق تحليل البيانات تعمل على أتمتة بناء نماذج تحليل بيانات تعرف بنماذج التحليل التنبؤية، يُعرف هذا النمط باسم نموذج التحليل التنبؤي أو التصنيفي، تتيح هذه النماذج للباحثين وعلماء البيانات والمحللين اتخاذ قرارات تمكن من الوصول لنتائج موثوقة.

يخضع بناء نموذج تحليل البيانات لعدة آليات أو طرق للوصول لنتائج دقيقة، حيث أن مرحلة بناء النموذج تستلزم تنفيذ عدد من الخطوات الرياضية أو المنطقية المرتبة والمتسلسلة للوصول لنموذج تحليلي مدرب يستطيع التنبؤ بقيم أقرب للحقيقة أو تصنيف البيانات بشكل دقيق، يطلق على هذه المجموعة من الخطوات اسم الخوارزمية، كما يطلق على نموذج تصنيف البيانات اسم المصنف.

وفي هذا البحث سنتناول استخدام نماذج خوارزميات تعلم الآلة في تصنيف همزتي الوصل والقطع وسيتم تطبيق هذه الخوارزميات على الكلمات العربية نسبة لأن اللغة العربية لم تنل حظًا من الدراسة والبحث كغيرها من اللغات الأخرى.

أهداف البحث:

- حل مشكلة الاستخدام الخاطئ لرسم همزتي الوصل والقطع في النصوص العربية.
- التعرف على أنسب خوارزميات التصنيف ضمن خوارزميات تعلم الآلة لتحديد المواضع الصحيحة لاستخدام همزتي الوصل والقطع في بداية الكلمة.

– قياس جودة أشهر خوارزميات التصنيف في التمييز بين همزتي الوصل والقطع في بداية الكلمة.

مشكلة البحث وأسئلته:

تكمن مشكلة البحث الأساسية في الاستخدام الخاطئ لرسم همزتي الوصل أو القطع في غير موضعها الصحيح، وذلك عند كتابة الكلمات المبدوءة بحرف الألف بصورة خاصة، حيث يخطئ الكثيرين في رسم الهمزة أو إسقاطها الشيء الذي يضعف قوة اللغة.

يسهم هذا البحث في تطوير اللغة العربية، وذلك عن طريق بناء نموذج تصنيف ذكي يمكن المحررين والممارسين، وكل مستخدمي اللغة من التصنيف الصحيح لهمزتي الوصل والقطع.

يسعى هذا البحث للإجابة عن الأسئلة التالية في ضوء الاستفادة من خوارزميات تعلم الآلة المستخدمة في التصنيف والتنبؤ:

– كيف يمكننا التمييز بين همزتي الوصل والقطع في النصوص العربية؟

– ما الفوائد التي نجنيها عند تصنيف همزتي الوصل والقطع في النصوص العربية؟

– ما خوارزميات تعلم الآلة المناسبة لتصنيف همزتي الوصل والقطع في النصوص العربية؟

الدراسات السابقة:

تعد هذه الدراسة من الدراسات الجديدة في حقل الدراسات النحوية والتقنية التي تتناول استخدام خوارزميات تعلم الآلة لتصنيف همزتي الوصل والقطع في النصوص العربية، وعلى حد علمي، ومن خلال بحثي، لم أجد أي دراسات تعرضت لهذا الموضوع من قبل نسبة لصعوبة استخدام المكتبات البرمجية في التعامل مع النصوص العربية بجدارة، ولكن هناك وجه شبه كبير وتداخل لدراسات أخرى ارتبطت بالبحث لاستخدام هذه الخوارزميات في تطبيقات مماثلة ومشابهة، وقد استفدت منها، وساعدتني كثيرًا في الولوج لبحثي، ومنها:

1. دراسة بعنوان: تنقيب الآراء في جمل المقارنة العربية [1]

تناولت هذه الدراسة مشكلة التعرف على جمل المقارنة في تنقيب الآراء المستخدمة في النص العربي، وقد ركزت الدراسة على استخلاص الآراء من جمل المقارنة وذلك بمعرفة المنتج الذي يفضله كاتب الرأي مقارنة مع منتج آخر أو أكثر.

وقد ذكر الباحث أن هناك بعض الأبحاث في هذا المجال بالنسبة لجمل اللغة الإنجليزية وغيرها من اللغات، ولكن بالنسبة للجمل العربية فهذه أول دراسة، كما أن الدراسة استخدمت تقنية تعتمد على التصنيف اللغوي وتقنية أخرى تعتمد على تعليم الآلة.

2. دراسة بعنوان: دراسة مقارنة لخوارزميات التنقيب في الآراء وتحليل العواطف وتطبيقاتها[2]

تناولت هذه الدراسة مشكلة تعدد وجهات نظر الزبائن المخزونة في مستودعات البيانات الخاصة بالأنترنت، الشيء الذي أعطى للتنقيب في البيانات وتحليل المشاعر اهتماماً في السنوات الأخيرة ، وقد ذكر الباحث أن الناس **قد اعتمدوا على الالة في تصنيف البيانات ومعالجتها**، اذ أن توافر كميات هائلة من وجهات النظر حول منتج واحد يساعد على التنبؤ بمشاعر الزبون وذلك عن طريق تحليل وجهات النظر التي تساعد ليس فقط في زيادة الأرباح ولكن أيضاً في تحسين المنتج، وقد قارنت هذه الدراسة التقانات المتوفرة حالياً والمستخدمة في التطبيقات المتعددة في مجال التنقيب في الآراء.

بعد استعراض الدراسات أعلاه نبعت فكرة تصنيف الهمزة انطلاقا من استخدام الباحثين في الدراستين أعلاه مفهوم التنقيب في النصوص العربية باستخدام طريق التنقيب في النصوص العربية مع اختلاف خوارزميات التصنيف المستخدمة.

منهجية البحث:

1- استخدمنا في هذا البحث أسلوب الاستبانة الرقمية، واقتصرت مهمة الاستبانة في تجميع أكبر قدر من الكلمات المبتدئة بالهمزة وتصنيفها (بواسطة مختصين في اللغة العربية) لهمزتي وصل أو قطع، وفقاً للقواعد النحوية المتبعة في تصنيف همزتي الوصل والقطع.

2- استخدمنا في هذا البحث أيضاً النهج الإحصائي الذي يعتمد على الاستبانة المصممة وعلى (مجموعة تدريب الكلمات) باستخدام نموذج التصنيف، ومن أجل هذا البحث سنستخدم خوارزميات تعلم الآلة وذلك بعد النظر في تكرار الشروط والخصائص (القواعد النحوية المرتبطة بالكلمات المبتدئة بالهمزة والمستخلصة عبر الاستبانة) لاستنتاج العناصر الأساسية في نموذج التصنيف.

هيكل البحث:

يتكون البحث من ثلاثة مباحث:

المبحث الأول (الإطار النظري للبحث): يتناول مفهوم الذكاء الاصطناعي، مفهوم تعلم الآلة، خوارزميات تصنيف البيانات وهمزتي الوصل والقطع.

المبحث الثاني (المعالجة والتطبيق): يتناول مراحل تطبيق نماذج المصنفات التي تشمل: جمع ووصف وتنظيف ومعالجة وترميز البيانات وتحديد المتغيرات المستقلة والتابعة في عينة البيانات كما يتناول أيضاً بناء نماذج المصنفات.

- المبحث الثالث (النتائج مخرجات البحث): يتناول اختبار نماذج المصنفات، قياس دقة النماذج، النتائج، ومخرجات البحث والفوائد المرجوة منه.

المبحث الأول: الإطار النظري للبحث

أولا: مفهوم الذكاء الاصطناعي:

يُعتبر علم الذكاء الاصطناعي (Artificial Intelligence) هو أحد فروع علوم الحاسب، وأحد ركائز صناعة التقنية في عصرنا الحديث، ويشار له بالاختصار (AI) ويُمكن تعريف علم الذكاء الاصطناعي على أنه قدرة الآلات وأجهزة الحاسب الآلي على القيام بمهام تحاكي إلى حد كبير ما يقوم به العقل البشري والذي يتميز بالذكاء، ويمكن تلخيص هذه المهام في القدرة على التفكير أو التعلم من تجاربه السابقة، إذن يمكننا القول بأن الذكاء الاصطناعي يهدف إلى الوصول إلى أنظمة تتصرف وتتعلم وتفهم كما يتصرف ويتعلم ويفهم البشر كما تمتلك خاصية الذكاء[3].

أنواع الذكاء الاصطناعي:

يُمكن تقسيم أنواع الذكاء الاصطناعي حسب قدراته إلى ثلاثة أنواع[4] كالآتي:

1- الذكاء الاصطناعي المحدود:

وهو من الأنواع التي تستطيع القيام بمهام مُحددة وواضحة، مثل تطبيقات السيارات ذاتيّة القيادة أو برامج التعرف على الكلام أو الصور، أو لعبة الشطرنج، ويُعتبر هذا النوع من أنواع الذكاء الاصطناعي أكثرها شيوعاً.

2- الذكاء الاصطناعي العام :

وهو من الأنواع التي لها قدرات تفكيرية تُشابه قدرة الإنسان، إذ أنه يجعل الآلة تكون قادرة على التفكير من تلقاء نفسها ويشابه بشكل كبير التفكير البشري، وفي الحقيقة لا توجد أيّ تطبيقات عمليّة لهذا النوع، بل توجد فقط دراسات بحثيّة تحتاج للكثير من الجهد لتحويلها إلى واقع، وتعد طريقة الشبكات العصبيّة أحد نماذج الذكاء الاصطناعي العام، إذ أنها تُعنى بإنتاج نظام شبكات عصبيّة للآلة تكون مُشابهة لتلك التي يحتويها العقل البشري[5].

3- الذكاء الاصطناعي غير المحدود:

يعتبر الذكاء الاصطناعي غير المحدود من النوع الذي قد يتجاوز مستوى الذكاء البشري، وقدرته على أداء المهام بصورة يمكن أن تكون أفضل من قدرة البشر المتخصصين ذوي المعرفة، وهذا النوع له العديد من الخصائص الضرورية، مثل: التعلم والتخطيط التلقائي والقدرة على التواصل وإصدار القرار المناسب، لكن مفهوم الذكاء الاصطناعي الفائق يعتبر مفهومًا افتراضيًا غير موجود في عصرنا. يمكن أيضًا تصنيف الذكاء الاصطناعي وفقًا للوظائف التي يمكن أن يؤديها إلى الأنواع الأربعة المختلفة الآتية:

1- الآلات التفاعلية:

هي أبسط أنواع الذكاء الاصطناعي نظرًا لأنها تفتقر إلى القدرة على التعلُّم من الخبرة السابقة لتطوير الأعمال المستقبلية، لذا فأنها سوف تتفاعل مع الخبرة الحالية لإنتاج أفضل طريقة ممكنة، من أمثلة هذا النوع معدات Deep Blue التي طورتها شركةIBM ونظام AlphaGo من شركة جوجل.

2- الذاكرة المحدودة:

يستطيع الذكاء الاصطناعي من فئة الذاكرة المحدودة (Limited Memory)تخزين بيانات تاريخية سابقة عن النظام الحالي لمدة زمنيّة مقيدة، ويُعد نهج القيادة الذاتيّ من أجدر الأمثلة على ذاك النمط، إذ يقوم بحفظ السرعة الأخيرة للسيارات الأُخرى، ومعدل المسافة الفاصلة بين تلك السيارات، والحد الأقصى للسرعة، وغيرها من المعلومات الأُخرى الأساسية للقيادة عبر طرق النقل[6].

3- نظريّة العقل

يُعنى ذلك النوع من أنواع الذكاء الاصطناعي باستيعاب الآلة للمشاعر البشرية، والتفاعل مع البشر والتواصل معهم، وتجدر الإشارة أنه لا نجد أيّة تطبيقات عمليّة حتى هذه اللحظة على ذاك النمط من الذكاء الاصطناعي.

4- الإدراك الذّاتي:

يُعتبر فئة الإدراك الذاتي (Self-Awareness) من التنبؤات المستقبلية التي يصبو إليها الذكاء الاصطناعي، ويعمل على مبدأ تقني وحسي (إدراكي) حديث للغاية حيث يمكن أن تتولد عند الآلة معرفة ذاتية وأحاسيس خاصة بها، الشأن الذي سيجعلها أكثر ذكاءً من الكائن البشري، وما يزال ذلك المفهوم غير حاضر في الواقع[7].

الحقول الفرعيّة للذكاء الاصطناعي:

يحتوي علم الذكاء الاصطناعي على العديد من المجالات الفرعية، مثل: التعلم الآلي، والذي يتضمن تمكين أجهزة الكمبيوتر من التعلم بشكل مستقل عن أي تجربة سابقة، حتى تتمكن أجهزة الكمبيوتر من التنبؤ لاتخاذ القرار المناسب بسرعة، يتم من خلال تطوير خوارزمية تسمح بهذا الموقف. وتجدر الإشارة إلى أن هذا المصطلح اقترحه آرثر صموئيل لأول مرة في عام 1959. سنشير أدناه إلى بعضاً من أشهر المجالات الفرعية للذكاء الاصطناعي على النحو الآتي:

- تنقيب البيانات:

يُقصد به البحث والتنقيب عن بيانات مُحددة وأنماط مُعينة ضمن مجموعة كبيرة من البيانات بواسطة برامج حاسوبيّة، إذ تستطيع الشركات الاستفادة من تنقيب البيانات في تطوير أدائها وزيادة مبيعاتها وتقليص تكاليف الإنتاج[8].

- استرجاع المعلومات والويب الدلالي:

يشير مفهوم استرجاع المعلومات إلى عملية البحث عن أي نوع من البيانات والمستندات التي قد تكون موجودة على الإنترنت من خلال مفهوم الويب الدلالي. تقوم خدمة الويب الدلالية بتحويل البيانات الموجودة على شبكة الويب إلى قاعدة بيانات عالمية للمعلومات المترابطة، بحيث يمكن للآلات أن تفهمها، ولا يقتصر استخدامها على البشر. بهذه الطريقة، فيمكن للآلة حجز التذاكر عبر الإنترنت، أو استخدام قاموس على الإنترنت، أو أشياء أخرى تتطلب في البداية استخدامًا يدويًا لإكمالها[9].

- تمثيل المعرفة :

يُعتبر تمثيل المعرفة أحد مجالات الذكاء الاصطناعي الذي يهتم بجعل الآلات تفكر وتتخذ القرار المناسب، حيث أنه يتم تجميع وتخزين المعارف المكتسبة بواسطة الآلة في قاعدة بيانات يتم استخدامها لتبادل المعرفة وإدارة مكوناتها، وتكون مرجعاً لاتخاذ أيّة قرارات ذكيّة تصدر عن الآلة في المستقبل.

- التفكير المنطقي والتفكير الاحتمالي

التفكير المنطقي في الذكاء الاصطناعي هو أحد أشكال التفكير المختلفة، لأن الحقائق يتم استنتاجها بناءً على البيانات المتاحة. يتوافق التفكير المنطقي مع ما يسمى بالتفكير الاحتمالي، والذي يستخدم مفاهيم الاحتمالية وعدم اليقين في المعرفة للتعامل مع جميع أوجه عدم اليقين في المستقبل لجميع الأحداث التي قد يشتبه في حدوثها[10].

- تعلم الآلة:

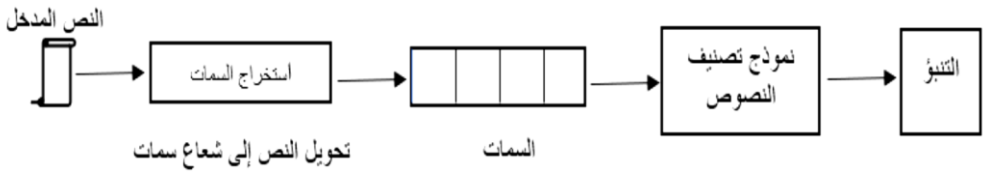
يعد التعلم الآلي أحد فروع الذكاء الاصطناعي، والذي يتضمن تصميم وتطوير خوارزميات وتقنيات تسمح لأجهزة الكمبيوتر بامتلاك خصائص "التعلم". وبشكل عام، ينقسم التعلم إلى مستويين هما: الاستقرائي والاستنباطي، حيث يقوم المنهج الاستنباطي باستنتاج القواعد والأحكام العامة من البيانات الضخمة.

ثانياً: مفهوم تعلم الآلة (Machine Learning):

تتمثل المهمة الرئيسية للتعلم الآلي في استخلاص معلومات قيمة من البيانات، لذا فهي قريبة جدًا من استخراج البيانات، يستخدم التعلم الآلي في مجال تحليل البيانات وهو طريقة لتطوير النماذج المعقدة والخوارزميات المناسبة لاستخلاص البيانات باستخدام عمليات تنبؤية، يسمى هذا التحليل بالتحليل التنبؤي. تتيح هذه النماذج التحليلية للباحثين والمحللين البيانات تعلم قرارات ونتائج موثوقة وتستطيع إدراك البيانات المخزنة وعلاقاتها.

كما يمكن تعريف نظم تعلم الآلة على أنها أنظمة تقوم بتنبؤات بناءً على ما تعلمته من المعطيات السابقة وتحتاج هذه الأنظمة إلى التدريب على العديد من أمثلة النصوص والتنبؤات (العلامات) المتوقعة لكل منها، وتسمى المعطيات المستخدمة للتدريب بمجموعة بيانات التدريب وتكون هذه المعطيات مصنفة مسبقاً بمميزات وكلما كانت مجموعة التدريب أكثر دقة والسمات المختارة مناسبة كانت تنبؤات المصنف أفضل، فعندما يتم تدريب مصنف بطريقة التعلم الآلي يجب تحويل بيانات التدريب إلى شيء يمكن أن تفهمه الآلة، حيث يتم استخراج السمات وتحويلها إلى أشعة (تمثيل النصوص بواسطة الأرقام) والتي سوف تساعده على التعلم من البيانات الموجودة ووضع التنبؤات حول النصوص القادمة[11].

يمكن للنموذج المدَّرب استخراج السمات من النص الجديد والتنبؤ أو تصنيف النصوص حسب خصائص محددة باستخدام خوارزميات تصنيف البيانات كما في الشكل (1-1) أدناه:



شكل (1-1): يوضح الية تصنيف النصوص باستخدام خوارزميات تصنيف البيانات

ثالثاً: خوارزميات تصنيف البيانات:

هناك عدة خوارزميات لتصنيف البيانات تناسب تطبيقات التنقيب في البيانات النصية بكل سهولة بعد معالجتها، كما أنيا سهلة التدريب سواءً مع الكميات الكبيرة أو الصغيرة من البيانات المقدمة، وفيما يلي سنستعرض أشهر خوارزميات تعلم الآلة الموجه لتصنيف البيانات النصية والتي تم استخدامها في هذا البحث:

1- خوارزمية آلة دعم المتجه SVM-Support Vector Machines:

تندرج هذه الخوارزمية المعروفة اختصارا ب بخوارزمية (SVM) تحت خوارزميات التعلم الموجه للآلة الذي يعتمد على مجموعة بيانات معروفة النتائج مسبقاً (مجموعة التدريب) في تدريب الخوارزمية حتى تتمكن من تحليل أي مجموعة جديدة من البيانات وتصنيفها أو تحديد ميولها، وضعت هذه الخوارزمية من قبل العالمان فلاديمير فابنك وأليكسى شيرفونينكيز عام 1963م ومن ثم طورها كل من كورينا كورتز وفابنك عام 1993م ونشرت عام 1995م [12].

تعد خوارزمية (SVM)من أشهر طرق التصنيف الآلي التي تعتمد على إيجاد منحنى أو مستوى فاصل، يفصل العينات التي تم إدخالها عن بعضها البعض، وتتميز الخوارزمية باستخدامها في تصنيف المسائل ذات الفئات الثنائية حصراً.

بصورة عامة يمكننا القول إن خوارزمية آلة دعم المتجه تستخدم للتصنيف وتمييز الأنماط. والهدف منها هو إيجاد أفضل دالة تصنيف وأيضاً تهدف إلى التمييز بين أعضاء فئتين من بيانات التدريب، وكما ذكرنا مسبقاً فإن الفكرة من الخوارزمية هي إيجاد مستوى مثالي يفصل بين الفئتين والذي يستخدم للتصنيف وتحديد كل نمط، ومن مميزاتها الدقة العالية في التصنيف وتطبق في مجالات واسعة منها تحديد فئات النص حسب تصنيف الصورة [13].

في حال استطاعت الخوارزمية إيجاد مستوى فاصل ببعد أقل بواحد من بعد متجهات النقاط يكون التصنيف خطياً، وإلا فإن التصنيف يعتبر غير خطي. وفي حالة وجود أكثر من فاصل خطى، يتم اختيار الفاصل الذي يحقق هامش أوسع بين أقرب نقطتين من نوعين مختلفين لبعضهم، وهو ما يسمى بالمستوى ذو الهامش الأكبر[14].

تحدد دقة الخوارزمية بقدرتها على فصل نوعين بحيث تكون أقرب نقطتين لبعضهما البعض أبعد ما يكون، ويمكننا تسمية المستوى الفاصل بالحافة أو هامش الفصل. وبصفة عامة كلما ازدادت الحافة أو هامش الفصل، كلما قل الخطأ عند التعميم على مجموعة بيانات جديدة.

2- خوارزمية نايف بيز : (Naive Bayes)

تعتبر خوارزمية نايف بيز(NB) من خوارزميات التعلم الموجه للآلة أيضاً، وتعتمد على قواعد الاحتمال الشرطي التي صاغها العالم توماس بيز[15]، حيث تحسب الاحتمال باستخدام عدد التكرارات للقيم وتكرارات وتركيبات القيم في البيانات المعروفة النتائج مسبقاً (بيانات التدريب).

يُعرف مصنف نايف بيز كمجموعة من المصنّفات الاحتمالية البسيطة القائم على فرضية عامة مفادها أن جميع السمات مستقلة عن بعضها البعض وفقاً للصنف المحدد، ولسهولة تطبيق هذا المصنف وسرعته فهو يعتبر خط الأساس في تصنيف النصوص ويعتبر فعالا في العديد من المجالات بالرغم من وجود عدد من المصنفات الأخرى بدقة، أعلى مثل نموذجSVM، حيث أن نموذج Naive Bayes يقوم بتوزيع النصوص لكل صنف باستخدام نموذج احتمالي مع افتراضات مستقلة، هذه الطريقة شائعة جدا في مجال تصنيف النصوص، حيث إن المصنف الثنائي واحداً من أفضل الطرق المعروفة لنموذج Naive Bayes الذي استخدم تمثيل شعاعي ثنائي القيمة للنصوص.

تم إجراء العديد من التحسينات لمصنف بيز منها تعديل حساب الاحتمالات وتقليل السمات وقليل من الخصائص الأخرى، وحيث إن نظرية بيز تبحث عن احتمال وقوع حدث ما علما بأن حدث آخر قد وقع مسبقاً.

كما تعتبر خوارزمية نايف بيز واحدة من أهم خوارزميات تعلم الآلة الموجه لعدة أسباب منها: سهولة بناء المصنف، كما أنها لا تحتاج لاستخدام ما يسمى بمخططات التخمين estimation schemes لأي متغيرات تكرارية معقدة ، وقد تطبق بسهولة على مجموعة بيانات ضخمة والهدف من الخوارزمية هو بناء قاعدة تسمح بتخصيص هياكل مستقبلية إلى صنف معين وذلك بإعطاء متجهات للمتغيرات التي تصف ذلك الهيكل ، ويمكن بواسطتها إجراء المستخدم للعديد من الإحصاءات لسهولة استخدامها[16].

3-خوارزمية الجار الأقرب K-Nearest Neighbor

يمكن استخدام خوارزمية الجار الأقرب (KNN) على أنها مصنف بسيط وفعال لتصنيف النصوص، يحتوي المصنف KNN على عيبين هما: التعقيد الحسابي في حال تشابه العينات كبير جدا، وأنه يتأثر أداءها بسهولة في حال كانت العينات التدريبية فردية.

يمكن تقليل تعقيد KNN من خلال استخدام ثلاث طرق: إما بالحد من أبعاد المتجه الممثل للنص، أو بالحد من كمية العينات التدريبية أو بالحد من إيجاد أقرب الجيران أي قيمة الـ K.

تستخدم KNN في تصنيف النصوص عن طريق حساب المسافة بين النص وكل النصوص في مجموعة البيانات التدريبية باستخدام مقياس للاختلاف أو التشابه فيما بينها، ثم العثور على أقرب K مجاورة بين جميع نصوص التدريب ويتم تحديد صنف النص إلى الصنف الذي يضم أكبر عدد من النصوص الموجودة في أقرب الجيران من النصوص وكباقي الخوارزميات فقد تم التحسين عليها بأكثر من طريقة [17].

رابعاً: همزتا الوصل والقطع

في هذا الجزء من الجانب النظري سنتناول همزتي الوصل والقطع في أول الكلمة تحت المسميات الآتية:

. معنى الهمزة (لغةً):

ورد في كثير من معاجم اللغة ومنها لسان العرب لابن منظور إن الهمزة في اللغة بمعنى الغمز واللمز والضغط[18].

. معنى الهمزة (اصطلاحا):

الهمزة في اصطلاح اللغويين وردت لها عدة تعريفات نورد منها تعريف الأزهري في كتابه تهذيب اللغة حيث يقول: "اعلم أن الهمزة لا هجاء لها، وإنما تُكتبُ مَرةً ألفاً، ومَرة واواً، ومرة ياءً، والألف اللينةُ لا حرف لها، وإنما هي جزء من مدةٍ بعد فتحة، والحروف ثمانية وعشرون حَرفاً مع الواو والألف والياء، وتتم بالهمزة تسعة وعشرين حرفاً، والهمزة كالحرف الصحيح، غير أن لها حالات من التليين والحذف والإبدال والتخفيف... وليست من حروف الجوف، إنما هي حَلقية من أقصى الفم ".

أقسام الهمزة في بداية الكلمة:

تنقسم الهمزة في بداية الكلمة إلى قسمين:

أ/ همزة الوصل:

- تسميتها:

سميت همزة الوصل بهذا الاسم، لأنها يتوصل بها إلى النطق بالحرف الساكن الواقع في ابتداء الكلام، حيث أن القاعدة المهمة في ذلك أنه لا يبتدأ بساكن ولا يوقف بحركة.

- تعريفها:

هي الهمزة التي تثبت في الابتداء، وتسقط في حالة الوصل، وذكر ابن مالك: "إذا سمي بما أوله همزة وصل قطعت الهمزة إن كانت في منقول من فعل مثل: (رأيت اعلم)" يعني إذا سمي به شخص[19].

- حركة همزة الوصل:

الأصل في همزة الوصل أن تكون مكسورةً مثل: انطلق، ماعدا ألف التعريف تكون مفتوحة، وكذلك ألف أيمن مثل قولك: "أيمن الله بفتح الألف" [20].

- مواضع همزة الوصل:

لهمزة الوصل مواضع ثلاث كما يلي:

أ/ الأسماء:

تقع همزة الوصل في أول الكلمة في عشرة أسماء في اللغة هي (اسم، ابن، ابنة، اثنان، اثنتان، امرؤ، امرأة، ابنم، ايم الله، است) ولقد ورد عدد من هذه الأسماء في القرآن الكريم.

ب/ الأفعال:

همزة الوصل في الأفعال المبدوءة بها هي فعلان:

- الفعل الماضي: وتكون همزة الوصل في ماضي الخماسي مثل: انفتح، وكذلك ماضي السداسي مثل: استغفر.
- فعل الأمر: وتكون همزة الوصل في فعل الأمر في المواضع التالية:
  - أمر الفعل الثلاثي مثل: نصر -> انصر.
  - أمر الفعل الخماسي مثل: انتصر للحق.
  - أمر الفعل السداسي مثل: استغفر لذنبك.

ج/ الحروف:

لا توجد همزة الوصل في الحروف إلا في حرف واحد وهو (ال) التعريفية أو ما يعرف باللام الشمسية مثل: الشمس، وكذلك في اللام القمرية مثل: القمر[21]

أ/ همزة القطع:

- تسميتها:

سميت همزة القطع بهذا الاسم لأنها تقطع وتحجز ما قبلها عمَّا بعدها من الحروف مثل قولك: درس –أدرس، فالهمزة قطعت الدال عن الهمزة التي قبلها وكذلك نصر– أنصر، فالهمزة حجزت ما بعدها وهو حرف النون عن الهمزة التي قبلها.

- مواضع همزة القطع:

لهمزة القطع مواضع ثلاث كما يلي:

أ/ الأسماء:

تقع همزة القطع في كل الأسماء في اللغة العربية مثل: (أحمد – أمجد – أكرم – أسعد) وكذلك الأسماء الستة فكل ما يبدأ فيها بهمزة فهي همزة قطع مثل: (أخوك – أبوك)، ماعدا الأسماء العشرة التي تكون همزتها همزة وصل.

ب/ الأفعال:

تقع همزة القطع في الأفعال كما يلي:

- ماضي الفعل الثلاثي مثل: أخذ– أكل.
- ماضي الفعل الرباعي مثل: أكرم – أنجز.
- أمر الفعل الرباعي مثل: أَكرِم.
- المصدر الرباعي مثل: إنذار من أنذر.

ج/ الحروف:

كل الحروف في اللغة العربية همزتها همزة قطع ماعدا (ال) التعريفية مثل: (إنَّ – أنَّ – إذا – إلى)[22].

- الفرق بين همزتي الوصل والقطع:

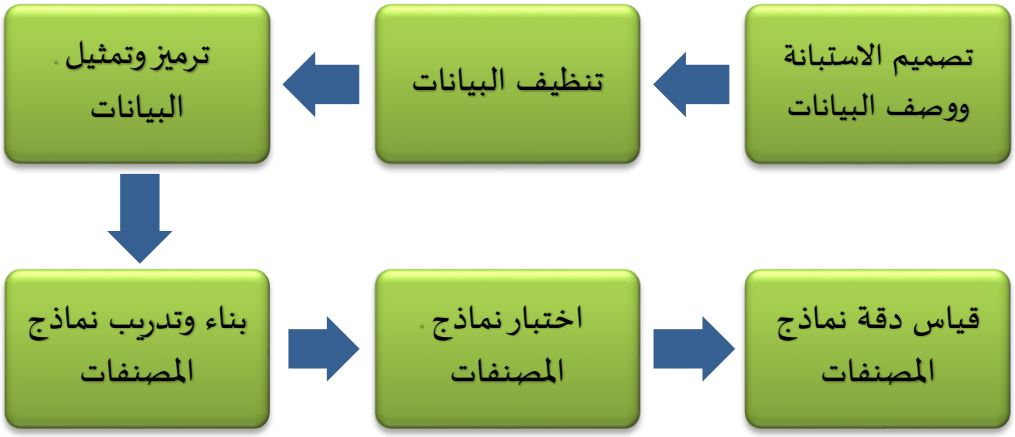
لكي نميز ونفرق بين همزة الوصل وهمزة القطع لابد من معرفة الآتي:

- تأتي همزة القطع ساكنة أو متحركة.
- همزة الوصل لا تأتي إلا متحركة.
- همزة القطع تثبت في بداية الكلام وفي أثنائه وفي طرفه مثل: سأل في وسطها وقرأ في طرفها.
- همزة الوصل تسقط أثناء الكلام بينما تثبت في البدء فقط.
- همزة الوصل لا تكون إلا زائدة بينما همزة القطع تكون أصلية أو زائدة[23].

المبحث الثاني: التطبيق والتنفيذ

تمهيد:

إن البيانات في يومنا هذا تتضخم يوما بعد يوم وتتعدد مصادرها، وهذا يقود إلى تعرض هذه البيانات إلى الكثير من المشاكل التي تقلل من جودة البيانات مثل كثرة البيانات المفقودة وعدم تناسق البيانات، لذلك قمنا في هذا البحث بتقسيم مراحل تطبيق نماذج مصنفات البيانات إلى ستة مراحل بدءاً بمرحلة تصميم الاستبانة وجمع البيانات وانتهاءً بمرحلة قياس دقة نماذج المصنفات، كما في الشكل (1-2) التالي:



شكل رقم (1-2): مراحل تطبيق نماذج المصنفات

في هذا المبحث سنتناول الأربعة مراحل الأُوَل من مراحل تطبيق نماذج المصنفات وسيتم تناول المرحلتين المتبقيتين في المبحث الثالث.

- مرحلة تصميم الاستبانة ووصف البيانات

تم تجميع الكلمات العربية المبتدئة بالهمزة وذلك عن طريق تصميم استبانة رقمية انظر (ملحق أ) وقد تم تصميمها باستخدام نماذج جوجل وتم تسميتها (استبانة تصنيف الكلمات المبتدئة بهمزة وصل أو قطع)، وقد اقتصرت مهمة الاستبانة في تجميع أكبر قدر من الكلمات المبتدئة بالهمزة بواسطة متخصصين في اللغة العربية وقد قام المتخصصون بتصنيفها لهمزتي وصل أو قطع وفقاً للقواعد النحوية المعروفة، كما تم نشر الاستبانة على عنوان الويب التالي:

https://docs.google.com/forms/d/e/1FAIpQLSdtwnflWQ7hoTdVbKfZUHDVix4fmXPE3grXMiQGIDem9Dsc0Q/viewform

تمت تعبئة بواسطة 50 متخصص في النحو بدرجات علمية متفاوتة بهدف الحصول على عينة متجانسة، وقد بلغ عدد الكلمات الكلية المصنفة 400 كلمة.

تم التصنيف للكلمات المستخلصة من نصوص العينة إلى كلمات تبدأ بهمزة وصل أُشير إليها في هذا البحث بـ (Wasl) وكلمات تبدأ بهمزة قطع أُشير إليها في هذا البحث بـ (Gtaa) وقد تم تحديد المتغير التابع (Val) وفقاً للقيمتين (Wasl/Gtaa) أما المتغيرات المستقلة (الخصائص) فقد تم تقسيمها لثلاث خصائص طبقاً للقواعد النحوية المشار إليها لاحقاً في الجدول (2-1).

تم اختيار صيغة ملف مجموعة البيانات (Dataset) من نوع (csv) وقد احتوى على الخصائص (المتغيرات المستقلة) وهي:

- المتغير (diacritic): متغير رقمي بطول (1) رقم يشير إلى حركة حرف الألف في بداية الكلمة (فتحة، كسرة، ضمة) تم تمثيلها بالقيم الرقمية (1، 2، 3) على التوالي.
- المتغير (Count): متغير رقمي يشير لعدد حروف الكلمة حيث إن الكلمة التي تحتوي على حرفين، ثلاثة أحرف أو أربعة ...إلخ، يتم تمثيلها رقمياً بالقيم (2، 3 أو 4 ...10) على التوالي.
- المتغير (morphological): متغير رقمي يشير للوزن الصرفي للأفعال والمصادر (فعل، أفعل، أنفعل، .... فعال) يتم تمثيلها رقمياً بالقيم (0، 1، 2، 3،...11) على التوالي.
- المتغير (noun): متغير رقمي يشير إلى أن الكلمة (ليست اسم، اسم عادي، من الأسماء العشرة، من الأسماء الستة، اسم موصول، اسم فاعل) وتم تمثيله رقمياً بالأرقام (0، 1، 2، 3، ...، 6) على التوالي.
- المتغير (verb): متغير رقمي يشير إلى أن الكلمة (ليست فعلاً، فعل ماضٍ، فعل مضارع، فعل أمر) وتم تمثيله رقمياً بالأرقام (0، 1، 2، 3) على التوالي.

- المتغير (adjective): متغير رقمي يشير إلى أن الكلمة (ليست صفة، صفة) وتم تمثيله رقمياً بالأرقام (0، 1) على التوالي.
- المتغير (letter): متغير رقمي يشير إلى أن الكلمة (ليست حرفاً، حرف) وتم تمثيله رقمياً بالأرقام (0، 1) على التوالي.
- المتغير (the): متغير رقمي يشير إلى أن الكلمة (غير معرفة بـ آل، معرفة بـ آل) وتم تمثيله رقمياً بالأرقام (0، 1) على التوالي.

تم تسمية ملف مجموعة البيانات (Dataset) بالاسم (Arabic Words)، الجدول (2-1) يوضح جانب من توصيف بيانات الملف (Arabic_Words.csv) كما يلي:

191	افتعل	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
192	افتقد	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
193	اكتفى	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
194	امتحن	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
195	انتصر	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
196	انتقى	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
197	انتكس	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
198	انتمى	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
199	انتهى	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
200	اكتشف	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
201	التف	5	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افتعل	غير معرفة بـال	كسرة	Wasl
202	أحمد	4	اسم عادي	ليس فعلاً	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
203	أيمن	4	من الأسماء العشرة	ليس فعلاً	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
204	أثنى	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
205	أبعد	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
206	أبلغ	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
207	أثرى	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
208	أحرم	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
209	أدرج	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
210	أسرى	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa
211	أسلم	4	ليس اسماً	فعل ماضي	ليس صفة	ليس حرفاً	افعل	غير معرفة بـال	فتحة	Gtaa

- مرحلة تنظيف البيانات

وبالتأكيد عندما تكون جودة البيانات منخفضة فهذا سيؤثر حتما على نتائج التحليل. في هذا البحث استخدمنا عدة أساليب لتنظيف البيانات (Cleaning Data) على النصوص المجمعة، وقد اشتملت مرحلة تنظيف البيانات على المراحل الآتية:

-التعامل مع البيانات المفقودة

- حذف البيانات المكررة.

بعد إجراء عمليات تنظيف البيانات (Cleaning Data) على النصوص المجمعة، وبعد معالجة واستبعاد الكلمات المكررة وكذلك إكمال البيانات الناقصة أصبح عدد الكلمات المستبعدة 101 كلمة، حصلنا على عدد 299 كلمة صالحة لتطبيقها على نموذج المصنف.

- مرحلة ترميز وتمثيل البيانات:

بعد دراسة القواعد النحوية التي توضح مواضع همزتي الوصل والقطع في أول الكلمة تم تحديد الأسس والخصائص (Features) التي يمكن الاعتماد عليها في تحديد قيمة المتغير التابع (Outcome)، الجدول (2-2) يوضح القواعد المتبعة لتحديد مواضع همزتي الوصل والقطع.

اسم Noun		فعل verb		حرف letter		صفة adjective		حركة الهمزة diacritic		الميزان الصرفي morphology		عدد الحروف Count		ال التعريف the	
خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز
ليس اسماً	0	ليس فعلاً	0	ليس حرفاً	0	ليست صفة	0	فتحة	1	فعل	1	حرفين	2	غير مضافة	0
اسم عادي	1	ماضي	1	حرف	1	صفة	1	كسرة	2	افعل	2	ثلاثة	3	مضافة	1
الأسماء العشرة	2	مضارع	2					ضمة	3	انفعل	3	أربعة	4		
الأسماء الستة	3	امر	3							افتعال	4	خمسة	5		
أسماء الموصول	4									افاعيل	5	ستة	6		
أسماء الأفعال	5									افعال	6	سبعة	7		
										افتعل	7	ثمانية	8		
										استفعل	8	تسعة	9		
										استفعال	9	عشرة	10		
										انفعال	10				
										فعال	11				

الجدول (2-2) يوضح القواعد المتبعة لتحديد مواضع همزتي الوصل والقطع

تم ترميز البيانات الموجودة في جدول (1-2) وتحويل القيم الحرفية إلى قيم رقمية حتى تتمكن الخوارزميات من التعامل معها وفقاً للجدول (2-2) الذي يوضح القواعد المتبعة لتحديد مواضع همزتي الوصل والقطع وتصبح البيانات بعد ترميزها كما في الجدول (3-2) الآتي:

	Outcome	diacritic	the	morphological	letter	adjective	verb	noun	Count	Word
171	Gtaa	1	0	2	0	1	0	0	4	أصفر
172	Gtaa	1	0	2	0	1	0	0	4	أعذب
173	Gtaa	1	0	2	0	1	0	0	4	أعزب
174	Gtaa	1	0	2	0	1	0	0	4	أعور
175	Gtaa	1	0	2	0	1	0	0	4	أسود
176	Gtaa	1	0	2	0	1	0	0	4	أعزل
177	Gtaa	1	0	2	0	0	0	1	4	أحمد
178	Gtaa	1	0	2	0	0	1	0	4	أثنى
179	Gtaa	1	0	2	0	0	0	2	4	أيمن
180	Wasl	2	0	3	0	0	1	0	5	انحسر
181	Wasl	2	0	3	0	0	1	0	5	انحصر
182	Wasl	2	0	3	0	0	1	0	5	انطلق
183	Wasl	2	0	3	0	0	1	0	5	انفعل
184	Wasl	2	0	4	0	0	0	1	6	اجتهاد
185	Wasl	2	0	4	0	0	0	1	6	احتيار
186	Wasl	2	0	4	0	0	0	1	6	ابتداء
187	Wasl	2	0	4	0	0	0	1	6	ابتدار
188	Wasl	2	0	4	0	0	0	1	6	ابتسام
189	Wasl	2	0	4	0	0	0	1	6	ابتهال
190	Wasl	2	0	4	0	0	0	1	6	اجتهاد

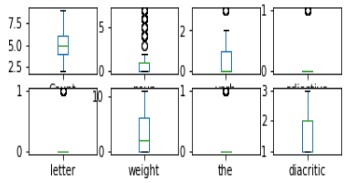
جدول (2–3) ترميز البيانات في ملف مجموعة البيانات (arabic.csv)

بعد اكتمال عملية ترميز البيانات يتم تمثيل البيانات باستخدام لغة بايثون Python واستيراد مكتبات (numpy- scipy- sklearn) باستخدام محرر (Jupyter) كما في الشكل (2–2) الآتي:

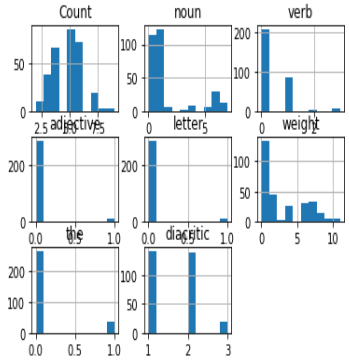
4. Data Visualization

4.1 Univariate Plots

In [20]:
```
# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(4,4), sharex=False, sharey=False)
pyplot.show()
```



In [21]:
```
# histograms
dataset.hist()
pyplot.show()
```



شكل (2–2): تمثيل البيانات باستخدام محرر (Jupyter Notebook)

- مرحلة بناء وتدريب نماذج المصنفات:

تم تقسيم البيانات الممثلة إلى بيانات تدريبة بنسبة (67%) وبيانات تجريبية بنسبة (33%) من أصل البيانات الكلية (299) سجل تمهيداً لبناء نموذج تدريب البيانات باستخدام خوارزميات التصنيف المختارة (خوارزمية الة دعم المتجه SVM – خوارزمية نايف بيز NB – خوارزمية الجار الأقرب KNN) وقد تم اختيارها نسبة لتناسب حجم العينة وقيم البيانات مع الخوارزميات أعلاه وقد تم بناء النموذج باستخدام دوال مكتبة (Sk-Learn) كما في الشكل (2–3) أدناه:

5.2 Build Models

```
In [35]: # Spot Check Algorithms
models = []

models.append(('KNN', KNeighborsClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    # Test options and evaluation metric
    kfold = StratifiedKFold(n_splits=10, random_state=None)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```

```
KNN: 0.940000 (0.037417)
NB: 0.540000 (0.053852)
SVM: 0.955000 (0.041533)
```

شكل (2–3): بناء نماذج التصنيف باستخدام محرر (Jupyter Notebook)

المبحث الثالث: النتائج ومخرجات البحث

أولاً: اختبار نماذج المصنفات

بعد اكتمال مرحلة بناء نماذج المصنفات باستخدام خوارزميات (SVM, NB, KNN) انتقلنا لمرحلة اختبار هذه النماذج حتى نتمكن من قياس جودتها فيما بعد، وقد تم اختبار النماذج بإجراء تنبؤات على النماذج بعد تدريبها باستخدام دوال مكتبة (sklearn)، الشكل (3–1) يوضح اختبار نماذج المصنفات باستخدام دالة التنبؤ (predict) كما يلي:

6. Make Predictions

6.1 Make Predictions

```
In [47]: # Make predictions on validation dataset For SVM
         model1 = SVC(gamma='auto')
         model1.fit(X_train, Y_train)
         predictions1 = model1.predict(X_validation)
```

```
In [51]: # Make predictions on validation dataset FOR Naive Bayes
         from sklearn.naive_bayes import MultinomialNB
         model2 = MultinomialNB()
         model2.fit(X_train, Y_train)
         predictions2 = model2.predict(X_validation)
```

```
In [76]: # Make predictions on validation dataset FOR Naive Bayes
         from sklearn.naive_bayes import MultinomialNB
         model3 = KNeighborsClassifier(n_neighbors = 8)
         model3.fit(X_train, Y_train)
         predictions3 = model3.predict(X_validation)
```

شكل (3–1) يوضح اختبار نماذج المصنفات باستخدام دالة التنبؤ (predict)

ثانياً: قياس دقة نماذج المصنفات

تعتبر مرحلة تقييم نتائج نماذج التنقيب في البيانات من المراحل المهمة التي تمكننا من تعريف النموذج الأكثر فعالية، وتقاس فعالية النموذج من خلال دقة الخطة المعمول بها وتلعب طبيعة البيانات المستخدمة في بناء النماذج دورًا أساسياً في فعاليتها، ويوجد العديد من الطرق الإحصائية التي تختبر نماذج التصنيف نبين أهمها فيما يلي:

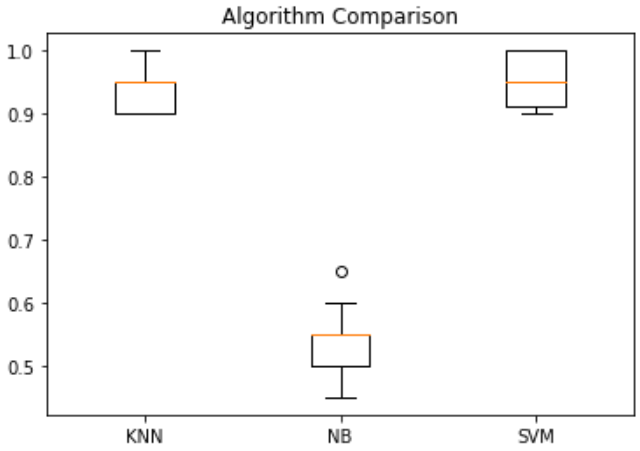
-حساب متوسط الدقةِ Average Accuracy

هو المتوسط الحسابي لنسب دقة التوقعات الصحيحة لكل فئة يقدمها النموذج إلى عدد التصنيفات الفعلية لهذه الفئة في مجموعة بيانات الاختبار، وقد بلغ متوسط دقة التوقعات للخوارزميات المستخدمة كما هو مبين في الجدول (3–1) الذي يبين أن خوارزمية (SVM) قد حصلت على أعلى نسبة عند حساب متوسط الدقة حيث بلغت دقتها (94%) بينما حصلت خوارزميتي (NB) و (KNN) على متوسط دقة (54%) و (85%) على الترتيب كما يلي:

الخوارزمية	متوسط الدقة
SVM	%94
NB	%54
KNN	%85

جدول (3–1): متوسط دقة الخوارزميات

الشكل (3–2) يوضح رسم بياني لمتوسط دقة خوارزميات (SVM) و (NB) و (KNN)



شكل (3–2): رسم بياني لمتوسط دقة الخوارزميات

- حساب الدقة الإجمالية Total Accuracy

هي نسبة مجموع التوقعات الصحيحة المقدمة من النموذج إلى مجموع التصنيفات الفعلية في مجموعة بيانات الاختبار، وقد بلغت الدقة الاجمالية للخوارزميات المستخدمة كما هو مبين في الجدول (3-2):

الخوارزمية	الدقة الإجمالية
SVM	%92
NB	%78
KNN	%91

جدول (2-3): الدقة الاجمالية للخوارزميات

من الجدول أعلاه يتضح أن خوارزمية KNN حصلت على دقة إجمالية بلغت 91% مقارنة بخوارزمية SVM التي حصلت على أعلى معدل دقة إجمالية بلغت 92% بينما حصلت خوارزمية NB على أقل معدل دقة اجمالية بنسبة 78%.

- إيجاد مصفوفة الارتباك Confusion Matrix

تظهر مصفوفة الارتباك عدد الحالات المتوقعة بشكل صحيح وعدد الحالات المتوقعة بشكل خاطئ في مجموعة الاختبار لكل صنف من الأصناف مقارنةً مع عدد الحالات الفعلية لتلك الأصناف. رتبة هذه المصفوفة هي $n \times n$ حيث n هي عدد الأصناف في عمود المتغير الهدف (المتغير التابع)، وقد تم إيجاد مصفوفة الارتباك للخوارزميات المستخدمة لتصنيف المتغير التابع (Outcome) إلى صنفين (همزة وصل Wasl وهمزة قطع Gtaa) كما في الشكل (3-3) التالي:

Confusion matrix for SVM:

	: Gtaa	pred: Wasl
true: Gtaa	**49**	**5**
true: Wasl	**3**	**42**

confusion matrix for NB:

	pred: Gtaa	pred: Wasl
true: Gtaa	**41**	**13**
true: Wasl	**9**	**36**

confusion matrix for KNN:

	pred: Gtaa	pred: Wasl
true: Gtaa	**49**	**5**
true: Wasl	**4**	**41**

شكل (3-3): مصفوفة الارتباك لخوارزميات (SVM, NB, KNN) المستخدمة في تصنيف همزتي الوصل والقطع

ثالثاً: النتائج

وقد قمنا باستخدام خوارزمية آلة دعم المتجه (SVM) وخوارزمية نايف بيز (NB) وخوارزمية الجار الأقرب (KNN) لتصنيف همزتي الوصل والقطع (في بداية الكلمة) وقد تم اختيار هذه الخوارزميات بناء على عوامل عدة أبرزها تناسب حجم مجموعة البيانات مع هذه الخوارزميات وكثرة الخصائص المعتمد عليها في عملية التصنيف، تم عمل نموذج تدريب لجميع هذه الخوارزميات وقد اتضح من خلال حساب دقة النموذج لكل خوارزمية تفوق خوارزمية ألة دعم المتجه (SVM) حيث حصلت على أعلى نسبة دقة للنموذج بلغت 92%، وهي نسبة عالية وكافية لحل مشكلة البحث الأساسية، وقد كان السبب الأساسي في حصول الخوارزمية على هذه النسبة العالية هو كثرة عدد الخصائص المستخدمة في التصنيف (تم استخدام 8 متغيرات مستقلة) مما أسهم في جعل نموذج التصنيف أكثر واقعية.

رابعاً: مخرجات البحث والفوائد المرجوة منه:

كان الهدف من هذا البحث المنجز هو بناء نموذج ذكي يعمل على تصنيف همزتي الوصل والقطع (في بداية الكلمة) من خلال استخدام خوارزميات تصنيف البيانات من أجل تأسيس معايير دقيقة وصحيحة في كتابة النصوص العربية بصورة دقيقة لتسهم وتساعد في تطويع التقنية لخدمة اللغة العربية، كما هدف البحث أيضاً لقياس جودة أشهر خوارزميات التصنيف في التمييز بين همزتي الوصل والقطع في بداية الكلمة، ويمكن لهذا النموذج الذكي الذي تم تصميمه أن يسهم في تطوير اللغة العربية كما يلي:

- النموذج الجديد يمكن استخدامه في مراجعة البحوث العلمية والتأكد من رسم الهمزة في الكلمات بصورة صحيحة مما يسهم في اكتمال عناصر البحث العلمي.
- النموذج الجديد يمكن استخدامه في مراجعة المقالات الإخبارية المنشورة في وسائط التواصل الاجتماعي المختلفة من خلال تثبيته في جزء الإضافات الخاصة بالمستعرض للتأكد من رسم الهمزة بشكل صحيح في جميع كلمات المقالات المنشورة مما يسهم في تطوير الإعلام العربي الجديد.

- النموذج الجديد يمكن استخدامه مع تطبيقات الجوالات واعتباره كجزء أصيل من أنظمة تشغيل الجوالات للتأكد من كتابة الأوامر المترجمة للغة العربية والمبتدئة بهمزة بشكل صحيح.
- النموذج الجديد يمكن دمجه مع أنظمة الترجمة المستخدمة في المؤتمرات والطائرات والقطارات حيث يساعد في عرض الهمزة بصورة صحيحة.

خلاصة القول يمكننا استخدام النموذج الجديد كأداة برمجية يمكن أن يتم دمجها مع كافة الأنظمة البرمجية والتقنية التي تعرض نصوص عربية مقروءة أو مكتوبة باستخدام أنظمة برمجية لتصنيف الهمزة في الكلمات المبدوءة بهمزة وصل أو قطع.

التوصيات

1- تطبيق خوارزميات تصلح للتعامل مع الأغراض ذات فئات التصنيف المتعددة باستخدام خوارزميات فهرسة أخرى غير التي تعالج الأغراض ذات الفئات الثنائية فقط.
2- تطوير مكتبات sklearn لتدعم اللغة العربية بشكل كامل.
3- تطوير النموذج ليعمل كمصنف للنصوص العربية المسموعة.

الهوامش:

[1] – د. علاء مصطفى الهليس: تنقيب الآراء في جمل المقارنة العربية، المجلة العربية الدولية للمعلوماتية، المجلد الثاني، العدد الرابع، 2013م.

[2] د. رنا زهير عبد الغني العبيدي، د. غيداء عبد العزيز الطالب: دراسة مقارنة لخوارزميات التنقيب في الآراء وتحليل العواطف وتطبيقاتها، مجلة الرافدين لعلوم الحاسوب والرياضيات المجلد (12)، العدد الثاني، 2018م.

[3] امل كاظم ميرة & ,تحرير جاسم كاطع. (2019). تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر تدريسي الجامعة.

[4] امل كاظم ميرة & ,تحرير جاسم كاطع. (2019). تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر تدريسي الجامعة.

[5] د. م. مصطفى عبيد، (2018).التحليل المتقدم وتنقيب البيانات، دار الفكر العربي، الطبعة الأولى، المجلد الأول، القاهرة.

[6] Burkov, A. (2019). The hundred-page machine learning book (Vol. 1). Canada: Andriy Burkov.

[7] Florin, G. (2011). Data mining: concepts, models and techniques. Springer-Verlag. Berlin Heidelberg.

8 Ian, H. W and Eibe, F. (2005). Data Mining: practical machine learning tools and techniques, Second Edition. Elsevier Inc. San Francisco: USA.

[9] Jiawei, H., Micheline, K. and Jian P. (2012). Data mining: concepts and techniques, Third edition. Elsevier Inc: USA.

[10] Kalyani, G. and Jaya, A. Lakshmi Performance assessment of different classification techniques for intrusion detection. Journal of Computer Engineering (IOSRJCE).

[11] Kalyani, G. and Jaya, A. Lakshmi Performance assessment of different classification techniques for intrusion detection. Journal of Computer Engineering (IOSRJCE).

[12] Michael, J.A. B. and Gordon, S. L. (2004). Data mining techniques for marketing, sales, and customer relationship management, Second edition. Wiley Publishing, Inc. Indianapolis, Indiana: USA.

[13]امل كاظم ميرة & ,تحرير جاسم كاطع. (2019). تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر تدريسي الجامعة.

[14] د. م. مصطفى عبيد، (2018).التحليل المتقدم وتنقيب البيانات، دار الفكر العربي، الطبعة الأولى، المجلد الأول، القاهرة.

[15]العالم الإحصائي الإنجليزي توماس بيز(Thomas Bayes) عاش خلال الفترة (1701–1761م). هو من قام بصياغة حالة خاصة من النظرية المشهورة والتي تحمل اسمه وهي نظرية بيز (Bayes' theorem)رغم أنها لم تنشر في حياته وإنما نشرت بعد وفاته بواسطة ريتشارد برايس (Richard Price) .

[16] د. م. مصطفى عبيد، (2018).التحليل المتقدم وتنقيب البيانات، دار الفكر العربي، الطبعة الأولى، المجلد الأول، القاهرة.

[17] امل كاظم ميرة & ,تحرير جاسم كاطع. (2019). تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر تدريسي الجامعة.

[18] محمد بن مكرم بن على، أبو الفضل، جمال الدين ابن منظور الأنصاري الرويفعى الإفريقي. (1993). لسان العرب. دار صادر – بيروت الطبعة: الثالثة (17/1).

[19] محمد بن عبد الله، ابن مالك الطائي الجياني، أبو عبد الله، جمال الدين. تحقيق: عبد المنعم أحمد هريدي. (د.ت). شرح الكافية الشافية. جامعة أم القرى مركز البحث العلمي وإحياء التراث الإسلامي كلية الشريعة والدراسات الإسلامية مكة المكرمة الطبعة: الأولى (1466/3).

[20] أبو الفتح عثمان بن جني الموصلي ، تحقيق: فائز فارس. (د.ت) اللمع في العربية. دار الكتب الثقافية – الكويت (226–225).

[21] محمد رفيق مؤمن الشوبكي. (2015) اللآلئ الذهبية في شرح المقدمة الجزرية. غزة– فلسطين الطبعة: الأولى (77).

[22] عبد الله محمد النقراط ، تحقيق: محمد خليل هراس. (2003) الشامل في اللغة العربية. دار الكتب الوطنية – لبنان الطبعة: الأولى (178–162).

[23] الطاهر بن محمد زواوي البيريني الجزائري ،الشيخ. (2019). الميسر المفيد في فن التلاوة والتجويد من قراءة نافع المدني وعاصم الكوفي ومن رواية ورش وقالون. دار الكتب العلمية – بيروت – لبنان.

المصادر والمراجع

أولاً: المصادر والمراجع العربية

[1] محمد بن مكرم بن على، أبو الفضل، جمال الدين ابن منظور الأنصاري الرويفعى الإفريقي. (1993). لسان العرب. دار صادر – بيروت الطبعة: الثالثة (17/1).

[2] محمد بن أحمد بن الأزهري الهروي، أبو منصور. (2001) تهذيب اللغة. دار إحياء التراث العربي – بيروت الطبعة: الأولى (33/1).

[3] محمد بن عبد الله، ابن مالك الطائي الجياني، أبو عبد الله، جمال الدين. تحقيق: عبد المنعم أحمد هريدي. (د.ت). شرح الكافية الشافية. جامعة أم القرى مركز البحث العلمي وإحياء التراث الإسلامي كلية الشريعة والدراسات الإسلامية مكة المكرمة الطبعة: الأولى (1466/3).

[4] أبو الفتح عثمان بن جني الموصلي ، تحقيق: فائز فارس. (د.ت) اللمع في العربية. دار الكتب الثقافية – الكويت (226-225).

[5] محمد رفيق مؤمن الشوبكي. (2015) اللآلئ الذهبية في شرح المقدمة الجزرية. غزة– فلسطين الطبعة: الأولى (77).

[6] عبد الله محمد النقراط ، تحقيق: محمد خليل هراس. (2003) الشامل في اللغة العربية. دار الكتب الوطنية – لبنان الطبعة: الأولى (178-162).

[7] الطاهر بن محمد زواوي البيريني الجزائري ،الشيخ. (2019). الميسر المفيد في فن التلاوة والتجويد من قراءة نافع المدني وعاصم الكوفي ومن رواية ورش وقالون. دار الكتب العلمية – بيروت – لبنان.

[8] م. د. نعيم سلمان البدري. (2008) همزة الوصل في اللغة العربية. جامعة واسط – مجلة كلية التربية – لبنان المجلد: الأول (29-15).

[9] الهليس ، علاء مصطفى. (2013). تنقيب الآراء في جمل المقارنة العربية Arab . International Informatics Journal, 178(2117), 1-14

[10] د. رنا زهير عبد الغني العبيدي، د. غيداء عبد العزيز الطالب: دراسة مقارنة لخوارزميات التنقيب في الآراء وتحليل العواطف وتطبيقاتها، مجلة الرافدين لعلوم الحاسوب والرياضيات المجلد (12)، العدد الثاني، 2018م.

[11] امل كاظم ميرة & ,تحرير جاسم كاطع. (2019). تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر تدريسي الجامعة.(22) ,Psychological Science .

[12] د. م. مصطفى عبيد، (2018).التحليل المتقدم وتنقيب البيانات، دار الفكر العربي، الطبعة الأولى، المجلد الأول، القاهرة.

ثانياً: المصادر والمراجع الأجنبية:

[13] Burkov, A. (2019). The hundred-page machine learning book (Vol. 1). Canada: Andriy Burkov.

[14] Florin, G. (2011). Data mining: concepts, models and techniques. Springer-Verlag. Berlin Heidelberg.

[15] Ian, H. W and Eibe, F. (2005). Data Mining: practical machine learning tools and techniques, Second Edition. Elsevier Inc. San Francisco: USA.

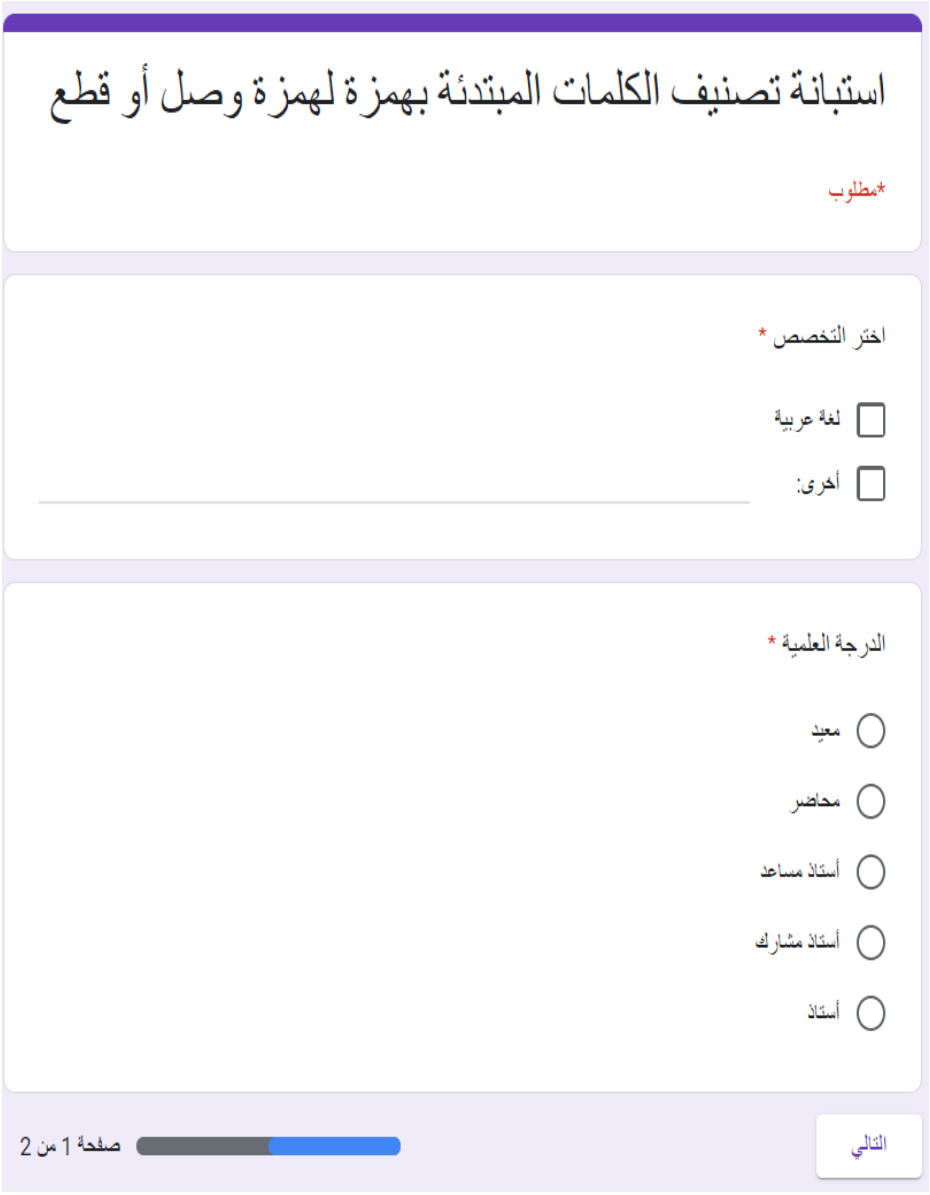
[16] Jiawei, H., Micheline, K. and Jian P. (2012). Data mining: concepts and techniques, Third edition. Elsevier Inc: USA.

[17] Kalyani, G. and Jaya, A. Lakshmi Performance assessment of different classification techniques for intrusion detection. Journal of Computer Engineering (IOSRJCE).

[18] Michael, J.A. B. and Gordon, S. L. (2004). Data mining techniques for marketing, sales, and customer relationship management, Second edition. Wiley Publishing, Inc. Indianapolis, Indiana: USA.

ملحق (أ)

نموذج استبانة تصنيف الكلمات المبتدئة بهمزة وصل أو قطع



استبانة تصنيف الكلمات المبتدئة بهمزة لهمزة وصل أو قطع

*مطلوب

اختر التخصص *

- لغة عربية
- أخرى:

الدرجة العلمية *

- معيد
- محاضر
- أستاذ مساعد
- أستاذ مشارك
- أستاذ

التالي

صفحة 1 من 2

عليك كتابة كلمة تكون مبتدئة بهمزه وصل أو قطع

الاسئلة متعلقة بالكلمة التي اخترتها

أكتب الكلمة *

إجابتك

عدد حروف الكلمة *

إجابتك

هل الكلمة اسم *

- اسم عادي
- من الأسماء الستة
- من الأسماء العشرة
- اسم موصول
- اسم اعجمي
- ليس اسماً

هل الكلمة فعل *

- ليس فعلاً
- فعل ماضي
- فعل مضارع
- فعل أمر

هل الكلمة حرفاً *

- ليس حرفاً
- حرف

هل الكلمة صفة *

- ليست صفة
- صفة

حدد حركة الهمزة *

- فتحة
- كسرة
- ضمة

هل الكلمة معرفة بـ (ال) *

- غير معرفة بـال
- معرفة بـال

حدد الميزان الصرفي للكلمة *

- فعل
- افعل
- انفعل
- افتعال
- افاعيل
- افعال
- افتعل
- استفعل
- استفعال
- انفعال
- فعال
- غير محدد

حدد نوع الهمزة للكلمة التي كتبتها *

- همزة وصل
- همزة القطع

رجوع إرسال صفحة 2 من 2