مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(9) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (9) 2022

تحسين دقة البيانات المكتوبة باللغة العربية باستخدام تقنيات تنظيف البيانات

د.م. ماهر ابراهيم*

ريم قصي محمود**

(تاريخ الإيداع 4/20 /2022 . قُبِل للنشر في 2022/10/3)

□ ملخّص□

أصبح تحسين دقة البيانات قضية حاسمة للعديد من الشركات والمؤسسات لأن دقة البيانات العالية تحسن الأداء التنظيمي إذ تؤدي إلى نتائج تحليل أفضل . يعتبر تنظيف البيانات مرحلة أساسية من مراحل المعالجة المسبقة للبيانات و التي تهدف إلى تجهيز البيانات لعمليات التحليل و التنقيب لاستخراج نماذج معرفية تفيد في عمليات اتخاذ القرار و الهدف من هذه العملية هو إيجاد طريقة لزيادة دقة البيانات أي الحصول على بيانات صحيحة و دقيقة دون التلاعب بالبيانات المتاحة . تنظيف البيانات هي عملية تحديد البيانات الخاطئة وتصحيحها و تعتبر عملية معقدة خاصة فيما يخص اللغة العربية. يوفر هذا البحث نهج هجين لتنظيف البيانات عن طريق دمج تقنيات محسنة قادرة على التعامل مع اللغة العربية بهدف اكتشاف و تصحيح معظم أنواع الأخطاء التي قد تحدث في مرحلة جمع البيانات لضمان نتائج دقيقة لعمليات التحليل. في هذا البحث تمت معالجة أخطاء التناسق و القيم المتناقضة و الأخطاء الإملائية العربية حيث تم اقتراح خوارزمية جديدة لقياس مسافة التحرير التي تأخذ بعين الاعتبار تشابه الأحرف على لوحة المفاتيح حيث تم إجراء عدة اختبارات لقياس دقة البيانات في عدة مراحل و أظهرت النتائج دقة عالية للتقنيات المقترحة .

الكلمات المفتاحية: تنظيف البيانات ، دقة البيانات ، اللغة العربية ، الأخطاء الإملائية ، خوارزمية مسافة التحرير

* مدرس في قسم هندسة تكنولوجيا المعلومات– كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

** طالبة ماجستير في قسم هندسة تكنولوجيا المعلومات– كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(9) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (9) 2022

Enhancing the accuracy of written Arabic data by using data cleaning techniques

Dr. Maher Ibrahim*
Reem Mahmoud**

(Received 20/4/ 2022 . Accepted 3/10/ 2022)

□ ABSTRACT

sImproving data accuracy has become a critical issue for many companies and organizations because higher data accuracy improves organizational performance as it leads to better analysis results. Data cleaning is an essential stage of data pre-processing, which aims to prepare data for analysis and exploration processes to extract cognitive models that are useful in decision-making processes. The goal of this process is to find a way to increase the accuracy of the data, that is, to obtain correct and accurate data without manipulating the available data . Data cleaning is the process of identifying and correcting wrong data, and it is considered a complex process, especially with regard to the Arabic language. This research provides a hybrid approach to data cleaning by incorporating improved techniques capable of dealing with the Arabic language , with the aim of discovering and correcting most types of errors , that may occur in the data collection stage to ensure accurate results for the analysis processes. In this research, consistency errors, contradictory values, and Arabic spelling errors were addressed, a new algorithm has been proposed to measure the edit distance that takes into account the similarity of characters on the keyboard. Several tests have been conducted to measure the accuracy of the data in several stages, and the results showed a high accuracy of the proposed techniques.

Key Words: Data cleaning, Data accuracy, Arabic language, Spelling errors, editing distance

*Lecturer, InformationTechnology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.
** Master student, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

1. المقدمة

سجلت تكنولوجيا المعلومات نمواً هائلاً خلال العقد الماضي مما أدى بدوره إلى نمو متناسب في حجم البيانات. أصبحت معه الإدارة الفعالة للحجم الكبير والتشغيلي للبيانات في غاية الأهمية. تعتبر عملية تنظيف البيانات من أهم مراحل تحضير البيانات لعمليات التحليل والتنقيب والسبب في ذلك أن التضخم الكبير الذي أضحت عليه قواعد البيانات في هذا العصر يجعلها عرضة لاحتواء الكثير من البيانات غير الصحيحة أو المتناقضة أو غير المتناسقة أو حتى فقدان بعض البيانات الموجودة فيها وذلك بسبب ضخامتها و تدفقها من مصادر متعددة أو الإدخال اليدوي الخاطئ.

يتمثل الهدف الأساسي لتنظيف البيانات في إيجاد طريقة لزيادة دقة البيانات إلى أقصى حد دون حذف المعلومات بالضرورة أي الحصول على بيانات صحيحة خالية من الأخطاء يمكن استخدامها كمصدر موثوق للمعلومات ، و إنشاء مجموعات بيانات موحدة ومتسقة للسماح لأدوات ذكاء الأعمال وتحليل البيانات بالوصول إلى البيانات المناسبة لكل استعلام والبحث عنها بسهولة.

في الوقت الحالي تعمل العديد من الأبحاث على حل مشاكل تنظيف البيانات و تحسين جودة البيانات و تناسقها و تركز بشكل أساسي على الأخطاء الإملائية التي تسبب مشاكل في عمليات التحليل ، إذ قامت العديد من الدراسات بالعمل على اكتشاف الأخطاء الإملائية و تصحيحها في اللغة الانكليزية ، بالمقابل هناك مجموعة محددة من الأساليب التي تم تطبيقها لحل مثل هذه الأخطاء في اللغة العربية لأنها معقدة للغاية حتى أنه لا توجد مجموعة بيانات عربية معيارية تم العمل عليها [1].

تنقسم الأبحاث التي تم الاعتماد عليها إلى دراسات تتعلق بتقنيات تنظيف البيانات التي تحوي لغة عربية و لكنها محدودة المجال، و دراسات تتعلق بتنظيف البيانات الكمية التي تحمل قيماً محددة و أخرى تتعلق بأنظمة تصحيح الأخطاء العربية:

قدم [1] Al-Hagery, M. A., Alreshoodi, L. A., Almutairi, M. A., Alsharekh, S. I., & Alkhowaiter, E. S. خوارزمية هجينة تستخدم نموذج DTI(Decision Tree Induction) الذي يعتمد على أشجار القرار لملء القيم المفقودة و نموذج AMDCM(Arabic Misspelling Detection,Correction Model) و هو نهج هجين لاكتشاف و تصحيح الأخطاء الإملائية العربية عن طريق استخدام آلية البحث في القاموس لاكتشاف الأخطاء و خوارزمية مسافة التحرير لتصحيحها. يمكن أن يحل مشاكل أخرى مثل القيم المفقودة ، والقيم الوهمية و هي تسلسل غير منطقي من الأحرف ، وتوحيد أنماط التسمية المختلفة لقيم سمة معينة . قدم Hamad, M. M., & Jihad, A. A. [2] حلاً للتعامل مع الأخطاء في البيانات الكمية و أي بيانات لها قيماً محددة. توفر هذه الخوارزمية تفاعلاً للمستخدم من خلال تحديد القواعد و المصادر والأهداف المرجوة في محاولة لحل معظم المشاكل و الأخطاء منها: القيم المخالفة، التكرار، أخطاء تنسيق المجال... تم مقارنة هذه التقنية مع مجموعة من الأعمال السابقة من عدة جوانب أهمها: التفاعل، سهولة التطبيق، التوسع، و الأداء و أثبتت كفاءة عالية . و اقترح [3] Ghafour, H. H. A., El-Bastawissy, A., & Heggazy, A. F. A. خوارزمية جديدة لمطابقة السلاسل العربية و هي خوارزمية(Arabic Edit Distance Algorithm) AEDA التي تأخذ في الاعتبار الميزات الفريدة للغة العربية ومستويات التشابه المختلفة للحروف العربية معتمدةً على خوارزمية تحرير المسافة التقليدية. أعطت الخوارزمية المقترحة نتائج أكثر دقة خاصة عند مقارنة السلاسل التي تحتوي همزة أو نقاط . كما و ناقش [4] Alotaibi, S. B.

مشاكل جودة البيانات وقدم تقنية ذكية يمكنها تحديد معظم أخطاء البيانات (القيم المفقودة ، الأخطاء الإملائية ، توحيد التسمية) و تدعم اللغة العربية. قامت بالعمل على البيانات التي تم جمعها من مصادر متعددة قبل التحميل إلى مستودع البيانات. تم اختبار هذه التقنية و التحقق منها باستخدام حالات من بيانات حقيقية.

2. أهمية البحث و أهدافه:

يقدم هذا البحث مساهمة جديدة في مجال معالجة البيانات العربية ، من حيث اقتراح خوارزمية هجينة لتنظيف البيانات تهدف إلى التعامل مع اللغة العربية. يوفر هذا البحث طريقة هجينة تعتمد على دمج عدة تقنيات لتنظيف البيانات العربية و ذلك نظراً لوجود ملايين السجلات التي تكتب يومياً باللغة العربية و الحاجة المستمرة لوجود تقنيات تتعامل مع تعقيد هذه اللغة بهدف زيادة دقة هذه البيانات و بالتالي زيادة دقة نتائج عمليات التحليل و التنقيب عليها. يقدم هذا البحث تقنيات لمعالجة مجموعة من الأخطاء التي قد تحدث عند قيام المستخدم بإدخال البيانات باللغة العربية و أهمها : القيم غير المتناسقة (الصيغ المختلفة لسمة واحدة)، القيم المتناقضة (تعارض القيم في السمات المرتبطة) ، الأخطاء الإملائية التي تم التركيز بشكل أساسي على معالجتها و اقتراح آلية لتحسين عمليات الكشف عن هذه الأخطاء و تصحيحها. و تبرز أهمية البحث في تجميع مجموعة بيانات عربية جديدة لاختبار التقنيات المقترحة عليها ، وتجميع قاموس كبير الحجم للكلمات العربية الحديثة بهدف زيادة دقة عمليات تصحيح الأخطاء الإملائية.

3. طرائق البحث و مواده:

مجموعة البيانات هي جزء أساسي يتم استخدمه في أبحاث تنظيف البيانات حيث يعتمد البحث على :

3-1- مجموعة البيانات و طريقة تجميعها و سماتها:

لا يوجد مجموعة بيانات عربية معيارية لتنظيف البيانات و بالتالي فإن مجموعة البيانات المستخدمة في هذا البحث و التي سيتم تطبيق تقنيات التنظيف المقترحة عليها تم جمعها عن طريق نشر استبيان على مواقع التواصل الاجتماعي و تمت الإجابة عليه باللغة العربية. تتكون مجموعة البيانات هذه من 12 سمة كما هو موضح في الجدول(1):

الجدول (1): سمات مجموعة البيانات

Description	Attribute
Name of Student	Nname
Name of Student's mother	Mname
University	Univ
Department	Dept
Grade Point Average	GPA
Birth date	Bdate
Address	Addr
Male or Female	Gender
Age of student	Age
Marital Statue	MS
Number of children	Nchild
Current Job	Job

تحتوي مجموعة البيانات هذه على أخطاء مختلفة مثل قيم غير متناسقة و أخرى غير متسقة و قيم مفقودة و قيم وهمية و أخطاء إملائية.

3-2- القاموس المستخدم في عملية المعالجة:

من المكونات الأساسية لكشف و تصحيح الأخطاء الإملائية هو تجميع قاموس كبير الحجم يمكن استخدامه لتغطية معظم الكلمات العربية من أجل اكتشاف الكلمات التي تحتوي أخطاء إملائية و لبناء هذا القاموس تم الاعتماد على دمج :

3-2-1- KACST Arabic corpus: يقدم موقع المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية (المدونة العربية) KACST Arabic corpus و هي عبارة عن مجموعة بيانات عربية كبيرة و متنوعة ذات معايير تصميم محددة ، تتكون من أكثر من 700 مليون كلمة من عصر ما قبل الإسلام و حتى يومنا هذا (و هي فترة تغطي أكثر من 1500 عام) تم جمعها من 10 وسائط متنوعة كما هو موضح في الجدول (2). يمكن استخدام هذه المجموعة في اهتمامات بحثية مختلفة، بدءاً من الدراسات اللغوية على مستويات مختلفة وتمتد إلى معالجة اللغات الطبيعية. [5]

الجدول (2): توزع محتوى المجموعة على الأنواع العشرة من الوسائط

نوع الوسائط	نسبته من محتوى المجموعة
الصحف	37.7%
المجلات	12.9%
الكتب	14.2%
المناهج الدراسية	1.2%
الرسائل الجامعية	3%
الدوريات المحكمة	2.7%
الإصدارات الرسمية	0.5%
وكالات الأنباء	1%
الانترنت	1.3%
المخطوطات المحققة	25.5%

3-2-2- نظراً لاهتمامنا بالبيانات و الأسماء الشخصية كان من المهم التركيز بشكل أساسي على جمع:

- الأسماء العربية الأصل الشائعة الاستخدام.
- الأسماء الأعجمية التي أصبحت شائعة الاستخدام في المجتمعات العربية الحديثة (مثل : ليا ، لودي ، ميراي).
- أسماء الدول و العواصم و المحافظات و المناطق الأساسية و خاصة السورية و ذلك بهدف توحيد تنسيقها و مطابقتها في جميع السجلات.

3-3- أنواع أخطاء البيانات:

يمكن أن تحدث أنواع مختلفة من الأخطاء أثناء إدخال البيانات من قبل المستخدم و التي تُفقد هذه البيانات جودتها و اتساقها:

3-3-1- الأخطاء الإملائية: في هذا البحث تم العمل على الكلمات المعزولة غير المرتبطة بسياق النص و أهم أنواعها:

3-3-1-1- أخطاء الكلمة الواحدة Single word errors: تحدث عند حذف حرف أو إضافته أو استبداله أو تكراره في كلمة معينة عن طريق الخطأ . يوضح الجدول (3) أمثلة عن أخطاء الكلمة الواحدة .

الجدول (3): أمثلة عن أخطاء الكلمة الواحدة

"سلييم"بدلاً من "سليم"	تكرار
"مرت" بدلاً من "مرتب"	حذف
"خريفن" بدلاً من "خريف"	إضافة
"قروب" بدلاً من "قريب"	استبدال

3-3-1-2- حذف أو إضافة مسافات بيضاء White spaces deletion / insertion error:
تحدث عندما ينسى المستخدم وضع مسافات بين الكلمات أو عندما يضيف مسافة أو أكثر في الكلمة الواحدة كما هو موضح في الجدول (4) .

الجدول (4): أمثلة عن أخطاء حذف و إضافة مسافة

"صلاحالدين"بدلاً من "صلاح الدين"	حذف مسافة
"سلي مان" بدلاً من "سليمان"	إضافة مسافة

3-3-2- **القيم المتناقضة:** تحدث عندما تكون قيمة في سمة معينة، مرتبطة بقيمة في سمة أخرى، مثلاً: الحالة الاجتماعية و عدد الأولاد، إذ من غير الممكن أن تكون القيمة في حقل (عدد الأولاد) 2 عندما تكون القيمة في حقل (الحالة الاجتماعية) أعزب.

3-3-3- **القيم غير المتناسقة** (القيم المتكافئة): تحدث عندما تكون أنماط التسمية مختلفة ، مثلاً : في حقل (المدينة) يمكن أن تكتب اللاذقية بعدة أشكال (اللاذقية ، اللاذقيه ، لاذقية) و جميعها تشير إلى الكائن نفسه . تسبب مثل هذه الأخطاء مشاكل عدة في عمليات تحليل البيانات.

3-3-4- **القيم المفقودة:** تحدث عندما تكون قيمة الحقل في سمة معينة فارغة

3-3-5- **القيم الوهمية** : هي القيم التي يتم إنشاؤها بشكل كبير على الانترنت مثل عنوان البريد الالكتروني أو الحسابات على وسائل التواصل الاجتماعي ، أو من الممكن أن تكون تسلسل من الأحرف المتتالية ليس له أي دلالة أو معنى ، مثلاً كأن تكون القيمة في حقل (الاسم) بببب . حتى الآن لا يوجد أي خوارزمية قادرة على تحديد القيم الوهمية ، إنما يتبع كل مجال طريقة معينة لتحديد القيم الوهمية و تصحيحها .

4. الخوارزمية المقترحة :

تم اقتراح منهجية تعتمد على التنظيف متعدد المستويات و الخطوات كانت بالشكل :

4-1 معالجة القيم الكمية المتناقضة و غير المتناسقة :

أنماط التسمية المختلفة و الصيغ المتعددة تسبب العديد من المشاكل في تحليل البيانات مثلاً : صيغة التاريخ ، المعدل ، العمر الذي لا يتوافق مع تاريخ الميلاد ...

و بالتالي تم فرض مجموعة من القيود و القواعد على السمات التي تأخذ قيماً محددة في مجموعة البيانات وبالتالي تضمن تناسق القيم في هذه السمات و توحيد صيغتها ، هذه القواعد موضحة في الجدول(5): [2]

الجدول (5): القيود على السمات

السمة	السمة المتعلقة بها	القيود
GPA	–	نسبة مئوية %
Bdate	_	DD/MM/YYYY
Age	Bdate	العمر = التاريخ الحالي – تاريخ الميلاد قيمة موجبة 120>العمر>0
NChild	MS	قيمة موجبة الحالة الاجتماعية لا تساوي أعزب
Gender	Name	يستخدم الاسم لتحديد الجنس ذكر/أنثى

4-2 معالجة الأخطاء الإملائية :

تتكون من مرحلتين رئيسيتين :

4-2-1 : كشف الأخطاء الإملائية (Spelling Error Detection) :

و هي عملية الكشف عن الكلمات غير الصحيحة إملائياً . تم الاعتماد على آلية البحث بالقاموس لأنها الأسرع في التعامل مع الكلمات الشائعة و الأكثر استخداماً في العديد من التطبيقات لاكتشاف الأخطاء الإملائية حيث يتم مطابقة الكلمة المدخلة بالكلمات الموجودة في القاموس ، في حال وجدت الكلمة تعتبر صحيحة ، و إلا تعتبر كلمة خاطئة إملائياً . [1]

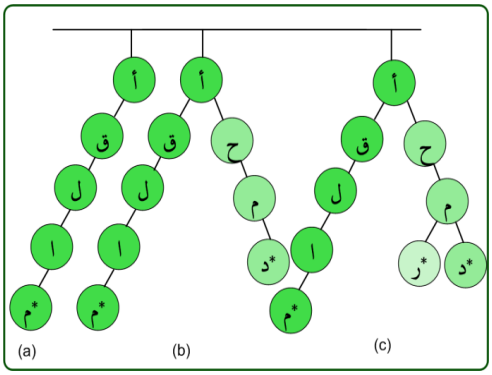
عامل الدقة في هذه المرحلة هو القدرة على التعرف على الكلمات الصحيحة و يتوقف بشكل أساسي على حجم القاموس المستخدم في عملية تحديد الأخطاء الإملائية ، حيث كلما زاد حجم القاموس زادت الدقة لأنه يغطي عدد أكبر من الكلمات [1] . عملية البحث الخطي في القاموس تأخذ زمناً كبيراً جداً و بالتالي لتحسين عملية البحث تم الاعتماد على خوارزمية Radix و هي عبارة عن بنية بيانات تعتمد على الأشجار في تخزين السلاسل المحرفية [6]، حيث:

يتم **بناء الأشجار** عن طريق أخذ كل كلمة من القاموس و تقسيمها إلى أحرف . الحرف الأول هو جذر الشجرة إذ يتم تخزينه مرة واحدة فقط و يتم تخزين باقي أحرف الكلمة كأبناء لهذا الجذر حتى تنتهي الكلمة . يتم تخزين الحرف الأخير مع علامة (*) تشير إلى انتهاء الكلمة . و بتكرار العملية السابقة على جميع كلمات القاموس سيتشكل 28

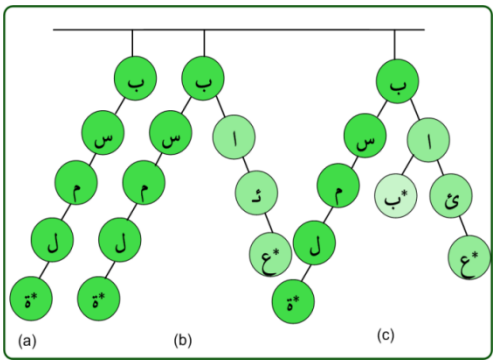
شجرة منفصلة جذر كل منها حرف من أحرف اللغة العربية . في الشكل (1) و (2) مثال عن مراحل بناء شجرة حرف (أ) و (ب) بالترتيب .

و بذلك **عملية البحث** تتم عن طريق أخذ الكلمة المراد البحث عنها و استخراج الحرف الأول منها ، و البحث عن هذا الحرف في جذور جميع الأشجار لتحديد الشجرة التي سيتم البحث فيها ، ثم استمرار البحث عن طريق تتبع أحرف الكلمة في أبناء هذا الجذر حتى الوصول إلى الحرف الأخير مع علامة (*) و بذلك تعتبر كلمة صحيحة ، إذا لم يتم إيجاد الكلمة تعتبر كلمة خاطئة إملائياً .

خوارزمية Radix تساهم في تقليل مساحة التخزين لأنه يتم تخزين الجذر و الأحرف المشتركة بين الكلمات مرة واحدة فقط و تقلل زمن البحث حيث يتم البحث ضمن مسارات الشجرة التي جذرها أول حرف من الكلمة و بلغ **أداء الخوارزمية و نجاح استخدامها** في عملية البحث 100% بالنسبة للبيانات المستخدمة في الدراسة [6] .



الشكل (1) : مثال بناء شجرة حرف (أ)



الشكل (2) : مثال بناء شجرة حرف (ب)

4-2-2 : تصحيح الأخطاء الإملائية(Spelling Error Correction):

بعد تحديد الكلمات الخاطئة إملائياً من المرحلة السابقة ، يتم الانتقال إلى مرحلة تصحيح الأخطاء الإملائية .

في العديد من الأبحاث تم الاعتماد على خوارزمية **مسافة التحرير** التي تستخدم البرمجة الديناميكية لإيجاد الكلفة الأقل بين كلمتين. و تشير الكلفة إلى الحد الأدنى لعدد عمليات التحرير اللازمة لتحويل السلسلة X (الكلمة الخاطئة إملائياً) إلى السلسلة Y (الكلمة ذات الكلفة الأقل المقابلة لها من القاموس).

يتم حساب المسافة $D(i,j)$ بين السلسلتين $X : x_1 , x_2, .. x_i$ و $Y : y_1 , y_2 , ...y_j$ من خلال العلاقة التالية:

$$D(i,j)= Min (D (i-1 , j)+ 1 ,$$
$$D (i , j-1)+ 1,$$
$$D (i-1 , j-1)+ cost \quad)$$

$$with : \quad cost = \begin{cases} 0 & if\ x_i = y_j \\ 1 & otherwise \end{cases}$$

لكن المشكلة في هذه الخوارزمية أنها تعطي نفس الكلفة لجميع عمليات التحرير و لا تولي أي اهتمام إلى تشابه الأحرف .

مثلاً : مسافة التحرير تساوي 1 عند المقارنة بين (كحمد ، أحمد) و أيضاً بين (كحمد ، محمد) . و لحل هذه المشكلة قمنا بإضافة بارامتر آخر إلى هذه الخوارزمية و هو مستوى التشابه بين الأحرف العربية على

لوحة المفاتيح أي قيمة التجاور بين الأحرف على لوحة المفاتيح و احتمالية إدخال محرف بالخطأ ، و ذلك لأن معظم الأخطاء قد تحدث نتيجة الإدخال الالكتروني الخاطئ من قبل المستخدم فعلى الرغم من أن الأبجدية العربية تحتوي على 28 حرفاً ، إلا أننا بحاجة إلى 35 مفتاحاً لكتابة النص العربي (أحرف الأبجدية العربية + ئ ، ء ، ؤ ، ى ، لا ، ة ، المسافة البيضاء) ، كما هو موضح بالشكل (3) .

الشكل (3) : لوحة مفاتيح الأحرف العربية

و عليه قمنا بالتالي :

- حساب التشابه بين الأحرف العربية من خلال العلاقة (1):[2]

$$Sim_{KB}(\mathrm{a},\mathrm{b}) = 1 - \frac{\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}}{\varphi} \ldots\ldots\ldots (1)$$

حيث : $Sim_{KB}(\mathrm{a},\mathrm{b})$: قيمة التشابه بين الحرفين a و b

x_a ، x_b: إحداثيات الحرفين a و b على محور الفواصل

y_a، y_b : إحداثيات الحرفين a و b على محور التراتيب

φ : أكبر مسافة بين حرفين على لوحة المفاتيح و تأخذ القيمة 12

إذ قمنا بتحويل لوحة المفاتيح إلى شبكة إحداثيات لتحديد مواقع الأحرف عليها كما هو موضح بالشكل (4):

Y														
3	ذ													
2		ض	ص	ث	ق	ف	غ	ع	ه	خ	ح	ج	د	
1		ش	س	ي	ب	ل	ا	ت	ن	م	ك	ط		
0		ئ	ء	ؤ	ر	لا	ى	ة	و	ز	ظ			
	0	1	2	3	4	5	6	7	8	9	10	11	12	X

الشكل (4) : إحداثيات الأحرف على لوحة مفاتيح الأحرف العربية

مثال : حساب قيمة التشابه على لوحة المفاتيح بين الحرفين (ع) و (م) و الحرفين (ر) و (م):

$$Sim_{KB}(\text{م},\text{ع}) = 1 - \frac{\sqrt{(9-7)^2 + (1-2)^2}}{12} = 0.81$$

$$Sim_{KB}(\text{م},\text{ر}) = 1 - \frac{\sqrt{(9-4)^2 + (1-0)^2}}{12} = 0.57$$

كما نلاحظ المسافة بين الحرفين (ع ، م) على لوحة المفاتيح أصغر من المسافة بين(ر ، م) و بالتالي قيمة التشابه بينهما أكبر.

- بناء مصفوفة التشابه :

تم تكرار الخطوة السابقة لقياس التشابه بين جميع الأحرف و بالتالي تشكل لدينا مصفوفة 33*33 ، جزء منها موضح في الجدول (6) :

الجدول (6) : مصفوفة التشابه بين الأحرف العربية على لوحة المفاتيح

	ا	ب	ت	ث	ج	ح	خ
ا	1	0.83	0.92	0.74	0.57	0.65	0.74
ب	0.83	1	0.75	0.88	0.41	0.49	0.57
ت	0.92	0.75	1	0.65	0.65	0.74	0.81
ث	0.74	0.88	0.65	1	0.33	0.42	0.5
ج	0.57	0.41	0.65	0.33	1	0.92	0.83
ح	0.65	0.49	0.74	0.42	0.92	1	0.92
خ	0.74	0.57	0.81	0.5	0.83	0.92	1
د	0.49	0.32	0.57	0.25	0.92	0.83	0.75
ذ	0.47	0.62	0.39	0.74	0.07	0.16	0.24
ر	0.81	0.92	0.74	0.81	0.39	0.47	0.55
ز	0.74	0.57	0.81	0.47	0.76	0.81	0.83
س	0.67	0.83	0.58	0.88	0.24	0.32	0.41
ش	0.58	0.75	0.5	0.81	0.16	0.24	0.32

- الطريقة المقترحة :

الهدف من هذه الطريقة هو تحسين عملية تصحيح الأخطاء الإملائية العربية عن طريق أخذ مستوى تشابه الأحرف على لوحة المفاتيح بعين الاعتبار و إدخال هذه القيمة إلى خوارزمية مسافة التحرير التقليدية . و عليه يتم حساب المسافة ***M(i,j)*** بين السلسلتين $X : x_1 , x_2, .. x_i$ و $Y : y_1 , y_2 , ... y_j$ من خلال العلاقة التالية :

$$M(i,j)= Min (M (i-1 , j)+ \mathbf{cost} ,\ M (i , j-1)+ \mathbf{cost},\ M (i-1 , j-1)+ \mathbf{cost})$$

$$\text{with : } \mathbf{cost} = \begin{cases} 0 & \text{if } x_i = y_j \\ 1- \mathbf{Sim}(x_i , y_j) & \text{otherwise} \end{cases}$$

حيث أصبحت الكلفة (***cost***) تساوي قيمة تشابه الأحرف على لوحة المفاتيح ***Sim*** (x_i , y_j) بدلاً من أن تأخذ القيمة 1 .

5. النتائج و المناقشة :

تم استخدام مجموعة البيانات التي تم جمعها لتطبيق تقنيات التنظيف المقترحة حيث تتكون مجموعة البيانات المستخدمة في الاختبار من 100 سجل و تحتوي مجموعة مختلفة من الأخطاء . تم تطبيق الآليات و الخوارزميات المقترحة بالاعتماد على لغة البرمجة Python باستخدام بيئة العمل jupyter notebook و المكتبات المتوفرة لمعالجة اللغات الطبيعية.

تم إجراء تجربتين لدراسة تأثير النهج المقترح على تحسين القدرة على التعرف على الكلمات الصحيحة و تصحيح الكلمات الخاطئة :

- التجربة الأولى : ارتباط عامل الدقة بحجم القاموس المستخدم للكشف عن الأخطاء الإملائية :

لدراسة تأثير حجم القاموس المستخدم في عملية البحث و إضافة الأسماء الأعجمية و الكلمات الحديثة في اللغة العربية على عملية الكشف عن الأخطاء الإملائية ، تم إجراء عملية البحث عن هذه الأخطاء باستخدام :

- القاموس (1) : و هو القاموس الجزئي المتضمن الأسماء العربية الأصل و KACST Arabic corpus
- القاموس (2) : و هو القاموس الكلي المقترح

تم قياس قدرة تصنيف الكلمات الصحيحة و تقييم النتائج من خلال العلاقة :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots (2)$$

TP(*True Positive*): الكلمة صحيحة إملائياً و تم تحديدها على أنها صحيحة

TN(*True Negative*): الكلمة خاطئة إملائياً و تم تحديدها على أنها خاطئة

FP(*False Positive*) : الكلمة خاطئة إملائياً و تم تحديدها على أنها صحيحة

FN (*False Negative*) *:* الكلمة صحيحة إملائياً و تم تحديدها على أنها خاطئة

الجدول (7) يوضح المقارنة بين القاموس1 و القاموس 2 من حيث دقة التعرف على الكلمات الصحيحة إذ تم حساب هذه القيمة من العلاقة (2) و كانت النتائج على الشكل التالي :

الجدول (7) : مقارنة نتائج الدقة للقاموس 1 و القاموس 2

	TP	FP	TN	FN	ACC%
القاموس (1)	797	6	120	200	81.65%
القاموس (2)	997	6	120	0	99.46%

من خلال هذه النتائج تبين أن القاموس المستخدم في هذا البحث و إضافة الكلمات الحديثة و الجديدة في اللغة العربية كان لها تأثير في زيادة قيمة الدقة لتصل إلى **99.46% و بالتالي تم تأكيد أنه كلما زاد حجم القاموس زادت دقة التعرف على الكلمات الصحيحة .**

- التجربة الثانية : تأثير إضافة تشابه لوحة المفاتيح إلى خوارزمية مسافة التحرير على دقة تصحيح الأخطاء الإملائية :

تمت المقارنة بين الطريقة المقترحة و خوارزمية مسافة التحرير التقليدية المستخدمة في الأبحاث السابقة ، تم قياس التشابه و المسافة بين عينة عشوائية من الكلمات التي تحوي خطأً إملائياً و التي تم اكتشافها في المرحلة السابقة و الكلمات الصحيحة المقابلة لها و نتائج المقارنة موضحة في الجدول (8) :

الجدول (8) : مقارنة بين خوارزمية edit distance و الخوارزمية المقترحة من حيث المسافة و نسبة التشابه

No.	String		Edit distance		Proposed	
	S	T	Dist	Sim%	Dist	Sim%
1	عاي	علي	1	67%	0.08	97.4%
2	تصرين	تشرين	1	80%	0.12	97.6%
3	سلليمنان	سليمان	2	75%	0.08	99%
4	ميسونم	ميسون	1	84%	0.08	98.7%
5	ريمم	ريم	1	75%	0	100%
6	غاننم	غانم	1	84%	0	100%
7	مظفة	موظفة	1	80%	0.12	97.6%
8	نو	نور	1	67%	0.33	89%

من الجدول السابق نلاحظ أن الخوارزمية المقترحة تقدم نسبة تشابه أعلى بين الكلمات الخاطئة و مقابلاتها الصحيحة أي أنها تعالج أنواع الأخطاء الإملائية المختلفة بدقة أعلى من الخوارزمية التقليدية .تم حساب الدقة الإجمالية للخوارزمية للحل المقترح الأول و الثاني الذي تقدمه من خلال العلاقة (3):

$$SR = \frac{CW}{N} \% \dots \dots \dots \dots (3)$$

SR: نسبة نجاح الخوارزمية في تصحيح الكلمات الخاطئة إملائياً

CW: عدد الكلمات التي تم معالجتها بشكل صحيح

N: عدد الكلمات الكلية

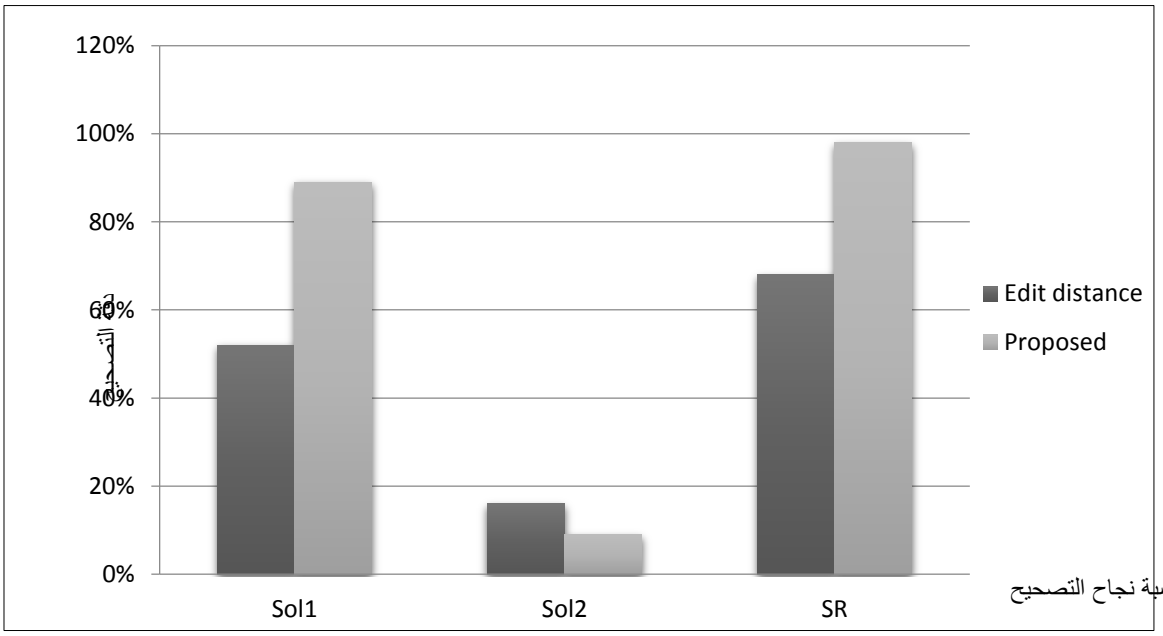
يبين الجدول (9) نتيجة المقارنة بين الطريقة المقترحة و خوارزمية مسافة التحرير من حيث الدقة في عملية تصحيح الكلمات الخاطئة بالنسبة للحل المقترح الأول و الثاني الذي تقدمه كل من الطريقتين إذ تم تصحيح جميع الكلمات الخاطئة المكتشفة من المرحلة السابقة و حساب قيمة الدقة من خلال العلاقة (3):

الجدول (9) : مقارنة بين خوارزمية edit distance و الخوارزمية المقترحة من حيث نسبة النجاح للحل المقترح الأول و الثاني

Proposed	edit distance	
89%	52%	Sol1
9%	16%	Sol2
98%	68%	SR

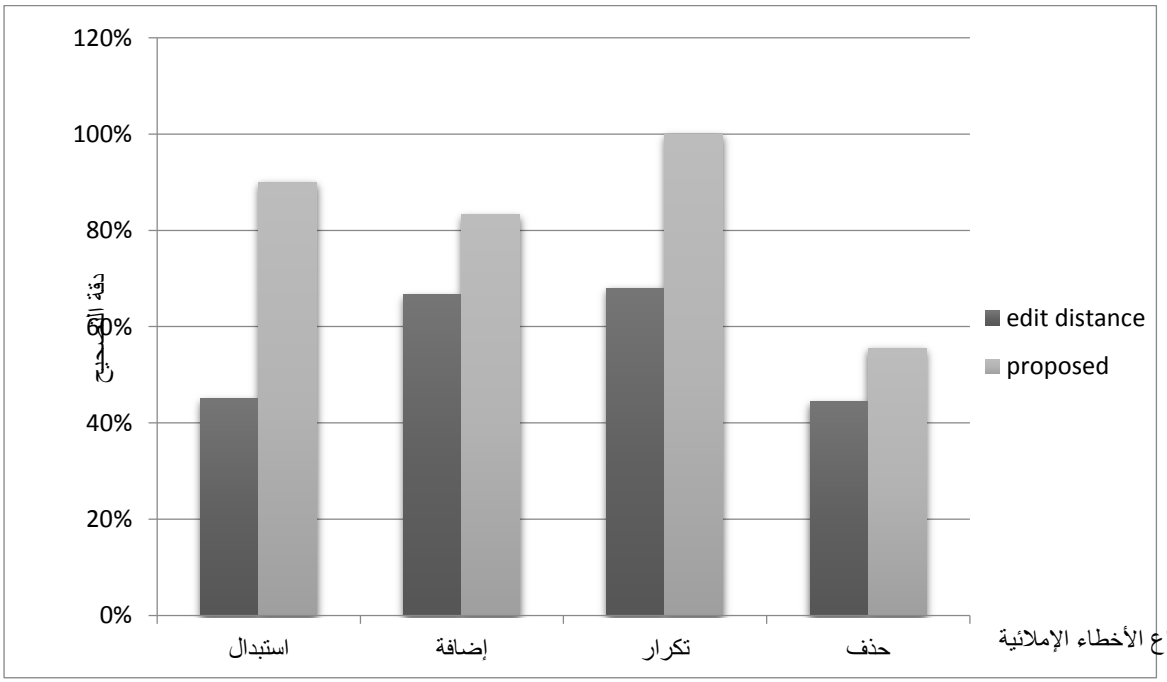
من الجدول السابق نجد أن الطريقة المقترحة قامت بتحسين دقة عملية التصحيح بنسبة 30% .

الشكل (5) مخطط توضيحي يبين المقارنة التي تمت في الجدول (9) بين الطريقة المقترحة و خوارزمية مسافة التحرير التقليدية



الشكل (5) : مقارنة بين خوارزمية edit distance و الخوارزمية المقترحة من حيث نسبة نجاح عملية التصحيح

و بدراسة تفصيلية لأنواع الأخطاء الإملائية الأربعة التي نقوم بمعالجتها (استبدال ، حذف ، إضافة ، تكرار) ، و قدرة الخوارزمية المقترحة على تصحيحها ، تم حساب دقة الطريقة المقترحة في تصحيح كل نوع من الأخطاء السابقة و مقارنتها بخوارزمية مسافة التحرير كانت النتائج موضحة بالشكل(6) :



الشكل (6) : مقارنة بين خوارزمية *edit distance* و الخوارزمية المقترحة بالنسبة لكل نوع من الأخطاء

النتائج :

مما سبق نجد أن :

- ساهمت عملية فرض القيود و القواعد على السمات بتوحيد و تنسيق صيغ و قيم هذه السمات
- حسّن القاموس المستخدم في عملية الكشف عن الأخطاء الإملائية دقة البحث و أظهر نتائج أفضل و يعود ذلك إلى استخدام قاموس كبير الحجم تم تجميعه بحيث يغطي عدداً أكبر من الكلمات و الأسماء العربية و الأعجمية و خاصة الحديثة منها حيث بلغة دقة البحث في القاموس المقترح 99.46%
- الطريقة المقترحة لتصحيح الأخطاء الإملائية قامت بتصحيح الأنواع المختلفة من الأخطاء الإملائية بشكل أكثر كفاءة من خوارزمية مسافة التحرير التقليدية و يعود ذلك إلى إضافة بارامتر تشابه لوحة المفاتيح إلى الخوارزمية حيث حسنت دقة التصحيح بنسبة 30% و بلغت الدقة الإجمالية للطريقة المقترحة 98% ، إذ تم اختبار هذه الطريقة مع أخطاء الكلمة الواحدة فقط و لكن لم يتم اختبارها مع أنواع أخطاء أخرى مثل إضافة أو حذف المسافة البيضاء بين كلمتين و بالتالي لم يتم التأكد من فعاليتها مع هذه الأنواع من الأخطاء الإملائية .

التوصيات المستقبلية :

في هذا البحث تم اقتراح خوارزمية هجينة تعمل على معالجة مجموعة من المشاكل و الأخطاء التي قد تحدث في مجموعات البيانات العربية و في الأبحاث القادمة يمكن أن يتم :

- تطوير الخوارزمية المستخدمة في عملية كشف الأخطاء الإملائية بهدف تقليل زمن العمليات و الحصول على أداء أفضل
- تحسين الخوارزمية المستخدمة في تصحيح الأخطاء الإملائية لتشمل أنواعاً أخرى من الأخطاء الإملائية العربية
- زيادة حجم القاموس المستخدم في عملية الكشف عن الأخطاء الإملائية بحيث يشمل عدداً أكبر من الكلمات
- تحسين النهج بحيث يشمل تصحيح لأنواع أخرى من أخطاء البيانات مثل السجلات المكررة و القيم المفقودة .
- تحسين هذا النهج ليشمل لغات متعددة

المراجع:

[1] Al-Hagery, M. A., Alreshoodi, L. A., Almutairi, M. A., Alsharekh, S. I., & Alkhowaiter, E. S. (2019). A hybrid Technique for Cleaning Missing and Misspelling Arabic Data in Data Warehouse

[2] Hamad, M. M., & Jihad, A. A. (2011, December). An enhanced technique to clean data in the data warehouse. In *2011 Developments in E-systems Engineering* (pp. 306-311). IEEE.

[3]Ghafour, H. H. A., El-Bastawissy, A., & Heggazy, A. F. A. (2011, December). AEDA: Arabic edit distance algorithm Towards a new approach for Arabic name matching. In *The 2011 International Conference on Computer Engineering & Systems* (pp. 307-311). IEEE.

[4] Alotaibi, S. B. (2017, August). ETDC: An efficient technique to cleanse data in the data warehouse. In *Proceedings of the International Conference on Advances in Image Processing* (pp. 135-138).

[5] Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, *49*(3), 721-751.

[6] Al-tarawneh, R., Hamatta, H. S., & Muiadi, H. (2014). Novel approach for Arabic spell-checker: based on radix search tree. *International Journal of Computer Applications*, *95*(7).

[7] H. M. Noaman, S. S. Sarhan, and M. A. A. Rashwan, "Automatic Arabic Spelling Errors Detection and Correction Based on Confusion Matrix- Noisy Channel Hybrid System," *J. Theor. Appl. Inf. Technol.*, vol. 40, no. 2, pp. 54–64, 2016.

[8] Mr. Maher Ibrahim & Dr. S.A.M. Rizvi, A Framework for Building Standard System Data Dictionary, proceedings of the first international conference on emerging technologies and applications in engineering technology and sciences (ICETAETS), volume 3, paper number 3, Computer Science Department, Saurashtra University, 13-14 January,2008. Rajkot, Gujarat, India.

[9] Dr. S.A.M. Rizvi & Mr. Maher Ibrahim, The Chart Representation of the System Data Dictionary, invertis international journal for science and technology, 2008