مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

تطوير نظام مجيب آلي على رسائل الزبائن النصية باللغة العربية باستخدام تقنيات الذكاء الاصطناعي

د. جعفر سلمان*

م. بشار علي علي**

(تاريخ الإيداع 2022/9/29 . قُبِل للنشر في 2022/12/21)

□ ملخّص□

المجيب الآلي هو النظام الذي يُمكن المؤسسات من التعامل مع الاستفسارات وتصنيفها والرد علها بشكل ألي، حيث أن تكرار الإجابة على نفس الاستفسارات يعتبر أمر مجهد ويتطلب وقت كبير ويحتاج إلى توظيف عدة أشخاص للإجابة كلٌ حسب اختصاصه، مما أظهر الحاجة إلى تطوير حلول تستخدم تقنيات الذكاء الاصطناعي للاستعاضة عن الجزء البشري.

يعتمد الحل المقترح على تطوير نظام مجيب آلي يتعامل مع الاستفسارات المكتوبة باللغة العربية ويتكون من ثلاث مراحل وهي:

في البداية يتم استقبال رسائل الزبائن النصية عن طريق البريد الالكتروني وتحليل محتوياتها، ثانياً، يتم تصنيفها باستخدام خوارزميات تصنيف النصوص، ثالثاً، يتم اقتراح الإجابة من خلال مقارنة الرسالة مع الاستفسارات الموجودة في قاعدة البيانات باستخدام تقنيات تشابه النص، والرد عبر الإيميل بالإجابة المناسبة في حال وجود تشابه أو توجيهها إلى الشخص المختص للإجابة عنها وذلك في حال كان الاستفسار غير شائع، ومن ثم يتم تخزين الاستفسار الجديد مع الإجابة في قاعدة البيانات.

وبالنتيجة فإن النظام المقترح سيقوم بتحسين آلية الإجابة من خلال تسريع العملية، وتقليل الوقت والجهد المطلوبين.

الكلمات المفتاحية: المجيب الآلي، التنقيب عن النص، تصنيف النصوص، قياس التشابه، اقتراح الإجابة.

*مدرس في قسم تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس .

**طالب ماجستير في قسم تكنولوجيا المعلومات – كلية هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس .

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(12) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (12) 2022

Developing an automated answering system for customers Arabic text messages using artificial intelligence techniques

Dr.Jaafar Salman*
Eng.Bashar Ali Ali**

(Received 29/9/ 2022 . Accepted 21/12/ 2022)

□ ABSTRACT

Automated answering system is the system that enables institutions to deal with inquiries categorize and respond to them automatically, as repeating answering the same inquiries is considered stressful and requires a lot of time and needs to employ several people to answer each according to his specialization, this demonstrated the need to develop solutions that use artificial intelligence techniques to replace the human part.

Where the proposed solution depends on the development of an automated answering system that deals with inquiries written in Arabic and it consists of three stages:

At first customers' text messages are received by e-mail and their contents analyzed, secondly, they are classified using text classification algorithms, and third the answer is suggested by comparing the message with the inquiries in the database using text similarity techniques, and replying via e-mail with the appropriate answer if there is a similarity or directing it to the competent person to answer it in case the Inquiry is not common, and store it with the answer in the database.

As a result, the proposed system will improve the response mechanism by speeding up the process, and reducing the time and effort required.

Keywords: Automated answering system, data mining, categorization, similarity, Answer Suggestion.

*teacher, Information Technology Engineering Department, Information and Communication Technology Engineering, Tartous University, Syria.
**Student Master, Information Technology Engineering Department, Information and Communication Technology Engineering, Tartous University, Syria.

1- مقدمة:

علم البيانات data science هو عبارة عن علم يجمع بين مجموعة من التخصصات التي ترتبط ارتباط وثيق مع البيانات والتكنولوجيا وتطوير الخوارزميات لحل المشكلات المعقدة بطريقة تحليلية.

وقد تزايد الاهتمام بعلم البيانات في السنوات الاخيرة بشكل ملحوظ من قبل الشركات من جهة ومن قبل الافراد الراغبين بمعرفته وتعلمه من جهة اخرى، إذ يركز بشكل أساسي على معرفة وفهم البيانات التي تمتلكها أي شركة أو مؤسسة ومن ثم استخراج المعرفة منها، ويعتبر من أكثر المجالات شهرة في الوقت الحالي ويهدف علم البيانات إلى معرفة أسباب المشكلة ومعرفة متطلبات نجاح العمل، والاستفادة من تحليل البيانات وتعلم الآلة لحل هذه المشكلة وبالتالي الحصول على تصور للنتائج، بالإضافة إلى تقديم المقترحات والأفكار.

ويعتبر علم البيانات أحد أهم فروع الذكاء الصنعي artificial intelligence وهو عبارة عن أنظمة تُحاكي الذكاء البشري لأداء المهام والتي لديها القدرة على تحسين نفسها باستخدام المعلومات التي تجمعها.

تعتمد أنظمة الإجابة التقليدية على الفهم البشري للرسائل المستلمة، لذلك هناك جزء بشري مسؤول عن جميع مراحل عملية الإجابة في حين يعتبر التنظيم اليدوي لعدد كبير من الاستفسارات أمر صعب للغاية، فهو يستغرق الكثير من الوقت ومكلف وغالباً ما يؤدي إلى عدم رضا صاحب الاستفسار.

تتمثل المشكلة الرئيسية لتلك الأنظمة بالوقت المطلوب لمعالجة الاستفسارات التي تؤثر على وقت الاستجابة وجودة الخدمات بالنسبة لآلية تقديم الاستفسارات بالكامل، وأحياناً يتم الحصول على تصنيف خاطئ أو إرسال الاستفسارات إلى شخص خاطئ، في حين أن وقت الإجابة على السؤال يؤثر مباشرة على رضا المستخدم [11].

لتحسين جودة الخدمة نحتاج إلى تقليل وقت المعالجة إلى الحد الأدنى عن طريق استبدال الأطراف البشرية بأنظمة تلقائية تعتمد تقنيات الذكاء الصنعي كالتصنيف التلقائي واقتراح الإجابات والرد الآلي، وتحسين نوعية وعملية الإجابة عن الاستفسارات.

2- هدف البحث وأهميته:

يهدف البحث إلى تطوير نظام عالي الجودة لإدارة الاستفسارات يتغلب على محدودية الأنظمة التقليدية ويساهم في تقليل عدد الموظفين المطلوبين لإدارة النظام وتقليل الجهد والوقت المطلوبين لمعالجة الاستفسارات، بحيث يحقق النظام ما يلي:

- استخدام تقنيات الذكاء الاصطناعي لإدارة الاستفسارات المكتوبة باللغة العربية.
- تحسين جودة الخدمة وتوفير الزمن من خلال السرعة في الرد على الاستفسارات.
- تقليل الجهود البشرية.
- تحسين نوعية ودقة التصنيف والإجابة عن الاستفسارات.

3- طرق البحث ومواده:

تم إجراء البحث بالاعتماد على الاستفسارات الخاصة بكلية هندسة تكنولوجيا المعلومات والاتصالات في جامعة طرطوس حيث تم جمع الاستفسارات من عدة مصادر منها موقع جامعة طرطوس /تبويب الأسئلة الأكثر تكراراً/ ومن البوت الخاص بالهيئة الطلابية في الجامعة على التيليجرام والذي يجيب على العديد من الاستفسارات المتنوعة ومصادر

أخرى مثل الاستبيانات وكتاب دليل الطالب، حيث قام الباحث بتطوير نظام متكامل باستخدام منصة RapidMiner وهي من أشهر المنصات البرمجية التي تعنى بعلوم البيانات والذكاء الاصطناعي، ولاستقبال الاستفسارات تم إنشاء حساب بريد الكتروني خاص بالكلية، ثم تم بناء نظام تصنيف بالاعتماد على هذه البيانات لتصنيف الرسائل الواردة إلى ثلاث أصناف رئيسية (شؤون الطلاب – الامتحانات – الدراسات العليا).

يتكون نظام المجيب الآلي المقترح من عدة مراحل، في البداية يتم استقبال الاستفسارات الواردة ويتم تطبيق خطوات معالجة النص عليها لتصبح جاهزة للمعالجة ثم يتم تحليل الرسالة وفهم محتوياتها ثم إرسالها إلى المصنف لتصنيفها إلى أحد الأقسام الرئيسية ثم ترسل إلى وحدة اقتراح الإجابة حيث يتم مقارنتها مع الاستفسارات المخزنة في قاعدة البيانات من أجل تحديد الاستفسار المشابه باستخدام خوارزميات قياس التشابه ومن ثم يتم إرسال الإجابة إلى بريد الشخص المرسل أو إرسال الرسالة إلى الشخص المختص للإجابة عنها في حال عدم العثور على استفسار مشابه [1].

3-1 المعالجة المسبقة للنصوص Process Documents:

لاستخدام التنقيب عن النص Text mining نحتاج إلى إعداد بياناتنا لتكون جاهزة لتطبيق تقنيات التنقيب، وذلك بهدف تحويل المستندات النصية العربية إلى نموذج مناسب لخوارزميات استخراج بيانات التصنيف، حيث أن المعالجة المسبقة تتكون من الخطوات التالية:

1) التقطيع Tokenization: هو عملية تقسيم الجمل إلى مجموعة كلمات tokens بالاعتماد على الفراغات بين الكلمات وعلامات الترقيم.

2) إزالة كلمات التوقف Stop word removal: في هذه الخطوة يتم إزالة الكلمات غير الهامة التي لا يؤثر حذفها على عمل المصنف مثل الضمائر (هو، هم، هؤلاء...) وكلمات الاستفهام (ماذا، لماذا...) كلمات غير هامة (بغض النظر، بالرغم، بالإضافة، بالنسبة...) والكلمات العامة والأرقام (الأيام والأشهر...) والعديد من الكلمات الأخرى.

3) التجذير Stemming: هنا يتم استنتاج جذر الكلمات أي ردها إلى أصلها (إزالة البادئات واللواحق) [9][3].

يبين الشكل (1) تطبيق عمليات المعالجة المسبقة للنص باستخدام منصة RapidMiner على الاستفسار التالي: "مرحبا ما الأوراق المطلوبة للحصول على مصدقة تخرج"

الشكل (1) عمليات المعالجة المسبقة للنص

ملاحظة: تحتوي أداة Filter Stopwords على قاموس بكل كلمات التوقف الشائعة في اللغة العربية والتي يمكن الإضافة عليها، مثلاً كلمة "الحصول" في المثال السابق تم إزالتها لأنها من ضمن الكلمات التي قام الباحث بإضافتها إلى قاموس كلمات التوقف نظراً لعدم أهميتها في عملية التصنيف.

وبتطبيق العملية نفسها على جميع الاستفسارات المخزنة في قاعدة البيانات الخاصة بكلية هندسة تكنولوجيا المعلومات والاتصالات المعتمدة في البحث، تم الحصول على قائمة بالكلمات المميزة word List [7]، (277 كلمة)

الجدول (1) يظهر جزء من الكلمات المميزة الناتجة مع تكراراتها:

الجدول (1) جدول الكلمات المميزة وتكرارها

الكلمة المميزة	عدد التكرارات	عدد المستندات الحاوية على الكلمة	عدد التكرارات في كل قسم		
			قسم الامتحانات	قسم شؤون الطلاب	قسم الدراسات العليا
وثق	3	3	0	3	0
وجد	10	10	1	4	5
ورق	9	9	2	5	2
وزع	2	2	2	0	0
وصف	1	1	0	1	0
وصل	6	5	0	2	4
وضع	4	4	1	1	2
وظف	2	2	1	1	0
وفر	1	1	0	0	1
وفق	2	2	0	1	1
وقع	6	6	1	2	3

3-2 حساب الأوزان Weights Account:

تم استخدام معامل Term frequency–Inverse document frequency (TF–IDF)

لحساب وزن الكلمات المميزة وهو مقياس إحصائي يستخدم لتقييم مدى أهمية وجود كلمة في مستند معين، حيث أن الأهمية تزداد نسبيا بزيادة عدد مرات ظهور الكلمة أو المصطلح في المستند، وتنقص مع زيادة تكرار الكلمة في المستندات، أي مع زيادة Document Frequency [3] [12]، ويعطى بالعلاقة:

$$W_{x,y} = tf_{x,y} * \log\left(\frac{N}{df_x}\right) \qquad (1)$$

حيث:

$W_{x،y}$: وزن المصطلح x في المستند y

$tf_{x،y}$: تردد المصطلح x في المستند y

N : عدد المستندات الكلي

df_x : عدد المستندات التي تحتوي المصطلح x

الجدول (2) يوضح الكلمات المميزة (الميزات) المستخلصة من الاستفسار السابق مع أوزانها:

الجدول (2) أوزان الميزات

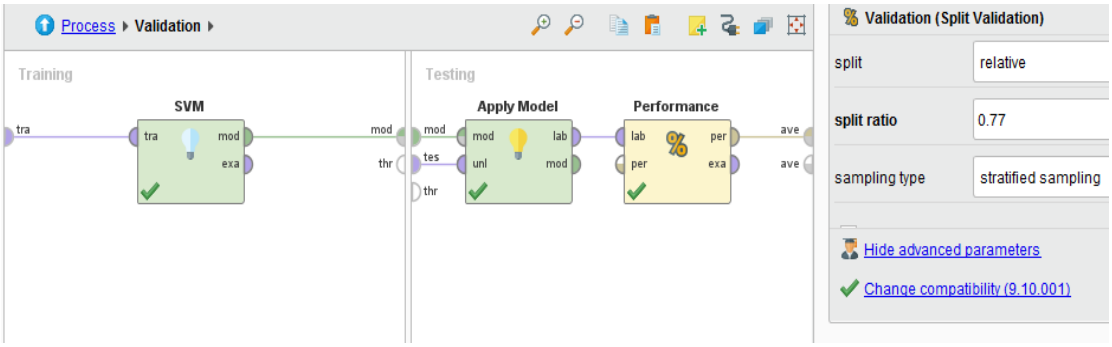
ورق	طلب	صدق	خرج	
0.483	0.208	0.631	0.472	" مرحبا ما الأوراق المطلوبة للحصول على مصدقة تخرج"

3-3 التصنيف CLASSIFICATION:

هي المهمة التي يتم فيها تصنيف النصوص (الاستفسارات) إلى فئات محددة مسبقًا بناءً على محتوياتها[5]، وقد استخدمنا في عملية التصنيف خوارزميات التصنيف التالية وقمنا بحساب الصحة Accuracy من أجل كل خوارزمية:

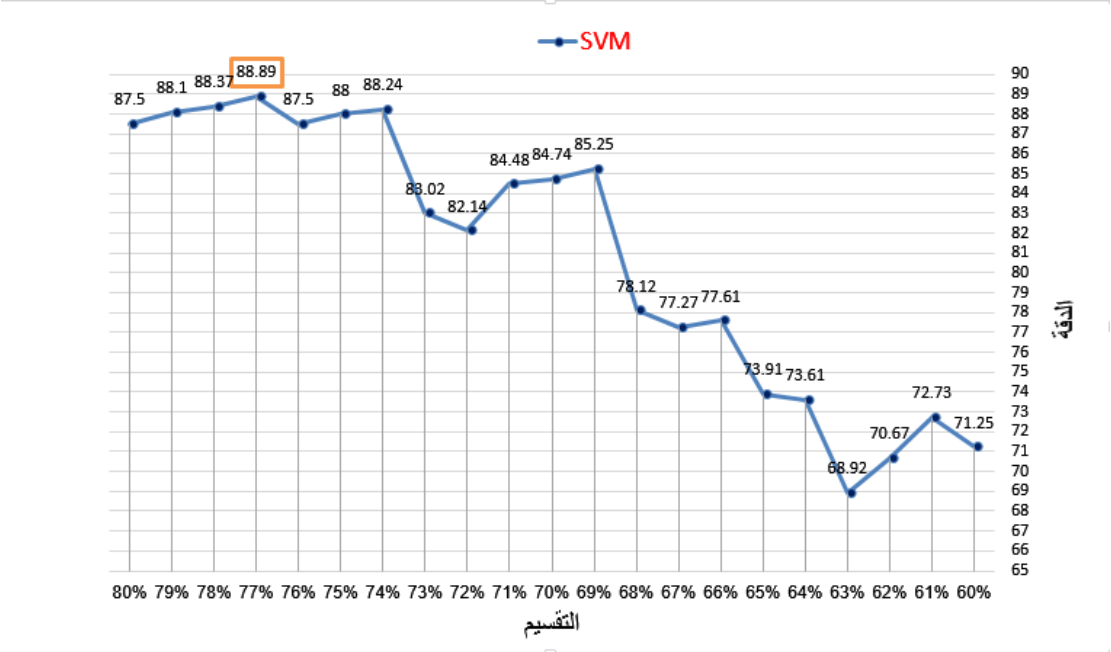
- Naïve Bayes (NB)
- K-Nearest Neighbors (KNN)
- Decision Tree Algorithm (DT)
- Support vector machines (SVM)

الشكل (2) يوضح تطبيق خوارزمية التصنيف SVM على الاستفسارات المخزنة في قاعدة البيانات والتي تنتمي إلى ثلاث أقسام رئيسية (شؤون الطلاب والدراسات العليا والامتحانات) باستخدام منصة RapidMiner بعد اعتماد نسبة تقسيم (0.77) أي 77% من الاستفسارات تم استخدامها للتدريب و23% لاختبار الخوارزمية [10]:



الشكل (2) تطبيق خوارزمية التصنيف SVM باستخدام منصة RapidMiner

حيث تم اعتماد نسبة التقسيم بناءً على المخطط البياني شكل (3) الذي يوضح العلاقة بين دقة التصنيف ومعدل التقسيم بالنسبة لخوارزمية svm حيث حققت أعلى دقة تصنيف وهي 88.89% عند معدل تقسيم بيانات 77%:



الشكل (3) العلاقة بين دقة التصنيف ومعدل التقسيم بالنسبة لخوارزمية svm

ونتيجة تطبيق الخوارزمية تم تصنيف الاستفسارات بدقة تصل إلى 88.89% مبينة بالجدول (3):

الجدول (3) حساب الصحة باستخدام خوارزمية svm

accuracy: 88.89%

	true دراسات عليا	true شؤون الطلاب	true الامتحانات	class precision
pred. دراسات عليا	15	2	0	88.24%
pred. شؤون الطلاب	2	14	0	87.50%
pred. الامتحانات	0	1	11	91.67%
class recall	88.24%	82.35%	100.00%	

حيث أن عدد الاستفسارات التي تم استخدامها للاختبار هي 45 استفسار (17 استفسار من قسم الدراسات العليا و17 استفسار من قسم شؤون الطلاب و11 استفسار من قسم الامتحانات).

بالنسبة لقسم الدراسات العليا تم تصنيف 15 استفسار بشكل صحيح واستفسارين بشكل خاطئ، وبالنسبة لقسم شؤون الطلاب تم تصنيف 14 استفسار بشكل صحيح وثلاثة استفسارات بشكل خاطئ، أما بالنسبة لقسم الامتحانات تم تصنيف جميع الاستفسارات (11 استفسار) بشكل صحيح، أي بالمجمل تم تصنيف 40 استفسار بشكل صحيح من أصل 45 استفسار، بالتالي فإن صحة خوارزمية التصنيف يعبر عنها بعدد الاستفسارات التي تم تصنيفها بشكل صحيح على عدد الاستفسارات الكلية المصنفة أي: Accuracy=40/45=88.89%

الجدول (4) يوضح دقة التصنيف التي حصلنا عليها من تطبيق خوارزميات التصنيف التالية:

Naïve Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree Algorithm (DT), and Support vector machines (SVM).

وذلك بعد اختيار أفضل معدل تقسيم split ratio لكل خوارزمية:

الجدول (4) مقارنة دقة التصنيف الناتجة بالنسبة للخوارزميات الأربعة

	Accuracy	Split Ratio
SVM	88.89%	77%
Naïve Bays	73%	75%
KNN	86%	75%
Decision Tree	79.71%	65%

ونتيجة ذلك نلاحظ أن خوارزمية SVM أعطت أفضل دقة ممكنة عند معدل تقسيم 77% لذلك قمنا باعتمادها في بناء المصنف من أجل تصنيف الاستفسارات الجديدة الواردة.

الجدول (5) يوضح نتيجة تصنيف بعض الاستفسارات الواردة حيث تم تصنيفها بشكل صحيح:

الجدول (5) تصنيف الاستفسارات

text	prediction(القسم)	From	To
أين يمكنني إيجاد نتائج موادي	الامتحانات	Zain Ali <zainaliali.7654321@gmail.com>	Tartous University <tartousuniversity1
ما الأوراق المطلوبة للحصول على مصدقة تخرج	شؤون الطلاب	Zain Ali <zainaliali.7654321@gmail.com>	Tartous University <tartousuniversity1
متى يمكننا الاستفادة من علامات المساعدة الامتحانية	الامتحانات	Zain Ali <zainaliali.7654321@gmail.com>	Tartous University <tartousuniversity1
ما الاختصاصات الموجودة في الجامعة	شؤون الطلاب	bashar ali <basharaliali503@gmail.com>	"tartousuniversity1@outlook.com" <tar

3-4 قياس التشابه Similarity:

في هذه الخطوة بعد أن يقوم المصنف بتصنيف الاستفسار الجديد إلى صنف معين يتم توظيف خوارزميات التشابه لمقارنة الاستفسار الجديد مع جميع الاستفسارات المخزنة في قاعدة البيانات التابعة لذلك الصنف، واختيار الاستفسار الأكثر شبهاً، حيث تم في هذا البحث اعتماد طريقتين لقياس التشابه وهما:

- ❑ خوارزمية Levenshtein لقياس التشابه.
- ❑ خوارزميّة تشابه-جيب التّمام Cosine Similarity.

3-4-1 خوارزمية Levenshtein:

تحسب هذه الخوارزمية المسافة Distance بين نصين، حيث تُقاس هذه المسافة بعدد التغييرات المطلوب إجراؤها على النص الأول ليصبح مطابقاً للنص الثاني، هذا التغيير يحدث بتبديل حرف مكان حرف أو حذف حرف أو

إضافة حرف، فإذا كانت المسافة بين النصين صفراً كان هذا معناه أنهما متطابقان وإذا كانت (1) فهذا يعني أن أحدهما يختلف عن الآخر بحرف (زيادة أو نقصان أو تغيير) [8]، وتعطى بالعلاقة:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \quad (2)$$

تستخدم هذه الخوارزمية القواعد التالية:

1. ترقيم أول حرف في النص هو 1.
2. أقصر مسافة ممكنة هي 0 (النصان متماثلان).
3. أطول مسافة ممكنة هي عدد حروف أطول نص من النصين (في حال كانت كل الحروف مختلفة أو أحد النصين فارغ).
4. يتم حساب المسافة بين الحرف رقم I في النص الأول والحرف رقم J في النص الثاني كالتالي:

- إذا كان I يساوي 0 فالمسافة هي J وإذا كان J يساوي 0 فالمسافة هي I وإذا كان كل منهما 0 فالمسافة هي 0.
- عندما تكون قيمة I و J أكبر من الصفر يتم حساب المسافة بأخذ التكلفة الأقل بين التكاليف الثلاثة (تكلفة حذف حرف، تكلفة إضافة حرف، تكلفة استبدال حرف).

وبتطبيق هذه الخوارزمية لحساب المسافة للاستفسار التالي: "كيف يمكن الاستفادة من علامات المساعدة الامتحانية" حيث تتم مقارنته مع جميع الاستفسارات في قاعدة البيانات التابعة لنفس الصنف (الامتحانات) وحساب المسافة كما يظهر في الجدول (6):

الجدول (6) حساب المسافة باستخدام خوارزمية Levenshtein

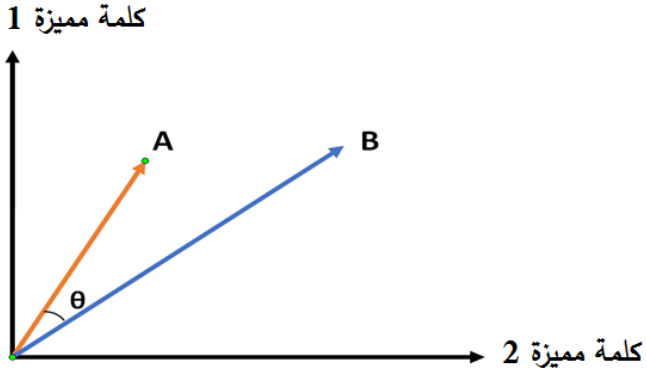
Row No.	استفسار جديد	استفسار	الإجابة	القسم	distance
1	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	متى يمكن الاستفادة من علامات المساعدة الامتحانية	يحصل الطالب على مساعدة امتحانية بم...	الامتحانات	3
2	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	ما قيمة علامات المساعدة الامتحانية	يحصل الطالب على مساعدة امتحانية بم...	الامتحانات	18
3	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	أين يصدر توزيع الطلاب على القاعات الامتحانية	يتم نشر توزيع الطلاب على القاعات الا...	الامتحانات	24
4	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	كيفة الحصول على النتائج الامتحانية	سيتم نشر النتائج الامتحانية في حال ص...	الامتحانات	24
5	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	كيف يمكنني تقديم طلب إعادة القسم العملي	آلية تقديم طلب #اعادة_القسم_العملي لم...	الامتحانات	28
6	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	كيف يتم حساب علامة العملي	علامة العملي هي مجموع علامة امتحا...	الامتحانات	29
7	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	متى يتم بدء تقديم طلبات إعادة العملي	يتم بدء تقديم طلبات إعادة العملي خلال ...	الامتحانات	29
8	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	هل يمكن تقديم طلب إلغاء إعادة العملي	يمكن تقديم طلب إلغاء إعادة العملي خلا...	الامتحانات	29
9	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	ما قيمة علامة العملي من العلامة الكلية	قيمة علامة العملي هي 40 درجة	الامتحانات	30
10	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	لم أجد اسمي في توزيع الطلاب على القاعات الامتحانية	يمكنك مراجعة رئيس الدائرة في حال ...	الامتحانات	30
11	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	كيف يمكن حساب تقدير النجاح بالأعتماد على المعدل	تمنح الإجازة من المراتب التالية:	الامتحانات	30
12	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	ما علامة العملي من العلامة الكلية	قيمة علامة العملي هي 40 درجة وقد ت...	الامتحانات	31
13	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	أين يمكنني إيجاد نتائج امتحانات السنة الأولى	سيتم نشر النتائج الامتحانية في حال ص...	الامتحانات	31
14	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	هل يمكن تقديم مواد من السنين الأعلى	لايمكن تقديم مواد من السنين الأعلى في ...	الامتحانات	31
15	كيف يمكن الاستفادة من علامات المساعدة الامتحانية	متى يمكن الحصول على مصدقة تخرج	يحصل الطالب على إشعار تخرج عند ان...	الامتحانات	32

حيث تم حساب المسافة بين الاستفسار الجديد وجميع الاستفسارات المخزنة في قاعدة البيانات التابعة لقسم الامتحانات وترتيبها تصاعدياً أي يظهر الاستفسار ذو المسافة الأقل في الأعلى (الاستفسار الأقرب إلى الاستفسار الجديد).

3-4-2 خوارزميّة تشابه-جيب التّمام Cosine Similarity:

تشابه جيب التّمام هو عبارة عن مقياس نستطيع من خلاله أن نحدد مدى تطابق المستندات بغض النظر عن حجمها بعد تحويلها إلى متجهات Vectors، ورياضياً هو عبارة عن مقياس جيب تمام الزاوية بين متجهين في فضاء ثنائي الأبعاد، ويمكن أن يكون هذان المتجهان عبارة عن بيانات رقمية أو نصية، حيث يتطلب حساب تشابه جيب التمام تحويل المستندات النصية إلى متجهات في الفضاء [2]، وذلك من خلال إجراء عمليات المعالجة الأولية للمستندات من ترميز وتجذير وتلخيص ومن ثم استنتاج الكلمات المميزة.

يبين الشكل (4) توضيح بسيط لتمثيل مستندين A و B على شكل متجهين في الفضاء بالاعتماد على الكلمات المميزة (هنا كلمتين مميزتين فقط يتم التعبير عنهما باستخدام محوري الإحداثيات) حيث أن شكل المتجه وميله يعتمد على احتواء المستند على الكلمة المميزة أو عدمه وعلى عدد مرات ظهور الكلمة في المستند.



الشكل(4) تمثيل المستندات على شكل متجهات

من خلال العلاقة التالية:B وAويمكن حساب تشابه جيب التمام بين متجهين

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

وتكون قيمة التشابه similarity بين مستندين ضمن المجال [0,1]، حيث أن القيمة (1) تدل على أن المستندين متطابقين [4].

يبين الجدول (7) حساب التشابه باستخدام خوارزميّة تشابه جيب التّمام بالنسبة للاستفسار السابق:

"كيف يمكن الاستفادة من علامات المساعدة الامتحانية"

الجدول (7) حساب التشابه باستخدام خوارزمية جيب التمام

Row No.	Request_ID	Document_ID	Similarity
1	1	155	0.800
2	1	154	0.783
3	1	166	0.544
4	1	182	0.475
5	1	171	0.235
6	1	158	0.194
7	1	162	0.192
8	1	177	0.170
9	1	156	0.167
10	1	153	0.163
11	1	185	0.162
12	1	167	0.154
13	1	160	0.154
14	1	175	0.138
15	1	163	0.137

حيث تم مقارنة الاستفسار الجديد Request مع جميع الاستفسارات Documents في قاعدة البيانات والتي تنتمي إلى قسم الامتحانات وحساب التشابه Similarity لكل منها وترتيبها تنازليا حيث يظهر أن الاستفسار 155 هو الاستفسار الأقرب، حيث تبلغ نسبة التشابه 0.8 أي 80%.

3-5 اقتراح الإجابة Answer Suggestion:

في هذه المرحلة يتم اقتراح الإجابة بعد مقارنة الاستفسار الجديد مع جميع الاستفسارات المخزنة في قاعدة البيانات التابعة لنفس صنف الاستفسار الجديد وتحديد الاستفسار الأقرب ومن ثم إرسال الإجابة الخاصة بالاستفسار المشابه إلى بريد الشخص المرسل، ويتم ذلك باستخدام منصة RapidMiner من خلال المعامل Branch الذي يختبر الشرط if (Similarity)>0.75

حيث تبين بالتجريب أن غالبية الاستفسارات التي تحقق نسبة التشابه (0.75) تكون الإجابة الخاصة بها منطقية ويمكن الاعتماد عليها في عملية الرد على الاستفسار، لكن يمكن زيادة النسبة قليلا من أجل زيادة عملية الدقة في الرد.

وبالتالي يقوم بتنفيذ إحدى الحالتين:

- ❖ في حال تحقق الشرط يقوم بإرسال رسالة تحتوي الإجابة عبر البريد الإلكتروني إلى عنوان الشخص المرسل (مع احتمال إرسال أكثر من إجابة)
- ❖ في حال لم يتحقق الشرط (عدم العثور على استفسار مشابه) يقوم بإرسال الاستفسار إلى بريد الشخص المختص (حسب نتيجة تصنيف الرسالة) الذي يقوم بدوره بالإجابة على الاستفسار وتخزينه مع الإجابة ضمن قاعدة البيانات.

4- النتائج:

تم من خلال هذا البحث تطوير نظام مجيب آلي متكامل يعتمد في عمله على تقنيات الذكاء الاصطناعي يقوم باستقبال الاستفسارات (رسائل بريد الكتروني) ويجري عليها عمليات معالجة النصوص الأساسية (الترميز، التجذير، إزالة كلمات التوقف) واختصارها لتصبح ملائمة لعمل المصنف، ثم يقوم المصنف بتصنيفها إلى أحد الأقسام التالية (شؤون الطلاب، الدراسات العليا، الامتحانات)، ثم يقوم النظام بمقارنة الاستفسار الجديد مع جميع الاستفسارات في قاعدة البيانات التي تنتمي إلى نفس القسم لاختيار الاستفسار الأقرب وذلك باستخدام خوارزميات التشابه، ومن ثم يقوم الجزء الخاص باقتراح الإجابة باسترجاع الإجابة الخاصة بأقرب استفسار وإرسالها إلى بريد الشخص المرسل في حال تحقق الشرط Similarity>0.75 وإلا فيتم إرسالها إلى الشخص المختص ليقوم بالإجابة عليها لمرة واحدة وتخزين الاستفسار الجديد مع الإجابة في قاعدة البيانات للاستفادة منها في حال ورود استفسار جديد مشابه.

وبذلك يكون النظام قادر على تطوير وبناء نفسه من خلال الاستفسارات الواردة.

بالنسبة لمرحلة التصنيف تم بناء عدة مصنفات باستخدام أشهر الخوارزميات المستخدمة في التصنيف وهي:

Naïve Bayes (NB), K–Nearest Neighbors (KNN), Decision Tree Algorithm (DT), and Support vector machines (SVM)

من أجل تقيم عمل Accuracy, Precision, F–measure Recall وحساب مقاييس الأداء التالية ، والتي تعطى بالعلاقات التالية:[6]خوارزميات التصنيف

$$Accuracy = \frac{Number\ f\ correct\ predictions}{Total\ Number\ of\ predicions\ made} \quad (4)$$

$$Precision = \frac{TP(True\ Positive)}{TP(True\ Positive) + FP(False\ Positive)} \quad (5)$$

$$Recall = \frac{TP(True\ Positive)}{TP(True\ Positive) + FN(False\ Negative)} \quad (6)$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

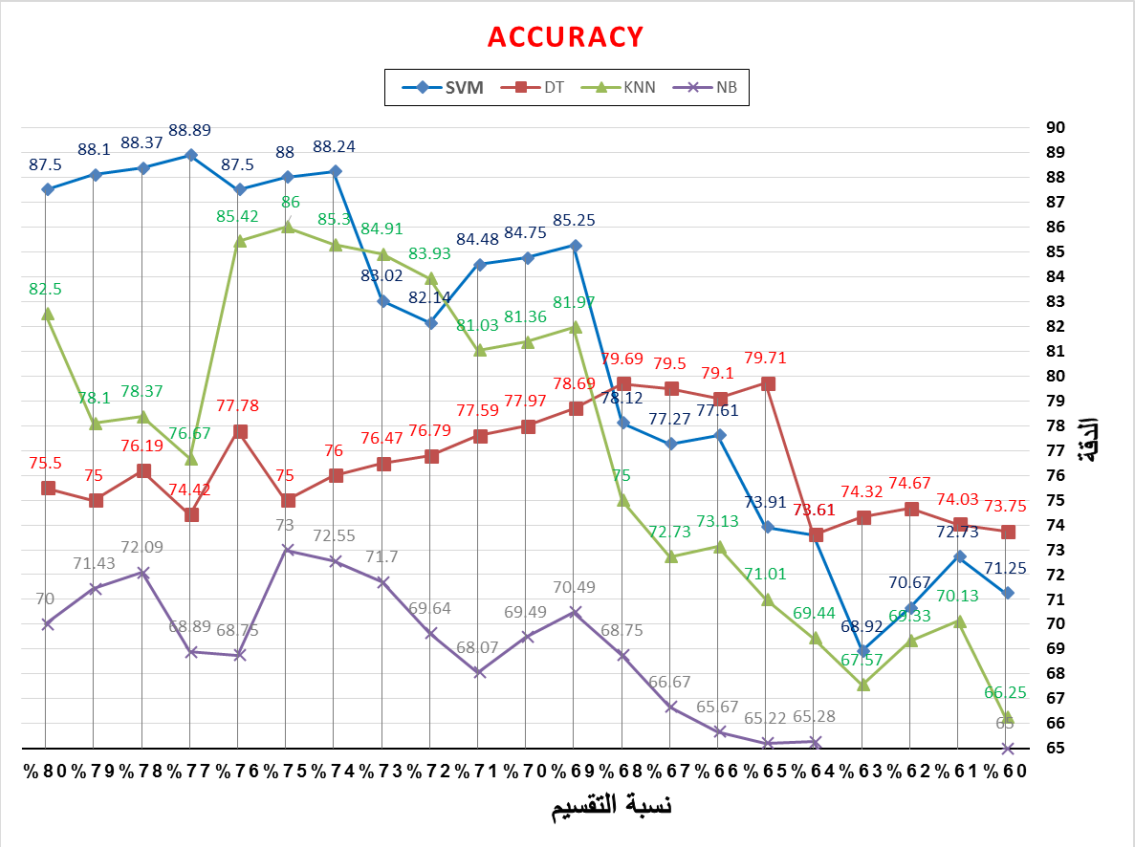
يوضح الجدول (8) مقارنة بين خوارزميات التصنيف باستعمال مقاييس الأداء:

الجدول (8) مقارنة بين خوارزميات التصنيف

Algorithm	Split Ratio	Accuracy	Precision	Recall	F-measure
SVM	0.77	88.89%	91.19%	89.13%	90.14%
DT	0.65	79.71%	77.3%	79.41%	78.43%
KNN	0.75	86%	86.69%	85.42%	86.05%
NB	0.75	73%	72.66%	70.77%	70.70%

من خلال الجدول السابق نلاحظ أن خوارزمية SVM حققت أعلى دقة تصنيف.

حيث تم اختيار معدل التقسيم لكل خوارزمية بالاعتماد على المخطط شكل (5) الذي يظهر العلاقة بين دقة التصنيف ومعدل التقسيم بالنسبة لخوارزميات التصنيف التي تم اختبارها:



الشكل (5) العلاقة بين دقة التصنيف ومعدل التقسيم بالنسبة لخوارزميات التصنيف التي تم اختبارها

بالنسبة لقياس التشابه استخدمنا طريقتين من أكثر الطرق استخداما في قياس التشابه وحساب المسافة وهما خوارزمية levenshtien وتشابه جيب التمام واستنتجنا بالتجارب أن خوارزمية تشابه جيب التمام أعطت نتائج أفضل والمثال التالي يوضح ذلك:

من أجل حساب المسافة بين الاستفسارين:

A="مرحبا هل يمكن الاستفادة من علامات المساعدة الامتحانية في حال الرسوب في مادة واحدة وماهي قيمتها"

B= "ما قيمة علامات المساعدة الامتحانية"

باستخدام تشابه جيب التمام علينا في البداية تقسيم النص إلى رموز باستخدام tokenize ثم إزالة الكلمات غير الضرورية باستخدام Stop word removal ورد الكلمات إلى أصلها باستخدام Stemming وبالتالي نحصل على قائمة الميزات.

الجدول (9) يوضح الميزات المستنتجة واستخدامها في حساب المتجهات:

الجدول (9) الميزات المستنتجة وحساب المتجهات

الميزات	علم	ساعد	امتحن	فيد	رسب	قيم
A	1	1	1	1	1	1
B	1	1	1	0	0	1

ونتيجة ذلك نحصل على المتجهين:

$\vec{A}$= (1, 1, 1, 1, 1, 1)

$\vec{B}$= (1, 0, 0, 1, 1, 1)

وبتطبيق العلاقة (3) نجد: similarty ≈0.8

بينما لو تم حساب المسافة باستخدام خوارزمية Levenshtein Similarity لحصلنا على distance=62 وهو رقم كبير بالرغم من تشابه النصين لذلك نلاحظ أن تشابه جيب التمام يقدم أداء أفضل في حساب التشابه، أما خوارزمية Levenshtein فهي جيدة في حال النصوص القصيرة المتقاربة.

وبالتالي فإن النظام المقترح يوفر الزمن من خلال الرد على الاستفسارات بسرعة كبيرة حيث أن الوقت اللازم لاستلام الرسائل ومعالجتها والرد عليها لا يتجاوز 20 ثانية، ويوفر الجهود البشرية من خلال قدرته على التعامل مع الاستفسارات بدون أي تدخل بشري إلّا في حال ورود استفسار جديد لا يحقق شرط التشابه Similarity>0.75 فإنه يقوم بإرساله للشخص المختص الذي يقوم بالإجابة عنه لمرة واحدة فقط.

5- المراجع:

[1] Das, D, J., & Kumar, S, A., & Reddy, B, R., & Prakash, S, S. (2011), *Web Data Refining*
Using Feedback Mechanism and k-mean Clustering. Journal of computing, 3(5), 2151-9617.

[2] Al-Anzi, F, S., & AbuZeina, D. (2017). *Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing*. Journal of King Saud University-Computer and Information Sciences, 29(2), 189-195.

[3] Saleh, S., & Shaheen, M., & Saqer, Z. (2013*), Arabic Document Classifier*.

[4] Alian, M., & Awajan, A. (2018, November). *Arabic semantic similarity approaches-review*. In 2018 International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE.

[5] Mesleh, A. (2007), *Chi Square Feature Extraction Based SVM Arabic Language Text Categorization System*. Journal of Computer Science, 3(6), 430-435.

[6] Song, M., & Brook, Y. (2009), *Handbook of research on text and web mining technologies, information science reference*, Volume I,329-345.

[7] Sobh, I., & Darwish, N., & Fayek. M. (2008), *Evaluation Approaches for an Arabic Extractive Generic Text Summarization System*, the Departement of Computer Engineering, Cairo University, Giza, Egypt.

[8] Black, P, E.(2008), *Levenshtein distance*, Dictionary of Algorithms and Data Structures
[online], U.S. National Institute of Standards and Technology.

[9] Mustafa, M., & Eldeen, A. S., & Bani-Ahmad, S., & Elfaki, A. O. (2017). *A comparative survey on Arabic stemming: approaches and challenges*. Intelligent Information Management, 9(02), 39.

[10] Scott. R. (2002). *Automatic text categorization applied to email*. NAVAL POSTGRADUATE SCHOOL Monterey, California.

[11] Alnajjar, M., & Abu-Naser, S, S. (*2015, May*). *Improving Quality of Feedback Mechanism in UN by Using Data Mining Techniques*. International Journal of Soft Computing, Mathematics and Control (IJSCMC), 7(2).

[12] Hmeidi, I., & Hawashin, B., & El-Qawasmeh, E. (2008). *Performance of KNN and SVM classifiers on full word Arabic articles*. Advanced Engineering Informatics, 22(1), 106-111.