المجلة الإلكترونية الشاملة متعددة التخصصات
العدد التاسع شهر (٣) \2019
www.eimj.org

التنقيب عن النصوص وتطبيقاته

الباحثة : نهى عبدالكريم النزاوي
كلية الجبيل الجامعية

المجلة الإلكترونية الشاملة متعددة التخصصات
العدد التاسع شهر (٣) \2019
www.eimj.org

ملخص:

تزايد الاهتمام في السنوات الأخيرة بعلم التنقيب عن المعلومات المهمة في النصوص الحية ويرجع ذلك الى التزايد المستمر في كمية النصوص التي تحتوي على كمية هائلة من المعلومات المهمة والتي تنشر بشكل مستمر في مختلف الكتب والمجلات بالإضافة إلى مواقع التواصل الاجتماعي المختلفة مثل المدونات وفيس بوك وتويتر الخ. فأصبح من الصعب متابعة ومعرفة كل المعلومات الخاصة بمجال معين وذلك لان هذه المعلومات موجودة بشكل غير منظم بين ملايين السطور من النصوص.

التنقيب عن النصوص هو العلم الخاص باستخراج معلومات مفيدة من النصوص الحية وتصنيفها حسب نوعها وتخزينها بشكل منظم داخل قواعد بيانات بحيث يسهل الوصول اليها واستخدامها لاحقا في تطبيقات أخرى.

تهدف هذه الورقة العلمية الى عرض ملخص عن مراحل التنقيب عن المعلومات المفيدة في النصوص الحية وأهم التقنيات التي تم استخدامها في هذا المجال بالإضافة الى اهم التطبيقات التي تم فيها استخدام التنقيب عن النصوص. فبالرغم من وجود العديد من الأوراق العلمية التي تلخص آخر التطورات والتقنيات المستخدمة في التنقيب عن المعلومات المفيدة واستخراجها وتصنيفها من النصوص الحية. لاتزال المكتبة العربية تفتقر الى الأبحاث العلمية في هذا المجال.

الكلمات الجوهرية. التنقيب عن البيانات، التنقيب عن النصوص، استرجاع المستندات، استرجاع المعلومات، تطبيقات التنقيب عن النصوص.

Abstract:

Over recent years, text mining has received a lot of attention due to the unprecedented increase in information, which is available textual sources such as books and magazines and through social media channels i.e. blogs, Twitter, Facebook etc. Text mining is the process of extracting and classifying useful information from unstructured text and storing the extracted information in a more organized form, in databases, in order to use it later in useful applications. This paper provides a brief overview of the current state of text mining and will place great emphasis on the latest techniques and methods which are used in this domain. It will also shed light on the most important applications of text mining.

Although there are many research papers that summaries the latest updates on the technologies and methods which are used for text-mining tasks, research papers in the Arabic language are still limited.

Keywords: data mining, text mining, document retrieval, information retrieval, text-mining applications.

١. مقدمة:

إن التطور السريع الذي يشهده العالم اليوم والتضخم الكبير في حجم المعلومات يتطلب وسائل فعاله وسريعة في تحليل النصوص واستخراج المعلومات المهمة وإدارتها وتخزينها بطريقه منظمة بحيث يسهل الوصول اليها وإيجاد الترابط فيما بينها مما يساعد لاحقا في استخدامها في أبحاث وتطبيقات لإثبات فرضيات محدده مسبقا او اكتشاف معلومة جديدة.

التنقيب عن البيانات هو استخراج العلاقات واستكشاف الأنماط فيما بين البيانات المنظمة في قواعد البيانات بينما التنقيب عن النصوص هو عبارة عن تحليل النصوص الغير مهيكله مثل الايميلات والمقالات العلمية في الكتب وفي مواقع الانترنت المختلفة والتي عادة تحتوي على معلومات كثيره غير معروفة مسبقا وتحويلها الى بيانات مهيكلة ومنظمة وذلك لاكتشاف الانماط اللغويه المستخدمة والعلاقات فيما بين المفردات ومن ثم استخدامها لاحقا في استخراج معلومات مفيدة ومهمه.

[جوبتا ومن معه،٢٠٠٩] [سكانيا وبيرونثا ومن معه،٢٠١٢].

ينقسم التنقيب عن النصوص لمهمتين رئيسيتين هما : استرجاع المعلومات وفيها يتم تحديد المستندات التي تحتوي على المعلومات المرغوب استخراجها و استخراج المعلومات وفيها يتم استخراج وتصنيف المعلومات المطلوب استخراجها من هذه المستندات.

٢. استرجاع المعلومات:

شهد مجال استرجاع المعلومات تطوراً ملحوظاً في العقود الأخيرة وذلك نظرا للتزايد الضخم في محتوى الشبكة العنكبوتية وحاجة المستخدمين الماسة للبحث في هذا الكم الهائل من المعلومات [السلمان والربيعه،٢٠١٠][الهليس،٢٠١٣].

ومن هنا نشأت الحاجة الملحة لإيجاد مجموعة المستندات التي يحتاجها المستخدم والتي يتناسب محتواها مع حاجة المستخدم وذلك لقراءتها والبحث فيها بدلا من البحث في مئات المستندات والتي قد لايكون لها علاقة في استعلام المستخدم.

من أهم التطبيقات لنظم استرجاع المعلومات هي محركات البحث مثل قوقل ، ياهو، سفاري الخ.

ويتكون نظم استرجاع المعلومات من ثلاث مراحل أساسية وهي كالتالي:

١- الفهرسة وفيها تتم فهرسة المستندات الموجودة باستخدام كلمات وعبارات تمثل هذه المستندات وتدل عليها.
٢- إعادة تكوين الاستعلام وفيه تتم مراجعة واعادة صياغة الاستعلام الذي يدخله المستخدم بحيث يتوافق مع الكلمات والعبارات التي تم استخدامها في عملية الفهرسة.
٣- المطابقة وفيه تتم مطابقة الاستعلام الذي أدخله المستخدم مع الفهرس الموجود ومن ثم استرجاع جميع المستندات التي لها علاقة باستعلام المستخدم وترتيبها تنازليا حسب درجة توافقها مع استعلام المستخدم.

وبعد ذلك يتم تحديد مجموعة المستندات التي تحتوي على المعلومات المرغوب استخراجها والتي تشكل ما يسمى بالمدونة التي تستخدم كمصدر للنصوص التي سيتم تطبيق تقنيات التنقيب عن النصوص عليها [سيمبسون ودينا ،٢٠١٢].

٢.١ ترميز تكويد المدونة:

وبعدما يتم تحديد المدونة والتي هي عبارة عن مجموعة المستندات التي تحتوي على المعلومات المهمة وقبل استخدامها في مرحلة استخراج المعلومات عادة تخضع هذه الوثائق الى عملية إضافة شرح للأسماء المهمة والتي تكون ذات صلة بالمهمة المراد تنفيذها من جمع هذه الوثائق. تتم إضافة المعلومات والشروحات للأسماء يدويا من قبل خبراء بالمجال. تمثل المعلومات التي يتم اضافتها الى الوثائق مستويات مختلفة من التحليل اللغوي مثل القواعد النحوية، الدلالات اللفظية ومعلومات أجزاء الكلام [سيمبسون ودينا،٢٠١٢].

هناك ثلاث طرق رئيسية لإضافة المعلومات والشروحات الى الوثائق وذلك لاستخدامها لاحقا لأغراض تدريب واختبار أنظمة استخراج المعلومات:

١- إضافة الشروحات يدويا من قبل خبراء بالمجال.
٢- إضافة الشروحات الى الوثائق باستخدام أدوات خاصة ومن ثم مراجعتها يدويا من قبل خبراء في المجال.
٣- إضافة الشروحات للوثائق بمساعدة مصادر معرفة معينة مثل الأنطولوجي بحيث يتم إضافة الشروحات فقط للأسماء والعلاقات المحددة في الأنطولوجي.

لكل طريقة من الطرق السابقة مميزات وعيوب فعلى سبيل المثال نتائج إضافة الشروحات الى الوثائق باستخدام أدوات خاصة يعاني من مشكلة الانحراف الى حقائق معينة ومعروفة.
كذلك الطريقة المعتمدة على الانطولوجي في إضافة الشروحات تعاني من انحراف الى حقائق معينة ومعروفة سابقا والتي تكون موجودة سابقا في الانطولوجي. بينما إضافة الشروحات الى المستندات باستخدام أكثر من خبير بالمجال يساهم في التغلب على مشكلة الانحراف الى حقائق معينة والتي تعاني منها طريقة الطريقة المعتمدة على الانطولوجي والطريقة التي تستخدم أدوات خاصة بإضافة الشروحات.

بعد تحديد المستندات التي يحتاجها المستخدم يأتي دور مرحلة استخراج المعلومات والتي فيها يتم تحليل النصوص ومن ثم استخراج المعلومات المهمة منها.

٣. استخراج المعلومات:

عملية استخراج المعلومات هي عباره عن استخراج واستكشاف كل المفاهيم المحددة مسبقا من قبل المستخدم والتي تخدم احتياجات المستخدم في تنفيذ مهمه معينه في مجال معين بمعنى آخر استخراج المعلومات هو عباره عن استخراج معلومات مهيكله ومنظمة من نص غير مهيكل.

استخراج المعلومات يشكل التقنية الأساسية لتطبيقات التنقيب عن النصوص الأكثر تعقيدا مثل ترجمة الاله ،الإجابة عن السؤال، اختصار النص والتنقيب عن الآراء [موني ومن معه ،٢٠٠٥].

تكمن الصعوبة الأساسية في استخراج المعلومات من النصوص الحية في كون هذه النصوص غير مهيكلة فهي مكتوبة بلغة طبيعية ومخصصة للقراءة والتفسير والتحليل من قبل البشر.

تنقسم مرحلة استخراج المعلومات الى مهمتين رئيسيتين هما: التعرف الاسماء الكينونية من النص الغير مهيكل واستخراج العلاقات فيما بين هذه الأسماء [دانيال وجيمس،٢٠١٧].

٣.١ التعرف على الأسماء الكينونيه:

وهي عبارة عن استخلاص الكلمات الجوهرية وهي المصطلحات والعناصر الهامة في النص مثل أسماء الأشخاص (ستيف جوبز)، أسماء الشركات (آبل)، الأماكن (البحر الاحمر)، الوقت (السادسة مساءا) والتاريخ (١ سبتمبر ٢٠١١) أو الكلمات الأكثر تحديدا والتي تشير الى أسماء الأمراض(انفلونزا) وأعراضها (التهاب الحلق) الأدوية (بروفين) [بيسكورسكي ومن معه،٢٠١٣].

تعتبر هذه المرحلة الخطوة الأولى من خطوات استخراج المعلومات وهي اهم مراحل التنقيب عن النصوص وذلك لأنها تشكل الأساس الذي تعتمد عليه المراحل الأخرى من التنقيب عن النصوص مثل استخراج العلاقات بين الأسماء التي تم تصنيفها.

هناك ثلاث طرق رئيسية لاستخراج الاسماء من النصوص الحية: الطريقة المعتمدة على استخدام القواميس والمعاجم ، طريقة استخدم القواعد وطريقة تعليم الآلة[سيمبسون ودينا،٢٠١٢].

وفيما يلي نستعرض كل طريقة من هذه الطرق بمزيد من التفاصيل:

٣.١.١ الطريقة المعتمدة على استخدام القواميس والمعاجم:

وفيها يتم استخدام قاموس خاص تتوفر فيه قائمة بالأسماء المراد استخراجها من النص مثل اسماء الاشخاص، الأماكن، الأدوية ، الأمراض على حسب المهمة المراد تنفيذها. وقد استخدمت هذه الطريقة لاستخراج الاسماء في بناء عدة انظمة لاستخراج الجينات والبروتينات .

من عيوب هذه الطريقة في استخراج الاسماء هي انها تستطيع التعرف ومن ثم استخراج وتصنيف الاسماء المدرجة في القاموس وبالتالي اي اسم جديد غير مضاف في القاموس لا تستطيع هذه الطريقة التعرف عليه. فعلى سبيل المثال تفشل هذه الطريقة في استخراج المرادفات اذا لم يكن منصوص عليها في القاموس مثلا اذا كان القاموس يحتوي على المصطلح الطبي للمرض "أنيميا" بينما ذكر المرض في النص باستخدام المرادف العلمي للمرض "فقر الدم" فإن هذه الطريقة لاستخراج الأسماء لا تنجح في استخراج "فقر الدم" وتصنيفه كمرض وذلك لأن هذا المرادف غير موجود في القاموس [بيسكورسكي ومن معه،٢٠١٣].

كذلك تفتقر هذه الطريقة الى التعرف على الاسماء المكتوبة بترتيب مختلف أو بصيغة مختلفة عن تلك الموجودة في القاموس فمثلا اذا كان القاموس يحتوي على المصطلح العلمي للمرض " ارتفاع ضغط الدم" بينما تم استخدام صياغة أخرى للمرض في النص مثل "ضغط الدم كان مرتفعا" فإن الطريقة المعتمدة على القواميس تفشل في التعرف على الأسماء والمصطلحات حينما تذكر بصياغة لغوية تختلف عن تلك الموجودة في القاموس [النزاوي،٢٠١٦].

هذه المشاكل والعيوب في الطريقة المعتمدة على القواميس جعلت الباحثين يلجئوا الى طريقة استخدام القواعد المكتوبة يدويا لاستخراج الاسماء والتي يمكن من خلالها التعرف على الأسماء مع مراعاة الصيغ المختلفة التي يمكن استخدامها في ذكر المصطلحات. وبالتالي ساعدت الطريقة التي تستخدم القواعد على التغلب على المشاكل التي تواجه الطرق المعتمدة على القواميس [سيمبسون ودينا ،٢٠١٢].

٣.١.٢ طريقة استخدام القواعد:

هذه الطريقة تعتمد على كتابة مجموعة من القواعد تشمل جميع أو أغلب الانماط والاشكال التي تظهر فيها الاسماء المراد استخراجها من النص. أهم ميزه في هذه الطريقة هي استخدام المعلومات اللغوية في انشاء القواعد وبالتالي فإنها قادرة على استخراج الاسماء من النصوص بشكل دقيق مع مراعاة الأشكال المختلفة التي ممكن أن تظهر فيها هذه الاسماء. من عيوب هذه الطريقة انها تحتاج لوقت وجهد كبير لإنشاء واختبار هذه القواعد ومن الصعب استخدامها في مهمه اخرى لاستخراج اسماء جديده تختلف عن الاسماء التي استخدمت لإنشاء هذه القواعد [قارتن ييل ومن معه،٢٠١٠][شعلان،٢٠١٤].

٣.١.٣ طريقة تعليم الآلة:

في هذه الطريقة يتم استخدام خوارزميات تتعلم اليا مواصفات الاسماء المراد استخراجها وباستخدام هذه المواصفات تستطيع هذه الخوارزميات التعرف على جميع الاسماء التي تتحقق فيها هذه المواصفات. هذه الطريقة لاستخراج الاسماء تتطلب استخدام مجموعة كبيره من الامثلة الصحيحة للأسماء المراد استخراجها وذلك ليتم تدريب خوارزميات تعليم الآلة على المواصفات والخصائص اللغوية لهذه الأسماء بحيث تستطيع هذه الخوارزميات عند تطبيقها على نصوص جديده من استخراج الأسماء التي تتحقق فيها المواصفات التي تدربت عليها من أمثلة سابقه [سيمبسون ودينا،٢٠١٢].

٣.٢ استخراج العلاقات:

بعد الانتهاء من استخراج وتصنيف الاسماء الكينونية من النص يمكن استخراج وتصنيف العلاقات بين هذه الاسماء على سبيل المثال:

- **موظف في** هي علاقة بين اسم الشخص "ستيف جوبز" و "شركة ابل" تم استخراجها من الجمله "ستيف جوبز يعمل في آبل".
- **تقع في** هي علاقة تم استخراجها بين اسم الشخص "الدكتور سميث" والموقع مدينة "نيويورك" تم استخراجها من الجملة "الدكتور سميث ألقى محاضرة في مؤتمر في نيويورك".

بالرغم من ان العلاقات بين الاسماء قد تكون كثيره وغير محدودة الا انها غالبا تكون معروفة ومحدده مسبقا في توصيف المهمة الخاصة باستخراج المعلومات [بيسكورسكي ومن معه،٢٠١٣].

كما هو الحال في استخراج الأسماء من النصوص هناك ثلاث طرق رئيسية لاستخراج العلاقات بين الأسماء وفيما يلي نستعرض كل طريقة بمزيد من التفاصيل:

٣.٢.١ طريقة الأسماء التي تكرر ظهورها مع بعض:

تعتمد هذه الطريقة على الفرضية التالية إذا تكرر ظهور اسمين او مصطلحين مع بعضها البعض في نص معين مثل ظهورهما في نفس الجملة اوفي نفس الفقرة أو في نفس المقال فإن هناك فرصة كبيره بأن تكون هناك علاقة بين هذين الاسمين. والجدير بالذكر هو أن تكرار ظهور الأسماء مع بعض لا يضمن وجودة علاقة بينهم ولذلك فإنه ليس بالضرورة ان تكون هذه الفرضية صحيحه في جميع الحالات. لذلك كان من الضروري استخدام طريقة لتصفية العلاقات المتوقعة بين اسمين باستخدام هذه الطريقة وذلك للتقليل من عدد نتائج الإيجابية الخاطئة [سيمبسون ودينا ،٢٠١٢].

٣.٢.٢ طريقة استخدم القواعد:

وفيها يتم إنشاء قواعد تصف الأنماط اللغوية التي تستخدم في التعبير عن علاقة معينة. من مميزات هذه الطريقة هي الحصول على نتائج دقيقة للعلاقات التي تنطبق عليها الأنماط التي تم استخدامها في انشاء القواعد ولكن في المقابل هذه القواعد لا تستطيع استخراج وتصنيف العلاقات التي تحتوي على أنماط لغوية تختلف عن تلك التي تم استخدامها في إنشاء القواعد [سيمبسون ودينا ،٢٠١٢].

٣.٢.٣ طريقة تعليم الآلة:

هذه الطريقة تعتمد على استخدام خوارزميات تتعلم المواصفات اللغوية مثل الهياكل النحوية للعلاقات بين الأسماء آليا. وباستخدام المواصفات اللغوية للعلاقات تستطيع هذه الطريقة استخراج العلاقات بين الأسماء والتي تتوافر فيها المواصفات اللغوية التي تعلمتها أثناء مرحلة التدريب.

٤. تطبيقات التنقيب عن النصوص:

بعد الانتهاء من استخراج المعلومات وتصنيفها يتم تخزينها بشكل منظم في قواعد بيانات وذلك لتكون جاهز للاستخدام في تطبيقات كثيرة من أهمها الاختصار، التنقيب عن الآراء، الإجابة عن السؤال. هذه التطبيقات تستخدم لأغراض متعددة في مجالات مختلفة مثل ذكاء الأعمال، المعلوماتية الحيوية، الطب الشخصي [سيمبسون ودينا،٢٠١٢].

٤.1 الاختصار:

هو عباره عن اختصار لمحتوى مستند واحد او أكثر وإعطاء ملخص لأهم المعلومات والحقائق الموجودة في المستند/ المستندات ككل وفي هذا اختصار كبير لوقت القارئ او الباحث [زويقن بوم ومن معه،٢٠٠٩].

بشكل عام هناك نوعين للاختصار هما: الاستخراج والتلخيص. تعتمد طريقة الاستخراج على استخراج جمل وعبارات مباشرة من المستند واستخدامها في انشاء ملخص لمحتوى المستند.

بينما تعتمد طريقة التلخيص على فهم محتوى النص وإنشاء مختصر عن ذلك المحتوى بشكل عام هذه الطريقة تستخدم كلمات وجمل قد لا تكون موجودة في النص الأساسي [اليجوليف،٢٠٠٧] [ماني ومارك ،١٩٩٩].

٤.2 إجابة السؤال:

تعتبر الإجابة على السؤال أحد اهم تطبيقات التنقيب في النصوص والتي تعتمد بشكل كبير على مخرجات أنظمة استخراج المعلومات [سيمبسون ودينا،٢٠١٢].

يمكن النظر الى أنظمة الإجابة على السؤال كحالة خاصة من استدعاء المعلومات. والذي يميز الإجابة على السؤال عن استدعاء المعلومات هو الدقة في اعطاء المعلومة التي يحتاجها المستخدم [زويقن بوم ومن معه،٢٠٠٩].

فبدلا من إعطاء قائمة طويلة بكافة المستندات التي لها علاقة باستفسار المستخدم كما يحدث مع استدعاء المعلومات أنظمة الإجابة عن السؤال تعطي إجابة دقيقه ومختصره تجيب تماما عن سؤال المستخدم لذلك تعتبر أنظمة الإجابة على السؤال الجيل القادم لمحركات البحث [دينيشيا كارال ومن معه،٢٠٠٦].

تتكون أنظمة الإجابة على السؤال من مرحلتين رئيسيين وهما: مرحلة معالجة السؤال ومرحلة معالجة الإجابة.

أهم وأول خطوة من خطوات نظام الإجابة على السؤال هي معالجة السؤال وفيها يتم التحليل اللغوي وتصنيف السؤال لتحديد نوعية السؤال المطروح وبالتالي نوع الإجابة المتوقعة. بعد ذلك يتم استخدام السؤال لإنشاء استفسار والذي يستخدم كمدخل لمرحلة معالجة المستندات. في مرحلة معالجة المستندات يتم استخدام الاستفسار في محركات البحث وذلك لاستدعاء كل المستندات التي لها علاقة بالسؤال والتي يتم استخدامها لاستخراج كل الجمل والمقاطع التي تمثل إجابة محتمله للسؤال.

في مرحلة معالجة الإجابة يتم ترتيب الإجابات حسب الأولوية من حيث درجة مشابهتها لنوع السؤال والذي تم تحديده في مرحلة معالجة السؤال وبالتالي يكون المخرج من أنظمة الإجابة على السؤال هو الإجابات التي حصلت على اعلى درجة من حيث ملائمتها للإجابة على نوعية السؤال.

معظم الدراسات السابقة لتطبيق أنظمة الإجابة على السؤال تركز على إعطاء إجابة قصيره لسؤال معين مثل الإجابة عن سؤال التعريف باختصار معين مثل: ما هو اليونيسيف؟ أو السؤال عن منصب شخص معين مثل: من هو ستيف جوبز؟.

وفي المجال الطبي تلعب أنظمة الإجابة على السؤال دورا مهما وكبيرا في تطور العديد من التقنيات الطبية الحديثة مثل الطب المعتمد على الدليل والذي يساعد الطبيب في اتخاذ القرار الصحيح بخصوص تصنيف وعلاج حالة المريض وفقا لتاريخه المرضي وملفه الشخصي [ريتشاردسون ومن معه،١٩٩٥] [زويقن بوم،٢٠٠٦].

٤.3 التنقيب عن الآراء:

مع الانتشار الواسع لاستخدام المصادر التي تحتوي على النصوص التي تعكس اراء ومدى رضا الاخرين مثل نماذج التقييم والاستبانات الموجودة على الانترنت والانتشار الواسع لاستخدام وسائل التواصل الاجتماعي في التعبير عن الآراء بشكل عام تجاه موضوع معين ازداد الاهتمام بنتائج أنظمة التنقيب عن الآراء وذلك لاستخدامها في التسويق وخدمة العملاء ومعرفة مدى رضا المستخدمين تجاه منتج أو خدمة معينة. فأصبح العميل وقبل شراء أي منتج يبحث عن التقييمات السابقة لهذا المنتج عبر الانترنت ليعرف ملخص تجارب الاخرين وكذلك مميزات وعيوب ذك المنتج. لذلك أصبح أصحاب الشركات وموردين المنتجات يهتمون بشكل كبير برضا العميل لأنهم يعلمون بأن رضا العميل سيؤثر بشكل إيجابي على عملية التسويق والمبيعات[ويتن،٢٠٠٤].

التنقيب عن الآراء او التعبير عن وجهة النظر غالبا يطبق على النصوص والتعليقات التي تمت كتابتها من قبل العملاء مثل استبانات رضاء العميل. بعد ذلك يتم تحليل قطبية المشاعر للجمل واختبار احتوائها على كلمات ايجابيه او سلبيه باستخدام معجم خاص بالكلمات التي تعبر عن المشاعر الإيجابية والسلبية مثل جميل، قبيح، جيد، سيئ الخ. وبناءا على نتائج تحليل قطبية المشاعر يتم تصنيف الآراء الى إيجابية او سلبيه [اقراول وتانفير،٢٠٠٩]. ولكن هذا النوع من التحليل قد لا يكون دقيق بالشكل

المطلوب لتصنيف الآراء وذلك لأن بعض الكلمات قد تستخدم للتعبير عن الآراء السلبية والإيجابية على حد سواء على حسب السياق الذي استخدمت فيه مثل قصير، طويل الخ. لذلك يعد استخدام طرق تعليم الآلة حل مثالي للتغلب على مشاكل الطريقة السابقة والتي تعتمد على تحليل قطبية المشاعر حيث يتم تدريب خوارزميات تعليم الآلة على امثلة تتضمن آراء سلبية وإيجابية وبالتالي تتعلم هذه الخوارزميات آليا مواصفات الجمل والألفاظ التي تعبر عن رأي سلبي أو إيجابي بحيث أنه عند تطبيق هذه الخوارزميات على جمل جديدة تتمكن الخوارزميات من تصنيف الأراء السلبية والإيجابية بناءا على المواصفات التي تعلمتها خلال فترة التدريب [دينيك ،٢٠٠٨].

٥. الختام:

في هذه الورقة العلمية تم استعراض آخر التطورات والتقنيات المستخدمة في التنقيب عن النصوص المفيدة وتطبيقاتها. كذلك تم تلخيص اهم الطرق المستخدمة لترميز تكويد المدونة، استخراج الأسماء الكينونية واستخراج العلاقات بين الأسماء مع ذكر مميزات وعيوب كل طريقة. والجدير بالذكر ان اختيار الطريقة المستخدمة لاستخراج المعلومات يعتمد بشكل كبير على المهمة المراد تنفيذها والدقة في استخراج المعلومات المراد تحقيقها.

كذلك تم استعراض أهم تطبيقات التنقيب عن النصوص والتي تستخدم لأغراض مختلفة في مجالات متعددة من أهمها المجال الطبي حيث ساهمت تقنيات وتطبيقات التنقيب عن النصوص بشكل كبير في تطوير التقنيات المستخدمة لتشخيص حالة المريض وإعطاءه العلاج المناسب لحالته.

على الرغم من أن معظم أبحاث التنقيب عن النصوص والتطور في استخدام التقنيات المختلفة لاستخراج المعلومات يتم تطبيقه على النصوص الإنجليزية إلا ان الأبحاث في استخراج المعلومات من النصوص العربية وتطبيقاتها المختلفة هو في تطور مستمر.

٦. المصادر والمراجع:

[اليجوليف ، ٢٠٠٧]

Aliguliyev, Ramiz M. **Automatic document summarization by sentence extraction**. *Вычислительные технологии* 12.5 (2007).

[اقراول وتانفير، ٢٠٠٩]

Agrawal, Shaishav, and Tanveer j Siddiqui. **Using syntactic and Contextual Information for Sentiment Polarity Analysis**. (2009).

[السلمان والربيعه ،٢٠١٠]

Alsalman, Abdulmalik and Alrabiah Maha. **Recent Advances and trends in Arabic information extraction**. *Workshop on enriching Arabic digital contents*. Damascus, Syria, 2010.

[النزاوي ،٢٠١٦]

Alnazzawi, Noha Abdulkareem D. **Linking clinical records to the biomedical literature**. *Diss. University of Manchester*, 2016.

[المقرن ومن معه ، ٢٠١٧]

Almuqren, Latifah, et al. **A Review on Corpus Annotation for Arabic Sentiment Analysis**. *International Conference on Social Computing and Social Media*. Springer, Cham,

[الهليس ،٢٠١٣]

الهليس، علاء مصطفى. **"تنقيب الآراء في جمل المقارنة العربية."** (٢٠١٣).

[بيسكورسكي ومن معه ، ٢٠١٣]

Piskorski , Jakub, and Roman Yangarber. **Information extraction: Past, present and future**. *Multi-source, multilingual information extraction and summarization*. Springer, Berlin, Heidelberg, 2013. 23-49.

[جوبتا ومن معه ، ٢٠٠٩]

Gupta, Vishal, and Gurpreet S. Lehal. **A survey of text mining techniques and applications**. *Journal of emerging technologies in web intelligence* 1.1 (2009): 60-76.

[دانيال و جيمس، ٢٠١٧]

Daniel Jurafsky & James H. **Information Extraction**. *Proceeding of Speech and language processing*. 2017.

[دينيشيا كارال ومن معه،٢٠٠٦]

Denicia-Carral, Claudia, et al. **A text mining approach for definition question answering**. *Advances in Natural Language Processing*. Springer, Berlin, Heidelberg, 2006. 76-86.

[دينينك ، ٢٠٠٨]

- Denecke, Kerstin. **Using sentiwordnet for multilingual sentiment analysis**. *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*. IEEE, 2008.

[ريتشاردسون ومن معه ، ١٩٩٥]

Richardson, W. Scott, et al. **The well-built clinical question: a key to evidence-based decisions**. *ACP journal club* 123.3 (1995): A12-A12.

[زوبقن بوم ، ٢٠٠٦]

Zweigenbaum, Pierre. **Question answering in biomedicine**. *Proceedings Workshop on Natural Language Processing for Question Answering, EACL*. Vol. 2005. 2003.

زويقن بوم ومن معه ،٢٠٠٩

Zweigenbaum, Pierre, et al. **Frontiers of biomedical text mining: current progress**. *Briefings in bioinformatics* 8.5 (2007): 358-375.

[سكانيا وبيرونثا ،٢٠١٢]

Sukanya, M., and S. Biruntha. **Techniques on text mining**. *Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on*. IEEE, 2012.

[سيمبسون و دينا ، ٢٠١٢]

Simpson, Matthew S., and Dina Demner-Fushman. **Biomedical text mining: A survey of recent progress**. *Mining text data*. Springer, Boston, MA, 2012. 465-517.

[شعلان ، ٢٠١٤]

Shaalan, Khaled. **A survey of Arabic named entity recognition and classification**. *Computational Linguistics* 40.2 (2014): 469-510.

[قارتين ييل ومن معه، ٢٠١٠]

Garten, Yael, Adrien Coulet, and Russ B. Altman. **Recent progress in automatically extracting information from the pharmacogenomic literature**. *Pharmacogenomics* 11.10 (2010): 1467-1489.

[ماني ومارك ،١٩٩٩]

Mani, Inderjeet, and Mark T. Maybury. **Advances in automatic text summarization**. *MIT press*, 1999.

[موني ومن معه ، ٢٠٠٥]

-Mooney, Raymond J., and Razvan Bunescu. **Mining knowledge from text using information extraction**. *ACM SIGKDD explorations newsletter* 7.1 (2005): 3-10.

[ويتن ، ٢٠٠٤]

Witten, Ian H. **Text Mining**. (2004): 198.

٧. جدول الألفاظ:

انجليزي	عربي
Text Mining	التنقيب عن النصوص
Data Mining	التنقيب عن البيانات
Unstructured text	النصوص الغير مهيكله
Information Retrieval	استرجاع المعلومات
Information Extraction	استخراج المعلومات
Indexing	الفهرسة
Query Reformulation	تكوين الاستعلام
Matching	المطابقة
Corpus Annotation	ترميز تكويد المدونة
Corpus	المدونة
Part-of-Speech tag	معلومات أجزاء الكلام
Ontology	أنطولوجي
Structured Data	بيانات مهيكله
Machine Translation	ترجمة الاله
Question Answering	إجابة السؤال
Summarization	الاختصار
Opinion Mining	التنقيب عن الاراء
Named Entity Recognition	التعرف على الأسماء الكينونية
Relation Extraction	استخراج العلاقات
Dictionary-based Method	طريقة استخدام القواميس والمعاجم
Rule-based Method	طريقة استخدام القواعد
Machine Learning-based Method	طريقة تعليم الآلة
Co-occurrence Method	طريقة الأسماء التي تكرر ظهورها مع بعض
Name Features	مواصفات الأسماء
False Positive	الإيجابية الخاطئة
Grammar Structures	الهياكل النحوية
Business Intelligence	ذكاء الأعمال
Bioinformatics	المعلوماتية الحيوية
Personalised Medicine	الطب الشخصي
Extractive	الاستخراج
Abstractive	التلخيص
Evidence-based Medicine	الطب المعتمد على الدليل