مبادئ التصنيف الآلي للنصوص العربية

الأستاذ: العربي بوعمران بوعلام
جامعة خميس مليانة

ملخص:

لقد أدى تنامي البيانات الرقمية العربية على شبكة الانترنيت ومختلف المؤسسات الى وجود فائض هائل يصعب التعامل معه بمرونة لذا بدأ العمل على ايجاد حلول قصد تنظيم هذه المعلومات وتخزينها وايجاد طرق فعالة من أجل تسهيل عملية استرجاعها هذا ما ادى الى ضرورة العمل على تصنيف هذه المعلومات بمناهج علمية دقيقة ولقد ساعد بروز المستكشفات العلمية الحديثة كالحاسب الآلي الى تسهيل عملية جمعها وادخالها واسترجاعها وايجاد طرق آلية لتصنيفها فنتيجة لهذا التطور ظهرت العديد من برامج المعالجات اللغوية كالقارئ الآلي والمشكل الآلي والمصنف الآلي.

الكلمات المفتاحية: البيانات الرقمية، المعلومات، الحاسب الآلي، التصنيف الآلي، خوارزميات التصنيف.

Résumé :

La croissance des données numériques arabes sur internet et diverses institution a conduit à un nombre énorme est difficile à traiter avec souplesse, les sorte qu'il semblait être utile de trouver des solutions destinées à organiser et stocké cette information, en plus de trouver des moyens efficaces pour faciliter le processus de récupération qui a mené à la nécessite de travailler pour classer ces méthodes scientifiques d'information précises et a aidé l'émergence d'explorateurs scientifiques modernes, tels que les ordinateurs ,a facilité le processus de la collecte, entrée et récupération, et trouver des moyens automatique de les classer à la suite de ce développement de nombreux programmes de traitement langue et lecteur automatique et classeur automatique des textes.

أصبحت مجالات المعلوماتية والحوسبة حاليا من بين الميادين الأكثر فعالية ونشاطا ولاقى انتشارا واسعا في مختلف المجالات المعرفية، إذ تم تطوير العديد من الأدوات للمساعدة في إدارة المعلومات والتحكم بها خلال العقود القليلة الأخيرة، فالتطور التكنولوجي الصناعي وأدوات البحث (العلمي) قربت المسافة بين الروحاني والجسماني، وبين الروحاني والمادة، وجعلت الآلة، وكأنها أصبحت إنسانا جديداً وكشفت على وجود كوجيطو صنعي يقف مع الكوجيطو الذاتي على قدم المساواة من حيث المعرفة، والقدرة على الكشف والاختراع[1].

كما كان لابد من تطوير نظام لاسترجاع المعلومات التي تعمل وفق مجموعة من الأدوات التي تهتم بتمثيل وتخزين وترتيب والوصول إلى البيانات لذلك كان لا بد من إيجاد تقنيات وأساليب فعالة تعمل على تصنيف هذه الوثائق وتجميعها بطريقة منظمة، وقد ظهرت العديد من مناهج التصنيف الآلي في الآونة الأخيرة التي تعتمد على خوارزميات التعلم الآلي وإن كانت عملية التصنيف معروفة منذ القدم يدويا إلى أن تطور الوسائل التكنولوجية تولدت عنها تقنية التصنيف الآلي للنصوص.

تعريف التصنيف:

التصنيف وفق ما ورد في المعاجم هو تمييز الأشياء بعضها عن بعض، وصنف الأشياء أي قسمها وفق تشابهها إلى مجموعات تضم كل مجموعة وحدات تشترك في صفة أو خاصية واحدة على الأقل " فيقال صنف الشيء جعله اصنافا وميز بعضه عن بعض ولا يختلف معنى التصنيف في مفهومه العام عن معناه اللغوي ، وهو يعني جمع وترتيب الاشياء المتشابهة في أقسام تبعا للصفات المتشابهة"[2].

ويعرف أيضا " كلمة تصنيف من الفعل الثلاثي صنف والاسم منها ايضا صنف والصنف هو الشيء أو مجموعة الاشياء المتميزة عن غيرها بسمات معينة ... والغرض من ذلك هو تسهيل إدراك العلاقات بين الوحدات أو المجموعات المصنفة وحفظها في الذاكرة"[3].

---

[1] سامي أدهم، الذكاء الصنعي، ثنائية الآلة والدماغ، **مجلة كتابات معاصرة**،ع 28-29، دجنبر 1996، يناير 1997، ص.35.

[2] أحمد البدوي، فن تصنيف كتاب، دار الفكر العربي ، القاهرة، ط2، 1993، ص 09

[3] هاني العمد ، المعالجة الفنية للمعلومات، منشورات جمعية المكتبات الاردنية، عمان، 1975،ص60.

فلا يبتعد التعريف اللغوي عن الاصطلاح فالتصنيف هو عملية فرز المعلومات وإدارتها أي تنظيمها و وضعها ضمن وثائق أو مصنفات بطريقة أو منهجية مضبوطة يراعى فيها العديد من الضوابط سواء كان ذلك يدويا أو آليا أو عبارة عن عملية تجميع آلي لمجموعة من الوثائق في صنف أو عدة أصناف، وفقا لمعايير مختلفة كمحتواها النصي ونوع الوثيقة.

أو يمكن إعطاءه تعريفا آخر أكثر تحديدا إذ أن تصنيف النصوص هو تعيين النص بعلامة أو أكثر لفهرسة هذا المستند في مجموعة من الفئات محددة مسبقا، صممت في الأصل للمساعدة في أعمال الترتيب الوثائقي أو المقالات في المجالات التقنية أو العلمية، لذا فعملية تصنيف النصوص هي ضم وثيقة إلى فئة أو فئات محددة مسبقا سواء كان ذلك يدويا أو آليا.

التصنيف يتم فيه تحليل مجموعة من البيانات لتكوين من القواعد المتجمعة التي يمكن ان تستخدم لتصنيف بيانات المستقبل أي ايجاد المعلومات التي تتعلق بالخصائص المشتركة، وللتصنيف أدوات عديدة مثل شجرة القرار والمجاور الأقرب والانحدار"[1].

أما التصنيف الآلي للنصوص (Automatic Text Categorization) هي مهمة تصنيف المستندات النصية الإلكترونية اتوماتيكيا إلى أصنافها المعرفة مسبقا بحسب محتوياتها، بمعنى آخر تحديد الصنف الرئيسي الذي يندرج تحته النص أو المستند "سياسة ، اقتصاد ، رياضة، ... الخ".[2] يعلمك تصنيف النصوص كيفية التحكم في مجموعة من مواصفات التمييز للسماح بتخزين الوثيقة المعطاة في طبقات أو فئات موافقة لمحتواها.

أصبح التصنيف الآلي حاليا من بين الميادين الأكثر فعالية ونشاطا ولاقى اهتماما واسعا من قبل المتخصصين و الدارسين بمختلف المجالات العلمية، لذا ظهرت العديد من تقنيات التعلم الآلي في مجال تصنيف النصوص منها شجرة القرار والمجاور الأقرب

[1] زكريا الدوري داليا عبد الحسين،دور تنقيب البيانات في زيادة أداء المنظمة،مجلة العلوم الاقتصادية والاداريةّ،المجلد 13، العدد 2017،48

[2] بإستخدام تعليم بايزين الإحتمالي، بسام محمد واحمد السالمي، التصنيف الآلي للنصوص العربية 2011.

و الانحدار[1]،وأدوات أخرى كتقنية المكائن ذات الدعم الاتجاهي، وتقنية الاحتمالات المسماة بآيس خوارزمية (TF- IDF)، ومصنف الزناد [2]، إذ لوحظ في الآونة الأخيرة ظهور العديد من برامج التصنيف الآلي مزودة بالعديد من الادوات كبرنامج wekaو rabidminer الذي يعمل بلغة الجافا، وكلما سوفت لتصنيف النصوص العربية.

الإرهاصات الأولى لتصنيف النصوص آليا:

إن نظام تصنيف الموضوعات ظهر مع الأيام الأولى للتدوين من أجل إضفاء الطابع المؤسساتي للإسكندرية، مما استلزم ذلك إلى إنشاء نظام التصنيف العالمي (ديوي) عام 1876م الذي يتعلق بتصنيف الوثائق الشبيهة بالموسوعات و الكتب[3].

إلا أن فكرة إجراء تصنيف النصوص عن طريق الأجهزة يعود إلى أوائل الستينات، وقد حقق تقدما كبيرا مع بداية التسعينات مع ظهور العديد من الخوارزميات التي أصبحت أكثر كفاءة من ذي قبل، لذا حتى أوائل الثمانينات كان الواجب علينا لبناء مصنف تكريس الكثير من الموارد البشرية لهذه المهمة، فالعديد من الخبراء قاموا بنشر قواعد يدوية ووضعوا لها اختبارات ومع ظهور التعلم الآلي تم توفير الكثير من الجهد والوقت، هذه التطورات التكنولوجية والخوارزميات المتقدمة جعلت التصنيف اليوم أداة يمكن الاعتماد عليها بشكل كبير.

في بداية التسعينات بدأ البحث في المجمعات حول استرجاع المعلومات وتم وضع مناهج الرقمنة وخوارزميات التصنيف خاصة في المؤتمرات العلمية، كما اهتم مجمع التعلم الآلي بهذه المشكلة منذ أكثر من عشر سنوات مثل النظر في الخوارزميات للتعرف على الأشكال، وحاليا لاتزال مناهج رقمنة النصوص موجودة ومستوحاة إلى حد كبير من قبل البحث عن معلومات في حين أن المصنفات الأكثر أداء هي المتعلقة بالتعلم الآلي.

الحاجة إلى تصنيف النصوص:

---

[1] داليا عبد الحسين أحمد – زكريا الدوري، دور تنقيب البيانات في زيادة أداء المنظمة ، مجلة العلوم الاقتصادية والغدارية، المجلد 13، العدد 2007،48،ص 45

[2] مراد عباس ، كمال سماعيلي، داود بركاني: تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، (4/3)، العدد 2011،29، ص 185-186

[3] احمد بدر، التصنيف فلسفته وتاريخه، وكالة المطبوعات، الكويت، 1983 ، ط1، ص 24.

لقد اهتمت في السنوات الأخيرة الكثير من البحوث بتألية النصوص متعددة اللغات وذلك لعدة أسباب نذكر منها: انتشار استعمال الحاسوب وتزايد مجموعة البيانات على الشبكة العنكبوتية وانتشارها على نطاق واسع بمختلف الأصناف واللغات بالإضافة إلى تزايد حجم المدونات المستخدمة[1] لذا تولدت الحاجة إلى ضرورة تصنيفها آليا لأن المعالجة اليدوية لهذه البيانات قد تكلف الكثير من الوقت و الأفراد كما أنها غير مرنة وتعميمها على كل الميادين أمر شبه مستحيل

فالمعالجات اليدوية لهذه البيانات مكلفة للغاية في الوقت والأفراد، كما أنها ليست مرنة وتعميمها الى ميادين أخرى مستحيلة عمليا لذا لابد من السعي لتطوير الأساليب الآلية[2].

وبالتالي لا بد من استبدال العملية اليدوية بالتصنيف الآلي للنصوص من اجل تجنب العديد من النقائص التي يمكن أن يحدثها التصنيف اليدوي، وعليه فقدت ركزت البحوث في السنوات الأخيرة على التصنيف الآلي.

كيفية إجراء عملية التصنيف الآلي:

لتنفيذ عملية التصنيف الآلي للنصوص كما عرفناها سابقا لا بد من إتباع العديد من الخطوات المهمة:

المرحلة الأولى: تتعلق بإضفاء الصبغة الرسمية على النصوص المأخوذة كعينة الممثلة للفئة المراد تصنيف النصوص إليها وتحضيرها بحيث تكون مفهومة من قبل الآلة واستخدامها من قبل خوارزميات التعلم، ليتم بعدها معالجة هذه البيانات وإزالة الشوائب التي قد تؤثر على مهمة التصنيف الآلي للنصوص كإزالة حروف الجر، العطف، الضمائر، علامات الوقف والترقيم، علامات التشكيل والفراغات فهناك الكثير من الكلمات غير المفيدة بل تؤدي الى تدني كفاءة المصنفات فعلى سبيل المثال أدوات اللغة المتمثلة في حروف الجر ظرفي الزمان والمكان أسماء الاشارة وغيرها هي غير مرغوب فيها ويجب ازالتها من المدونة وكذلك الكلمات التي لا يتجاوز تكرارها قيمة معينة في أغلب

---

[1] محمد المهدي بودبوس، التلخيص الآلي للنصوص العربية، Communication of the Arabe Computer Society ,Vol.4,No.2,2011 P14

[2] عبد الملك أمين / مقاربة لتحديد اللغات تلقائيا في مدينة نصوص متعددة اللغات،المجلة العربية الدولية للمعلوماتية، المجلد الثاني، العدد الرابع، 2013، ص 29.

الاحيان[1] يلي ذلك تقسيم النصوص الى مجموعتين احداهما تستخدم لتدريب البرنامج والثانية لاختباره.

أما المرحلة الثانية: تتعلق بتصنيف الوثائق هذه المرحلة هي الحاسمة لأنها تسمح أو لا تسمح باستخدام تقنيات التعليم، لإنتاج تعميم جديد للمصنفات ولتحسين أداء النماذج يتم تقييم جودة المصنفات ومقارنة النتائج المقدمة من طرف مختلف النماذج.

يشمل مجال المعالجة الذكية للبيانات النصية جميع الأدوات والأساليب الفعالة التي تعمل على استخراج المعلومات من النصوص المكتوبة باللغة الطبيعية، إذ يوجد مجالين مهمين لمعالجة هذه المشكلة ولكل مجال أساليبه الخاصة به، أولا تعمل المناهج على تحليل البيانات وإعداد دراسة إحصائية خاصة كما أنها تسعى لتقديم أدوات للإحصائيين واللغويين لتمكينهم من تحليل قواعد البيانات النصية ذات الحجم الكبير، من خلال توفير معلومات موجزة عن المدونة، كما تقدم برامج تحليل المدونات قوائم تردد الكلمة وتمثيلها البياني من خلال تحليل جزء من هذه المصنفات، كما تقدم هذه المناهج أنظمة وطرق تعالج الوثائق بصفة آلية تلقائية .

تطبيق تقنيات التصنيف الآلي على النصوص العربية:

تتميز اللغة العربية عن بقية اللغات بأنها تكتب وتقرأ من اليمين إلى اليسار، كما أن حروفها تكتب بأشكال مختلفة تبعاً لموقعها والحروف المجاورة لها، وتختلف طريقة نطق الحرف وبالتالي معنى الكلمة وموقعها الإعرابي بناءً على حركة التشكيل الموجودة عليه، بالإضافة إلى أن العربية لغة اشتقاقية وليست إلصاقية، حيث يعد نظامها الصرفي من أكثر النظم الصرفية تقدماً، فهو مبني على تصريف الجذور وفقاً لمجموعة محددة من الأوزان للحصول على كلمات ذات دلالات مختلفة من نفس الجذر، وكل ما سبق ذكره يمثل تحديات لمقننة التحليل الصرفي والإعرابي والدلالي للغة العربية ومن ثم التصنيف الآلي لمجمل النصوص العربية وقد وجدنا عددا قليلا فقط من الأبحاث التي اهتمت بتصنيف الوثائق العربية لغرض استرجاع المعلومات.

حيث كانت هناك محاولات عديدة لبناء مصنف آلي للوثائق العربية مبنية على استخراج الكلمات التي تغطي المفهوم الأساسي لموضوع كل وثيقة، بحيث يتم حساب

[1] مراد عباس ، كمال سماعيلي، داود بركاني: تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، (4/3) العدد 29، 2011،ص 184

ISSN : 2353-046
EISSN : 2600-6421

وزن كل كلمة بناءً على مدى تكرار هذه الكلمة في الوثيقة وأماكن تواجدها، وقد وجد الباحثون أن استخدام خوارزمية التصنيف هذه قد زاد من كفاءة نظام استرجاع المعلومات[1].

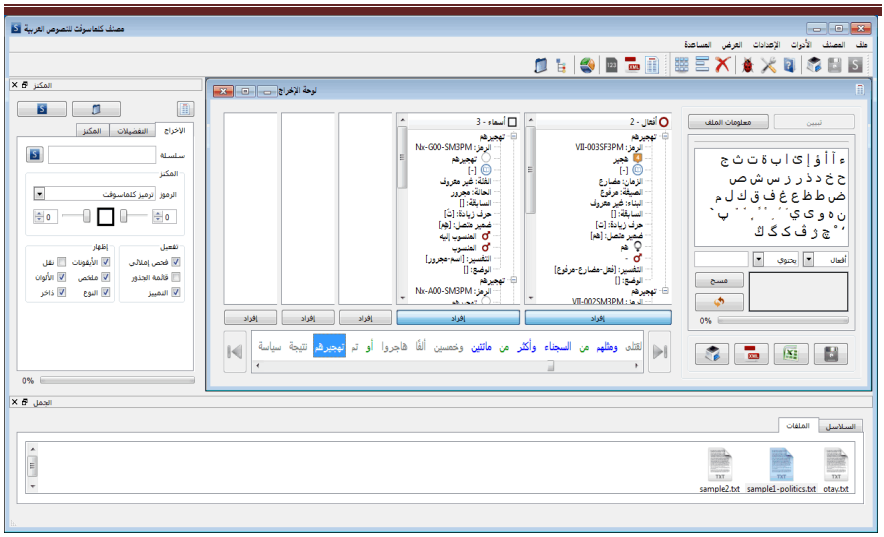
إلا اننا نجزم بالقول أن محاولة قولبة اللغة العربية في الحاسوب من أهم المشاكل التي تعترض طريق وضع المصطلحات العربية بدء بمحاولة ادخالها لبرامج التصنيف التي تعتمد على لغة الجافا . وجوب حصر الأوزان العربية حصراً دقيقاً وتحليلها وفق نظام تصنيفي معين، وهو ما سيمكن من وضع رموز رياضية لها في الحاسوب[2]. ثم يتم بعد ذلك تطوير برامج آلية يمكنها استيعاب القواعد النحوية العربية، بحيث يتمكن الحاسوب من تصويب الجمل الخاطئة عند قراءتها.

مع أنه يوجد العديد من برامج التصنيف الآلي للنصوص بمختلف اللغات إلا أنه لا يمكن تطبيقها بسهولة على اللغة العربية اذ لا نجد برنامج تصنيف عربي واضح لذا ظهرت هناك العديد من المحاولات التي عملت على ايجاد مصنف عربي يتعامل مع البيانات العربية من حيث ادخالها وتصنيفها واسترجاعها مثل نظام التصنيف الموسع كلما سوفت "Deep PoS Tagger" الذي يعمل على تصنيف الكلمات إلى الأنواع النحوية بحيث يقترب من التصنيف الدلالي للكلمة وهو يختلف كثيراً عن المصنفات التقليدية المتوفرة للاستخدام المباشر، نظام التصنيف يميز كذلك الدخيل والأصيل باستعمال قواعد مبسطة لا تعتمد على التحليل النحوي، استعمالات النظام تظهر في تهيئة المدونات (Corpora) العربية وتصنيف مفرداتها لأن المدونة غير المصنفة لا فائدة ترجى من استعمالها[3].

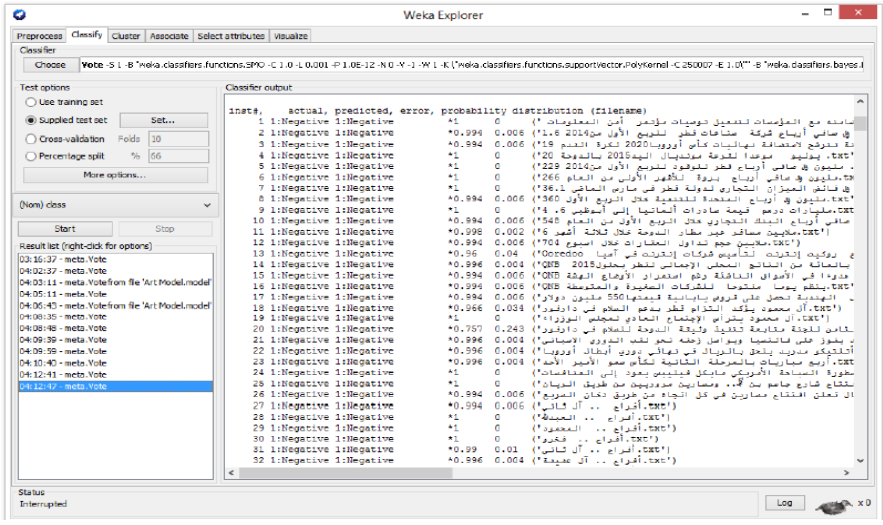
[1] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi and A. Ababneh, "Enhanced Arabic Information Retrieval System based on Arabic Text Classification", 4th International Conference on Innovations in Information Technology, pp.461 - 465, 2007.

[2] سعد بن هادي قحطاني، تحليل اللغة العربية بواسطة الحاسوب، مركز اللغة الانجليزية، معهد الادارة ، الرياض.

[3] www.Kalmasoft.com/papers/PA-13-ST01.pdf



كما قدم العديد من الباحثين دراسات حول تصنيف البيانات العربية معتمدين في ذلك على برنامج weka الذي يحوي مختلف أدوات وتقنيات التصنيف[1]



ظهرت محاولات كثيرة لوضع برامج آلية لتصنيف بيانات اللغة العربية بحيثُ يمكنها التعامل مع مختلف المشاكل و العوائق، خاصة فيما يتعلق التركيب الصرفي للغة العربية، بالإضافة إلى بعض هذه العوائق الناشئة عن عملية التعريب نفسها، أي تعدد الطرق المستخدمة في التعريب وتباينها فيما بينها بالإضافة إلى مشاكل أخرى نحصرها فيما يلي:

[1] https://fr.m.wikipedia.org>wiki>weka..

1- الطبيعة اللاصقة للكتابة العربية فتكون الكلمة العربية ملتصقة الحروف مشكلة كلمة خطية تحمل معلومات صرفية نحوية ودلالية وهذه الكلمة يمكن ان تترجم بالعديد من الكلمات في لغات أخرى هذا مما يولد صعوبة اثناء التصنيف اذ لابد من معرفة جذر الكلمة المركبة[1]

2- عدم وجود فوارق شكلية واضحة بين مكونات النص فبعض اللغات تميز أجزاء الكلام بأنواع مختلفة من الحروف و اللواصق إلا أن اللغة العربية لا نجد فيها أدوات تفرق بين الكلمات إذ أنها تعتمد على الأوزان و الحركات الإعرابية لذا لا بد من العودة إلى الحركات التي تعمل على إحداث فوارق بين الكلمات المختلفة.

3- افتقار اللغة العربية لمبدأ الوحدة الدلالية: تقوم اللغة العربية على عكس اللغات الاخرى على مبدأ الاشتقاق الذي ينبني اساسا على الجذر اذ يمكن لجذر ثلاثي ان تشتق منه العديد من المفردات التي تختلف في أكثرها دلاليا و هذا ما يعيق عملية التصنيف اذا ان الكلمات المشتقة من جذر واحد يمكن ان تحمل معان مختلفة عن بعضها البعض و هذا ما يمكنا تسميته بالتشتت الدلالي مثل : الجذر بلغ نشتق منه الكلمات التالية و التي تختلف دلاليا فيما بينها: بلوغ- مبالغة- بليغ- بلاغة- مبلغ- بلاغ، لذا كان لابد من استخراج جذور الكلمات وفصلها عن الضمائر المتصلة وأداة التعريف وغيرها من الزوائد يؤدي الى تحسين النتائج من خلال تزويد المصنف بالتكرارات الصحيحة للكلمات[2].

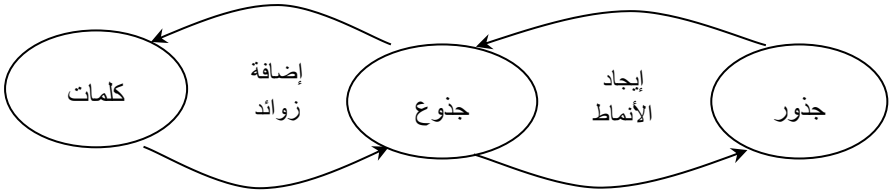
ثم إن معظم مفردات اللغات الأخرى المصنفة بالحاسوب مصنفة حسب الصيغة المبنية (أي صيغة الفعل، أو الاسم المجرد) وليس بالجذر كما في اللغة العربية. وتصنَّف السوابق واللواحق كمدخلات أساسية عكس اللغة العربية، وبما أن اللغة العربية تحتوي على صيغ صرفية داخلية (تحدث داخل الكلمة نفسها) وليست سوابق أو لواحق؛ فإنه يتحتم التعامل مع الجذر وليس مع كل صيغة على حدة. وهذه الخصائص التي تختص بها العربية تجعل من الصعب استقطاب البرامج الآلية الحديثة التي صممت أصلاً للتعامل مع الإنجليزية لذا ظهرت مجهودات عربية حاولت ايجاد برامج آلية مثل

[1] Abbes ,Ramzi –la conception et LA Réalisation d' un concordancier électronique pour l'arabe-thése de doctorat :informatique .-Lyon :Institut national des sciences appliquées ,2004.-P.26

[2] مراد عباس، كمال سماعيلي، داود بركاني : تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، 29(4/3) ،2011، ص 184 .

المعالج الصرفي لشركة صخر الذي يتعرف على اشكال جذر الكلمة بعد تجريدها من السوابق واللواحق[1].

والفرق الرئيسي بين اللغة العربية وغيرها من اللغات هي أنها اشتقاقية أما اللغات الأخرى فهي لصقية، الشكل التالي يوضح رسم تخطيطي ومثال على النظام العربي للاشتقاق:



شكل1: يوضح نظام الاشتقاق العربي

المشكلة الرئيسية للخوارزمية المعتمدة على الجذر في عملية التصنيف الآلي للنصوص هي أن العديد من التهجئات المختلفة للكلمة ليس لديها تفسيرات دلالية متشابهة، أي بالرغم من أن هذه الكلمات تنشأ وتنتج من نفس الجذر إلا أنها مختلفة في المعنى لذا استخدام الخوارزميات المعتمدة على الجذر في التصنيف تزيد من غموض الكلمة [2].

4- الاستخدام المفرط للأساليب البيانية (المجاز- الكناية – الاستعارات): تستخدم اللغة العربية هذه الأساليب بشكل واسع جدا بغية تحسين الخصائص البلاغية للنص أو التأثير العاطفي للمتلقي و بإمكان المتلقي أن يستوعب مضمون الأساليب البيانية بالاعتماد على خبراته المرجعية إلا أن جهاز الكمبيوتر لن يكون بالإمكان استيعابها لأنه يعتمد في مجمل عملياته على التفسير المنطقي المباشر و بالتالي الأساليب البيانية تشكل صعوبة أمام التصنيف الآلي للنصوص العربية.

5- عدم وجود علامات التشكيل: تعتمد اللغة العربية بالأساس على التشكيل والتنقيط وعادة ما تسمى تلك الحركات بالصوائت، ونجد أغلبية الوثائق أو النصوص الالكترونية لا تتضمن تشكيلا و ربما السبب يعود في ذلك إلي عدم مرونة لوحة مفاتيح عند

[1] انظر موقع شركة صخر. www.sakhr.com

[2] Kareem Darwish. "Building Shallow Arabic Morphological Analyzer in One Day", Association for Computational Linguistics. 40th Anniversary Meeting. July 6-12, 2002. pp. 47-54. University of Pennsylvania.

استخدام علامات التشكيل إذ نجد مفتاحين يعملان معا لكل علامة تشكيل، ونجد أغلبية النصوص المتعلقة بالأخبار و الروايات لا تعتمد على التشكيل.

6- الأخطاء اللغوية الشائعة: يتأثر التصنيف الآلي بهذه الأخطاء التي تخرج عن القاعدة اللغوية فالأفعال السداسية والخماسية في اللغة العربية تبدأ بهمزة وصل إلا أن بعض النصوص نجدها مكتوبة بهمزة قطع مثل: استخرج وإستخرج، وحتى بالنسبة للكلمات المبدوءة بهمزة قطع تكتب بهمزة وصل مثل: أنباء وانباء، أيضا بالنسبة للكلمات التي النصوص تنتهي بياء نجدها في بعض الأحيان تكتب بياء مقصورة، مثل: على وعلي، وبعض اللهجات تضيف همزة وصل في بداية أسماء الأعلام مثلا محمد تكتب امحمد، فهذه الأخطاء تمثل عائق كبير جدا في عملية التصنيف الآلي إذ لا بد من تصحيحها يدويا أو بمصحح آلي بالإشراف للتأكد من أخذ الكلمة الصحيحة ثم تصنيفها.

7- بالرغم من هذه الأخطاء التي تعيق عملية تطبيق أهم التقنيات التكنولوجية على اللغة العربية وأهمها عملية التصنيف الآلي للنصوص أو البيانات إلا أن البحوث مستمرة و هناك العديد من البحوث التي قدمت تقنيات حاسوبية (آلية) حاولت أن تعطي حلولا قيمة لعملية حوسبة اللغة العربية وكذلك تم تطويع العديد من المناهج الغربية و الخوارزميات حتى تناسب اللغة العربية .

خلاصة:

إن التعامل اليدوي مع هذا الكم الهائل من البيانات والنصوص العربية دون استخدام تقنيات حديثة يبعدنا عن التطور و الارتقاء إلى مستويات أداء أفضل إذ لا يكفي مجرد إدخال الآلات إلى العمل، بل من الأفضل استخدام تقنيات وبرمجيات تخدم آلية تصنيف البيانات وتقدم لها ما يمكن أن تستفيد منه دون إضاعة الوقت والجهد، لذا فإن هذا المجال تطور بشكل كبير في العشر سنوات الأخيرة و لعل ذلك يعود إلى الطلب الواسع لمستخدمي هذه التكنولوجيا.

إن التصنيف الآلي لهذه للنصوص وفق تقنيات التعلم والخوارزميات يقدم الحل الأمثل لمشكلة التزايد الهائل للبيانات النصية فهي تكنولوجيا جديدة تهدف إلى تنظيم وتصنيف النصوص المتراكمة التي لا يمكن بأي حال من الأحوال معالجتها يدويا