

أهمية المدونات الحاسوبية في المعالجة الآلية للبيانات النصية

The importance of computer blogs in the automated processing of text data

العربي بوعمران بوعلام^{1*}، نعيمة عيوش²¹ جامعة خميس مليانة (الجزائر)، b.larbi-bouamrane@univ-dbk.m.dz² جامعة خميس مليانة (الجزائر)، n.ayouche@univ-dbk.m.dz

تاريخ الاستلام: 2022/10/25 تاريخ القبول: 2022/11/02 تاريخ النشر: 2023/05/04

الملخص:

لقد كانت محاولات ربط اللغة بالوسائل التكنولوجية الحديثة مخاضا لولادة العديد من الفروع التقنية اللغوية من بينها اللسانيات الحاسوبية، وكذلك المعالجة الآلية للغات الطبيعية والذكاء الاصطناعي، التنقيب في البيانات، الهندسة اللغوية، والتي أخذت على عاتقها استحداث برامج آلية تخدم اللغة ليس قصد تطويعها وإخضاعها وإنما لتطويرها، ولعل تزايد البيانات المكتوبة باللغات الطبيعية ومحاولة رقمنتها والاستعانة بالحواسيب لتنظيمها ولد العديد من التقنيات كالمترجم الآلي والملخص الآلي وطرق التعرف الموضوعي والتصنيف الآلي، كما أدى إلى ظهور العديد من خوارزميات التعلم كأشجار القرار والعقد والشبكات العصبونية التي سهلت سبل التعامل مع اللغات الطبيعية.

الكلمات المفتاحية: المدونات، الحاسوبية- المعالجة الآلية- التصنيف الآلي.

Abstract:

Attempts to link language with modern technological means have been the throes of the birth of many linguistic technical branches, including computational linguistics, as well as automatic processing of natural languages and artificial intelligence, data mining, linguistic engineering, which took upon themselves the development of automated programs that serve the language not with the intention of subjugating and subjugating it, but rather to develop it. Perhaps the increase in data written in natural languages and the attempt to digitize it and use computers to organize it gave birth to many technologies such as automatic translator,

* المؤلف المرسل.

automatic summary, objective identification methods, and automatic classification.

Keywords: computer blogging - automated processing - automated classification

1. مقدمة:

مع تطور العلوم وتفرعها وظهور تكنولوجيات متعددة تميز عصرنا الحالي بتدفق كم هائل من المعلومات في شتى المجالات، مما أدى إلى صعوبة الوصول إلى المعرفة إذ لم تعد الطرق التقليدية قادرة على استخلاص المعلومات نظرا لضخامتها، هذه الأسباب وغيرها ولدت الحاجة إلى إيجاد تقنيات ذات فعالية تعمل على استكشاف المعرفة وإدارة هذه المعلومات، نتيجة لذلك ظهرت العديد من العلوم التي عملت على إيجاد تقنيات وأدوات وخوارزميات تسهل إجراءات الوصول للمعلومات، من خلال كمية هائلة من البيانات بطريقة سهلة وسريعة ودقيقة جدا، كما تعمل على بناء تنبؤات مستقبلية لحل المشكلات.

إن الكميات الضخمة من البيانات المتوفرة حاليا هي عبارة عن بيانات نصية والتي تتألف بدورها من مجموعة كبيرة من المستندات من مصادر مختلفة مثل مقالات من صحف إخبارية، بحوث إلكترونية، رسائل البريد الإلكتروني، صفحات الويب إذ نجد هذه البيانات النصية تنمو بشكل كبير نظرا للتزايد الهائل من البيانات الرقمية على شبكات الانترنت وبالتالي أصبح هناك ثغرة واسعة بين البحث والمعالجة لاسترجاع معلومة من هذه المستندات وبين كمية البيانات الهائلة لذا تولدت الحاجة إلى التنقيب في البيانات النصية (النصوص).

لذا أولت البحوث في السنوات الأخيرة الكثير من الاهتمام لمعالجة البيانات النصية وهذا عائد لعدة أسباب من بينها تزايد مجموعة البيانات على شبكات التواصل وتطوير البنية التحتية للاتصالات والانترنت مما أدى إلى الحاجة الماسة لتنظيم ومعالجة كميات ضخمة من البيانات إذ أن المعالجة اليدوية لهذه البيانات مكلفة للغاية في الوقت والأفراد كما أنها ليست مرنة وتعميمها إلى ميادين أخرى مستحيلة عمليا لذلك كان لابد من تطوير أساليب آلية.

يتم الاعتماد في المعالجة الآلية للبيانات النصية على المدونات الحاسوبية كونها توفر بيانات ضخمة قابلة للمعالجة السهلة وذلك بالاعتماد على مجموعة من الأدوات والبرامج الحاسوبية التي تتيح العديد من الخوارزميات التي تعمل على الجمع والاحصاء والتصنيف والاختبار والتنبؤ ويمكن كذلك أن تؤدي إلى استخلاص نتائج دقيقة ومنظمة.

2- ماهية المدونات الحاسوبية:

يمكن تعريف المدونة **corpus** بأنها قواعد بيانات تنظيمية ضخمة نسبيا لمجموعة من النصوص الإلكترونية، التي تجمع وتصمم وفق معايير محددة للاستخدام في البحث اللغوي، ويمكن تقسيم أنواع المدونات حسب التعداد

اللغوي، فقد تكون المدونة أحادية اللغة، أي تحتوي على نصوص بلغة واحدة فقط، كما قد تحتوي على لغتين فأكثر. (سلطان بن ناصر المحيول، 2016م، ص 170)

تتيح المدونات الحاسوبية تسهيلات ضخمة في مجال البحث اللغوي كما أنها توفر منهجية أكثر تفوقاً من أجل تحليل كثير من الظواهر اللغوية البارزة في العربية المعاصرة، كما أنها توفر بيانات هائلة قابلة للمعالجة السهلة بالاعتماد على الامكانيات والأدوات الحاسوبية في الجمع والاحصاء والتصنيف والاسترجاع، ويمكن كذلك أن تؤدي إلى استخلاص نتائج اختبارية موثوقة ومنظمة قابلة للتكرار والتراكم والمقارنة والتعميم، وهو ما يساعد على بناء النظريات واختبارها وتهذيبها على المدى الطويل، ويتيح التحليل الآلي للمدونات أنواعاً متعددة ومختلف من الدراسة اللغوية، التي يمكن أن تكون مفيدة في تحليل ثلاثة جوانب أساسية من الاستعمال اللغوي وهي: التكرار، والتنوع والسياق النصي. (عقيل بن حامد الشمري وآخرون، دت، ص 95)

3- طرق بناء المدونات الحاسوبية:

هناك العديد من المعايير المستخدمة لبناء المدونات كالتى تتمثل في حجم المدونة، ومدى تنوع اللغة المستخدمة، بالإضافة إلى خصائص النصوص وتختلف هذه المعايير فيما بينها، غير أنها تشترك في المجال الذي تستخدم فيه المدونة، والفترة الزمنية المغطاة، وطريقة استخدام هذه المدونة بشكل منهجي مقنن، كما تعتمد المدونة على جمع أكبر قدر ممكن من النصوص من مختلف المجالات، إلا أنه مؤخراً زاد الاحتياج إلى مدونات أصغر تهتم بجزء معين من النصوص، تنحصر في مجال محدد، ويطلق عادة على هذا النوع بالمدونات المختصة. (سلطان بن ناصر المحيول، 2016م، ص 95)

هناك العديد من الطرق المستخدمة في بناء المدونات والتي تركز بالأساس على المصادر التي يعتمد عليها مثل:

- استثمار ما تتيحه الشبكة العنكبوتية من بيانات نصية.
- الاعتماد على مدونات حاسوبية جاهزة.
- الجمع اليدوي للبيانات النصية وتحويلها لبيانات إلكترونية.

4- أساليب المعالجة الآلية للمدونات الحاسوبية:

إن أية معالجة لنص لغوي تعتمد اعتماداً كلياً على لغته لذا فمن الضروري أن تكون أدوات البحث قادرة على التعرف آلياً على لغات الوثائق أو النصوص التي تتعامل معها لذا ظهرت العديد من التقنيات في هذا المجال.

1- التجريد: يستخدم التجريد في معالجة اللغات الطبيعية واسترجاع البيانات وتصنيف المستندات إذ يعمل على تقليل عدد المزايا وبالتالي تقليل حجم المستند وزيادة سرعة التعلم والتصنيف خاصة فيما يتعلق بالمصنفات التي تعتمد على مجموعة بيانات نصية ضخمة.

هناك أربع طرق مختلفة للتجريد في اللغة العربية وهي

- التجريد الخفيف الذي يعمل على إزالة السوابق واللاحق من دون المساس بالجذع أو تحديد الوزن أو استنتاج الجذر.
- التحليل المورفولوجي لإيجاد الجذور: تعمل هذه المحلات الصرفية على إيجاد الجذور المحتملة للكلمة استنادا إلى مجموعة من القواعد التحليلية.
- التجريد الإحصائي : يحاول أن يجمع مختلف تشكيلات الكلمة باستخدام تقنيات العنقدة التي تعتمد بشكل أساسي على مبدأ التشابه الصرفي بين الكلمات
- بناء قواميس بشكل يدوي: تعطي لنا معلومات عن التجريد وهي عبارة عن قواميس للجذور والأوزان (مصعب شاهين، شادي صالح، 2011/2012، ص10)

2-مصنف الزناد: تعرف مجموعة زنادات كلمة ما بأنها الكلمات التي لديها ترابط قوي بينها وبين تلك الكلمة ويتم حساب هذه الزنادات باستعمال المعلومة المتبادلة المتوسطة لكل زوج من الكلمات التي تنتمي لمجموعة المفردات والزنادات الأكثر أهمية تلك التي تملك قيم كبرى للمعلومة المتبادلة المتوسطة الموافقة لكل زوج من الكلمات وبالتالي يصبح كل موضوع ممثلا بعدد من الزنادات المستخرجة من مدونة التدريب.(مراد عباس، 2011م، ص185)

3- تقنية TF-IDF: تقوم هذه التقنية على تمثيل كل وثيقة d بمتجه $D = (d1, d1, \dots, d/V)$ حيث يرمز $|V|$ إلى حجم مجموعة المفردات ويتم حساب مركبات المتجه عن طريق ضرب تكرار اللفظة $TF(w, d)$ الذي هو عبارة عن عدد المرات التي تظهر فيها اللفظة w في الوثيقة d بعكس تكرار الوثيقة $IDF(w)$ ويمثل تكرار الوثيقة $IDF(w)$ عدد الوثائق التي تظهر فيها اللفظة w مرة واحدة على الأقل . وتعرف القيمة di بوزن اللفظة wi في الوثيقة d وتعطى كالآتي:

$$di = TF(w, d) * IDF(w) \text{ مع } IDF(w) = \log(N/DF(w))$$

و N هو عدد الوثائق.

ولحساب التشابه $sim(Dj, Di)$ الموجود بين الوثيقة Di والموضوع Dj تستعمل المعادلة وتنسب الوثيقة إلى الموضوع الذي يحصل على أكبر قيمة تشابه $sim(Dj, D)$.

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|V|} (d_{jk})^2 \sum_{k=1}^{|V|} (d_{ik})^2}}$$

4-تقنية ن غرام: يستخدم هذا المصطلح في العديد من المجالات مثل أنظمة التعرف الآلي على الكلام بقيم نموذجية ل n تساوي 3 أو 4 وتستخدم أيضا في أنظمة المعالجة الآلية للغة في ميدان البحث عن المعلومات ويمكن

ل ن غرام أن تشير إلى حد سواء إلى ن غرام حروف أو ن غرام كلمات في هذا النموذج يتم تمثيل النصوص بمتجهات ال ن غرام عوضاً عن متجهات الألفاظ عناصرها تواترات ال ن غرام في النصوص المطابقة لها. و ال ن غرام هو سلسلة ألفاظ متتالية بالنسبة لأي نص مجموعة ال ن غرام التي يمكن استخراجها هي النتيجة التي نحصل عليها عن طريق تحريك خلال النص نافذة مكونة من ن مربعات وتتم هذه الحركة على مراحل مرحلة واحدة تقابل حرف واحد لكل ن غرام حروف وكلمة لكل ن غرام كلمات ثم نحسب ن غرام المتحصل عليها. (عبد المالك أمين، 2013، ص 37)

اقتصرنا على ذكر بعض التقنيات إلا أن هناك العديد من الدراسات التي بحثت في هذا المجال ووضعت تقنيات جديدة تتعامل بدقة مع النصوص بناء على لغتها.

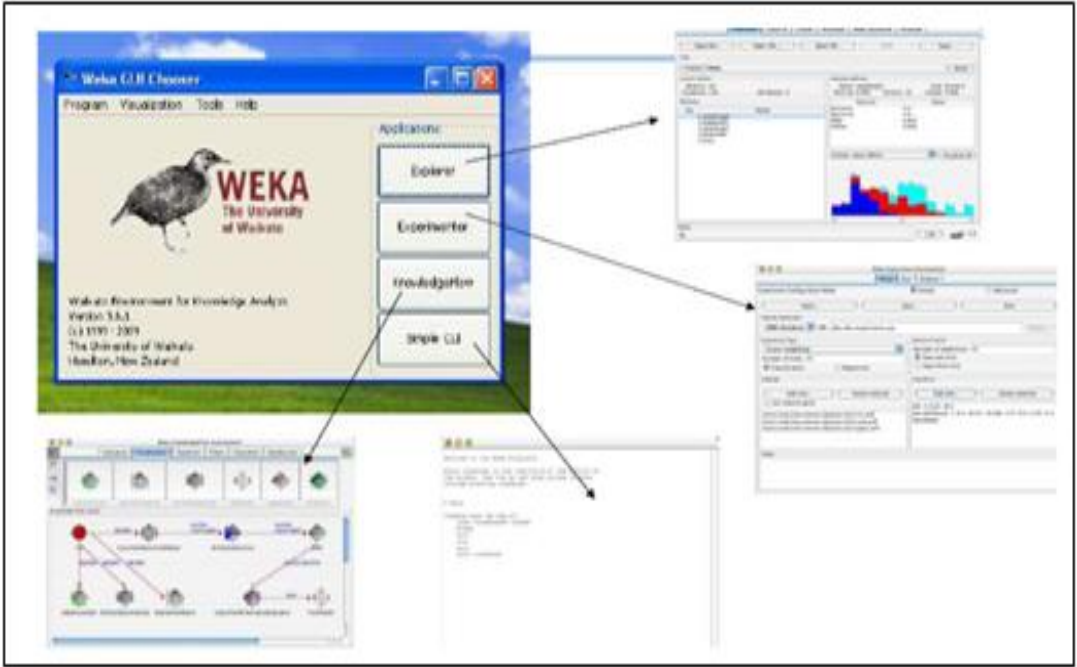
5- التصنيف الآلي للبيانات النصية بالاعتماد على المدونة الحاسوبية:

نسعى من خلال هذه الدراسة إلى بناء مدونة تحوي بيانات نصية عربية ارتأينا أن تكون نصوص شعرية عربية ونصوص روائية، ثم العمل على تصنيف هذه البيانات إلى فئتين تحدد الأساليب الخيرية والانشائية في هذه النصوص، وذلك قصد بناء نموذج استكشافي تنبؤي لتحديد الفئات التي تنتمي إليها هذه النصوص، إعتدنا على برنامج حاسوبي متطور يشمل مختلف الأدوات الاحصائية وكذا مجموعة من الخوارزميات المشهورة كما يتيح أدوات اختبار لهذه الخوارزميات، يتمثل في برنامج WEKA .

1- تطبيق WEKA (Waikato Environment for Knowledge Analysis):

تطبيق حديث ومتجدد يؤمن إجراء عمليات التنقيب في البيانات بسهولة، له العديد من الميزات التي تجعله أحد أهم التطبيقات المعروفة للتنقيب في البيانات في الوقت الحاضر من خلال تزويده بمجموعة ضخمة من خوارزميات التعلم الآلي التي يمكن أن تطبق مباشرة على مجموعة البيانات أو يتم استدعاؤها من التعليمات البرمجية الخاصة بالجافا، ويحتوي أيضا على أدوات تجهيز البيانات وتحويلها.

تم تطوير تطبيق WEKA في مختبر جامعة (Waikato) في نيوزلندا بالاعتماد على لغة البرمجة (JAVA) وكانت النسخة الأولى عام 1996 ويعتبر أحد أفضل برمجيات التنقيب في المعطيات المعروفة حتى الآن حيث يعمل بالتوافق مع كافة أنظمة التشغيل المعروفة (Windows, Linux , Macintosh) ويمكن الوصول والتعامل مع كافة عناصر هذا التطبيق من خلال لغة البرمجة جافا.



الشكل (01): صورة توضيحية لواجهة عمل تطبيق WEKA

تقوم الدراسة على مجموعة من الخطوات اللازمة نذكرها فيما يلي:

أ/ **جمع المدونة الحاسوبية:** قمنا في هذه الخطوة بجمع نصوص عربية مختلفة المصادر قسمت إلى ثلاث مجالات نصوص قرآنية تحوي شواهد آيات طويلة وقصيرة تتلاءم مع متطلبات الدراسة، أخذت هذه النصوص القرآنية من مصحف المدينة المنورة للنشر الحاسوبي حيث يتيح هذا المصحف مجموعة من الإمكانيات إذ يسمح للمستخدم من إضافة النص القرآني إلى الوثائق وملفات النصوص مع احتفاظها بخصائصها، كما يمكننا من البحث عن كلمة أو أكثر في نص القرآن الكريم لاستعراض المواضيع المختلفة التي جاءت فيها هذه الكلمة أو الجملة، وأيضا قابلية نسخ الآيات القرآنية إلى برنامج وورد، وسبب الاعتماد على النصوص القرآنية بالدرجة الأولى يرجع إلى كونه زاهر بالأساليب الإنشائية والخبرية.

البحال الثاني خصص للشواهد الشعرية لكون أن الأمة العربية هي أمة شاعرة بالدرجة الأولى، يتيح لنا هذا المصدر كم هائل من البيانات حاولنا أن نختار ما يتناسب مع الدراسة مع الأخذ بعين الاعتبار الاستشهاد بنصوص من عصور مختلفة بدء بالعصر الجاهلي حتى المعاصر وكذا من دول عربية مختلفة، اقتبست مجمل الشواهد الشعرية من موسوعة الشعر العربي الالكترونية (ديوان) والتي تضم حوالي 25693 قصيدة لـ 449 شاعر، نجد في هذه الموسوعة تسهيلات كثيرة للإبحار فيها بعد عملية الاشتراك والانضمام إليها.

أما المجال الثالث من المدونة خصصناه لمقتطفات من النصوص الروائية لكون أن كل من الشعر والرواية هي نصوص أدبية تحفل بمختلف الأساليب الفنية، مصدر هذه النصوص كان من الموقع الإلكتروني (حكايًا للرواية العربية) ، الذي يتيح العديد من الروايات والقصص والحكايات كما يسمح بتحميلها وطبعها ونسخها من دون قاعدة الاشتراك.

الطريقة التي اعتمدناها في جمع النصوص على شكل مدونة لها العديد من المزايا من بينها:

- أنها أكثر عملية كونها تحوي نصوص حقيقية.
- تحتوي العديد من النصوص التي تخضع للتطبيق مما يساعد في ظهور نتائج يقينية.
- تنوع النصوص وفقا لأسس علمية لتمثل كل مؤشرات الأساليب الخيرية والإنشائية مع مراعاة مختلف الاستعمالات اللغوية.

- الاستخدام الحاسوبي في التعامل مع نصوص المدونة "وهذا هو الغرض من الدراسة".

- الاعتماد على أساليب التحليل الكمي (إحصاء التكرارات).

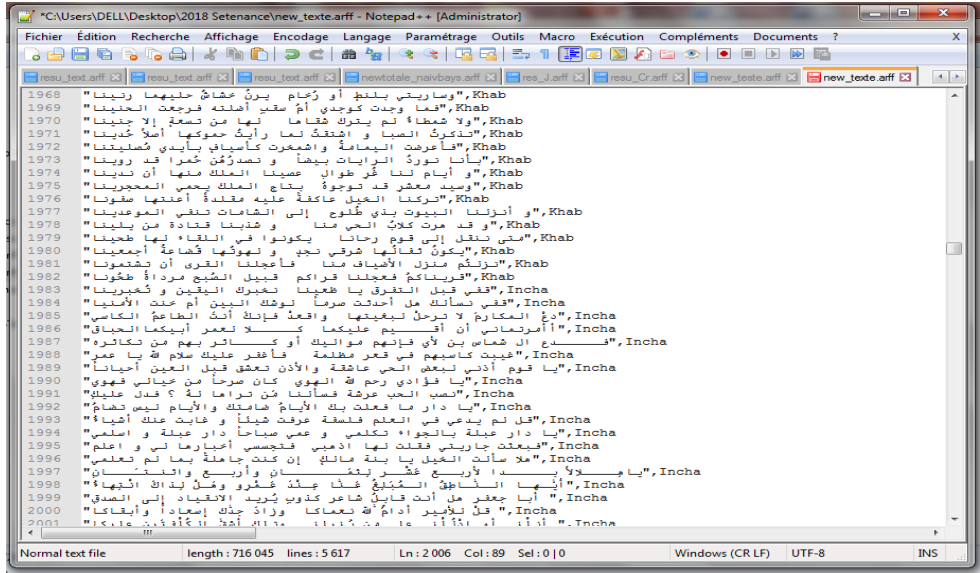
ب/ **تصفية بيانات المدونة:** تتعلق هذه المرحلة بعملية تنظيف البيانات وتنقيتها من الشوائب فبعض البيانات تكون فارغة كعلامات الترقيم " النقطة ، الفاصلة، علامة الاستفهام والتعجب، الأقواس، المزدوجتان، والأرقام فكلها ليست لها أهمية أثناء التصنيف وقد تؤدي إلى تدني كفاءة المصنفات، لذا يتم حذفها إما يدويا أو بطريقة آلية فمثلا في البرنامج الذي اعتمدناه في هذه الدراسة **weka** تكون عملية تصفية البيانات بطريقة آلية مباشرة بعد إدخالها عليه.

ومن مزايا هذه المرحلة:

- أنها تعمل على تخفيض كمية البيانات إلا أنها لا تمس بالبيانات الأساسية.

- إزالة الشوائب التي قد تعيق عملية تصنيف البيانات.

ج - **تحويل البيانات :** فالبرامج الآلية المعتمدة في عملية معالجة البيانات تستدعي التعامل مع نوع معين من الملفات، مما يقتضي ضرورة تحويل شكل البيانات بهدف إدخالها إلى البرنامج وطبعها مما هو معروف فإن برنامج **weka** يتعامل مع ملفات **Attribute Relation File Format (arff)** وبالتالي كان لا بد من تحويل جميع البيانات من صيغة (**txt**) إلى صيغة (**arff**) والبرنامج بدوره يوفر مجموعة من الصفوف أو الأعمدة تسهل عملية إدخال هذا الملف، إلا أنه أحيانا أخرى يتم الاستعانة ببرامج أخرى تعمل على تحويل البيانات مثل برنامج **Notepad++** . الشكل الموالي يوضح نص بصيغة **arff**:



الشكل (02): تحويل البيانات إلى صيغة ARFF

د/ التنقيب في البيانات: وهي مرحلة تأتي بعد إدخال البيانات إلى البرنامج لإيجاد آلية قادرة على تصنيف النصوص، إذ تم تغيير نوع البيانات واستخدام أكبر قدر من الخوارزميات المتاحة وقسمت البيانات إلى حزمتين بيانيتين الحزمة البينانية الأولى للتدريب وتحتوي على نصوص قرآنية وشعرية وروائية أي ما يعادل نسبة 80 % من البيانات، أما الحزمة البينانية الثانية فتحتوي على ما يعادل نسبة 20 % من البيانات، هذه الحزمة ستترك جانبا لاستخدامها في مرحلة التنبؤ، يتم في هذه المرحلة بناء نموذج التصنيف، تتم هذه العملية بخطوتين رئيسيتين هما: التصنيف وهو خطوة بناء واختبار نموذج التصنيف، والتنبؤ وهو استخدام النموذج ليتنبأ بفئات البيانات غير معروفة الفئة وبذلك تكون خطوات التصنيف كما يلي:

- يتم تدريب خوارزمية التصنيف (التعلم) على بيانات التدريب **Training Data Set** المحتوية على سجلات معروفة الفئة لبناء النموذج الذي يستخدم لاختبار حزمة بيانات التي تحتوي على سجلات غير معروفة الفئة .
- يتم تقييم أداء النماذج باستخدام خوارزميات التصنيف والهدف هو الحصول على أعلى دقة وأقل نسبة خطأ عند تطبيقها على بيانات الاختبار ومن خلال ذلك يتم بناء أفضل نموذج (**Model**) ويتم تنفيذ النموذج الذي تم بنائه على حزمة بيانات التطبيق **Apply Data Set** والتي تحتوي على النصوص غير معروفة الفئة وبعد إظهار النتائج يتم الاستفادة منها في التنبؤ المستقبلي.

هـ / مرحلة التقييم: بعد تحديد مجموعة التدريب وتحديد الفئات يأتي دور تقييم واختبار الفئة وهو مهم جدا لمعرفة هل هذه الفئة التي تم تدريبها يمكن استخدامها في تصنيف بيانات جديدة أم لا، ويمكن تحديد إمكانية إعادة

استخدام الفئة التي تم تدريبها من عدمه من خلال نسبة تصنيفه في مرحلة الاختبار، فكلما كانت عالية كانت إمكانية استخدامه أكبر، حتى يمكننا تقييم أداء مصنف لا بد من إجراء اختبار بإعطائه مجموعة من البيانات التي تم تدريبه عليها مسبقا ويتحدد ذلك من خلال النسبة المثوية للتصنيف، فكلما كانت النسبة عالية كانت دقة تصنيفه جيدة وتتم عملية التقييم واحتساب الدقة باعتماد أربعة طرق:

- **Using Training set**: تعتمد هذه الطريقة في عملية اختبار المصنف على البيانات التدريبية وتغطية وتحميل (label) الذي تم إعطائه للبيانات أي التسمية التي أعطيت للبيانات المرجو تصنيفها.

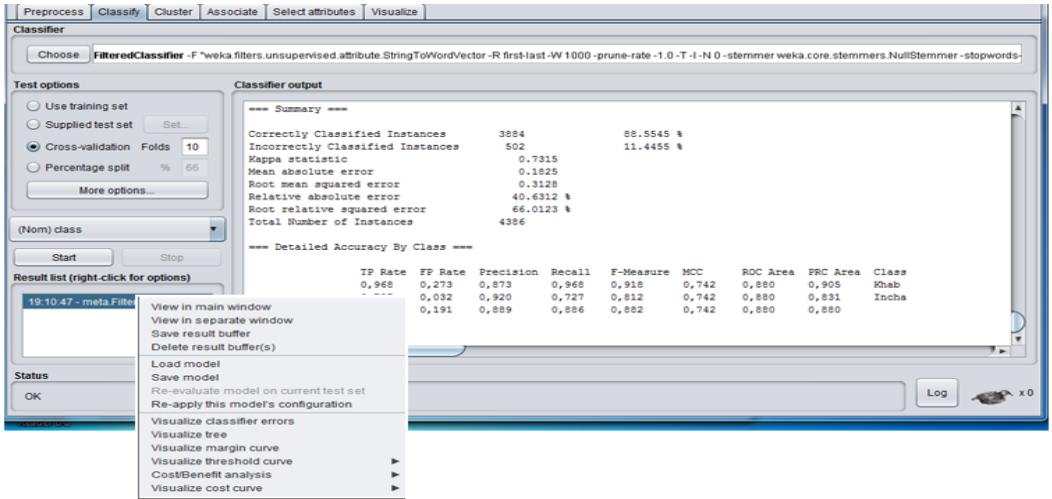
```

2072 Time taken to build model: 60.25 seconds
2073
2074 === Evaluation on training set ===
2075
2076 Time taken to test model on training data: 0.19 seconds
2077
2078 === Summary ===
2079
2080 Correctly Classified Instances      3971          90.5381 %
2081 Incorrectly Classified Instances    415           9.4619 %
2082 Kappa statistic                    0.7795
2083 Mean absolute error                 0.1629
2084 Root mean squared error             0.2854
2085 Relative absolute error             36.2852 %
2086 Root relative squared error         60.2387 %
2087 Total Number of Instances          4386
2088
2089 === Detailed Accuracy By Class ===
2090
2091      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
2092      0.976    0.231    0.891     0.976    0.932     0.788    0.911     0.923    Khab
2093      0.769    0.024    0.943     0.769    0.847     0.788    0.911     0.863    Incha
2094 Weighted Avg.    0.905    0.161    0.909     0.905    0.903     0.788    0.911     0.903
2095
2096 === Confusion Matrix ===
2097
2098      a  b  <-- classified as
2099      2823  70 |  a = Khab
2100      345 1148 |  b = Incha
2101

```

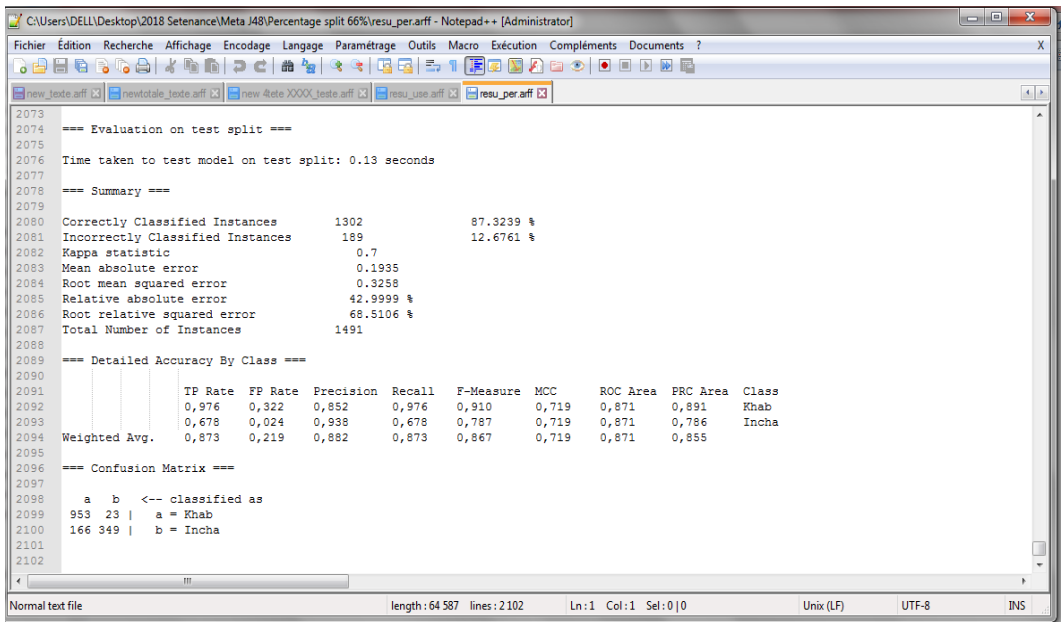
الشكل (03): اختبار Using training set للمصنف J48

- **Cross Validation**: تقوم هذه الطريقة على تقسيم البيانات حسب الرقم الذي يعطى (fold) ومن ثم يتدرب على 90% من هذه البيانات المقسمة و10% يتم اختبار المصنف عليها، مثلاً لدينا مجموعة تدريب عددها 100 يتم أولاً تقسيم البيانات حسب (fold) إلى عشرة مجموعات إذ يتم تدريب المصنف على تسع مجموعات والعاشرة يختبر عليها، ومن ثم يتم إجراء الاختبار على مجموعة أخرى وإدخال المجموعة السابقة التي اجري عليها الاختبار ضمن مجموعة التدريب وهكذا حتى يتم اختبار المصنف عشر مرات.



الشكل (04): اختبار Cross validation للخوارزمية J48 Meta

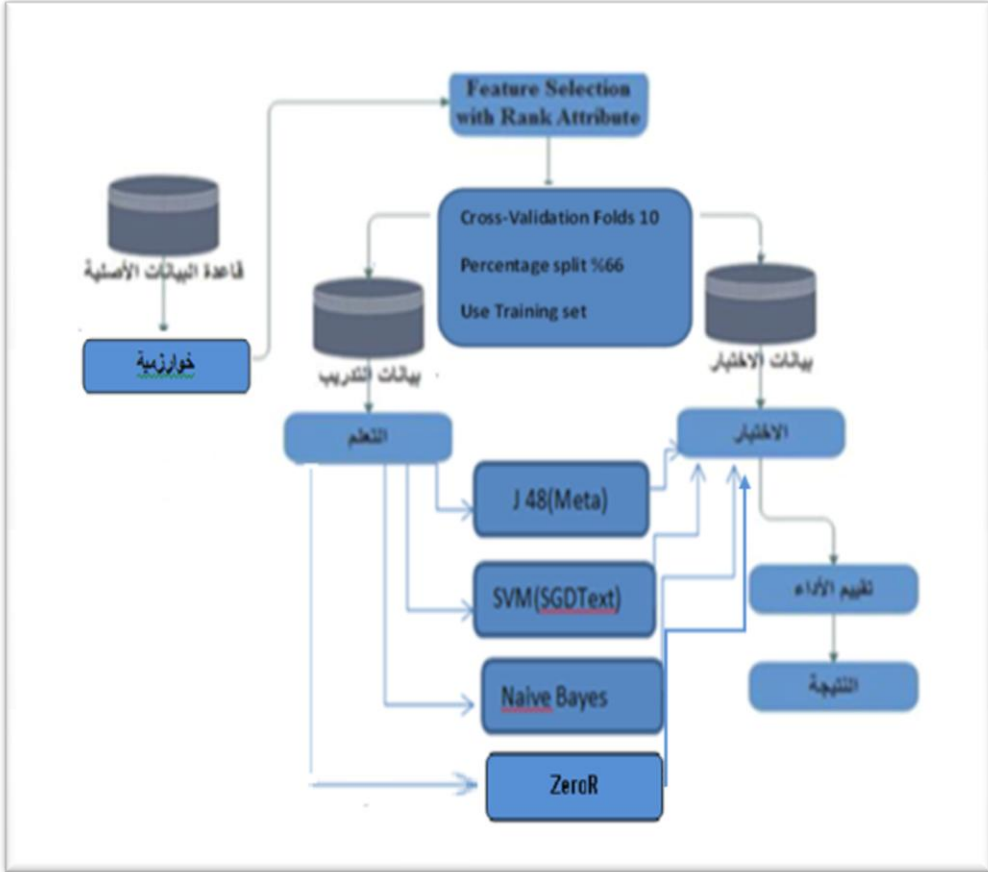
- **Percentage Split** : يتم في هذه الطريقة تقسيم مجموعة البيانات إلى قسمين مجموعة للتدريب ومجموعة للاختبار.



الشكل (05): اختبار Percentage split للمصنف J48

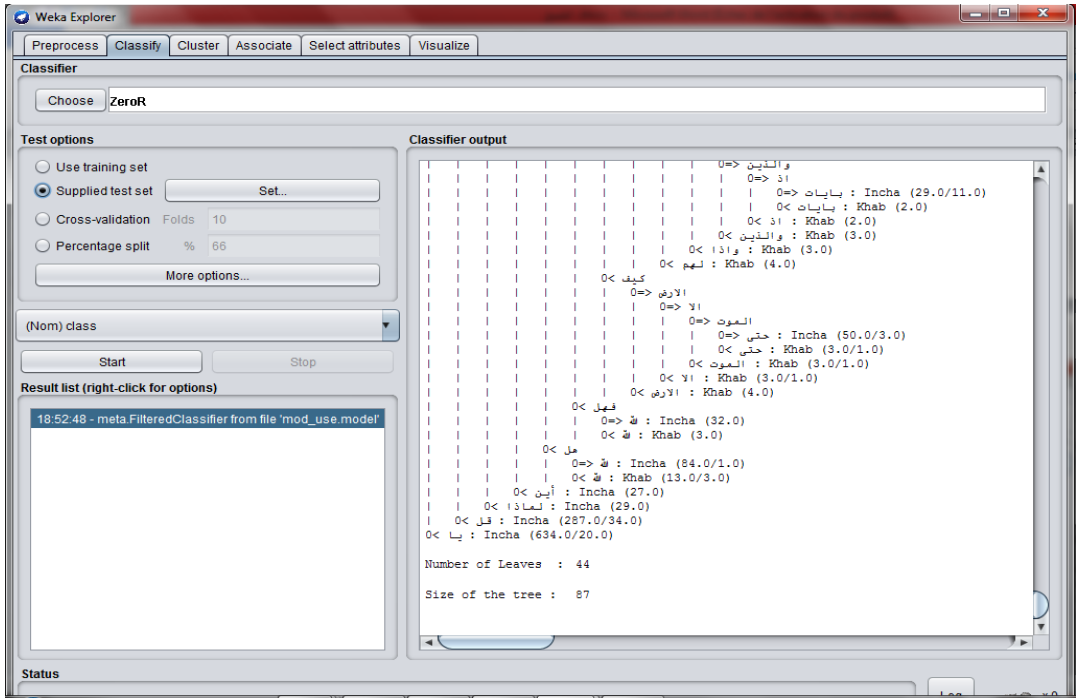
- **Supplied test Set**: تقوم هذه الطريقة على إدخال بيانات جديدة على المصنف ويتم الاختبار عليها وعادة ما تستخدم هذه الطريقة في عملية التنبؤ.

وفي عملنا هذا تم تجربة أربع خوارزميات لتحديد أفضلها بغية استخدامها في عملية تصنيف النصوص الأدبية (الخبرية والإنشائية)، ومن أجل معرفة وتحديد دقة تصنيف البيانات تم اعتماد ثلاث طرق للاختبار **PercentageSplit, Validation-Cross, UsingTrainigset**.



الشكل(06) : النموذج المقترح لمعالجة المعطيات التي تم جمعها

7- التنبؤ: الغرض من استخدام آلية التنبؤ على مدونة بيانات التطبيق هو الكشف أو التنبؤ بفئات البيانات غير معروفة الفئة (التصنيف المتوقع حدوثه)، وذلك من خلال تطبيق النموذج الذي تم بناءه خلال مرحلة التصنيف ويكون ذلك على حزمة البيانات الجديدة وللإختبار على الحزمة البينانية التي تم الاحتفاظ بها سابقا للإختبار والتي تقدر ب 20% من نصوص المدونة من أجل إستخدامها.



الشكل (07): استخدام طريقة Supplied Test Set للتنبؤ

6- خاتمة: تظهر هذه الدراسة أهمية المدونات الحاسوبية في إدارة البيانات الضخمة، وكذا فعاليتها في توجيه البحث اللغوي المعاصر، حيث أصبح أثر دقة وشمولية، نظرا للأدوات والبرامج التقنية التي تعدها للتنقيب في هذه البيانات وقد تطرقنا لذكر بعض منها في دراستنا هذه.

7. قائمة المراجع:

- سلطان بن ناصر المحيول، 2016م، لغويات المدونة الحاسوبية، ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، السعودية.
- عقيل بن حامد الشمري وآخرون، استخدام المدونات الحاسوبية في التحليل اللغوي، ورقة بحثية إلكترونية، الجمعية المصرية للقراءة والمعرفة، عضو الجمعية الدولية للمعرفة.
- مصعب شاهين، شادي صالح، 2012/2011، مشروع تخرج بعنوان: تطوير نظام لتصنيف المستندات العربية، إشراف الدكتور ناصر ناصر، جامعة تشرين سوريا، قسم البرمجيات ونظم المعلومات.
- مراد عباس، 2011م، تقييم طرق التعرف الموضوعي للنصوص العربية، مجلة الخليج العربي للبحوث العلمية، 29(4/3).

- عبد المالك أمين، 2013، مقارنة لتحديد اللغات تلقائيا في مدينة نصوص متعددة اللغات، المجلة العربية الدولية للمعلوماتية، المجلد الثاني، العدد الرابع.
- موقع موسوعة الشعر العربي (الديوان): <https://www.aldiwan.net>
- موقع حكايا للرواية العربية : <https://hakaya.com>