Istanbul Journal of Arabic Studies (ISTANBULJAS)
Volume 3, Issue 1, 2020-1, p. 89-111.

Automatic Classification of Arabic Text Data By Using Computer Programs "Weka Program"

التصنيف الآلي للبيانات النصية العربية باستخدام البرامج الحاسوبية "برنامج WEKA"

Larbi Bouamrane BOUALEM*

*Professor of Higher Education, Faculty of Arts and Languages, University of Khamis Miliana, Algeria, (b.larbi-bouamrane@univ-dbkm.dz), https://orcid.org/0000-0001-9778-2974

http://dergipark.org.tr/istanbuljas

Submission /Başvuru: 26 November/Kasım 2019
Acceptance /Kabul: 06 June/Haziran 2020
Article Type/Makalenin Türü: Research Article/ Araştırma Makalesi

Abstract: Manual handling of the vast amount of data without the use of modern techniques keeps us from evolving and upgrading to better performance. It is not enough just to introduce machines to work, but it is better to use techniques and software that serve the mechanism of data classification and provide them with what can benefit from it without wasting time and effort.Therefore, this field has evolved considerably in the last ten years due to the wide demand of users of this technology. The automated categorization of the texts according to learning techniques and algorithms offers the best solution to the problem of the huge increase of textual data.

Keywords: Algorithms, machine learning, classification, processing, data.

التصنيف الآلي للبيانات النصية العربية باستخدام البرامج الحاسوبية "برنامج WEKA"

العربي بوعمران بوعلام*

ملخص

إن التعامل اليدوي مع الكم الهائل من البيانات دون استخدام تقنيات حديثة يبعدنا عن التطور والارتقاء 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، ولعل ذلك يعود إلى الطلب الواسع لمستخدمي هذه التكنولوجيا. إن التصنيف الآلي للبيانات النصية 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.

الكلمات المفتاحية

خوارزميات، التعلم الآلي، التصنيف، المعالجة، البيانات.

* أستاذ التعليم العالي، كلية الآداب واللغات، جامعة خميس مليانة، الجزائر، (b.larbi-bouamrane@univ-dbkm.dz)

Extended Abstract

We all know that our time was marked by the development and branching of knowledge fields, accompanied by the emergence of many modern technologies that serve humanity. On the other hand, the world witnessed a tremendous growth in the amount of information flowing in various fields. Find mechanisms that can manage this information in terms of its classification and retrieval methods.

There are many sciences that have worked in this field such as data exploration, automatic processing of languages, language engineering, artificial intelligence, coincided with the emergence of many computer programs and machine learning algorithms that facilitate information access procedures, and work to build future predictions to solve problems.

So this study came to shed light on the field of automatic classification of textual data, based on the algorithms of machine learning that we find available on the WEKA program, which is a modern and renewed application that secures easy data mining operations, provided with a huge set of algorithms, and also provides a set of tests and measures that calculate the accuracy of works in percentages of classified cases.

The aim of this study is to know the efficiency and effectiveness of algorithms in the process of classifying texts, and work to build a predictive model for similar cases, and for that we have followed a set of important steps that we mention as follows:

- Defining the field of study and the problem to be researched and finding solutions to it.

- Collecting data by listing a large amount of Arabic texts in order to build a database or blog.

- Choosing the appropriate data for the study in order to process it.

- Filtering and purifying data from impurities and removing duplicate data, spaces and punctuation marks that may hinder classification.

- Converting the data format and encode it according to a set of programming codes, in order to facilitate its entry into the program.

- Data exploration, which is the stage where work begins by selecting the algorithm, defining the categories and beginning the process of training the algorithm in order to identify cases and classify them according to the specified categories.

- Evaluation: is the stage of evaluating the algorithm's performance with a set of tests and criteria that the program provides.

- Prediction: is the last stage in which we are working to build a model that predicts similar situations.

التصنيف الآلي للبيانات النصية العربية باستخدام البرامج الحاسوبية "برنامج WEKA"

مقدمة

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، 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 أدّى إلى ظهور العديد من خوارزميات التعلم كأشجار القرار (tree decision)، والعنقدة (clustrering algorithms)، والشبكات العصبونية (work net neural)، التي سهلت سبل التعامل مع اللغات الطبيعية، ولعل أكثر الأدوات فعالية في قدرتها على التعامل مع اللغة هي تلك البرامج الحاسوبية الشائعة مثل برنامج WEKA،Rapidminer،Tanagra، تلم هذه البرامج بعدد هائل من الخوارزميات والأدوات المستحدثة والتقنيات المتعددة، كالتنبؤ والتصنيف والإحصاء، كما تعمل على تحليل كميات ضخمة من البيانات بسرعة فائقة، والتوصل بالطبع إلى نتائج دقيقة.

رغم توافر العديد من الخوارزميات والبرامج الحاسوبية مازال يعترض الباحث العربي مجموعة من العقبات والصعوبات التي تحول دون إيجاد آليات تتناسب وخصائص اللغة العربية، وذلك ربما يعود لعدة أسباب قد تتعلق باللغة العربية نفسها.

نسلط الضوء في هذه الدراسة على التصنيف الآلي للنصوص الأدبية العربية، نحاول أن نقدم مجموعة من الآليات لتصنيف البيانات النصية باستخدام خوارزميات التعلم

الآلي، ويتمحور هدف الدراسة على بناء نموذج تنبؤي للأساليب الخبرية والإنشائية، استعنا بمجموعة من الدراسات والبحوث العربية من بينها:

- دراسة محمد سعيد الدسوقي (2014) بعنوان" تطبيق العنقدة المتعددة المستويات على نص القرآن الكريم".
- دراسة خلوف وآخرون (2009) بعنوان "استخدام آليات التنقيب في المعطيات للمساعدة في اكتشاف عمليات الاحتيال في البيئة المصرفية".
- دراسة مراد عباس وآخرون (2011) بعنوان "تقييم طرق التعرف الموضوعي للنصوص العربية".

أوجه صعوبة المعالجة الآلية للنصوص العربية:

مما لا شك فيه أن محاولة إخضاع اللغة للحاسوب لا بد وأن يعترضها العديد من الإشكاليات والعقبات، وعندما تتشابه العقبات في لغات عديدة فإنه بلا شك تتشابه طرق حلها، غير أن تحليل اللغة العربية بواسطة الحاسوب يكتنفه عقبات كثيرة، أكثر من أي لغة أخرى، ومعظم هذه المشاكل متعلقة بالجوانب التي تختلف فيها العربية عن اللغات الأوروبية، تلك اللغات التي صممت معظم البرامج الحاسوبية أصلاً لتحليلها.

ولا شك أن محاولة قولبة اللغة العربية في الحاسوب من أهم المشاكل التي تعترض طريق المعالجة الآلية واللسانيات الحاسوبية، وذلك لما تتميز به العربية عن بقية اللغات الأخرى 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[1]، ومن ثم التصنيف الآلي لمجمل النصوص العربية. يمكن أن نحصر أهم هذه المشاكل فيما يلي:

- الاستخدام المفرط للأساليب البيانية (المجاز– الكناية – الاستعارات).
- عدم وجود فوارق شكلية واضحة بين مكونات النص.
- افتقار اللغة العربية لمبدأ الوحدة الدلالية.
- عدم وجود علامات التشكيل.
- الأخطاء اللغوية الشائعة.
- تعقيد خوارزمية التعلم.
- التجانس اللفظي.
- التغيرات الصرفية.
- الكلام المركب.

بالرغم من هذه الصعوبات التي قد تعيق عملية تطبيق أهم التقنيات التكنولوجية على اللغة العربية، إلا أن البحوث مستمرة، وهناك العديد من البحوث التي قدمت تقنيات حاسوبية (آلية) حاولت أن تعطي حلولا قيمة لعملية حوسبة اللغة العربية، وكذلك تم تطويع العديد من المناهج الغربية والخوارزميات حتى تناسب اللغة العربية.

[1] زينب هاشم، "أثر البرمجيات الحديثة على اللغة العربية"، *مجلة العلوم الإنسانية*، 02 (2015)، ص 234.

التصنيف الآلي للبيانات النصية:

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 والإنترنت، 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 لا بد من تطوير أساليب آلية تعمل على إدارة هذه البيانات النصية (النصوص)، فظهر ما يسمى بالتصنيف الآلي للنصوص العربية.

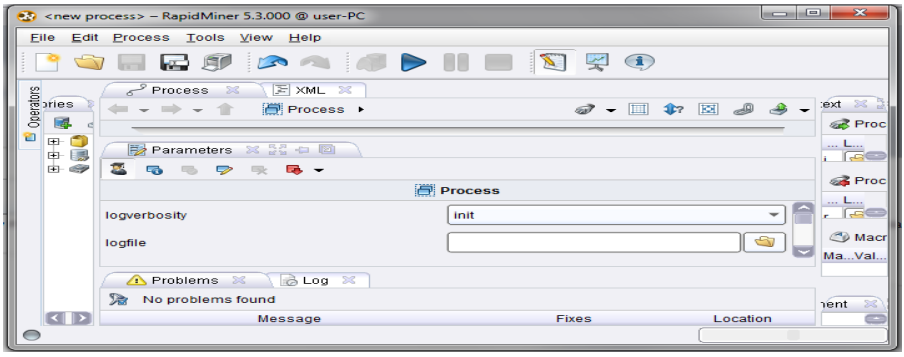
إن التصنيف الآلي وفق تقنيات التعلم والخوارزميات يقدم الحل الأمثل لمشكلة التزايد الهائل للبيانات النصية، فهي تكنولوجيا جديدة تهدف إلى تنظيم وتصنيف النصوص المتراكمة التي لا يمكن بأي حال من الأحوال معالجتها يدويا.

يعد التصنيف الآلي أحد فروع المعالجة الآلية للغة، وقد تزايد الاهتمام به في الآونة الأخيرة، نظرًا لتزايد حجم البيانات ذات المحتوى النصي، لذا ظهرت العديد من التقنيات والأدوات والخوارزميات التي تعمل على معالجة النصوص آليا، منها الربط بين الكلمات والمقاطع في النصوص، وتصنيف النصوص ضمن موضوعات محددة مسبقا، لذا يمكن تعريف التصنيف الآلي للنصوص (AutomaticTextCategorization) هي مهمة تصنيف المستندات النصية الإلكترونية اوتوماتيكيا إلى أصنافها المعرفة مسبقا بحسب محتوياتها، بمعنى آخر تحديد الصنف الرئيسي الذي يندرج تحته النص أو المستند "سياسة ، اقتصاد ، رياضة، ... الخ".

أدوات وخوارزميات التصنيف الآلي:

ظهرت العديد من البرامج والأدوات التي تقوم بعملية التصنيف الآلي التي تعمل على معالجة كميات ضخمة من البيانات بكفاءة ودقة عالية ومن هذه البرامج:

1-**برنامج Rapidminer:** يعتبر من البرامج المجانية مفتوحة المصدر صمم من قبل شركة Rapid-IGermany يعمل بلغة الجافا، يتوفر هذا البرنامج على واجهة رسومية سهلة الاستخدام مقارنة ببرامج أخرى، إذ لا يستلزم الأمر صعوبة في التعامل مع هذا البرنامج، يتيح هذا البرنامج جملة من الخوارزميات المعروفة لمعالجة كميات ضخمة من البيانات.



الشكل(1): الواجهة الرسومية لبرنامج Rapidminer

2- برنامج Clementine: صمم هذا البرنامج من قبل شركة (SPSS) يتوفر هذا البرنامج على مكتبات كاملة لتنقيب البيانات بواسطة مختلف خوارزميات التصنيف والتحليل العنقودي وقواعد اكتشاف العلاقات والارتباطات، يتصف هذا البرنامج بسهولة الاستخدام والتعلم.

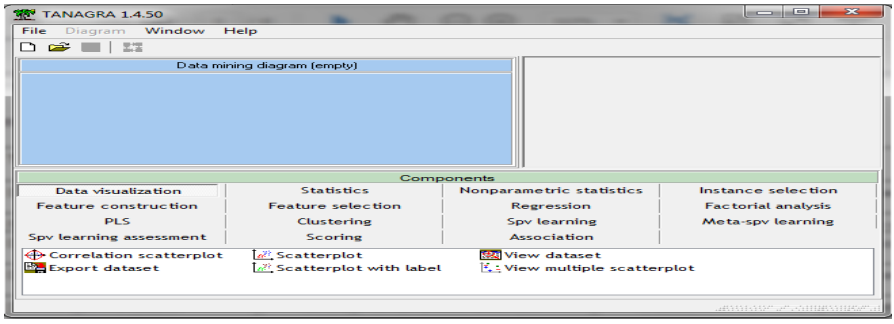
3- برنامج WEKA: يعتبر من البرامج المجانية مفتوحة المصدر، تم تصميم هذا البرنامج في جامعة ويكاتو بنيوزلندا جاء بهذا الاسم اختصارا لـ WekatoEnvernoment for the KnoweledgeAnalysis يعمل بلغة الجافا، يتميز بقدرته على معالجة كمية

هائلة من البيانات، يمدنا بمجموعة كاملة لمختلف الخوارزميات المعروفة في هذا المجال.

الشكل(2): الواجهة الرسومية لبرنامج WEKA

4- برنامج Rattle: يعتبر هذا البرنامج من البرامج مفتوحة المصدر، صمم من قبل شركة TogawareAustralia يعمل بلغة (R)، تنفرد هذه الأداة بتضمينها حجم كبير من البيانات، ما يؤخذ على هذا البرنامج عدم مرونته في التعامل مع البيانات.

5- برنامج Tanagra: يعتبر من البرامج مفتوحة المصدر صمم من قبل شركةLumière University Lyon- France يعمل بلغةC^{++}، سهل الاستخدام، يتوفر على مجموعة من الخوارزميات، إلا أن ما يعاب على هذا البرنامج هو أنه يعرض البيانات والنموذج بشكل ضعيف.



الشكل(3): الواجهة الرسومية لبرنامج Tanagra

تتوفر هذه البرامج على أهم الخوارزميات المعروفة والمستحدثة، منها مصنفات أشجار القرار والمصنفات القاعدية المعتمدة على القاعدة (rule-based)، والشبكات العصبية (neural net work)، ومكائن الإسناد الموجه (support vector machines)، ومصنفات بيز الاحتمالية (bues classifer).

تستخدم كل تقنية من التقنيات السابقة كخوارزمية تعلم، لتحديد نموذج يلائم العلاقة بين مجموعة الصفات ومؤشر الصنف لبيانات الإدخال، حيث يتم توليد النموذج من خلال خوارزمية تعلم، ويجب على كل من النموذج والخوارزمية أن يتلاءما مع البيانات المدخلة بصورة جيدة، والتنبؤ بصورة دقيقة لمؤشرات الصنف، لذلك فإن الهدف الرئيسي لخوارزمية التعلم هو بناء نماذج يمكن تعميمها، أي نماذج تتنبأ بشكل دقيق بتسميات أصناف سجلات غير معروفة مسبقا[2]. سنكتفي في هذه الدراسة بتسليط الضوء على خوارزمية **SVM** لأننا سنعتمدها في هذه الدراسة، ولكونها من بين أكثر الخوارزميات رواجا.

خوارزميةSVM (support vector machines):

تعد من أشهر طرق التصنيف الآلي، والتي تعتمد على إيجاد منحنى أو مستوى فاصل، يفصل العينات المدخلة عن بعضها البعض، وتتميز باستخدامها في تصنيف البيانات ذات الفئات الثنائية حصرا، تقوم الخوارزمية بحساب المستوى الفاصل أو مجموعة المستويات الفاصلة في بعد يختلف طوله عن طول بعد متجه خصائص البيانات المدروسة، وتحدد دقة الخوارزمية بقدرتها على الفصل بين النوعين، بحيث تكون أقرب عينتين من كلا النوعين أبعد ما يكون عن بعضهما البعض، وندعو هذا

[2] محمد حسن عبد الله، "تنقيب بيانات نتيجة التعليم الأساسي"، مذكرة ماجستير في تقانة المعلومات، كلية الدراسات العليا، جامعة النيلين، 2016، ص53.

المستوى الفاصل بالهامش، فكلما زاد هامش الفصل كلما قل الخطأ عند التعميم على مجموعة بيانات جديدة[3].

في دراستنا هذه سنحاول عرض أهم مراحل التصنيف الآلي من خلال تصنيف نصوص عربية أدبية، وفق فئتين؛ فئة إنشائية، والأخرى خبرية، بالاعتماد على خوارزمية SVM، والتي نجدها متاحة على برنامج WEKA، والهدف من هذه الدراسة معرفة كفاءة هذه الخوارزمية في التعرف على النصوص ذات الأساليب الخبرية والإنشائية، وبناء نموذج نعتمده في مختلف الدراسات المستقبلية.

الخطوات المتبعة في عملية التصنيف الآلي:

حتى نقوم بعملية التصنيف الآلي لا بد من إتباع مجموعة من الخطوات اللازمة، إذ قبل الشروع في العملية لا بد من تحديد مجال الدراسة، والمشكلة المراد بحثها، وإيجاد حلول لها، تأتي بعد ذلك مجموعة من المراحل المهمة نفصلها فيما يلي:

1- جمع البيانات: وهي مرحلة تجميع البيانات النصية الأدبية، وذلك بشكل عشوائي، والشروع في بناء قاعدة بيانات نستخدمها للتصنيف.

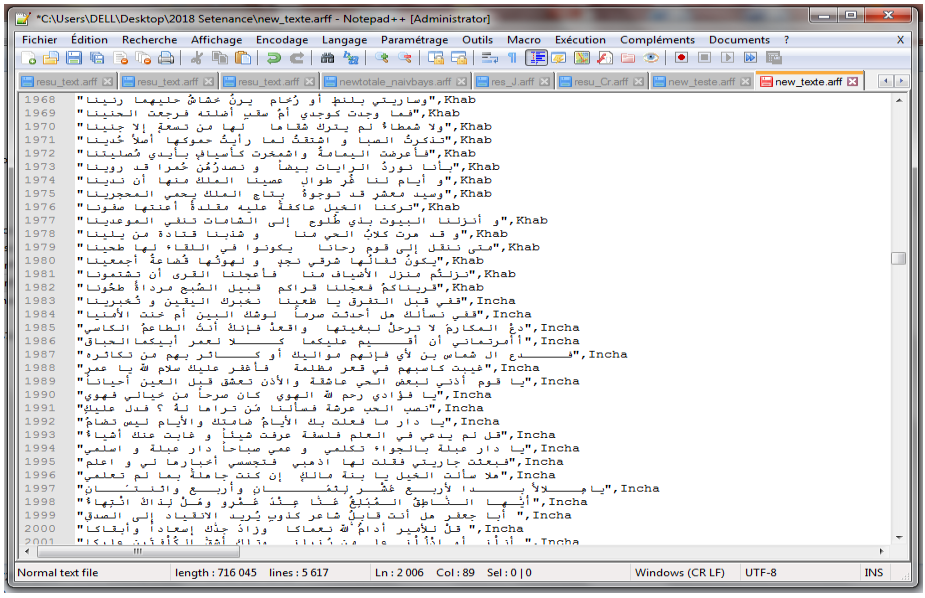
2- اختيار البيانات Data Selection: يتم في هذه المرحلة تعيين واختيار البيانات الملائمة من مجموع البيانات قصد معالجتها.

3- تصفية البيانات وتنقيتهاData Cleaning: يتم في هذه المرحلة حذف البيانات الزائدة التي لا تشكل أهمية أثناء الدراسة، وتشتمل على التخلص من الحقول المتكررة،

[3] بسام الديب، "تصنيف النصوص العربية باستخدام الخصائص الغرضية في قواعد البيانات"، *مجلة جامعة البعث*، 15، (2016)، ص 116.

إزالة البيانات المزعجة التي تعيق عملية التصنيف، تعيين البيانات غير المكتملة، تحديد الفراغات وإزالتها، حذف علامات الترقيم.

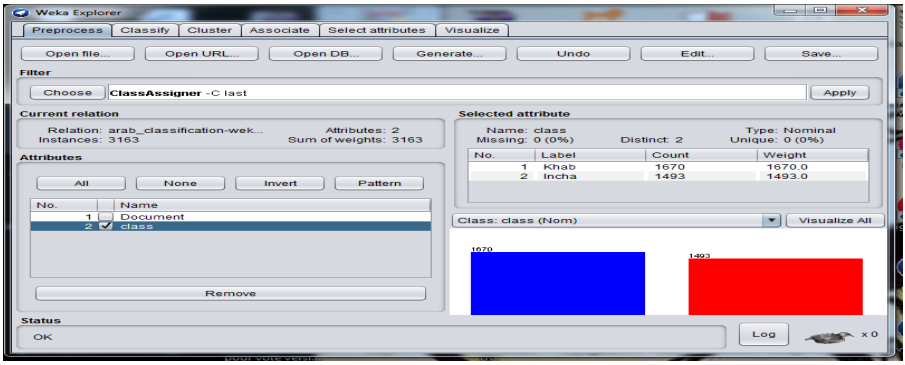
4- تحويل البيانات Data Transformation: يتم في هذه المرحلة تحويل صيغة البيانات من **(txt)** إلى صيغة **(arff)**، استندنا في هذه العملية إلى برنامج **notepad++**، إذ يعتبر من أفضل برامج تحرير النصوص وترميزها وفق مجموعة من الأكواد البرمجية، قمنا بترميز هذه البيانات بكود **UTF-8** حتى يسهل إدخالها إلى برنامج weka، إذ أن هذا البرنامج هو أجنبي لا يتعامل مع اللغة العربية، إذ لا بد من تشفير هذه النصوص العربية، ثم إدخالها إلى البرنامج.



الشكل(4): عينة من البيانات المأخوذة للتدريب بعد إدخالها برنامج notepad++

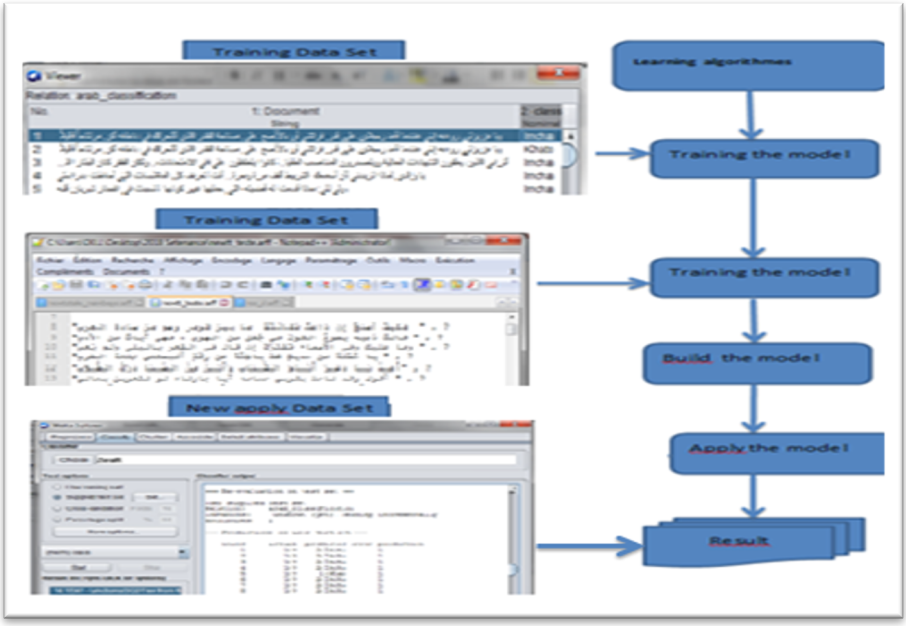
بعد إدخال البيانات إلى برنامج Weka نشرع في اختيار الخوارزمية المراد العمل بها SVM، بعد ذلك نختار تعليمة filters تظهر مجموعة من المؤشرات نكبس على unsupervised، بعد ذلك لا بد من تعيين الفئة class assigner ثم الضغط على تعليمة

Apply فتظهر أمامك الشاشة الموضحة للفئتين المدخلتين، كما هو مبين في الشكل التالي.



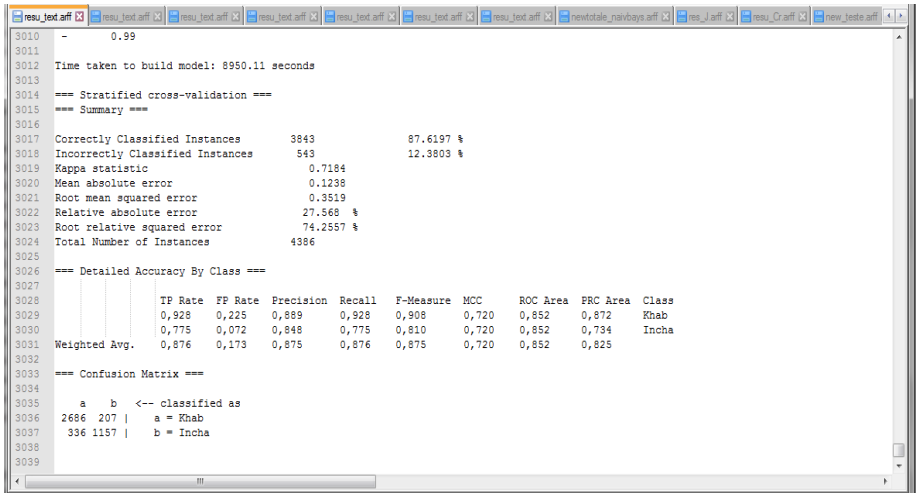
الشكل(5): تحديد الفئات من قبل البرنامج Weka

5-التنقيب في البيانات Data Mining: تعتبر هذه المرحلة الأهم حيث يتم فيها تنفيذ العمل وبناء النماذج للتنبؤ، إذ بعد تعيين الفئات من قبل الخوارزمية نقوم بعملية التدريب على حزمة بيانات التدريب



الشكل(6): خطوات بناء النموذج

6- التقييم Pattern Evaluation: يتم في هذه المرحلة تحديد النموذج النهائي وتطبيقه واستخراج النتائج، نقوم بتقييم أداء هذه الخوارزمية بمجموعة من الاختبارات والمعايير التي يتيحها لنا البرنامج، سنكتفي بعرض نتائج اختبار Cross Validation، ويتحدد ذلك من خلال النسبة المئوية للتصنيف، فكلما كانت النسبة عالية كانت دقة تصنيفه جيدة، سنقوم باستظهار نتائج اختبار هذه الخوارزمية من خلال الشكل التالي:



```
resu_text.arff | resu_text.arff | resu_text.arff | resu_text.arff | resu_text.arff | resu_text.arff | resu_text.arff | newtotale_naivbays.arff | res_J.arff | resu_Cr.arff | new_teste.arff

3010   -       0.99
3011
3012  Time taken to build model: 8950.11 seconds
3013
3014  === Stratified cross-validation ===
3015  === Summary ===
3016
3017  Correctly Classified Instances        3843               87.6197 %
3018  Incorrectly Classified Instances      543                12.3803 %
3019  Kappa statistic                         0.7184
3020  Mean absolute error                     0.1238
3021  Root mean squared error                 0.3519
3022  Relative absolute error                27.568  %
3023  Root relative squared error            74.2557 %
3024  Total Number of Instances             4386
3025
3026  === Detailed Accuracy By Class ===
3027
3028                  TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
3029                  0,928    0,225    0,889      0,928    0,908      0,720    0,852     0,872     Khab
3030                  0,775    0,072    0,848      0,775    0,810      0,720    0,852     0,734     Incha
3031  Weighted Avg.   0,876    0,173    0,875      0,876    0,875      0,720    0,852     0,825
3032
3033  === Confusion Matrix ===
3034
3035     a    b   <-- classified as
3036   2686  207 |    a = Khab
3037    336 1157 |    b = Incha
3038
3039
```

الشكل(7): اختبار Cross validation للمصنف SVM

من خلال قراءة الشاشة الموضحة في الشكل(7) والتي تظهر لنا مختلف النتائج المتوصل إليها بعد اختبار وتقييم أداء هذه الخوارزمية.

– السطر الأول والثاني من الشاشة يظهر عدد الحالات المصنفة في اختبار (Cross Validation) بشكل صحيح هو 3843 حالة بنسبة مئوية مقدارها 87,6197% وعدد الحالات المصنفة بشكل غير صحيح هو 543 حالة بنسبة 12,3803%.

– السطر الثالث يمثل مقياس لتصحيح احتمال الاتفاق بين التصنيفات الحقيقية إحصاء كابا (Kappa Statistiques) والتي كان مقدارها 0,7184 حيث $K = \frac{P0-Pe}{1-Pe}$

– السطر الرابع نجد Meanabsoluterror (الخطأ المطلق في المتوسط) ويستخدم معدلات الخطأ للتنبؤ الرقمي بدلا من التصنيف حيث ان التنبؤات ليست فقط الصحيحة و الخاطئة Meanabsoluterror=0,1238

– السطر الخامس Rootmeansquarederror جذر متوسط مربع الخطأ يساوي 0,3519.

- السطر السادس Relative absoluterror الخطأ المطلق النسبي يساوي 27,568%.

- السطر السابع Root relative squarederror هو الجذر التربيعي للخطأ النسبي يساوي 74,2557%.

مقاييس تقييم أداء الخوارزمية:

يتم معرفة أداء الخوارزمية من خلال مجموعة من مقاييس الأداء، التي تعمل على تحديد النسبة المئوية للحالات المصنفة بشكل صحيح، مع توضيح نسبة الحالات المصنفة بشكل خاطئ، تظهر مقاييس الأداء في شاشة Accuracy class التي توضح نتائج دقة الفئات المصنفة، وهي فئة الخبري والإنشائي، نقوم بعرض نتائج أهم المقاييس التي تظهر في الشاشة الشكل(7):

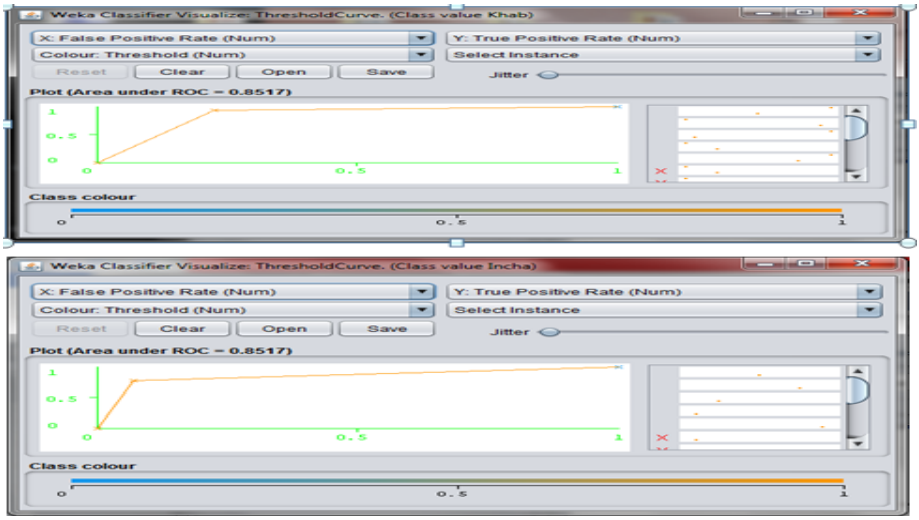
1- مصفوفة الشك (التشويش)Confusion matrix: تعتبر من أهم مقاييس الأداء، يكمن دورها في تقييم أداء المصنف بحساب عدد الحالات المتوقعة المصنفة بشكل صحيح، والمصنفة بشكل خاطئ، وهي عبارة عن جداول تحتوي على قيم التصنيفات الحقيقية والخاطئة للخوارزمية المعتمدة في الدراسة، والتي يظهرها الشكل التالي:

		التصنيف المتوقع	
		A	B
التصنيف الحقيقي	A	2686	207
	B	336	1157

Matrix Confusion الشك مصفوفة SVM Cross-validation(10)

الشكل (8): مصفوفة الشك بالاستعمال خوارزميةSVM

2- مقياس ROC: يعتبر من بين المقاييس المستخدمة بكثرة لمعرفة فعالية أداء المصنف من خلال مخطط يظهر معدل القيم الايجابية الصحيحة والخاطئة، بحيث يحوي المخطط على نقطة (0-1) كلما اقترب منحنى الحالات من النقطة 1 كان أداء المصنف مثالي، وكلما اقترب من 0 كان أداء المصنف ضعيفا، ومن خلال الدراسة التي أجريناها حاولنا استخراج معدلات قيم **ROC** لخوارزميةSVM لمعرفة فعاليتها في عملية تصنيف النصوص فكانت النسبة المتوصل إليها هي : 0,8517، والتي تظهر من خلال الشكل التالي:



الشكل(9): منحنى مقياس ROC لخوارزمية SVM

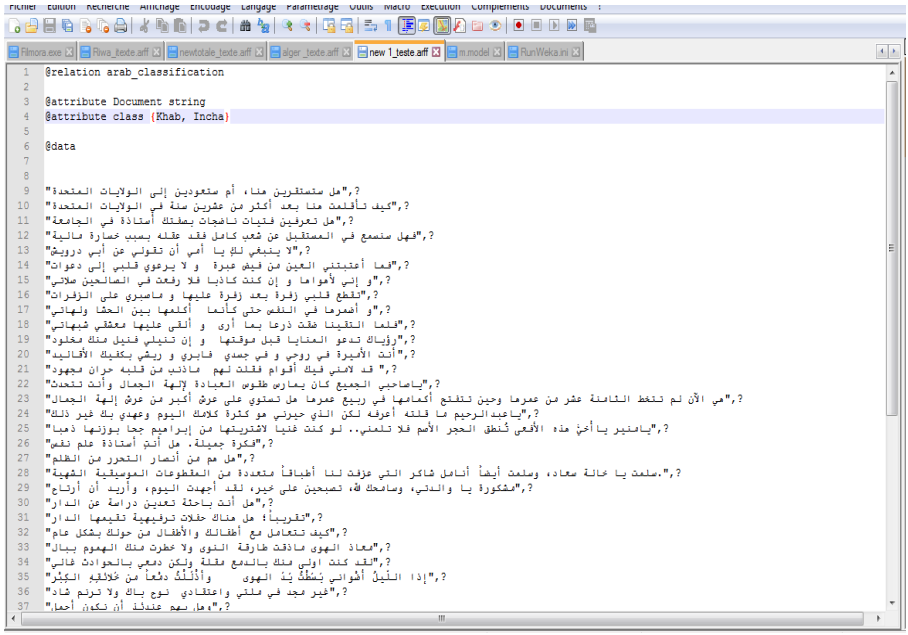
3- مقياس الدقة Recall: وهو تحديد النسبة المئوية للحالات الايجابية التي تم تصنيفها بشكل صحيح، ويتحقق من خلال المعادلة التالية: $Recall = \frac{TP}{TP+FN}$، وتوصلنا إلى النتيجة التالية: **0,876.**

4-مقياس F-Meausure هو مقياس لقياس دقة المصنف يعطى بالعلاقة التالية:

$F - measures = \frac{2*precision*recall}{(presision+recall)}$ وتوصلنا إلى النتيجة التالية: **0,875.**

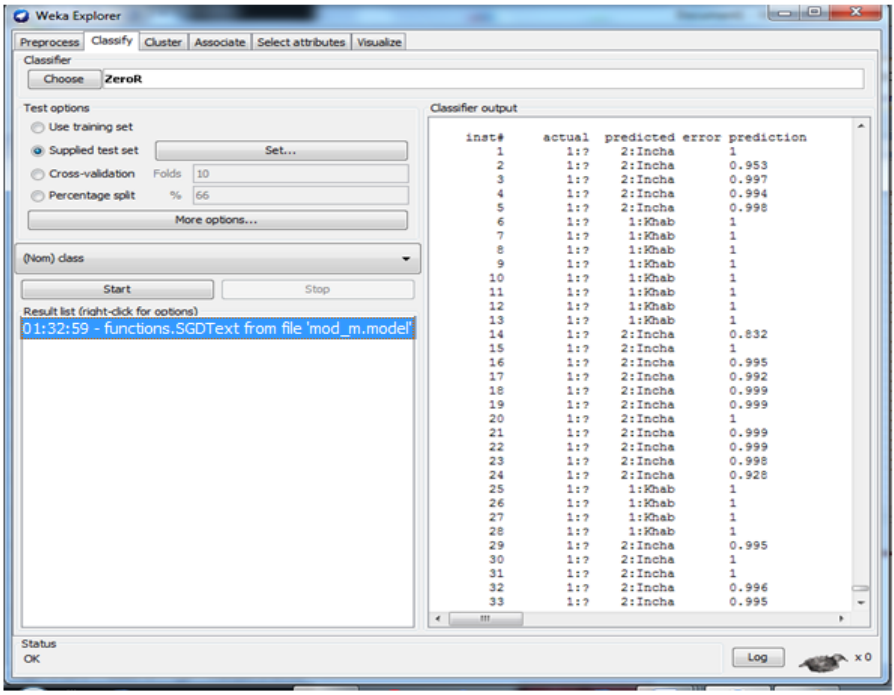
من خلال النتائج المتوصل إليها تبين أن خوارزمية SVM حققت نسبا جيدة في تصنيف الحالات الصحيحة، إذ ومن خلال العينة المختارة للتدريب لاحظنا تعرف هذه الخوارزمية على مختلف النصوص الخبرية والإنشائية، وتصنيفها في الفئات المحددة مسبقا، مع ظهور قصور طفيف في التعرف على النصوص المركبة من مؤشرات إنشائية ومركبات خبرية.

7- **التنبؤ:** الغرض من استخدام آلية التنبؤ على حزمة بيانات التطبيق هو الكشف أو التنبؤ بفئات البيانات غير معروفة الفئة، وذلك من خلال تطبيق النموذج الذي تم بناؤه خلال مرحلة التصنيف، ويكون ذلك على حزمة البيانات الجديدة، وللاختبار على حزمة البيانات التي تم الاحتفاظ بها سابقا، والتي تقدر بـ 20% من نصوص المدونة، من أجل استخدامها للتنبؤ، وجب أن تكون هيكليتها نفس هيكلية بيانات التدريب، لذا تم استخدام طريقة Supplied Test Set لتنفيذ هذه المهمة، وذلك من أجل معرفة دقة تنبؤ الخوارزمية المعتمدة.



الشكل(10): العينة التي تم تخصيصها لعملية التنبؤ

تظهر النتائج المتوصل اليها بعد عملية الاختبار في شاشة « Classifier output» تحت عنوان "Predictions on user test set" بمعنى التنبؤ بالبيانات المستخدمة للاختبار، كما هو مبين في الشكل (10) نتيجة حزمة البيانات التي تم استخدامها للتنبؤ خوارزمية SVM حيث يوضح نتيجة تطبيق هذه الخوارزمية على البيانات الجديدة، إذ يظهر الصف الحقيقي المحتمل (Actual class) وقيمته المجهولة والصف المتوقع (predicted class).



الشكل(11): **التنبؤ على البيانات بخوارزمية SVM**

يوضح الشكل(11) نتائج عملية التنبؤ لخوارزمية SVM والتي أسفرت عن أن تنبؤات هذه الخوارزمية صحيحة إلى حد ما، إذ يتضح لنا من خلال النتائج المتوصل إليها أن خوارزمية SVM حققت نتائج جيدة في عملية التنبؤ بالحالات الجديدة، إذ أن البيانات التي صنفت على أنها أساليب خبرية هي بالفعل خبرية، أما الأساليب الإنشائية فهي بالفعل إنشائية، يتضح ذلك من خلال المقارنة بين الشكل الذي يظهر عينة التطبيق والشكل الذي يظهر ناتج عملية التنبؤ.

إلا أننا لاحظنا قصورا جد طفيف في التعرف على بعض النصوص، إذ نجد المثالين رقم (16–17) الظاهران في العينة صنفتهما خوارزمية SVM في فئة الإنشائي، بحكم وجود مؤشرات إنشائية، وهي أدوات النداء "يا" إلا أننا لو تمعنا الأمر لوجدنا أن هذين

المثالين يحتويان على أكثر من جملة، والمرجح فيها الأسلوب الخبري أكثر من الإنشائي لذا الجائز أن تصنف في خانة الخبري.

ما نخلص إليه أن خوارزميات التعلم هي المنهج الصحيح الذي لا بد وأن ننتهجه في المعالجة الآلية للغة العربية، باعتبار أن هذه الخوارزميات تتيح إمكانات هائلة في التعامل مع اللغة، كما أنها قابلة للتطوير والتعديل ما يجعلها تتناسب مع مختلف الدراسات، بالإضافة إلى أننا نجدها متاحة في مختلف البرامج الحاسوبية.

الخاتمة:

توصلنا من خلال هذه الدراسة إلى مجموعة من النتائج الهامة:

- يعد التصنيف الآلي للنصوص إحدى تقنيات التنقيب في البيانات، يعتمد على مجموعة هائلة من البرامج الحاسوبية والخوارزميات المتطورة.
- يوفر برنامج WEKA سهولة في تصنيف النصوص العربية من خلال مجموعة من الخوارزميات التي يتيحها للمستخدم.
- يوفر برنامج WEKA عامل السرعة والجهد، حيث يمكن الوصول إلى فئات النصوص وأساليبها في غضون ثواني معدودة، مع السماح بتحديث النماذج باستمرار.

المصادر والمراجع

بسام الديب، "تصنيف النصوص العربية باستخدام الخصائص الغرضية في قواعد البيانات"، *مجلة جامعة البعث*، 15، (2016).

زينب هاشم، "أثر البرمجيات الحديثة على اللغة العربية"، *مجلة العلوم الإنسانية*، 02، (2015).

محمد حسن عبد الله، "تنقيب بيانات نتيجة التعليم الأساسي"، مذكرة ماجستير في تقانة المعلومات، كلية الدراسات العليا، جامعة النيلين، 2016.

References

Besām al-Deeb, "Taṣnīf al-nuṣūṣ al-ʿarabīya biʾstiḫdām al-ḫaṣāʾiṣ al-ġaradīya fī qawāʿid al-bayānāt", Majallat Jāmiʿat al-Baʿth, 15, (2016).

Zainab Hāshim, "Athar al-barmajīyāt al-ḥadīthīya ʿalā al-luġa al-ʿarabīya", Majallat al-ʿUlūm al-Insānīya, 02, (2015).

Muḥammad Ḥasan ʿAbdallāh, "Tanqīb bayānāt natījat al-taʿlīm al-asāsī" Mudhakkarat Mājestar fī Taqānat al-Maʿlūmāt, Kullīya al-Dirāsāt al-ʿUlyā, Jāmiʿat al-Neelain, 2016.