جامعة إب
مجلة الباحث الجامعي

إسنادية التأليف في الشعر العربي باستخدام تقنيات تنقيب النصوص

احمد الفلاحي[1]، محمد رمضاني*[2]، ومصطفى بلفقيه[3]

[1]قسم الحاسوب، كلية التربية النادرة، جامعة اب، اليمن

[2]قسم المعلوماتية، كلية العلوم والتقنيات المحمدية، جامعة الحسن الثاني الدار البيضاء، المغرب

[3]المعهد الوطني للبريد والمواصلات INPT، الرباط، المغرب

*E-mail: moha@fstm.ac.ma

الملخص:

يهدف هذا البحث إلى تقديم طريقة لمعرفة الإسناد التلقائي لنص شعري مجهول في الشعر العربي إلى شاعره الحقيقي، وأتَمَتَة هذه العملية باستخدام تقنيات تنقيب النصوص Text Mining، الجدير بالذكر أن هذا البحث تطوير لمشروع متكامل في عملية إسناد التأليف في الشعر العربي وهي عملية مهمة جدا في تنقيب النصوص العربية. في هذا البحث تمّ إدخال مجموعة من الدواوين الشعرية لعدد ستين شاعرًا من مختلف العصور في الشعر العربي الكلاسيكي بوصفهم مجتمعاً للتدريب وإدخال العدد نفسه لنصوص مجهولة المؤلف من نصوص مختلفة بوصفها عينة اختبار؛ ثم طُبقت خوارزميات Naïve Bayes (NB),Support Vector Machine (SVM) على تلك النصوص مع بارامترات ومتغيرات هي: القافية، الحرف، طول الكلمة، وطول الجملة الشعرية، الكلمة الأولى والبحر الشعري. وقد خرج البحث بنتيجة دقة وصلت إلى 96,667 %.

الكلمات المفتاحية Key words: الشعر العرب Arabic Poetry، إسناد التأليف Authorship Attribution خوازميات, SVM, NB.

1. المقدمة: Introduction

التعامل مع النص الشعري العربي باستخدام الآلة ليس مهمة سهلة، لما للشعر من سمات تختلف عن النصوص العربية الأخرى، ناهيك عن أن الشعر يتعرض للانتحال والسرقة والوضع أكثر من غيره[(1)].

لقد حظيت اللغات الأجنبية مثل الإنجليزية وغيرها بالدراسة والتمحيص باستخدام تقنيات الحاسوب بشكل متتابع؛ بينما اللغة العربية لم تنل حقها في هذا الجانب إذ إن الدراسات والأبحاث في هذا المجال لا تتجاوز أطروحة دكتوراه ومقالات علمية منشورة سوف نتناولها في الأدبيات السابقة؛ وقد تركز العمل في تلك الدراسات على النصوص العربية والقصيرة فقط، وتناول البعض منها صفحات الويب والبريد الإلكتروني، ويعدّ ذلك حافزا إضافيًا لمشروعنا، كوننا نتعامل مع نصوص شعرية بتركيبة خاصة[(2)].

يهدف هذا البحث إلى الإسناد التلقائي لنص شعري مجهول أو متنازع عليه في الشعر العربي إلى شاعره الحقيقي، وأتَمَتَة هذه العملية باستخدام تقنيات تنقيب النصوص Text Mining، ويمثل موضوع هذه البحث مرحلة متقدمة من مشروع متكامل لإسنادية التأليف في الشعر العربي[(3)].

يتضمن هذا البحث المحاور الآتية :

إسنادية التأليف، إسنادية التأليف في اللغة العربية، الدراسات، السابقة، المنهجية، وآلية العمل وتطبيق الخوارزميات، فالنتائج والتوجهات المستقبلية، والمراجع.

2. إسنادية التأليف Authorship Attribution

هي عملية إسناد نص مجهول أو مختلف عليه إلى مؤلفه الحقيقي، وهذا هو التعريف البسيط لإسنادية التأليف.

على أن عملية الإسناد تبدأ بدراسة خصائص النص من أجل استخلاص استنتاجات خاصة بتأليفه، وقد نشأ أساسًا عن علم قياس الأسلوب، وهو فرع من علم اللسانيات يطبق القياسات الإحصائية على الأسلوب الأدبي[4]

إن عملية إسنادية التأليف تتمركز حول إظهار أنساق التشابه لمؤلف، كتب نصاً ما بواسطة فحص أعماله الأخرى؛ حيث يتمحور فعل تحديد بصمة المؤلف حول استخلاص مجموعة من السمات للنص والتي تبقى ثابتة نسبيًا في مجموعة كتاباته، والتقاط أسلوبه فيها[5].

3. إسنادية التأليف في اللغة العربية Authorship Attribution in Arabic

إسناد النص إلى مؤلفه في اللغة العربية باستخدام الحاسوب وأتَمَتَة هذه العملية؛ يعد عملا جديدا، و يتطلب تطبيقه التعامل مع التحديات والصعوبات الناتجة عن بعض المواصفات الخاصة بهذه اللغة مثل: اشتقاق المفردات وطبيعة تلك المفردات وتجذيرها وتشكيل الكلمات، وطول الكلمة والحروف وطول الجملة والسمات النحوية والمعجمية [6].

4. خصائص الكتابة والسمات الاسلوبية Stylistic Features

السمات ونمطية الكتابة في نص ما تساعد على اكتشاف بصمة المؤلف وهويته؛ و تنقسم إلى أربع مجموعات:

السمات المعجمية: تكون على مستوى الأحرف ومستوى الكلمات، **السمات النحوية**: يتم التعامل من خلالها مع الأشكال الداخلة في بنية الجملة، **السمات التركيبية أو الهيكلية**: تعكس العادات الخاصة للكاتب والمؤلف في التخطيط والتنظيم لكتاباته، **وسمات خاصة:** ترتبط بمضمون النص مع الكلمات المفتاحية في موضوع معين أو نطاق محدد يتخذه المؤلف منهجًا في نصوصه وكتاباته. ومن أجل إتمام عملية التصنيف تُستخدم طرق و أدوات مختلفة من أبرزها طرق تقنيات تعليم الآلة، الشبكات العصبية، الاحتمالات، التحليل الإحصائي، وشجرة اتخاذ القرار[7].

أ. السمات المعجمية :Lexical features

هي من أكثر السمات المستخدمة لإسناد النص إلى مؤلفه، وتعتمد على طول الكلمة، وطول الجملة، وعدد تكرار الكلمة، ووفرة المفردات، إلا أن هناك مشكلة رئيسة في هذا النوع من السمات في بعض اللغات الشرقية تتمثل في أنه لا توجد حدود فاصلة بين الكلمات، ويصعب تطبيق هذه السمات دون الحاجة إلى أدوات خاصة مساعدة باستخدام الحاسوب.

ب. الحرف :Character

في هذا النوع من السمات يتم الاستناد إلى الأحرف في معالجة النصوص باستخدام تسلسل الأحرف، ويأخذ نوع الحرف Character Type ، ترددات الحرف Character Frequency وتجاور الحرفCharacter N-gram، و يمكن تطبيقها بسهولة في أية لغة دون الحاجة إلى أية أدوات خاصة.

ج. السمات النحوية Syntactic Features

تستخدم السمات النحوية من قبل المؤلفين دون وعي؛ ما يجعلها أكثر موثوقية من السمات المعجمية، وفيها يتم استخدام تدابير مختلفة في الدراسات النحوية من أجل عملية الإسناد بما في ذلك الجزء من الكلام

(POS: Part-Of-Speech) والتكرار والأخطاء النحوية والكلمات الوظيفية Functional Words، وهذه السمات تتطلب تقنيات حاسوبية لاستخراجها.

د.**السمات الدلالية** Semantic Features

تشمل هذه السمات المتعلقات الدلالية والمترادفات اللغوية و(SFL: Systemic Functional Linguistics)، التي تحدد الكلمات الوظيفية Functional Words مع ميزات POS.

هـ. السمات التركيبية Structural Features

هذه السّمات تلتقط عادات المؤلف عند تنظيم النص وبنائه، والأمثلة على هذه التدابير: طول الفقرة، وطول الجملة، واستخدام التوقيع، ولون الخط وحجمه، على أن السّمات التركيبية لا تظهر بوضوح في النصوص القصيرة؛ لأنه من الصعب التقاط الخصائص الأسلوبية للنص وتظهر جلية في النصوص الأطول [8].

5. الأدبيات السابقة Related work لم نجد أي عمل يتناول مشكلة إسنادية النص الشعري إلى مؤلفه بشكل عام، [1] والعربي بشكل خاص، ولكن هنالك بعض الأعمال ذات الصلة[9] ،وسيتم تصنيفها إلى مجموعتين:

الدراسات التي تعاملت مع الشعر العربي Works with Arabic poems: الهدف الأساسي من هذه الدراسات هو التصنيف والتحقق أو استنباط النص الشعري من النصوص المكتوبة.

الدراسات التي تعاملت مع النص العربي Works with Arabic text: هدف هذه الدراسات هو التعامل مع النص العربي باستخدام تقنيات التصنيف.

أ. تصنيف الشعر العربي Arabic poems classification

• في دراسة (2013).Alnagdawi قام ببناء نظام يقوم على اكتشاف البحر الشعري بالاعتماد على علم العروض، ويوفر طريقة لتصنيف القصائد العربية عروضيا إلى ستة عشر بحرا، والنظام يقوم بمساعدة المستخدم للعثور على اسم البحر لأية قصيدة شعرية مدخلة إلى النظام باستخدام السياق النحوي Context Free Grammar (CFG)، وقد ناقش الحلول لبعض المشاكل في البداية باستخدام التعبير العادي Regular Expressionو CFG، وتحصل على نتيجة تصل إلى 75%.[10]

• تناولت دراسة (2009) .AbdulBaki, Iqbal الشعر العربي الكلاسيكي من حيث تصنيفه إلى مدح، وهجاء، وذم وغزل وغيرها، وذلك باستخدام خوارزمية Naïve Bayesللتصنيف مع استخدام عملية Stemming التجذير بوصفه متغيرًا وحيدًا وحققت نسبة نجاح وصلت إلى 90 %.[11]

• في دراسة(2008).Al Hichri قدم نظامًا خبيرا لتصنيف القصائد العربية اعتمادا على بنية الأوزان للمقاطع القصيرة والطويلة، هذه الأوزان هي بحور الشعر، كما أنه استخدم خوارزمية تعتمد على بعض القواعد وتم تطبيقها في الحالات العامة، وتحويل سلسة النتائج إلى النظام الثنائي، واحتساب المسافة بين الأنماط الثنائية في قصيدة شعرية مدخلة ومقارنتها بالسلسلة الثنائية المستخرج من بحور الشعر.[12]

• في دراسة (2013).Almuhareb, Abdulrahman تناول الشعر العربي الكلاسيكي، وبنى أنموذجا يقوم باكتشاف البيت الشعري في صفحات الويب، بحيث يستخدم الطريقة الكلاسيكية للتعرف على القصيدة العربية من حيث شكل الأبيات، والقافية. والطريقة المقترحة حققت نسبة دقة، وصلت إلى 96.94 % من خلال محرك بحث للشعر الكلاسيكي.[13]

ب. تصنيف النص العربي Arabic text classification

- في دراسة Altheneyan, Ala.(2014) طبق خوارزمية Naïve Bayes classifiers على نظام إسنادية التأليف للنص العربي لعملية اختبار ومقارنة أربعة نماذج مختلفة من خوارزمية NB هي: MNB ، MBN ، وMPNB، MBNB واحتمال التقدير يعتمد على وجود سمة من عدمها، في حين تعتمد MPNB و MNB على احتمال تكرار السمة؛ وتم تقييم الأداء على حجم كبير من أربع مجموعات من البيانات المختلفة، ودرس التأثير الناجم عن عملية الإسناد؛ وتُظهر النتائج الإجمالية لديه أن النموذج MBNB يوفر أفضل النتائج بين كل نماذج خوارزمية NB، وكان قادرًا على تحديد سمة مؤلف النص بمتوسط دقة، وصلت إلى 97.43%[14].
- اقترح Baraka, Rebhi.(2014) أنموذجًا لإسنادية التأليف من خلال تصنيف مجموعة من الوثائق والنصوص العربية القصيرة مع مؤلفين غير معروفين والتعرف على أسلوب كل مؤلف من خلال خصائص مستخرجة من النص والنموذج اعتمد على خوارزمية (SVM)حيث قام بالعديد من التجارب على الوثائق التي تحتوي على نصوص عربية مأخوذة من نطاقين: التحليل السياسي والمقالات الأدبية، وتمثل الوثيقة نوعين من السمات المعجمية والنحوية واستخدم word bi-gramsعلى السمة المعجمية و-N (POS) Tags grams,على السمة النحوية .ولضمان دقة المصنف أجرى تجربتين منفصلتين؛ حيث تم إجراء الأولى على مجموعة بيانات من سمة word bi-grams فقط لكل من النطاقين، وأجريت التجربة الثانية على مجموعة البيانات التي تجمع بين الميزتين، وكانت دقة اختبار مجموعة البيانات التي تجمع بين الميزتين لكل من النطاقين أعلى مما كانت عليه عندما استخدم ميزة واحدة، وهذا يدل على أن الجمع بين أكثر من ميزة يعزز من عملية التصنيف وكانت النتائج التي حققها تصل إلى 100% ويعزى ذلك لقلة النصوص المستخدمة وقصرها[15].
- أطروحة Shaker, Kareem.(2012) تعد أول دراسة في مجال إسنادية التأليف للنصوص العربية؛ حيث استخدم فيها 54 كلمة من كلمات العطف المشتركة وحروف الجر، وتعامل في أطروحته مع خصائص الكلمات الوظيفية Function Wordsفي اللغة الإنجليزية وما يقابلها بالعربية؛ ولم ينطلق من خصائص اللغة العربية بشكل صرف. وبنى أنموذجًا هجينًا أسماه (Hybrid EA Approach)، وفيه توصل إلى دقة 100% وتعزى تلك النسبة أيضًا إلى قلة النصوص و واحدية المتغير[15].

6. طريقة ومنهجية العمل Methodology

يستند البحث الحالي في معالجة إسنادية النصوص الشعرية لأصحابها على طريقة تصنيف النصوص بالاعتماد على خوارزميتين NB,SVM، وهذه الطريقة تمر بعدة خطوات متتابعة وهي: جمع النصوص وتحضيرها Text Preprocessing، تمثيل النصوص Texts Representation، استخراج السمات Features Extraction واختيار السمات Features Selection، وذلك من أجل الكشف عن بصمة الشاعر وأسلوبه الخاص في الكتابة.

هذا البحث اعتمد على عملية إسنادية التأليف بطريقة تطبيق عملية التصنيف في تحضير النصوص، وتمثيلها؛ حيث: تمّ تجميع مجموعة البيانات لعدد ستين شاعرا وتقسيمها إلى مجموعتين هما مجموعة لبيانات التدريب ومجموعة أخرى للاختبار. الخطوة الأولى: يتمّ فيها استخراج السمات من البيانات الكلية بعد إجراء عملية التدريب عليها ثمّ تنجز عملية الاختبار على أساس هذه السّمات. الخطوة الثانية: يتم فيها بناء أنموذج لبيانات

التدريب واختباره على مجموعة بيانات الاختبار مجهولة المؤلف. تدل حالات التدريب والاختبار المتعددة على العديد من السّمات على أسلوب المؤلف الذي سيتمّ إسناد النص إليه، وذلك عبر تعليم الآلة طريقة لاستخلاص السمات من بيانات التدريب، وعملية التصنيف التى اعتمدها البحث تعدّ أفضل وسيلة لفحص القدرة على التكيّف والتعلم في عملية تصنيف النصوص(16).

أ. تحضير النصوص Text preprocessing

في هذه الخطوة جُمعت عينة الدراسة من الموسوعات الشعرية ومواقع الإنترنت، حيث أدخلت نصوص شعرية (دواوين) لستين شاعرًا من مختلف العصور، وتم اختيارهم عشوائيا، وأدُخل الجزء الأكبر من أشعارهم بوصفه مجتمعا يحتوي على بيانات (dataset) للتدريب، والجزء المتبقي بوصفه عينة اختبار، وأدُخلت مجموعة تتألف من ستين مؤلفًا مجهولاً Unknown Author تمثل قصائدهم نصوصًا مجهولة الشاعر، متفاوتة بعدد أبياتها بوصفها عينة للاختبار. وكل النصوص الشعرية التي تم اختيارها هي من الشعر العمودي الذي يحتوي على الوزن والقافية والبيت الشعري. كما أن عينة الدراسة ليست نقية تمامًا فبعضها يحتوي على: alphanumeric وعلامات الترقيم punctuation، لكنها تخضع بعد التهيئة إلى عملية تمثيل النصوص Texts Representation والتي بدورها تتضمن مجموعة من الخطوات المهمة التي يجب إجراؤها على النص(17):

التصفية: فيها تحذف الحروف الخاصة alphanumeric وعلامات الترقيم التي لا تعطي أي مؤشر دلالي للنص الشعري.

التقطيع: هي عملية تجزئة النص إلى كلمات ليسهل معالجتها آليًا.

التجذير: يتم إعادة الكلمات إلى جذورها.

حذف الكلمات الزائدة: الكلمة الزائدة التي لا تعطي أي معنى مميز للنص مثل الروابط بين الكلمات التي ليس لها معنى مستقل بذاته، بل تأخذ معناها من الارتباط مع الكلمات الأخرى مثل أدوات الجر وغيرها؛ ويمكن إجراء ذلك بمقارنة كل كلمة مع قائمة محضرة مسبقًا تضم الكلمات الزائدة المعروفة.

حذف التشكيل: يتم في هذه المرحلة حذف الحركات مثل الفتحة والضمة والسكون والتنوين.

ب. استخراج السمات Extracting Features

تُعد عملية استخراج السمات مرحلة حرجة في نمطية التأليف؛ وطريقة اكتشاف أسلوب الكتابة لدى مؤلف ما تتم من خلال مجموعة السمات الواضحة في نمطية التأليف التي يتبعها المؤلف، هذا الافتراض يعني: أن لكل مؤلف له ميزات معينة في الكتابة، يمكن أن تكون متاحة للتعرف عليه؛ ومن خلال السمات الأسلوبية Stylomatric يمكن التعرف على الكثير من السمات المحتملة مثل: السمات المعجمية Lexical، والحرفCharacter، والنحوية Syntactic ، والسمات الدلاليةSemantic.

وفي هذه البحث جرت الاستفادة من السمات المعجمية والحرف؛ لأنها الأكثر دلالة من السمات الدلالية، حيث أُستخدمت مجموعة من السمات الموضحة في الجدول (1)، وهي: الحرف وطول الجملة الشعرية، وطول الكلمة، والقافية، والكلمة الأولى في الجملة الشعرية والبحر الشعري(16).

ج.اختيار السمات Selection Features

بعد استخراج كافة السّمات قيد الدراسة، يتم اختيار سمة محددة من المجموعة ذات الصلة المحتملة، واختيار هذه السمة هو جزء اساس من إسنادية التأليف التي تبدأ بدراسة مجموعة واسعة من السمات، وتهدف إلى تحديد أكثر

السّمات صلة بالمؤلف. وللقيام بهذه العملية يتم ملاحظة التكرار لسمة ما لأي نوع سواء كانت: الحرف، أم المفردات، أم السمات النحوية أم الدلالية، وهذا التكرار هو المعيار الأقوى لاختيار سمات التأليف في أيّ نصّ لإسنادها إلى المؤلف. ولاختيار سِمة ما في هذا البحث، تمّ اختيار السّمة المتكررة بشكل ملحوظ في نصوص قيد الدراسة أثناء عملية التدريب والاختبار، ولحساب الاحتمال لهذا النموذج تمّ الاعتماد على المتوسط والانحراف المعياري للسمات، وهناك تقنيتان أُستخدمتا لهذا الغرض هما: Chi-squared و Information Gain (IG)، وذلك لأجل الحصول على معلومات عن السمات التي تم اختيارها في هذا البحث وكانت النتيجة جيدة جدًا[18].

جدول (1)
نتائج التطبيق على السمات بشكل مستقل

Features	Total correct		Accuracy Percent		Recall	
	NB	SVM	NB%	SVM%	NB	SVM
Character	57	54	95	90	0,95	0,9
word length	51	56	85	93,33333	0,85	0,933333
Sentences length	43	44	71,66667	73,33333	0,716667	0,733333
First word length	48	40	80	66,66667	0,8	0,666667
Metre	53	45	88,33333	75	0,883333	0,75
Rhyme	47	48	78,33333	80	0,783333	0,8
Average	49,83333	47,83333	83,05556	79,72222	0,830556	0,797222

د. التجربة Experience

لاستخراج السمات المقترحة في هذه البحث استخدمت أداة برمجية، تشمل مكتبة ويكاWeka، وهي مفتوحة المصدر، ثم قمنا بفرز السمات إلى ست مجموعات، المجموعة F1 الاحرف و F2 طول الكلمة و F3 طول الجملة الشعرية وF4الكلمة الأولى في الجملة، و F5 البحر الشعري، و F6 القافية.

قُسمّت النصوص الشعرية إلى مجموعتين لعدد ستين شاعرًا، المجموعة الأولى للتدريب وتحتوي على الجزء الأكبر من نصوص الشعراء والمجموعة الثانية تحتوي على نصوص مجهولة المؤلف للعدد نفسه.

الخطوة الأولى: هي العمل مع مجموعات السمات المستخرجة بشكل منفصل تمامًا، وكل سمة مستقلة لها حالة تطبيق للخوارزميات NB,SVM, منفصل، و خلاصة النتائج لتلك العملية مبينة في الجدول (1).

الخطوة الأخرى: استخدام مجموعة السمات المستخرجة بشكل متفاوت الارتباط كالاتي: F1+F1 معا وطبقت التجربة؛ ثمF1+F2+F3 مع بعض وطبقت الإجراءات نفسها، يلي ذلك F1+F2+F3+F4 ثم F1+F2+F3+F4+F5، وأخيرًا تم التطبيق على كل السمات مع بعض F1+F2+F3+F4+F5+F6؛ الجدول (2) يشتمل على خلاصة نتائج تلك العملية.

7. الختام:

بعد إجراء كافة العمليات و تطبيق التجربة حصلنا على النتائج المبينة في الجدولين (1)، (2)، وبالنظر إلى الجدول (1) الذي يعرض نتائج تطبيق إسنادية النص للمؤلف باستخدام تقنياتSVM,NB، وعلى السمات المختارة وتطبيقهما بشكل منفصل في كل مرة نجد أن: أعلى نسبة دقة ظهرت هي 95% كانت على سمة الحرف بتطبيق خوارزمية NB بينما خوازمية SVM كانت نسبتها 90% على السمة نفسها، واختلفت النسبة على سمة طول

الكلمة عند تطبيق SVM حيت كانت النتيجة 93.3%، بينما حصلت NB على نسبة 85% عند تطبيقها على السمة نفسها، يعزى هذا الاختلاف إلى نوع السمة وطريقة التعامل معها من خلال طبيعة الخوارزمية.

إن أدنى نسبة دقة هي 66.66% ظهرت في جدول(1) لخوارزمية SVM على سمة طول الكلمة الأولى في الجملة، بينما ظهرت النسبة 80% عند تطبيق NBعلى السمة نفسها.

إن السمتين الشعريتين البحر والقافية حصلتا على نسبٍ متقاربة عند تطبيق SVM,NB، عليهما، فكان نصيب سمة البحر الشعري بتطبيق خوارزمية NB هو 88.33% وبتطبيق SVM على السمة نفسها كانت النسبة 75%، وهي نسب ليست متدنية، لكنها ليست كافية ليظهر تأثيرها على تحديد هوية الشاعر؛ ما يعني أن البحر الشعري في الشعر الكلاسيكي ليس مؤشراً جيدا في تحديد هوية المؤلف، و يعزى ذلك إلى تكرار البحور الشعرية وتشابهها لدى الشعراء. بينما سمة القافية حصلت هي الاخرى على نسب دقة متقاربة عند تطبيق الخوارزميتن.

لقد حققت SVM نسبة دقة وصلت 80.33% عند تطبيقها على سمة القافية، بينما حصلت خوارزمية NB على نسبة دقة وصلت إلى 78.33% على السمة نفسها؛ و هذا الانخفاض مقارنة بسمة البحر يرجع إلى أن القافية أكثر احتمالية في الظهور بعدد يساوي حروف الأبجدية، أما البحور الشعرية فلا يتجاوز عددها ستة عشر بحرا، وعلى الارجح لا يُستخدم الا العدد القليل منها.

ولا يمكن تجاوز تلك النتائج السابقة مقارنة بطول الجملة الشعرية وطول الكلمة البادئة التي كنا نطمح إلى أن تكون ذات مؤشر جيد لتحديد أسلوبية الشاعر؛ كونها نادرة التكرار لدى الشعراء إلا أن النتائج كانت ذات نسب متدنية وصلت إلى 66.67% لسمة طول الكلمة البادئة عند تطبيق SVM و80% عند تطبيق NB على السمة نفسها.

و في الجدول نفسه يلاحظ أن طول الجملة الشعرية حققت نسبة دقة قيمتها 73.33% عند تطبيق SVM ونسبة قدرها 71.66% عند تطبيق NB، ويعزى ذلك الانخفاض إلى أن طول الجملة الشعرية في الشعر العمودي مرهون بالبحر الشعري وعدد التفاعيل في كل بيت شعري.

من الجدول (1) نلاحظ أيضا أن معدل الدقة لخوارزمية NB قيمته 83.055% و معدل Recall يساوي 0.8305، أما خوارزمية SVM فإن معدل الدقة فيها وصل إلى نسبة 79.722%، ومعدل Recall يساوي 0.7972 وهذا يعني أن NB كانت أكثر كفاءة عند تطبيقها على السمات المنفصلة من SVM.

في الجدول (2) تم تطبيق الخوارزميات على سمات متعددة أُخذت معًا وحذفنا بعض الحالات الأخرى من الجدول؛ لأنها لم تعط مؤشراً ذا دلالة مميزة.

جدول (2)
نتائج التطبيق على السمات بشكل متفاوت الارتباط

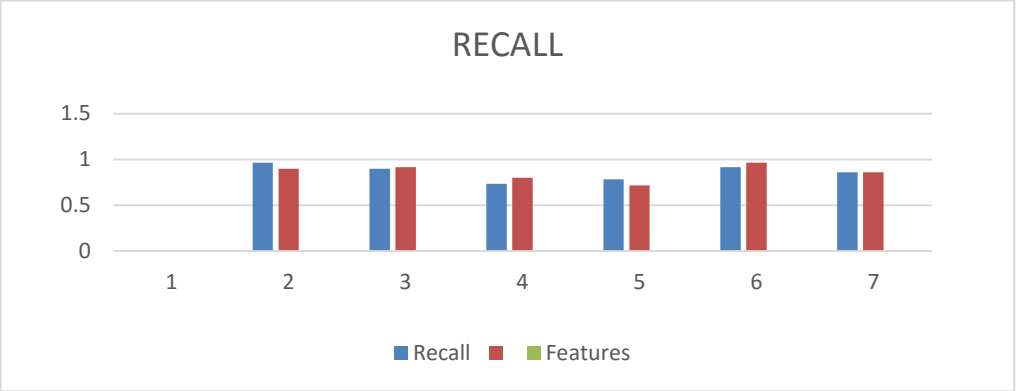
Factures	Total correct		Accuracy Percent		Recall	
	NB	SVM	NB%	SVM%	NB	SVM
F1,F2	58	54	96,66667	90	0,966667	0,9
F1,F2,F3	54	55	90	91,66667	0,9	0,916667
F1,F2,F3,F4	44	51	73,33333	85	0,733333	0,85
F1,F2,F3,F4,F5	46	48	76,66667	80	0,766667	0,8
F1,F2,F3,F4,F5,F6	55	58	91,66667	96,66667	0,916667	0,966667
Average	51,4	53,2	85,66667	88,66667	0,856667	0,886667

إن أعلى نسبة دقة تم الحصول عليها هي 96.67%، ناتجة عن تطبيق خوارزمية NB على F1+F2 و النسبة نفسها عن تطبيق خوارزمية، SVM على F1+F2+F3+F4+F5+F6، بالمقابل فإن أدنى نسبة دقة ظهرت لإسناد التأليف هي 73,333% ناتجة عن تطبيق خوارزمية NB على السمات F1+F2+F3+F4، ونسبة دقة وصلت إلى 76.667% على السمات F1+F2+F3+F4+F5 للخوارزمية نفسها؛ وعند تطبيق SVM تقاربت النسبة على السمات نفسها حيث كانت النسبة 80% على F1+F2+F3+F4+F5 وارتفعت إلى 85% على F1+F2+F3+F4.

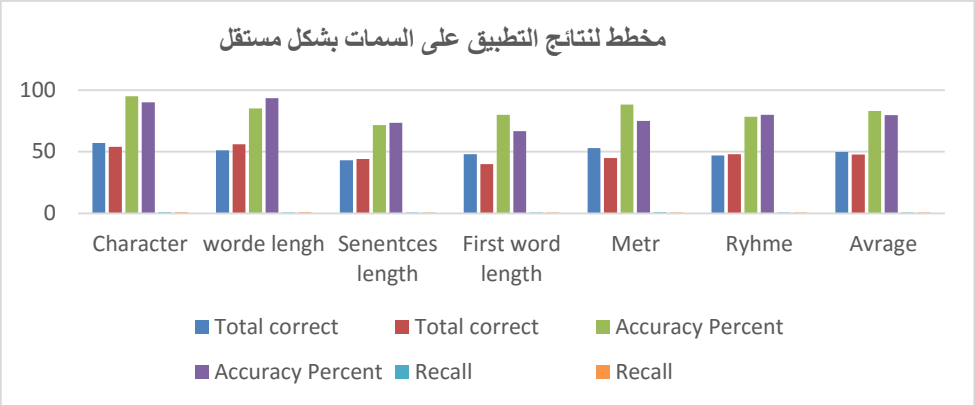
كما نجد أن السمتين F1+F2 الحرف وطول الكلمة أخذتا نسبة دقة عالية على مستوى خوارزمية NB، بينما انخفضت النسبة إلى 90% عند إضافة F3 طول الجملة الشعرية عند تطبيق NB، والنسبة نفسها 90% عند تطبيق SVM على السمتين F1+F2، كما ازادت النسبة إلى 91.667% بتطبيق SVM عند إضافة F3 مقارنة بالسطر الاول من الجدول (2)، ويعزى ذلك إلى طبيعة السمة F3 وتشابهها لدى الشعراء في الشعر العربي الكلاسيكي، هذا التشابه كان نتيجة لثبات طول الجملة الشعرية التي تعتمد على عدد التفاعيل في البيت الشعري، بينما نجد أن خوارزمية SVM حققت نسبة أعلى من البقية على السمات نفسها وهذا بسبب طبيعة الخوارزمية وطريقة معالجتها للمقاطع النصية.

في السطر الثالث من الجدول عند إضافة السمة F4 وإضافة F5 في السطر الرابع فإن النسبة انخفضت عندNB، وبإضافة سمة البحر الشعري F5 عند تطبيق خوارزمية SVM، كانت النسبة 80% وهي أدنى قيمة لهذه الخوارزمية في الجدول بشكل عام؛ ما يعني أن البحر الشعري لم يكن ذا دلالة مميزة بتطبيق الخوازميات، وبالعموم فإن جميع السمات التي أُخذت معًا في السطر الأخير قد حققت نسبة عالية عند تطبيق SVM التي وصلت إلى 96,667% ونسبة عالية أيضًا وصلت إلى 91,667% بتطبيق NB على جميع السمات، وهذا يعني أن إضافة السمات الشعرية مع بقية السمات يحقق نسبة أفضل بكثير مما كانت عليه في جدول(1) بشكل منفصل.

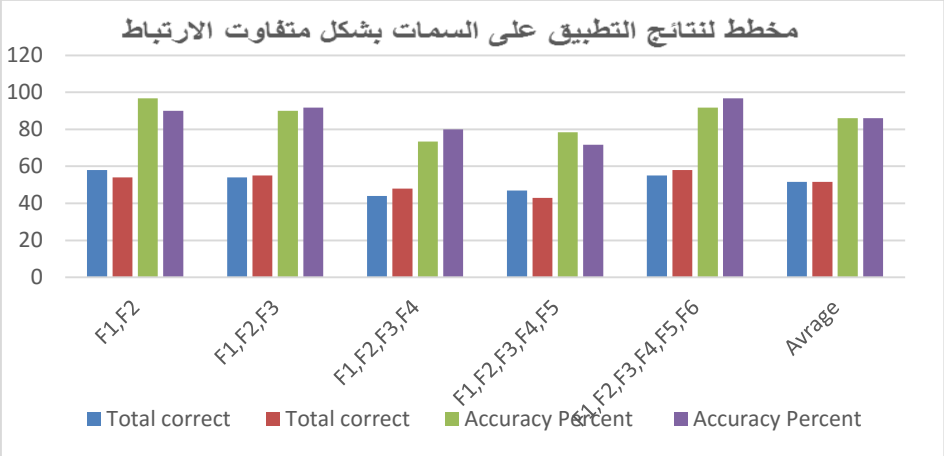
في جدول (2) نلاحظ أن معدل الدقة لخوارزمية SVM كانت قيمته 88,667%، و معدل Recall يساوي 0.8866، أما خوارزمية NB فإن معدل الدقة فيها وصل إلى نسبة 85,667% ومعدل Recall يساوي 0,8566 وهذا يعني أن SVM كانت أكثر كفاءة عند تطبيقها على السمات المرتبطة معًا.



مخطط 1. يوضح recall للنتائج



مخطط 2. لنتائج التطبيق على السمات بشكل مستقل



مخطط 3. لنتائج التطبيق على السمات بشكل متفاوت الارتباط

مما سبق نستطيع القول: إن النتائج جاءت مطابقة للتوقعات؛ حيث وصلت أعلى نتيجة لخوارزمية SVM 96.667 % على جميع السمات في الجدول(2) وبمعدل دقة 88.667% إلا أننا نطمح إلى المزيد من الدقة في الأعمال اللاحقة.

ولتجاوز تلك العقبات نقترح إدخال متغيرات أخرى خاصة بالشعر مثل الكلمات النادرة واستخدام الوزن، المترادفات، وبعض الخصائص الشعرية الصِّرْفة، و يمكن أخذ ثيمات خاصة بالشعر والعصر الذي ينتمي إليه.

ومن التوجهات المستقبلية فإننا نقترح زيادة بعدد العينة والنصوص الشعرية مع الأخذ بعين الاعتبار توحيد أطوال نصوص الاختبار، وتطبيق نفس المعايير عليها، كذلك تطوير تقنية ناتجة من تهجين الخوارزميتين معا و استخدام تقنيات أخرى من تقنيات تنقيب النصوص وتعليم الآلة، ومقارنة النتائج الجديدة مع هذه النتائج.

الهوامش:

(1) هناك دراسة صدرت حديثا عن إسناد التأليف في الشعر Punjabi Poetry في 2015 بعد دراستنا التي قدمناها في مؤتمر دولي حول اللغة العربية 2014 المغرب.

المراجع:

1. القرشي، أبو زيد محمد بن أبي الخطاب، "جمهرة أشعار العرب"، تحقيق :محمد علي الهاشمي، دمشق :دار القلم، ط2:، 1986، ص 146وما بعدها، وابن فارس، أبو الحسن أحمد، الصاحبي في فقه اللغة وسنن العربية في كلامها، تحقيق :عمر الطباع، بيروت :مكتبة المعارف، ط1:، 1993، 468 – 465

2. بوتبيا، الحسن، "المفاضلة بين النظم والنثر وأشكال التداخل بينهما في العصر العباسي"، مراكش :المطبعة والوراقة الوطنية، ط1: 2002، ص.67 - 66.

3. الفلاحي، أحمد، "نحو معالجة آلية للشعر العربي : عملية الإسناد التلقائي لنص شعري مجهول إلى شاعره،" المجلة الدولية للتطبيقات الإسلامية في علم الحاسب والتقنية -إجازات .vol. 3, NO, no. 2289 - 4020, p. 8, 2015.

4. D. Foster, Author Unknown: On the Trail of Anonymous. Henry Holt and Company, 2014.

5. E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem," Inf. Process. Manag., vol. 44, no. 2, pp. 790–799, Mar. 2008.

6. Shaker, Kareem.(2012)."Investigating Features and Techniques for Arabic Authorship Attribution", PhD. Thesis Of Computer Science, Department Of Computer Science School of Mathematics and Computer Science, Heriot-Watt University, March 2012.

7. F. Howedi and M. Mohd, "Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data," Computer Engineering and Intelligent Systems, vol. 5, no. 4. pp. 48–56, 2014.

8. A. Abbasi, "Applying Authorship Analysis to Extremist-Group Web Forum Messages," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 67–75, Sep. 2005.

9. "Authorship Attribution of Punjabi Poetry using SVM Classifier." [Online]. Available: http://www.ijarcsse.com/docs/papers/Volume_5/5_May2015/V5I5-0545.pdf. [Accessed: 17-May-2016].

10. M. a Alnagdawi, H. Rashideh, and A. Fahed, "Finding Arabic Poem Meter using Context Free Grammar," vol. 3, no. 1, pp. 52–59, 2013.

11. I. A. Mohammad, "NAIVE BAYES FOR CLASSICAL ARABIC POETRY," vol. 12, no. 4, pp. 217–225, 2009.

12. A. M. A. Alhichri H. S., "Expert System for Classical Arabic Poetry (ESCAP)," in in proceedings of International Conference on APL,Toronto, Ontario, Canada, 2008.

13. A. Almuhareb, I. Alkharashi, L. AL Saud, and H. Altuwaijri, "Recognition of Classical Arabic Poems," Proc. Work. Comput. Linguist. Lit., pp. 9–16, 2013.

14. A. Altheneyan and M. Menai, "Naïve Bayes classifiers for authorship attribution of Arabic texts," J. King Saud Univ. ..., 2014.

15. R. Baraka, S. Salem, M. Abu, N. Nayef, and W. A. Shaban, "Arabic Text Author Identification Using Support Vector Machines," J. Adv. Comput. Sci. Technol. Res., vol. 4, no. 1, pp. 1–11, 2014.

16. K. Luyckx, Scalability Issues in Authorship Attribution. Asp / Vubpress / Upa, 2011.

17. M. S. Desouki and A. Al-abdo, "Experiments in Mining Arabic Texts ' محاولات للتنقيب في النصوص العربية ,'" vol. 2, no. 1, pp. 14 -18, 2012.

18. E. Stamatatos, "A survey of modern authorship attribution methods," J. Am. Soc. Inf. …, 2009.

Authorship Attribution in Arabic Poetry using Text mining Techniques

Al-Falahi Ahmed [1], Ramdani Mohamed [*2] and Bellafkih Mostafa[3]

[1]*Department of computer science, Faculty of Education –Naderah, University of Ibb, Yemen*

[2]*Department of Computer Science, Faculty of Science and Technology Mohammedia(FSTM) , Hassan II University, Casablanca, Morocco*

[3] *National Institute of Posts Telecommunications (INPT), Rabat, Morocco*

[*]*E-mail: moha@fstm.ac.ma*

Abstract

In this paper, we present the Arabic poetry as an authorship attribution task. Several features such as Characters, Sentence length; Word length, Rhyme, Metre and First word in sentence are used as input data for NB,SVM methods. The data is filtered by removing the punctuation and alphanumeric marks that were present in the original text. The data set of experiment was divided into two groups: training dataset with known authors and test dataset with unknown authors. In the experiment, a set of Fortieth poets from different eras have been used. The Experiment shows interesting results with classification precision of 96.667%.

Keywords: Authorship attribution, Arabic Poetry, Text Classification, NB,SVM