نحو معالجة آلية للشعر العربي:

عملية الإسناد التلقائي لنص شعري مجهول إلى شاعره

احمد الفلاحي[1]، محمد الرمضاني[2]، مصطفى بلفقيه[3] ، محمد الصارم[4]
جامعة اب،اليمن [2] كلية العلوم والتقنيات، المحمدية [1]
الرباط [4] جامعة طيبة-المدينة المنورة -**INPT**[3]
[3] bellafki@inpt.ac.ma ،[2] moha@fstm.ac.ma ،[1]flahi79@gmail.com

الخلاصة: يمثل موضوع هذه الورقة مرحلة متقدمة من مشروع يقوم بالإسناد التلقائي لنص شعري مجهول في الشعر العربي إلى شاعره الحقيقي وأتمَتَة هذه العملية باستخدام تقنيات التنقيب عن النصوص Text Mining، الجدير بالذكر أن عملية إسناد المؤلف في الشعر العربي عملية مهمة جدا في التنقيب عن النصوص خصوصاً لأولئك الذين يدرسون الأسلوبية في الشعر العربي، ومساعدة الشعراء في إثبات حقهم الإبداعي ومعرفة النصوص المنتحلة من غيرها. بيد أن عملية الإسناد التلقائي لنص شعري مجهول إلى شاعره تتمّ على أساس استخراج خصائص عديدة من النص المجهول وأسلوبيته لمطابقتها مع أسلوب الشاعر والذي يتمّ استخرجه من نصوص معلومة له باستخدام تقنيات التنقيب المتوائمة مع بيئة النص الشعري. في هذا البحث تمّ إدخال مجموعة من الدواوين الشعرية لأربعين شاعراً من مختلف العصور في الشعر العربي الكلاسيكي كمجموعة للتدريب وإدخال مجموعة اربعين نصاً مجهولة المؤلف من نصوص مختلفة كمجموعة اختبار تم جمعها من الموسوعات الشعرية والمواقع الالكترونية بعد ذلك طبقت خوارزميات NB,SVM,M.C على تلك النصوص مع بارامترات ومتغيرات هي: القافية، الحرف، طول الكلمة، وطول الجملة الشعرية، وحصلنا على أعلى نتيجة لسلسلة ماركوف وصلت إلى 97.5 %.

الكلمات الجوهرية: الشعر العربي، إسناد التأليف، سلسلة ماركوف.

1. المقدمة

النص الشعري من النصوص المهمة في العربية وغيرها وقد احتل منزلة عالية لا منافس لها عند العرب في الجاهلية، وكان الشعراء بمنزلة الأنبياء لدورهم المهم وحاجة الناس إليهم، فهم الذين يقيدون مآثرهم ويعلون من شأنهم، ويخوّفون أعداءهم فتبوؤا منزلة عالية وذلك بسبب ما يقدمه الشعراء من منافع عامة يجنيها المجتمع القبلي القائم على حروب وعداوات متصلة بالشعر [1]. إن الشعر يتعرض للانتحال والسرقة وغيرها على مرّ العصور حتى يومنا هذا، ويحظى الشعر بالدراسة والتمحيص المستمر من الباحثين في الأدب واللغة وهذا ما يجعل النص الشعري موضوعاً خصباً في التنقيب عن النصوص والبيانات وهندسة اللغة العربية وأتْمَتِها. لقد صنف الباحثون في الأدب، الشعرَ إلى أصناف حسب العصور وهي: الشعر الجاهلي وشعر المخضرمين والشعر في عصر صدر الإسلام وفي العصر الأموي والعباسي و.... والشعر الحديث كما صُنّف الشعر العربي حسب البناء الشعري إلى شعر عمودي (الكلاسيكي) وشعر التفعيلة وقصيدة النثر [2.]

لقد ظهرت تقنيات إسناد التأليف غير التقليدية في عام 1887، عندما ابتدع Mendenhallلأول مرة فكرة حصر السمات (Counting Features)في النص مثل طول الكلمة لتدل على شخصية المؤلف. وتبع ذلك العمل لاحقاً ما قام به كل من : Yule (1938)و Morton (1965) باستخدام أطوال الجمل للحكم على هوية المؤلف[3].

توالت الدراسات على اللغة الإنجليزية وبشكل متتابع وعلى بعض اللغات الأخرى، وإلى وقت كتابة هذه الورقة فإن الدراسات والأبحاث في هذا المجال باللغة العربية لا تتجاوز أطروحة دكتوراه وورقات عمل منشورة نتناولها في الأدبيات السابقة، وتركز العمل في تلك الدراسات على النصوص العربية القصيرة فقط، وتناول البعض صفحات الويب والبريد

الالكتروني ويعدّ ذلك حافزا إضافياً لمشروعنا هذا كوننا نتعامل مع نصوصٍ شعريةٍ بتركيبة خاصة. يهدف هذا البحث إلى عملية الإسناد التلقائي لنص شعري مجهول في الشعر العربي إلى شاعره الحقيقي وأتمَتَة هذه العملية باستخدام تقنيات التنقيب عن النصوصText Mining ، ويمثل موضوع هذه البحث مرحلة متقدمة من مشروع متكامل لإسنادية التأليف في الشعر العربي.

تتناول هذه الورقة إسنادية التأليف بصورة عامة وإسنادية التأليف في اللغة العربية، والدارسات السابقة، وتشرح الورقة أيضا المنهجية وآلية العمل وتطبيق الخوارزميات، ونختم ذلك بالنتائج والتوجهات المستقبلية والمراجع .

2. إسنادية التأليف Authorship Attribution

تعرّف على أنها مشكلة إسناد المؤلف الحقيقي إلى نصّ مجهول أو مختلف عليه وهذا هو التعريف البسيط لإسنادية التأليف. إن عملية الإسناد تتمحور حول دراسة خصائص النص من أجل استخلاص استنتاجات خاصة بتأليفه، وقد ظهر هذا العلم مستفيدا من قياس الأسلوبية في علم اللسانيات الذي يستخدم الأدوات الإحصائية على الأسلوب الأدبي [4].
ويمكن استخدام إسنادية التأليف النص في عدد من التطبيقات مثل: تحقيق الكتب المجهولة والمخطوطات بعد تحويلها إلى نصوص والوثائق المتنازع عليها مصدريا وفضح السرقة والانتحال بهدف إظهار هوية المؤلف الحقيقي، أيضا يساهم في التتبع القانوني لغرض التأكد من مؤلفي المجموعات الإخبارية، الرسائل الإلكترونية والمدونات. ويستفاد ايضا من هذه التطبيقات في نطاقات مختلفة مثل: نظام تتبع الجرائم، القانون الجنائي، القانون المدني والاستخبارات [5].
عملية إسنادية التأليف تتركز حول إظهار التشابه لمؤلف كتب نصاً ما بواسطة فحص أعماله الأخرى؛ حيث يتمحور فعل تحديد بصمة المؤلف حول استخلاص مجموعة من السمات للنص والتي تبقى ثابتة نسبياً في مجموعة كتاباته، والتقاط أسلوب كتابته [6].

3. إسنادية التأليف في اللغة العربية Authorship Attribution in Arabic:

اللغة العربية هي من اللغات المهمة في العالم كونها لغة القران ولشريحة كبيرة من المسلمين، وعملية الإسناد تناولها بشكل كبير علماء الحديث للتدقيق في نصوص الأحاديث النبوية وكذلك الباحثون في اللغة والأدب لتتبع الانتحال والسرقات الأدبية، لكن في مجال استخدام الحاسوب وأتمَتَة هذه العملية فهو أمر نادر، كما تجدر الإشارة إلى أن أغلب البحوث التي تمّ إجراؤها في مجال إسنادية مؤلف النص كانت على اللغة الانجليزية، ويتطلب تطبيقها على اللغة العربية التعامل مع التحديات والصعوبات الناتجة عن بعض مواصفات خاصة بهذه اللغة مثل اشتقاق المفردات وطبيعة تلك المفردات وتجذيرها وتشكيل الكلمات، طول الكلمة والحروف وطول الجملة والسمات النحوية والمعجمية [7].

4. خصائص الكتابة والسمات الاسلوبية stylistic Features:

السمات ونمطية الكتابة في نص ما والتي تساعد على اكتشاف بصمة وهوية المؤلف تقسم إلى أربع مجموعات: السمات المعجمية وتكون على مستوى الأحرف ومستوى الكلمات، السمات النحوية ويتم التعامل خلالها مع الأشكال الداخلة في بنية الجملة، السمات التركيبية او الهيكلية وتعكس العادات الخاصة للكاتب والمؤلف في التخطيط والتنظيم لكتاباته، وسمات خاصة مرتبطة بمضمون النص مع الكلمات المفتاحية في موضوع معين أو نطاق محدد يتخذه المؤلف منهجاً في نصوصه وكتاباته. ومن أجل عملية التصنيف تُستخدم أدوات مختلفة من أبرزها طرق تقنيات تعليم الآلة، الشبكات العصبية، الاحتمالات، التحليل الإحصائي وشجرة اتخاذ القرار [8].

1) السمات المعجمية Lexical features:
وهي من أكثر السمات المستخدمة لإسناد النص إلى مؤلفه، وتعتمد على طول الكلمة، وطول الجملة، وعدد تكرار الكلمة، ووفرة المفردات، إلا أن المشكلة الرئيسية في هذا النوع من السمات في بعض اللغات الشرقية حيث لا توجد حدود فاصلة بين الكلمات، ويصعب تطبيق هذه السمات دون الحاجة إلى أدوات خاصة مساعدة.

2) الحرف Character:
في هذا النوع من السمات يتم الاستناد إلى الأحرف في معالجة النصوص باستخدام تسلسل الأحرف، ويأخذ نوع الأحرف، ترددتها وCharacter N-gram ويمكن تطبيقها بسهولة في أي لغة دون الحاجة إلى أي أدوات خاصة.

3) السمات النحوية Syntactic

وتستخدم السمات النحوية من قبل المؤلفين دون وعي، مما يجعلها أكثر موثوقية من السمات المعجمية. وفيها يتم استخدام تدابير مختلفة في الدراسات النحوية من اجل عملية الإسناد بما في ذلك الجزء من الكلام (POS: Part-Of-Speech)، التكرار والأخطاء النحوية وFunction words. هذه السمات تتطلب أدوات لاستخراجها.

4) السمات الدلالية Semantic

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

5) السمات التركيبيةStructural

هذه السّمات تلتقط عادات المؤلف عند تنظيم النص وبنائه والأمثلة على هذه التدابير طول الفقرة، وطول الجملة، واستخدام التوقيع ولون الخط وحجمه. السّمات التركيبية لا تظهر بوضوح في النصوص القصيرة لأنه من الصعب التقاط الخصائص الأسلوبية للنص وتظهر جلية في النصوص الأطول [9].

5. الادبيات السابقة Related work:

حتى تاريخ كتابة هذا البحث لم نجد أي عمل يمس مشكلة اسنادية النص الشعري لمؤلفه ومع ذلك هناك بعض الأعمال ذات الصلة تجعل مهمتنا أسهل وسيتم تصنيفها إلى مجموعتين:

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

الاعمال التي تعاملت مع النص العربي :Works with Arabic text أيضا هذه المجموعة من الأعمال تتعامل مع النص العربي كمهمة التصنيف.

أ. تصنيف الشعر العربيArabic poems classification

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

• AbdulBaki,Iqbal. (2009) تناولت دراستها الشعر العربي الكلاسيكي من حيث تصنيفه إلى مدح وهجاء وذم وغزل وغيرها وذلك باستخدام خوارزمية Naïve Bayes للتصنيف، عملها كان مقتصرا على ذلك مع استخدام عملية Stemmingالتجذير كمتغير وحيد وحققت نسبة نجاح 90 % [11].

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

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

ب. تصنيف النص العربيArabic text classification

• Altheneyan,Ala.(2014) طبق خوارزمية Naïve Bayes classifiers على نظام إسنادية التأليف للنص العربي لعملية اختبار ومقارنة أربعة نماذج مختلفة من خوارزمية NB هيMNB :، MBNB، وMPNB، MBNBان احتمال التقدير يعتمد على وجود ميزة من عدمها، في حين تعتمد MPNB و MNBعلى احتمال تكرار الميزة .وتم تقييم الأداء على حجم كبير من أربع مجموعات من البيانات المختلفة ودرس التأثير الناجم على عملية الإسناد وتظهر النتائج الإجمالية لديه أن النموذج MBNB يوفر أفضل النتائج بين كل نماذج خوارزمية NB فقد كان قادرا على تحديد سمة مؤلف النص بمتوسط دقة و97.43% [14].

• Baraka,Rebhi.(2014) اقترح نموذجاً لإسنادية التأليف حيث صنف مجموعة من الوثائق والنصوص العربية القصيرة مع مؤلفين غير معروفين والتعرف على أسلوب كل مؤلف من خلال خصائص مستخرجة من النص، النموذج أعتمد 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-grams على السمة المعجمية و ,Tags N-grams (POS)على السمة النحوية. ولضمان دقة المصنف أجرى تجربتين منفصلتين حيث تم إجراء الأولى على مجموعة بيانات من ميزة 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 واعتبر أن النتائج التي حققها تصل إلى 100% وقد نعزي ذلك لقلة النصوص المستخدمة وقصرها[15].

• Shaker,Kareem.(2012) تعتبر أول دراسة في مجال اسنادية التأليف للنصوص العربية حيث استخدم فيها 54 كلمة من كلمات العطف المشتركة وحروف الجر، وتعامل مع خصائص Function Words في اللغة الإنجليزية وما يقابلها بالعربية ولم ينطلق من خصائص اللغة العربية بشكل صرف. وبنى نموذجاً هجيناً اسماه (Hybrid EA Approach) وفيه توصل إلى دقة 100% ونعزي تلك النسبة أيضا لقلة النصوص وأحادية المتغير[7].

6. الطريقة Methodology:

ان العملية تتم بعدة خطوات هي :جمع النصوص وتحضيرهاText Preprocessing ، تمثيل النصوص Texts Representation واستخراج السمات Features Extraction واختيار السمات Features Selection وذلك من اجل الكشف عن بصمة الشاعر وأسلوبه الخاص في الكتابة.

في هذه الورقة نقدم عملية اسنادية التأليف بطريقة تطبيق عملية التصنيف في تحضير النصوص وتمثيلها حيث تمّ تجميع مجموعة البيانات لأربعين شاعرا وتقسيمها إلى مجموعتين هما مجموعة بيانات للتدريب ومجموعة أخرى للاختبار. في الخطوة الأولى يتمّ استخراج سمات من البيانات بعد إجراء التدريب عليها ثمّ عملية الاختبار على أساس هذه السّمات. أما الخطوة الثانية فيتم فيها بناء نموذج بيانات التدريب واختباره على مجموعة بيانات الاختبار غير معروفة المؤلف. حالات التدريب والاختبار المتعددة على عديد السّمات تدلّ على أسلوب المؤلف الذي سيتمّ إسناد النص اليه وذلك عبر تعليم الآلة طريقة استخلاص السمات من بيانات التدريب، وإجراء عملية التصنيف تعدّ أفضل وسيلة لفحص القدرة على التكيّف والتعلم في عملية تصنيف النصوص[16].

تحضير النصوصText perprocessing

جمعنا عينة الدراسة من الموسوعات الشعرية ومواقع الأنترنت حيث أدخلنا نصوصاً شعرية (دواوين) لأربعين شاعراً من مختلف العصور واختيارها عشوائيا، إن الجزء الأكبر من أشعارهم تمّ إدخاله كمجموعة بيانات dataset للتدريب والجزء المتبقي كمجموعة اختبار، أدخلنا مجموعة اربعين Unknown Author كقصائد مجهولة الشاعر ومتفاوتة بعدد الأبيات للاختبار. النصوص الشعرية المختارة هي من الشعر العمودي الذي يحتوي على الوزن والقافية والبيت الشعري .ان عينة الدراسة ليست نقية تماماً فبعضها يحتوي على alphanumeric وعلامات الترقيم.punctuation لكنها خضعت بعد عملية التهيئة إلى عملية تمثيل النصوصTexts Representation وتتضمن مجموعة من الخطوات الهامة التي يجب إجراؤها على النص [17]:

التصفية :وفيها تحذف المحارف الخاصة وعلامات الترقيم التي لا تعطي أي مؤشر دلالي للنص الشعري .

التقطيع :وهي عملية تجزئة النص إلى كلمات .

التجذير: وفيه يتم إعادة الكلمات إلى جذورها .

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

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

استخراج السماتExtracting Features

تعد عملية استخراج السمات مرحلة حرجة ، ومن ثم طريقة اكتشاف أسلوب الكتابة لدى مؤلف ما من خلال مجموعة السمات الواضحة في نمطية التأليف التي يتبعها، هذا الافتراض يعني ان لكل مؤلف له نمط كتابة وميزات معينة يمكن أن تكون متاحة للتعرف عليها من خلال السمات الاسلوبية Stylomatric ويمكن التعرف على الكثير من السمات المحتملة بداية من :السمات المعجميةLexical ، والحرفCharacter ، والنحويةSyntactic ، والسمات الدلالية .Semantic

في هذه الورقة جرت الاستفادة من السمات المعجمية والحرف، لأنها أكثر دلالة من السمات الدلالية، استخدمنا مجموعة السمات كما في الجدول رقم (1) وهي الحرف وطول الجملة الشعرية؛ طول الكلمة، والقافية، والكلمة الأولى في الجملة الشعرية.[16]

اختيار السمات Feature selection

عند استخراج كافة السّمات يتم اختيار سمة من المجموعة ذات الصلة المحتملة، واختيار السمة هو جزء أساسي من إسنادية التأليف التي تبدأ بدراسة مجموعة واسعة من السمات، وتهدف إلى تحديد أكثر السّمات صلة بالمؤلف. وللقيام بهذه المهمة يلاحظ تكرار سمة لأي نوع سواء كان الحرف، المفردات، السمات النحوية أو الدلالية، هذا التكرار هو المعيار الأقوى لاختيار سمات التأليف في أيّ نصّ لإسنادها للمؤلف، ولاختيار سِمة ما في عملنا هذا يتمّ اختيار السّمة المتكررة بشكل ملحوظ في نصوص قيد الدراسة أثناء عملية التدريب والاختبار ولحساب الاحتمال لهذا النموذج يتمّ الاعتماد على المتوسط والانحراف المعياري للسمات، وهناك تقنيتان استخدمناهماChi-squared و Information Gain (IG) للحصول على معلومات عن السمات التي تم اختيارها في عملنا هذا وكانت النتيجة جيدة جدا [18] .

اسنادية التأليف باستخدام سلسلة ماركوف Markov Chain Authorship Attribution

ذكرنا سابقا انه يمكن تحديد أي مؤلف من خلال نص مجهول وإسناد هذا النص إلى مؤلفه المرشح من بين مجموعة من المؤلفين. هذه العملية يمكن أن تصاغ رياضيا على النحو الاتي[19]:

نفرض أن X هو النص الذي كتب بواسطة المؤلف A فان احتمال أن يكون A هو كاتب النص X اكثر من اي مؤلف آخر يحسب وفقا للمعادلة (1).

$$\mathcal{O}_{A,G}(X) = |e_z(X)|^{-1} \log\left(\frac{p_A(e_z(X))}{p_G((e_z(X)))}\right) \quad (1)$$

حيث z ينتمي إلى مجموعة سمات وخصائص النص، و$e_z(X)$ تعطي مجموعة مرتبة من عناصر النصX و القيمة المطلقة $|e_z(X)|^{-1}$ تستخدم لعملية تطبيع الدالة. بينما $p_A(e_z(X))$ and $p_G(e_z(X))$هي تقديرات احتمال المؤلف Aالذي كتب النص X .

اذا كان t يرمز للعتبة فان$f_A(X)$ هي دالة تشير إلى ان النص X كُتب بواسطة المؤلف A ام لا.

$$f_A(X) = \begin{cases} 1, & \mathcal{O}_{A,G}(X) > t \\ 0, & \text{if else} \end{cases}$$

ان احتمال اسناد النص X مع سلسلة العناصر $X = (i_1, i_2, \dots, i_x)$
للمؤلف A يمكن أن يحسب من خلال حساب الاحتمال المشترك لكل m-th في سلسلة ماركوف[19].

$$p_A(X) \approx \prod_{j=(m+1)}^{|X|} p_A\left(i_j \middle| i_{j-m}^{j-1}\right) \quad (2)$$

بحيث i_{j-m}^{j-1} هو اختزال ل$j-m, \dots, j-1$ و m تشير إلى طول القيود او السجلات.

إن اقصى احتمال يمكن أن يقدر لتعين m في سلسلة ماركوف على اي مجموعة تدريب لإيجاد المؤلف A يحسب من المعادلة (3)

$$p_A^{ml}\left(i_j \middle| i_{j-m}^{j-1}\right) = C\left(i_{j-m}^{j} \middle| X_A\right) / C\left(i_{j-m}^{j-1} \middle| X_A\right) \quad (3)$$

حيث $C\left(i_{j-m}^{j-1}\middle|X_A\right)$ هو عدد عناصر السلسلة i_{j-m}^{j} التي تحدث في X_A ولأجل تجنب الاحتمال الصفري في المعادلة(1) يمكن استخدام تقنية التنعيم خلال التدريب[20].

جدول (1):نتائج التطبيق على السمات بشكل مستقل

Features	Total correct			Accuracy Percent		
	NB	SVM	M.C	NB%	SVM%	M.C%
Character	38	37	38	95	92.5	95
word length	34	38	39	85	95	97.5
Sentences length	29	28	33	72.5	70	82.5
First word length	24	25	30	60	62.5	75
Rhyme	23	25	27	57.5	62.5	67.5
Average	29.6	30.6	33.4	74	76.5	83.5

التجربة

بعد استخراج السمات باستخدام RapidMiner واداة تشمل مكتبة ويكا Weka قمنا بفرز السمات إلى خمس مجموعات، المجموعة F1 الاحرف و F2 طول الكلمة و F3 طول الجملة الشعرية وF4 الكلمة الأولى في الجملة و F5 القافية . قمنا بتقسيم النصوص الشعرية إلى مجموعتين لأربعين شاعرا تمثل المجموعة الأولى للتدريب وتحتوي على الجزء الأكبر من نصوص الشعراء والمجموعة الثانية تحتوي على نصوص مجهولة المؤلف لنفس العدد .الخطوة الأولى هي العمل مع مجموعات السمات المستخرجة بشكل منفصل تماما كل سمة مستقلة لها حالة تطبيق للخوارزميات الثلاث NB,SVM,M.C وكانت خلاصة النتائج لتلك العملية في الجدول (1). الخطوة الأخرى استخدمنا مجموعة السمات المستخرجة بشكل متفاوت الارتباط كالاتي F1+F1 معا وطبقنا التجربة ثم F1+F2+F3 مع بعض وطبقنا نفس الإجراءات، تلى ذلك F1+F2+F3+F4 وأخيرا كل السمات مع بعض F1+F2+F3+F4+F5 الجدول (2) يحتوي على خلاصة نتائج تلك العملية.

جدول (2): نتائج التطبيق ومقياس الدقة على السمات معا

Features	Total correct			Accuracy Percent		
	NB	SVM	M.C	NB%	SVM%	M.C%
F1,F2	38	37	38	95	92.5	95
F1,F2,F3	33	35	36	82.5	87.5	90
F1,F2,F3,F4	25	27	35	62.5	67.5	87.5
F1,F2,F3,F4,F5	38	38	39	95	95	97.5
Average	33.5	34.25	37	83.75	85.62	92.5

7. الخاتمة

بعد إجراء تجربتنا واخراج النتائج كما في الجدولين وبالنظر إلى الجدول رقم (1) نلاحظ ان النتائج كانت متفاوتة تماماً حيث كانت أعلى نسبة هي 97.5 % لخوارزمية سلسلة ماركوف MC على سمة طول الكلمة وادنى نسبة في الجدول هي 57.5 % لخوارزمية NB على سمة القافية. من الجدول أيضا نلاحظ أن النسب المنخفضة جدا كانت في سطر القافية وتلا ذلك سطر الكلمة الاولى مما يعني ذلك أن القافية لم تكن ذات تأثير كبير في إسناد النص بشكل منفصل ويرجع ذلك لكون القافية تتكرر كثيرا عند أغلب الشعراء. ومن الجدول أيضا يتضح أن أعلى معدل دقة كان من نصيب خوارزمية ماركوف بينما تلتها SVMو .NBفي الجدول رقم (2) والذي يحتوي على نتائج تطبيق نفس الخوارزميات مع أكثر من سمة في كل مرة، نلاحظ أن أعلى نسبة هي 97.5 % عندما تمّ أخذ جميع السّمات معاً لسلسلة ماركوف بينما تساوت NB,SVM بنسبة صحة وصلت إلى95 % وهذا يعني أن جميع السمات أظهرت نتائج ممتازة مجتمعة مع كل الخوارزميات. غير أننا لو نظرنا إلى الجدول (2) السطر الأول لنتائج السّمتين F1,F2 نجذ أن سلسلة ماركوف و Naïveأخذت نفس النسبة بينما انخفضت عند SVMلتصل إلى 92.5 وهذا بطبيعة الحال يرجع إلى التقدير الصحيح لاختيار التقنية مع تلك السمات. السمات F1,F2 حصلت على قيم دقة عالية بعد السمات الكلية F1,F2,F3,F4,F5 وعند إضافة طول الجملة الشعرية إلى الحرف وطول الكلمة انخفضت النسبة في جميع الخوارزميات في السطر الثاني حيث كانت أعلى قيمة هي 90 % لسلسلة ماركوف ويرجع ذلك لكون طول الجملة الشعرية في الشعر الكلاسيكي متشابها عند جميع الشعراء إذ يعتمدون على البحر الشعري والذي بدوره يلزم الشاعر بتفعيلات خاصة تتحكم في طول البيت الشعري. في السطر الثالث كانت أدنى القيم للجميع حيث تمّ إضافة سمة الكلمة التي في أول الجملة الشعرية وهذا يعود إلى طبيعة الشعر العمودي القاسية التي تلزم الشاعر والمؤلف بانتقاء الكلمات المتناغمة مع الوزن. السطر الأخير والذي أضيفت إليه القافية والتي كانت غير ذات دلالة بشكل مستقل في جدول (1) إلا أنها هنا رفعت من أدنى القيم في السطر الثالث إلى أعلى القيم في السطر الرابع وهذا يعني دورها بمعية السّمات الاخرى. مما سبق نستطيع القول أن النتائج جاءت مطابقة للتوقعات حيث وصلت أعلى نتيجة لسلسلة ماركوف هي 97.5 % مع جميع السمات ونفس النسبة مع طول الكلمة المستخدمة وبمعدل وصل إلى 92.5 % إلا أننا نطمح إلى المزيد في الأعمال اللاحقة. وإلى تجاوز تلك العقبات نقترح إدخال متغيرات أخرى خاصة بالشعر مثل البحر والكلمات النادرة واستخدام الوزن، المترادفات، وبعض الخصائص الشعرية الصِّرْفة ويمكن اخذ سمات خاصة بالشعر والعصر الذي ينتمي إليه. كما نقترح زيادة عدد العينة والنصوص الشعرية مع الأخذ بعين الاعتبار توحيد أطوال نصوص الاختبار، وتطبيق نفس المعايير عليها، كذلك استخدام تقنيات أخرى من تقنيات التنقيب عن النصوص وتعليم الآلة، ومقارنة النتائج الجديدة مع هذه النتائج.

8. المراجع

[1] بوتبيا، الحسن، "المفاضلة بين النظم والنثر وأشكال التداخل بينهما في العصر العباسي"، مراكش: المطبعة والوراقة الوطنية، ط:1، 2002، ص66-67.

[2] القرشي، أبو زيد محمد بن أبي الخطاب، "جمهرة أشعار العرب"، تحقيق: محمد علي الهاشمي، دمشق: دار القلم، ط:2، 1986، ص146 وما بعدها، وابن فارس، أبو الحسن أحمد، الصاحبي في فقه اللغة وسنن العربية في كلامها، تحقيق: عمر الطباع، بيروت: مكتبة المعارف، ط:1، 1993، 465-468.

[3] E. Stamatatos, "A survey of modern authorship attribution methods," J. Am. Soc. Inf. .Sci. Technol., vol. 60, no. 3, pp. 538–556, 2009

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

[5] .P. Juola, Authorship Attribution. Now Publishers Inc, 2008

[6] 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

[7] 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

[8] F. Howedi and M. Mohd, "Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data," Comput. Eng. Intell. Syst., vol. 5, no. 4, pp. .48–57, 2014

A. Abbasi and H. Chen, “Applying authorship analysis to Arabic web content,” Intell. Secur. Informatics, 2005. [9]

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

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

H. S. Al hichri, “ref,” in Expert System for Classical Arabic Poetry (ESCAP). [12]

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

A. S. Altheneyan and M. E. B. Menai, “Naïve Bayes classifiers for authorship attribution of Arabic texts,” J. King Saud Univ. - Comput. Inf. Sci., vol. 26, no. 4, pp. 473–484, 2014. [14]

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. [15]

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

M. S. Desouki and A. Al-abdo, “Experiments in Mining Arabic Texts ” محاولات في التنقيب عن النصوص“,” vol. 2, no. 1, pp. 14–18, 2012. [17]

E. Stamatatos, “A survey of modern authorship attribution methods,” J. Am. Soc. Inf. …., 2009. [18]

C. Sanderson and S. Guenter, “On authorship attribution via Markov chains and sequence kernels,” Proc. - Int. Conf. Pattern Recognit., vol. 3, pp. 437–440, 2006. [19]

S.F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 13:359–394. [20]

9. جدول الألفاظ

English	عربي
Abstract	الملخص
Key words	الكلمات الجوهرية
Authorship Attribution	اسنادية التأليف
stylistic Features	خصائص الكتابة
Markov Chain	سلسلة ماركوف
References	المراجع

10. الخلاصة باللغة الانجليزية

Logical Inference in the Holy Quran

Al-Falahi Ahmed [#1], Ramdani Mohamed [*2], BellafkihMostafa[#3], Al-Sarem Mohammed [#4]

[1,2#]Département d'informatique,[#1,2]FSTM Université Hassan II Mohammedia,
[3#]Institut National des Postes etTélécommunications,[#3] INPT-Rabat,
4#College of Computer, Taibah University
[1,3#]Rabat-Morocco,[*2]Mohammediah-Morocco,[#4] Medina-KSA
[1]flahi79@gmail.com
[3]bellafki@inpt.ac.ma
*Mohammediah-Morocco
Université Hassan II Mohammedia
[2*]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, and First word in sentence are used as input data for Markov Chain 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 97.5%.