# نظام مصادقة التأليف للنصوص العربية باستخدام خوارزميات التعلم الآلي

رسالة مقدمة الى كلية الهندسة جامعة الفيوم
ضمن متطلبات الحصول على درجة الماجستير فى العلوم الهندسية

**قسم الهندسة الكهربية**
(تخصص هندسة الحاسبات و النظم)

إعداد

**م/ مي شعبان احمد محمد براني**

إشراف

**أ.د/ رانيا أحمد عبد العظيم أبو السعود**
أستاذ الاشارات الرقمية بقسم الهندسة الكهربية
كلية الهندسة جامعة الفيوم
(المشرف الرئيسي)

**د/ أحمد سلامة إسماعيل أمين**
مدرس بقسم نظم المعلومات
كلية الحاسبات والمعلومات جامعة الفيوم
(المشرف المشارك)

كلية الهندسة- جامعة الفيوم
الفيوم- جمهورية مصر العربية
أبريل، 2024

# ملخص البحث

ظ ً ا للاستخدام العالمي السريع للإنترنت، تلعب مصادقة التأليف اليوم دور ً ا حاسمً ا في تحديد الهوية الحقيقية للمؤلف أو أصل العمل المكتوب. هي الطريقة التي يحاول اللغوي من خلالها معرفة من كتب عملاً مجهولاً بناءً على اللغة المستخدمة والأسلوب اللغوي للكاتب. لقد تم بذل جهد بحثي كبير لمحاولة حل هذه المشكلة. في البداية كانت هذه الجهود مبنية على أنماط إحصائية، وفي الآونة الأخيرة ركزت على مجموعة من التقنيات بدءً ا من الذكاء الاصطناعي. تم تحقيق اختراق مبكر مهم من قبل موستلر ووالاس في عام 1964 [1]، اللذين كانا رائدين في استخدام "الكلمات الوظيفية" - عادة الضمائر، وحروف العطف، وحروف الجر - باعتبارها السمات التي يرتكز عليها اكتشاف أنماط الاستخدام ذات الصلة بمؤلفين محددين.

كان التركيز الرئيسي هو دراسة اللغة الإنجليزية والإسبانية والألمانية. ونظر ً ا لصعوبة الجمل العربية وطولها، فقد أولى الأكاديميون اهتمامً ا أقل باللغة. على مدى العقدين الماضيين، توسعت النصوص العربية على شبكة الإنترنت بشكل كبير، مما جعلها رابع أكثر اللغات استخدامً ا. ومن هنا، هناك طلب كبير على التصنيف الفعال للنص العربي.

يعد تصنيف النص مجالاً كبير ً ا في الذكاء الاصطناعي التطبيقي. باستخدام التعلم الآلي والتكنولوجيا المتقدمة الحديثة، يمكن أتمتة مهمة التصنيف مما يؤدي إلى معالجة فائقة السرعة وفعالة. في مهام تصنيف النص، تعد المعالجة المسبقة لمجموعة البيانات وعملية اختيار الميزات مهمة جدًا لتحقيق أفضل نتيجة. قد يكون أداء المصنف الضعيف الذي يتم تغذيته بميزات ذات معنى أفضل من المصنف القوي الذي يتم تغذيته بميزات منخفضة الجودة.

في هذه االرسالة، تمت مناقشة مشكلة توثيق التأليف فيما يتعلق بالعديد من الاهتمامات العربية باستخدام مجموعة بيانات مكونة من عشرة مؤلفين مختلفين (الفارابي، الغزالي، الجاحظ، المسعودي، المقريزي، الطبري، التوحيدي، ابن الجوزي، ابن رشد، وابن سينا). يحتوي كل كاتب على 60 ملفًا نصيًا، لذا تحتوي مجموعة البيانات على 600 ملف. يبدأ هيكل النهج الخاص بنا بالمعالجة المسبقة لمجموعة البيانات التي تتضمن إزالة كلمات الإيقاف والتطبيع والترميز والإنشاء لدراسة تأثير المعالجة المسبقة على أداء المصنفات المختلفة.

لقد اختبرنا تقنيات مختلفة لاستخلاص الميزات مثل (TF-IDF) - (BOW) بالاضافة الي ميزات البنية و المعجم و بناء الجملة والدلالات واسلوب الكتابة لدراسة تأثيرها علي اللغة العربية. العملية التي تتبع استخراج الميزات هي اختيار الميزات، والمعروف أيضًا باسم اختيار السمات. يسعى إلى استخراج مجموعة فرعية من الميزات الأكثر أهمية من مساحة الميزات المتفرقة دون الإضرار بأداء المصنف. بعد

ذلك، يتم تقييم متجهات الميزة المحددة باستخدام عدة طرق تصنيف. قمنا بالتحقق من أداء عشر خوارزميات تصنيف،

نقوم stackingبما في ذلك (GBC) -(BC)- (DT)- (SGD)- (KNN)- (MNB)- (LR)- (RF)- (SVC)-(MLP) .
أيضًا بوصف وتقييم النهج المختلط الذي يجمع بين المصنفات باستخدام مصنف
لتحسين الأداء التنبئي العام بشكل فعال. أظهرت النتائج أن المصنف Stacking باستخدام LR
كمصنف تعريفي مع مصنفيين اساسيين قد حقق دقة قدرها 96% واستخدام Stacking باستخدام LR
باعتباره المصنف التعريفي مع استخدام ثلاث مصنافات اساسية قد حقق دقة بنسبة 96.67%.
بالاضافة الي ذلك حقق Stacking باستخدام RF كمصنف تعريفي مع مصنفيين اساسيين دقة قدرها 97%
في حين حقق Stacking باستخدام RF كمصنف تعريفي مع ثلاث مصنفات اساسية دقة قدرها 98.33%.
تم ايضا استخدام اختيار ميزة Chi-squaredلتعزيز فعالية الميزات المحددة و تحسين اداء
نموذج Stacking.