إسناد المؤلف في المدونات العربية المصغرة

إعداد

أحلام محمد العمري

مستخلص لبحث مقدم لنيل درجة الماجستير في العلوم
(نظم معلومات الحاسوبية)

إشراف

د.معظم أحمد صديقي

كلية الحاسبات وتقنية المعلومات
جامعة الملك عبد العزيز
جدة – المملكة العربية السعودية
شعبان ١٤٣٩ هـ – مايو ٢٠١٨ م

إسناد المؤلف في المدونات العربية المصغرة

أحلام محمد العمري

المستخلص

تعتبر اسناد المؤلف هي عملية تعرف على مؤلف نص معين. وتعتمد العملية على استخراج خصائص النص بأسلوب المؤلف. ويستخدم المؤلفون دون وعي أنماطًا نحوية في الكتابة مما يخلق بصمةً خاصً بهم. وأدى توافر النصوص الإلكترونية إلى زيادة التطبيقات المحتملة لإسناد المؤلف في مختلف المجالات مثل الذكاء وطب الحاسوب الشرعي والقانون الجنائي والقانون المدني والبحوث الأدبية. ويقدم هذا البحث إطارًا لإسناد المؤلف في المدونات العربية ليدرس مشكلة تتبع الهوية وقوة الأنماط والخصائص المستخدمة من قبل الكاتب. حيث يتم استخراج خصائص النص بواسطة أداة مادا أميرا (MADAMIRA) بثلاث مستويات وهي مستوى الحرف والكلمة والقواعد النحوية والتي يتم استخدمها لبناء نموذج تصنيف. وفي هذا النموذج يتم دمج هذه الخصائص تدريجياً من عدة مؤلفين واختبارها باستخدام ثلاث خوارزميات وهي خوارزمية الجار الأقرب (KNN), وخوارزمية شعاع الدعم الآلي (SVM), وخوارزمية الغابات العشوائية (RF).

Authorship Attribution in Arabic Microblog

By

Ahlam Mohammed Al-Amri

An abstract of thesis submitted for the requirements for the Degree of Master of Science in Computer Information System

Faculty of Computing and Information Technology

King Abdulaziz University

Jeddah – Saudi Arabia

Sha'aban 1439 H – May 2018 G

Authorship Attribution in Arabic Microblog

Ahlam Mohammed Alamri

ABSTRACT

Authorship attribution refers to the task of identifying the author of given text. The process extracts characteristics of that text which is perceived to have the author's style. Specifically, authors unconsciously use syntactic patterns in writing which create a personal writing fingerprint. The availability of electronic text has increased the potential applications of authorship attribution in various fields such as intelligence, computer forensics, criminal law, civil law and literary research. This research provides a framework of authorship attribution of Arabic microblogs to address the identity-tracing problem. This framework examines the power of stylometric features and scalability of authors using three different classifiers of Arabic microblog. The stylometric features that applied to the text of each user this framework which are; character, lexical and, syntactic. These features extracted by MADAMIRA tool and used to build features-based classification model. In this model the features are combining gradually over a different size set of authors using three classification techniques: Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and, Random Forest (RF).