Evaluating the Effect of Hybrid Arabic Classification Techniques Based on Naïve Bayes Algorithm

By:
Somaya Zacout

Supervised by:
Professor Mohammed Otair

Abstract

Text classification is a major data mining application that holds a high weight in the modern digitized world; it assists in many areas such as email filtering, digital libraries and security threats among many other areas. The research efforts in the Arabic text classification is still limited which creates a huge opportunity for new work, Arabic language is one of the top five most spoken languages in the world therefore, the availability of Arabic text is far from limited. In this study three of the most studied and used algorithms (SVM, ANN, and J48) are combined with the Naïve Bayes algorithm to create new hybrid algorithms namely (Vote NBSVM, Vote NBANN, Vote NBJ48) all of which are combined using the majority voting method in WEKA tool, another hybrid technique is used also based on the Naïve Bayes and combined with J48 algorithm named New NBJ48 algorithm. All the above mentioned algorithms were applied on three standard Arabic text datasets

with a total of 32262 documents and their performance was measured and compared.

Results show that the voting method in combining Naïve Bayes with other algorithms resulted in degrading performance of the algorithms when combined with Naïve Bayes algorithm using the voting method their accuracy levels dropped by 2.25% for the ANN algorithm, 6.62% for the SVM algorithm and by 2.52% for the J48 algorithm. However, the new algorithm NBJ48 showed superior outstanding results that are highly competitive and increased the accuracy of the J48 by 5.72%, this is especially important taking in fact that this algorithm has never been applied on Arabic text.

تقييم تأثير خوارزمية NAÏVE BAYES الهجينة على تصنيف النصوص العربية

إعداد
سمية ابراهيم زقوت

إشراف
الاستاذ الدكتور محمد عطير

الملخص

يعتبر تصنيف النصوص من أهم التطبيقات في مجال التنقيب عن البيانات و التي لها وزنًا كبيرًا في العالم الرقمي الحديث، حيث أنها تستخدم في العديد من التطبيقات مثل فلترة البريد الالكتروني، تصنيف المكتبات الالكترونية ، و كذلك تحديد التهديدات الأمنية في العديد من المجالات الاخرى.تعد الجهود البحثية المبذولة في مجال تصنيف النصوص العربية محدودة مما يفتح المجال للعمل على بحوث جديدة ومتنوعة، حيث تعد اللغة العربية من أكثر خمسة لغات متحدث بها عالميا، مما يدل على توافر النصوص العربية الالكترونية بكثرة.في هذ الرسالة تم تطبيق مجموعة من الخوارزميات الهجينة هيSVM) ، ANN، و (J48 لمعرفة تأثير خوارزمية Naïve Bayes على أداء هذه الخوارزميات عند استخدامها مع كل واحدة منها على حدى. و من خلال عملية الدمج فقد تم الحصول على ثلاثة خوارزميات هجينة سميت بـ ,Vote NBSVM) :
Vote NBANN, Vote NBJ48)و قد دمجها من خلال طريقة التصويت الموجودة في الويكا. كما تم استخدام تقنية هجينة اعتمادا على الـ Naïve Bayes مع خوارزمية J48 وسميت الخوارزمية الجديدة باسم NBJ48.وتم استخدام ثلاث قواعد بيانات قياسية لنصوص عربية تحتوي على 32262 مستند عربي في

المجمل لتقييم عمل هذه الخوارزميات. تم مقارنة اداء الخوارزميات المدموجة مع خوارزمية Naïve Bayes باستخدام طريقة الانتخاب، وتم تطبيق خوارزمية جديدة تم دمجها مع Naïve Bayes في اداة Weka .

أظهرت النتائج أنه عند دمج طريقة الانتخاب لل Naïve Bayes مع الخوارزميات الأخرى، فإن ذلك أدى إلى التقليل من الأداء في الخوارزميات جميعها. فقد انخفضت مستويات الدقة بها بنسبة 2.25٪ لخوارزمية ANN، و 6.62٪ لخوارزمية SVM و 2.52٪ لخوارزمية .J48 ولكن النتائج أظهرت تفوق الخوارزمية الجديدة NBJ48 بنتائج ممتازة ذات تنافسية عالية وزادت من دقة J48 بنسبة 5.72٪، وهذا مهم جدا خاصة أن هذه الخوارزمية لم يتم تطبيقها أبداً على النصوص العربي.