نهج التعويض عن القيم المفقودة في تصنيف التغريدات غير المرغوبة

وفاء حسين إسماعيل دفع

المستخلص

تعد مشكلة البيانات المفقودة تقريباً مشكلة لا يمكن تجنبها في الأبحاث التي تعتمد على جمع كمية كبيرة من البيانات. وقد تكون المعلومات مفقودة عندما لا يرغب المستخدمون في تقديم بياناتهم الشخصية بسبب مخاوف الخصوصية أو غياب الحافز. خاصةً بالنسبة للبيانات الاختيارية التي تطلبها بعض الأنظمة مثل نظام تويتر (Twitter). يمكن أن يؤثر إهمال أو حذف البيانات المفقودة سلبًا على القوة الإحصائية لدراسة ما ، كما يمكن أن يُنتج تقديرات متحيزة لفئات ما على حساب فئات اخرى ، مما يؤدي إلى استنتاجات غير صحيحة. وهدفت هذه الدراسة إلى تحسين جودة تصنيف البيانات من خلال استخدام نماذج متعددة للتعويض عن البيانات المفقودة والتي تم جمعها من حسابات المغردين على تويتر، ولتحسين فهم أسلوب عمل مرسلي الرسائل غير المرغوب فيها. وقدم البحث دراسة تحليلية للتعويض عن بيانات تويتر المفقودة باستخدام طريقتين: طريقة التعويض المتعدّد من خلال المعطيات المتسلسلة الكاملة (MICE) وطريقة زيادة التوقعات (EM) وتم اختبار نتيجة أثر البيانات الكاملة بعد عملية التعويض من خلال طريقتين: التحليل الإحصائي وخوارزميات التصنيف. وقد أظهرت نتائج النموذج المتبنى بإستخدام تقنية التعلم الآلي (مصنف الغابات العشوائية) دقة مقدارها (٩٦,٢%) بإستخدام عملية التعويض المتعدد من خلال المعطيات المتسلسلة الكاملة (MICE) كما أظهرت الدراسة أيضًا أن عدد إشارة إعلام المستخدمين (@mention) وعدد عناوين URL في كل تغريده هي أكثر ميزتين محتملتين يمكنهما اكتشاف الرسائل غير المرغوب فيها .

المشرف : أ.د. أميمة بامسق

Missing Values Imputation Approach in Spam Twitter Classification

Wafaa Hussein Ismail Daffa

ABSTRACT

Missing data is almost unavoidable problem in research that depends on collecting large amount of data. Information may be missing when users are unwilling to provide their personal data due to privacy concerns or absence of motivation. This is mostly true for elective data requested by the systems like Twitter. Omitting missing data can negatively affect the statistical power of a study and can produce biased estimates, leading to invalid conclusions. Our aim of this study is to improve the quality of data classification through using multiple imputation model on incomplete data collected from Twitter microblogging accounts; and to better understand the mode of operation of spammers when identifying their posts. Our research presents a case study of data analysis using two methods for dealing with Twitter missing data: multiple imputation (MI) and Expectation Maximization (EM). The impact of resulted imputed data is then tested through two ways: statistical analysis and classification algorithms. Our developed predictive model reported an evaluation result of 96.2% accuracy using Random Forest classifier from Multiple Imputation by chained equations (MICE) complete dataset. The study also identified that the number of user mentions and number of URLs in each tweet are the most two potential features that can detect spam in posts at most.

Supervised By
Prof. Omaimah Bamasag