The Effect of Dimensionality-Reduction Methods of Feature Selection on Fake News Detection based on Machine Learning Techniques

By

Marwan Ghazwan Mitab

Supervisor

Prof. Mohammad Otair

Abstract

News online has evolved as the main basis of data for individuals, specifically in the political globe. Much of the data that finds on social is existing faked, either consciously or unintentionally. The capability to recognize, assess, and process this data is of great importance. The fake news detection standards deal with the issue as a binary category study. This study explores the effect of feature selection methods Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on machine learning techniques such as (Support Vector Machine (SVM), Decision Tree (DT), Random Forest Tree (RFT), and Logistic Regression (LR)) to create a model of a proposed machine learning algorithm, that can

classify fake news as true or false, considering accuracy, f1-score, precision etc., as the evaluation metrics of proposed methodologies. The results presented that using ML without reducing dimensionalities is more accurate than using it, and SVM has more accuracy is 93%, but it takes more time processing than another that reached to 30s. In this study, it turns out that the detection of fake news depends on the features, as the greater the number of features, the more accurate the detection of news. However, this does not apply to all problems, as reducing the features is a necessary aspect in other problems to give better and more accurate results.

Keywords: Fake news detection; Machine learning; Feature Selection.

تأثير طرق تقليل الأبعاد لاختيار الميزات على اكتشاف الأخبار الزائفة استنادًا إلى تقنيات التعلم الآلي

إعداد

مروان غزوان متعب

إشراف

الأستاذ الدكتور محمد عطير

الملخص

تطورت الأخبار عبر الإنترنت كمصدر رئيسي للبيانات للأفراد ، وتحديداً في العالم السياسي. الكثير من البيانات التي يتم العثور عليها على مواقع التواصل الاجتماعي مزيفة ، إما عن قصد أو عن غير قصد ومن هنا القدرة على التعرف على هذه البيانات وتقييمها ومعالجتها لها أهمية كبيرة. حيث تتعامل معايير الكشف عن الأخبار المزيفة مع الموضوع كنوع من الفئة الثنائية.

تستكشف هذه الدراسة تأثير طرق اختيار الميزات ، تحليل المكونات الرئيسية (PCA) والتحليل التمييزي الخطي (LDA) على تقنيات التعلم الآلي مثل (Support Vector Machine (SVM) ، Decision Tree (DT)، Random Forest Tree (RFT) ، و الانحدار اللوجستي (LR)) لإنشاء نموذج لخوارزمية التعلم الآلي المقترحة ، والتي يمكنها تصنيف الأخبار المزيفة على أنها صحيحة أو خاطئة ، مع مراعاة الدقة ، ودرجة f1 ، والدقة وما إلى ذلك ، كمقاييس

تقييم للمنهجيات المقترحة. أظهرت النتائج أن استخدام ML دون تقليل الأبعاد هو أكثر دقة من استخدامه ، وأن SVM لديه دقة أكبر بنسبة 93٪ ، ولكنه يستغرق وقتًا أطول للمعالجة مقارنة بآخر يصل إلى 30 ثانية. في هذه الدراسة اتضح أن اكتشاف الأخبار المزيفة يعتمد على الميزات ، فكلما زاد عدد الميزات ، زادت دقة الكشف عن الأخبار. ومع ذلك ، هذا لا ينطبق على جميع المشاكل ، حيث أن تقليل الميزات هو جانب ضروري في المشاكل الأخرى لإعطاء نتائج أفضل وأكثر دقة.

الكلمات المفتاحية: كشف الأخبار الكاذبة. التعلم الالي؛ اختيار ميزة.