

TOWARDS MULTILINGUAL FEATURE ENGINEERING FOR SMS SPAM DETECTION

ABSTRACT

Billions of money is lost by mobile phone users every day due to SMS spam, a social engineering skill attempting to obtain sensitive information such as passwords, Personal identification numbers and other private data by masquerading as a trustworthy entity through Short message service. The design of efficient feature engineering techniques is the key to reducing these financial losses. Most machine learning classifiers solutions today produce less accurate predictions and are inefficient due to the dynamic nature of spamming. It is in this background that the study proposes an ensemble feature engineering techniques for SMS spam, that can be used for multilingual natural language processing, data training, validation and testing of a model. The contributors of data include the UCI database and local repositories that contain a mixture of English and Swahili messages. Machine learning and data mining experiments are conducted using the WEKA tool and the results and discussions are presented in the form of descriptive statistics. This novel approach recorded an overall satisfiable accuracy of 99%.

KEYWORDS: Algorithm, Detection, ensemble, Feature engineering, Machine learning, Sms.

Primary authors: Mr KIPKEBUT, ANDREW; Dr THIGA, Moses; Dr OKUMU, Elizabeth

Presenters: Mr KIPKEBUT, ANDREW; Dr THIGA, Moses; Dr OKUMU, Elizabeth

Track Classification: Masters and Doctoral Colloquium Abstracts