Title

An Advanced Semantic Feature-Based Cross-Domain PII Detection, De-Identification, and Re-Identification Model Using Ensemble Learning.

Authors

Kulkarni, Poornima; N. K., Cauvery; R., Hemavathy

Abstract

The digital data being core to any system requires communication across peers and human machine interfaces; however, ensuring (data) security and privacy remains a challenge for the industries, especially under the threat of man-in-themiddle attacks, intruders and even ill-intended unauthorized access at warehouses. Almost all digital communication practices embody personally identifiable information (PII) like an individual's address, contact details, identification credentials etc. The unauthorized or ill-intended access to these PII attributes can cause major losses to the individual and therefore it is inevitable to identify and de-identify aforesaid PII elements across digital platforms to preserve privacy. Unfortunately, the diversity of PII attributes across disciplines makes it challenging for state-of-arts to perform PII detection by using a predefined dictionary. The model developed for a specific PII type can't be universally viable for other disciplines. Moreover, applying multiple dictionaries for the different disciplines can make a solution more exhaustive. To alleviate these challenges, in this paper a robust ensemble of ensemble learning assisted semantic feature driven crossdiscipline PII detection and de-identification model (EESD-PII) is proposed. To achieve it, a large set of text queries encompassing diverse PII attributes including personal credentials, healthcare data, finance attributes etc. were considered for training based PII detection and classification. The input texts were processed for the different preprocessing tasks including stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization. The tokenized text was processed for Word2Vec driven continuous bag-of-word (CBOW) embedding that not only provided latent feature space for analytics but also enabled de-identification to preserve security aspects. To address class-imbalance problems, synthetic minority over-sampling techniques like SMOTE, SMOTE-BL, SMOTEENN were applied. Subsequently, the resampled features were processed for the feature selection by using Wilcoxon Rank Sum Test (WRST) method that in sync with 95% confidence interval retained the most significant features. The selected features were processed for Min-Max Normalization to alleviate over-fitting and convergence problems, while the normalized feature vector was classified by using ensemble of ensemble learning model encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier. The proposed model performed a consensus-based majority voting ensemble to annotate each text-query as PII or Non-PII data. The positively annotated query can later be processed for dictionary-based PII attribute masking to achieve de-identification. Though, the use of semantic embedding serves the purpose towards NLP-based PII detection, de identification and re-identification tasks. The simulation results reveal that the proposed EESD-PII model achieves PII annotation accuracy of 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%.

Subjects

MACHINE learning; HUMAN-machine systems; DATA security; DATA privacy; DIGITAL communications

Publication

International Journal of Advanced Computer Science & Applications, 2024, Vol 15, Issue 12, p763

ISSN

2158-107X

Publication type

Academic Journal

DOI

10.14569/ijacsa.2024.0151277