ISSN: 1391-8796
Proceedings of 7th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
January 22, 2020

Gender, race and religion prediction of Sri Lankan personal names using machine learning techniques

Chathuranga P.D.T.*, Loresnsuhewa S.A.S. and Kalyani M.A.L.

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka

Prediction of identification details such as gender, race and religion of a person can help natural language processing related tasks to perform better. Also it can be used to speed up the existing digital application filling processes by providing suggestions. To the best of our knowledge, few researches were carried out on gender prediction and no research on race or religion prediction was carried out for Sri Lankan names. We performed gender, race and religion prediction based on Sri Lankan personal names, which were written using both Sinhala Unicode and English characters. Feature vectors were constructed as character n-grams and Multinomial Naïve Bayes & Support Vector Machine classification techniques were used for the prediction. Highest accuracies between 89% - 98% were obtained for all three predictions performed. Promising results demonstrated the possibility to use n-gram models with machine learning techniques to predict gender, race and religion of Sri Lankan names.

Keywords: natural language processing, machine learning, gender prediction and prediction of identification details

*Corresponding author: pdtcrc@gmail.com