Machine Learning Approach for Intelligent Product Recommendation System based on Product Reviews Given in Sinhala Language

H.K.S.K. Hettikankanama[1*] and S.M.S.R. Manage[2]

[1]Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka
[2] Department of Research and Development, Sixvertex Software, Colombo 03
**Corresponding Author E-mail: keshalasanduni@gmail.com, TP: +94702130268*

In this era of technology businesses all around the world including Sri Lanka are moving to the web. The increased amount of businesses and their products online made the process of selecting the most appropriate, suitable and trustworthy product online highly exhaustive. To mine this vastly available data and make a better choice takes lots of effort, cost and time for a person to perform. Therefore data mining techniques can be used to intelligently mine data and provide the best suggestions to the customers. When it comes to Sri Lankan online businesses most users use Sinhala Language to provide their reviews. Here in this research some machine learning techniques are used to train a model to understand user reviews that are in Sinhala and Singlish and suggest products with best overall reviews. To develop this model Python programming techniques are used. For the model dataset of 2100 separate phrases were manually scraped from some online business pages in a way that it contains negative and positive sentiment. Annotation for sentiments was given by 5 annotators for increased reliability. A specific data cleaning was done by removing garbage tests like html tags, special characters, pronouns and numeric data as the whole result relies on the words included in the dataset. Especially after the cleaning the cleaned text is converted to "unidecode" mode to make it viable to use this knowledge for future usage. Feature extraction functions like 'Count Vectorization' and 'Tf-Idf vectorization' used and features stored using "LexiconBuilder". Supervised and Unsupervised learning techniques are used. Ensemble, RNN (with GRU), RNN (with LSTM), Word2Vec (with CNN), Word2Vec, Decision Tree and AdaBoost algorithms are compared for their accuracy. Word2Vec (with CNN) gave 100% Precision and F1-Score which led model overfitting and unacceptable. As the solution the model with high precision, high F1-Score and least overfitting/under-fitting was chosen. Therefore, Ensemble Classifier which had 69% F1-score and 70% of precision was selected as the most suitable model finally.

Keywords: Intelligent Product Recommendation Systems; Machine Learning; Natural Language Processing; Sentiment Analysis; Sinhala Product Reviews