Classification of Oncogenic Compounds in Consumer Products Using ChemBERTa & OCR

Kundanika Pradhan
Department of Computing Technologies, SRM University, Chennai, Tamil Nadu, India

Muhammed Anish
Department of Computing Technologies, SRM University, Chennai, Tamil Nadu, India

Abstract

Guaranteeing consumer safety and regulatory compliance for the cosmetics and personal care market involves thoroughly examining product ingredients. Conventional methods proved inadequate in detecting hazardous materials accurately and with high efficiency. A new method of detecting oncogenic compounds from cosmetic and personal care products using OCR and deep learning is introduced in this paper. The OCR retrieves ingredient information from product labels and packaging and then analyzes it through ChemBERTa, a transformer model trained on chemical representations. A specially developed module retrieves the SMILES representation of each ingredient extracted via an API-based mechanism. ChemBERTa is compared with the usual machine learning classifiers, such as Support Vector Machine (SVM), Random Forest, Decision Trees, Bagging, and XGBoost, in this work. ChemBERTa is a better option than traditional classifiers, and it has better accuracy when predicting the carcinogenicity of chemical compounds. An interface that is easy-to-use has been deployed using Streamlit that combines ChemBERTa and Llama 3.2 to present an informative experience for users. These outcomes show us the promise of deep learning in enhancing harmful chemical detection and classification, offering a powerful tool for safer consumer goods.

Keywords

ChemBERTa, deep learning, oncogenic chemicals, optical character recognition.