ABSTRACT

Directorate General of Customs and Excise (DGCE), an Indonesian Government agency under the Ministry of Finance, is responsible for ensuring importer or exporter classify their declared goods based on the Harmonized System Code (HS Code). Customs officers determine HS Codes based on their expertise, which has different levels of expertise. It can allow for differences in determining the HS Code for the same type of goods, even for the same importer. A previous study indicated that approximately 30% of submitted HS codes were inaccurate.

This study aims to develop a method for creating machine learning workflows during the HS Code classification by the characteristics of the problems in DGCE. In solving multiclass problems, this study adopted the sampling technique from previous studies. After the research sample is determined, the preprocessing stage is implemented to treat abbreviations, size metrics, misspellings, language translation, eliminating the exact words on the same line, and eliminating lines with one remaining word. This study then combines Term Frequency - Inverse Document Frequency (TF-IDF) with bigrams during feature extraction and implementation of One Hot Coding (OHC) for nominal categorical variables.

As a result, our machine learning models show that Linear SVM got an F1-score of 82.43% when classifying the HS Code's first four digits, and Linear SVM got an F1-score of 74.93% when classifying the HS Code's entire digits. Compared to baseline paper, those scores are 12.04% and 14.93% higher, respectively.

Keywords : HS Code, machine learning, Indonesian Customs.

INTISARI

Direktorat Jenderal Bea dan Cukai (DJBC), lembaga Pemerintah Indonesia di bawah Kementerian Keuangan, bertanggung jawab untuk memastikan importir atau eksportir mengklasifikasikan barang mereka berdasarkan Harmonized System Code (HS Code). Keputusan penentuan HS Code dilakukan oleh Pejabat Bea dan Cukai yang memiliki tingkat keahlian yang berbeda. Hal ini dapat memungkinkan perbedaan dalam menentukan HS Code atas jenis barang yang sama, bahkan untuk importir yang sama. Hasil studi sebelumnya mengindikasikan bahwa sekitar 30% dari HS Code yang di *submit* tidak akurat.

Penelitian ini bertujuan mengembangkan metode untuk membuat *workflow machine learning* saat klasifikasi HS Code sesuai karakteristik permasalahan pada DJBC. Dalam menyelesaikan *multiclass problem,* penelitian mengadopsi teknik sampling dari penelitian sebelumnya. Setelah sample penelitian ditentukan, kemudian tahapan *preprocessing* diimplementasikan untuk perlakuan singkatan, metriks ukuran, salah eja, penerjemahan bahasa, hilangkan kata yang sama di baris yang sama, dan hilangkan baris dengan satu kata yang tersisa. Penelitian kemudian mengkombinasikan Term Frequency - Inverse Document Frequency (TF-IDF) dengan bigrams pada saat ekstraksi fitur serta implementasi One Hot Coding (OHC) atas *nominal categorical variables.*

Hasilnya, *workflow machine learning* yang dikembangkan menunjukkan bahwa Linear SVM memperoleh *F1-score* 82,43% saat mengklasifikasikan empat digit pertama HS Code, dan Linear SVM memperoleh *F1-score* 74,93% saat mengklasifikasikan seluruh digit HS Code. Dibandingkan dengan *baseline paper*, *F1-score* tersebut secara berturut-turut lebih tinggi 12,04% dan 14,93%.

Kata kunci – HS Code, machine learning, DJBC.