A Study using Machine Learning with N-Gram Model in Harmonized System Classification

Prihastuti Harsani¹, Adang Suhendra², Lily Wulandari², Wahyu Catur Wibowo³
¹Department of Computer Science, Universitas Paukuan, Indonesia
²Department of Computer Science and Information Technology, Universitas Gunadarma, Indonesia
³Department of Computer Science, Universitas Indonesia, Indonesia

Abstract: Harmonized System or commonly called HS is a list of classifications of goods made systematically with the aim of facilitating the taxing, trade transactions, transportation and statistics that have been improved from the previous classification system. In international trade (import/export) each item to be traded must be determined its HS Code based on the description that accompanies the goods. The description of imported goods in the form of text will be translated into the classification of imported goods regulated in the 2017 Indonesian Customs Tariff Book BTKI is the Indonesian Customs Tariff Book that contains the goods classification system applicable in Indonesia, including Provisions for Interpretation (KUMHS), Notes, and Goods Classification Structures compiled based on the ASEAN Harmonized Tariff Nomenclature (AHTN) Harmonized System. The classification of goods based on the HS code faces several challenges, including HS Complexity, Gaps in HS terminology, The amount of text in the goods description. This study conducted an experiment that applied machine learning in classifying imported goods. The focus of this research is the classification based on short text categorization. Documents compiled on pandek text in accordance with the characteristics of the description of the goods. The study conducted experiments with three methods, namely: Libshorttext, text categorization (Text) and topic modeling. Feature extraction methods used are Term Frequency - Index Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA). Classification is done based on the 8 digit HS system. The goods description that accompanies transaction data has an average number of words as many as 7. Classification of goods based on the HS code is a matter of categorizing short texts. The used feature is the nGram model. The method used is Libshort, Text Categorization and topic modelling. Evaluation shows that libshort has the highest accuracy and f-score value followed by text categorization and topic modeling. SVM and KNN give two different results on the classification. Based on the experimental results, it is not yet concluded whether an increase in N values on the N-Gram model will result in a better F-Score value on short texts.

Keywords: HS, LDA, TF-IDF, Ngram, SVM, KNN

Introduction

Harmonized System or commonly called HS is a list of classifications of goods made systematically with the aim of facilitating the taxing, trade transactions, transportation and statistics that have been improved from the previous classification system. In international trade (import/export) each item to be traded must be determined its HS Code based on the description that accompanies the goods.

HS code determination can be done through analysis of goods based on the characteristics of the goods obtained from the characteristics of the goods. Based on the export/import notification form, the item description is inputted by the exporter/importer in accordance with the fields provided. The text entered is an open sentence that is inputted in accordance with the interpretation of the exporter/importer. HS code is determined based on the description that has been entered. Guidelines used by importers in interpreting goods are contained in regulations issued by the Ministry of Finance of the Republic of Indonesia [1]. The accuracy of interpretation is influenced by the knowledge, skills and experience possessed by traders and customs officials which will affect the accuracy of determining HS codes [2].

Determination of HS codes based on the actual description is a matter of text categorization. The description of imported goods in the form of text will be translated into the classification of imported goods regulated in the 2017 Indonesian Customs Tariff Book BTKI is the Indonesian Customs Tariff Book that contains the goods classification system applicable in Indonesia, including Provisions for Interpretation (KUMHS), Notes, and Goods Classification Structures compiled based on the ASEAN Harmonized Tariff Nomenclature (AHTN) Harmonized System.

The classification of goods based on the HS code faces several challenges, including, 1). HS Complexity. HS is a structured multipurpose nomenclature, organized in 21 Sections and 98 Chapters. Classification that is done manually requires carefulness, experience and good knowledge. 2). Gaps in HS terminology. There is a gap between the description of goods entered by the importer and the description of goods in the HS nomenclature used by customs. A simple string search helps the importer to find the relevant HS code slightly because of the difference between the
structured description of the HS nomenclature and the text description during the trading process. 3) The amount of text in the goods description. The description of the goods provided by the importer / exporter is also an open, concise and unstructured sentence. Input and interpretation errors due to characteristics like this have a high probability. The process of identifying keywords, information or features in the item description text will be a challenge for this research to obtain an accurate classification model. If the three challenges that arise in the interpretation of HS cannot be overcome then there will be disharmony in the classification of goods. Disharmony good classification causes international trade will experience obstacles because the demand for a good from the importing country will be responded to by different coding or numbering by the exporting country. This is certainly not desirable in the international world. Some of the consequences that can occur as a result of this disharmony are delays in delivery, due to longer custom clearance and penalties due to mismatch postal rates.

This study conducted an experiment that applied machine learning in classifying imported goods. The focus of this research is the classification based on short text categorization. Documents compiled on short text in accordance with the characteristics of the description of the goods. The study conducted experiments with three methods, namely: Libshorttext, text categorization (Text) and topic modeling. Feature extraction methods used are Term Frequency - Index Document Frequency (TF-IDF) and Latent Dirichlect Allocation (LDA). Dataset is the description of goods in Chapter 64 and classification is based on an 8-digit HS code.

**Previous Work**

Classification of goods based on the HS code is one of the cases in the text categorization problem. The item description is read as a document that stores the features of each class. In the case of imported goods, classification of goods based on the HS code means placing the goods into a certain class (HS Code). Based on studies that have been carried out, the classification of export / import goods based on the HS code has used several approaches, including fuzzy logic, rule base and machine learning. The study results show that the combination of text and image features provides better classification accuracy [3].

The main characteristic of the description of imported and exported goods is the use of short words. The difference between the document / text categorization in general and the classification of imported goods is the description of goods contained in imported goods is a collection of text with a small amount (short text), open sentences and one word to another is unrelated. With these characteristics needed a method or technique that can translate the text description of goods into the HS code class.

Research related to the categorization of short documents has been conducted. Experiments were carried out on short documents (<30 words) and class determination was carried out based on information extracted on the document. Previous studies used short news datasets and product descriptions in e-commerce [4][5]. These research use feature extraction with Bag of Word (TF) method. Other research uses topic modeling to find features that are hidden in each document's topic. Research using the topic modeling in the problem of text categorization has been implemented by several authors, including news text categorization using KNN and SVM. The results of the research show that the topic of modeling provides quite good accuracy [6][7].

**N-gram Model**

An N-gram is a consecutive sequence of N characters. All of the n-grams present in a file, and then count their frequencies. To allow identification, space character replaced the by the character “_”. This technique, purely statistical, requires no knowledge of the language of the text. Another strength of the N-grams is the automatic identification of the most common roots. The resistance to spelling errors and deformations is also an important property. Finally, this strategy does not have to delete the stop words or continue with the selection process [8].

The n-gram model consists of two general stages. That stage is n-gram generation and selection of n-gram characteristics. This first step is to represent each category by a vector in which all the n-grams generated with their occurrence number are shown.

In the second step, a profile is created for each category. A category profile must include all the N-grams characterizing this category as opposed to the other categories. To build the profiles of the categories a term selection approach needs to be used. The z2 statistic measures the degree of association between a term and the category. Its implementation is based on the assumption that a word whose frequency depends heavily on the category in which it occurs is useful in distinguishing between the categories.

In text categorization and topic modeling, n-gram has been implemented and provides good performance [9][10][11][12][13].

**HS Code Classification**

Research related to product classification based on HS codes using machine learning has been carried out for several different customs. The first study using data mining was conducted by Bastac in 2012. This study detected smuggling based on transaction data. Research is able to provide notification to customs and related authorities [14]. This research was conducted at Turkish customs. The next study was conducted by Ding for Singapore customs. The method used is Background-Net (B-Net). This method uses incremental learning in item description [15]. The results of this study
indicate the accuracy of the classification of goods with the description of goods with short text is smaller than the description of goods with longer text. Subsequent research was carried out by Turhan and Li. Both of these studies combine images and text for the determination of HS codes. The HS code used is a 4 digit HS Code and a 6 digit HS Code. The method used in these research is topic modeling and Convolutional Neural Network (CNN) [16][17]. The results of both studies indicate a combination of text and image will provide a better f-score. Subsequent research also applied CNN in determining the HS code conducted by Luppes[18]. Luppes identified HS codes for 2 digit and 4 digit HS Codes in Chapter 64. The focus of this research is how to classify short texts. This study claims that the proposed approach can increase the value of accuracy when compared to other methods as a base line

**Generalized Additive Model for Location, Scale and Shape**

**Methodology**

The stages in this study are illustrated in Figure 1. Trade dataset is footwear data and its accessories (HS code 64) in 2018 and 2019 obtained from the Directorate General of Customs and Excise and the site that records international trade transactions, namely www.zauba.com. Preprocessing is carried out on a dataset to get a corpus that contains a list of terms of footwear and equipment. Document indexing is the stage to get feature representation using TF-IDF. Feature extraction is done by two approaches, which are based on frequency and Latent Dirichlet Allocation (LDA)). The feature model used is the n-gram model (unigram and bigram). Furthermore, to produce a classification model used 3 methods, namely libshort, conventional text categorization and topic modeling with SVM (Support Vector Machine) and KNN (K-Nearest Neighbour) classifier.

![Figure 1. Research Flow Diagram](image)

**Short Text**

Short texts have the following characteristics as compared with standard texts [11].
1. Sparseness. Product description like in HS Specification have very short length with no more than 100 words. This leads to feature sparseness and to extract correct and key features for classification learning is very difficult.
2. Irregular. Short-text grammar is generally informal. Spelling errors also occur, and use open sentences. These unusual words contain a lot of noisy features and thus increase the computers’ difficulty in extracting information.
3. Real-time. Short texts are constantly updated on the transaction, so the volume of short texts is very high. This means that the classification algorithm has to perform very well on time.

Some examples of goods descriptions on imports:
- **brown high top leather over knee shoes rubber rain boot hanz shiny HZ-W316**
- **Sandalswamiscolor Blk Lgr**
- **Steel Toe Caps Ref. 1130-X Wf Epoxy A Without Strip-Sz 10**

**TD-IDF**

Term Frequency-Inverse Document Frequency (TF-IDF) is one approach which has found both simple and efficient for features extraction. TF-IDF is an information retrieval technique that can be used to determine the relevance of terms in documents in relation to a query. In this case it can be used for extraction of the function by deciding which words are most distinctive for that element in a text. TF-IDF consists of two steps, first calculating the term frequency (TF), and then calculating the inverse document frequency (IDF). In this case it can be used for feature extraction by determining which terms in a document are most distinguishing for that document. This method can also be viewed as a form of Bag of Words model, since it does not take grammar or order into consideration.
A Study using Machine Learning with N-Gram Model in Harmonized System Classification

LibshoText

LibShortText [4][16], which offers multiclassification support. Bigram [25] is used in the processing of short text, do not remove the pause words and lexical filtering. In this study the experiment was conducted by doing preprocessing.

Support Vector Machine

Another state-of -the-art machine learning technique which is commonly used is supporting vector machine (SVM) [18]. It is mainly used for classification SVM functions according to the theory of measuring margins. It essentially, boundaries are drawn between classes. The margins are drawn in such a way that the distance between the margin and the classes is maximum and thus the classification error is kept to a minimum. Pseudocode of SVM

INPUT : S,λ, T, k
Initialize : Choose w_1, s.t ||w_1|| ≤ 1/√K
For t = 1,2, ..., T
    Choose A_t ⊆ S, where |A_t| = k
    Set A^+ = \{(x,y) ∈ A_t : y(w_t,x) < 1\}
    Set η_t = \frac{1}{K}
    Set w_{t+1} = (1 - η_t)w_t + \frac{η_t}{K} \sum_{(x,y) ∈ A^+} yx
Set w_{T+1} = \min \left\{ 1, \frac{1/√K}{||w_{T+1}||} \right\} w_{T+1}
Output = w_{T+1}

KNN

The basic idea of this algorithm for each text to be categorized is that considering the K texts in the training set with the most similarities to the text, the category of this text may be verified according to the category of these K texts [18][19].

1) The vectors of the training texts are re-described according to the feature set.
2) Segmentation of the text to be categorized based on the characteristics, and determination of the text's vector representation.
3) Selecting the K texts that are the most similar to the given text. One can measure the similarity by:

\[ Sim(\vec{a}_i, \vec{a}_j) = \frac{\sum_{k=1}^{n} W_{k}, x_{i}, x_{j}}{\sqrt{\sum_{k=1}^{n} W_{k}^{2}}, x_{i}, x_{j}} \]

where, K can be set an original value, and then adjust it according to the experiments.
4) Computing the weight of the text in these K neighbors belongs to every subclass, the formula is as follows:

\[ p = \sum_{i \in \text{class}} Sim(\vec{x}, \vec{a}_i) y(\vec{a}_i, \vec{c}_j) \]

\( \vec{x} \) is the feature vector of the given text, \( Sim(\vec{x}, \vec{a}_i) \) is the similarity can be computes by previous formula, \( y(\vec{a}_i, \vec{c}_j) \) is the attribute function, its values is as:

\[ b(\vec{a}_i, \vec{c}_j) = \begin{cases} 1 & \text{if} \; \vec{a}_i \; \text{belong to} \; \vec{c}_j \\ 0 & \text{otherwise} \end{cases} \]

5) Comparing the weight of the category and classifying the text with the largest weight in the category;

Latent Dirichlet Methods

LDA document typically consists of a series of topics which can be represented by a particular distribution of frequencies or probabilities [20] . It is presumed that there are several secret topics in a text document. Every topic obeys a distribution of probabilities over the words on the feature. Latent Dirichlet Allocation (LDA) is a probabilistic generative model, which is first introduced by Blei. Documents are represented in the LDA topic model as random mixtures over latent topics. topic contains a probabilistic distribution over terms. The generating process for a document of LDA can be presented as follows:
1. Sample $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\text{Dir}(\alpha)$ is the Dirichlet distribution for parameter $\alpha$
2. Sample $\varphi_j \sim \text{Dir}(\beta_j)$, where $j \in \{1, \ldots, K\}$
3. For each of the word position $i,j$ where $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, N\}$
    a. Sample a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$
    b. Sample a word $w_{ij} \sim \text{Multinomial}(\varphi_{z_{ij}})$

$w_{ij}$ is the only observable variable, and the other variables are latent variables. According to the model, the probability of the model is:

$$P(W, Z, \theta, \varphi, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \times \prod_{i=1}^{N} P(Z_{ij} | \theta_i) P(W_{ij} | \varphi_{z_{ij}})$$

At first, by integrating out $\theta$ and $\varphi$ the likelihood of a document is obtained as follows:

$$P(Z, W; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(W, Z, \theta, \varphi, \beta)$$

$$= \int_{\varphi} \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \times \prod_{i=1}^{N} P(Z_{ij} | \theta_i) P(W_{ij} | \varphi_{z_{ij}})$$

Gibbs Sampling method to extract topics in the corpus, which is also adopted for Gibbs Sampling is a special case of Markov-Chain Monte Carlo (MCMC).

After finishing Gibbs Sampling, the word-topic matrix $\phi$ and $\theta$ are calculated as follows:

$$\phi_m = \left( \sum_{i=1}^{N} n_{m,i} \right) \left( \sum_{j=1}^{K} n_{m,j}^{(i)} + \rho \right)$$

$$\theta_{m,k} = \left( \sum_{i=1}^{N} n_{m,i}^{(j)} \right) \left( \sum_{j=1}^{K} n_{m,j}^{(i)} + \sigma \right)$$

**Results and Discussion**

**Dataset Collection**

Data on footwear that were successfully obtained were 677,195. Filtering is done to remove products with different titles and the same description and items that do not have the initial two digit code 64. Based on the results of the selection, the dataset consists of 33 classes from all 52 classes in Chapter 64. This means that there are several classes that do not have a dataset. Data distribution for each class can be seen in Figure 2.

![Figure 1. Distribution of Each Class](image-url)
The total number of terms produced is 9,264,801. Based on the results of the calculation of the average number of words in one document is 6,998 \( \approx 7 \). This amount is in accordance with the characteristics of the data contained in the Export / Import Notification form submitted by the importer / exporter. Primary data obtained from the importers / exporters of the Directorate General of Customs and Excise have the same characteristics (Figure 3). The stages of data preprocessing carried out are case folding, stop word removal and stemming.

**Figure 3.** Frequency Top 50 Tokens

**Document Indexing and Feature Representation**

Document indexing produces several tokens that will be used for word processing. In this study the document indexing used was TF-IDF. Feature representation used in this study is the result of document indexing using TF-IDF and topic modeling. TF-IDF is one of the parameters that can be used to represent features. Topic modeling can be used to obtain features based on the topics found. The list of words contained in each topic can be used as a feature that represents a particular class. In this experiment the feature model that will be used is the n-gram model (unigram and bigram). Experiments using the N-gram model are carried out to see the effectiveness of the bigram feature as a characteristic in the classification. In the BoW or TF model, if each feature is shown by each term (token) then the model to be used is unigram. Based on the characteristics of the description of the goods that consist of several words (short text) there are not many repeated words and do not have meaning as complete sentences. This research also investigates whether solving a phrase will cause the phrase to lose meaning. For example, the phrase "water proof" will have different features than "water" and "proof". Therefore, bigram as a preliminary study needs to be considered as a feature. The use of the n-gram model for feature representation is expected to provide better classification performance \([8][11]\). For all the products in the experiment, there were total of 43483 unique unigrams and around 182000 unique bigrams. In order to reduce the number of features, feature selection was conducted by experimenting with various choices for bag of words by picking top words based frequency of occurrence from each category, and LDA top topic words.

**Training and Test Sets**

Short text data set can be created from a variety of sources, such as news post titles, web comments, user tweets, user review and product description. This study uses transaction goods dataset from Indonesia customs. For training the classifier 80% of the data set was used for testing the rest of the 20% of the data set was used. From each product category of data 80% was used as a training data set and 20% as test data. The reason the percentages for each category were separately derived is to avoid the chance of over and under representation of data from a certain category (imbalanced classes). The description of the data set is for HS Code 4 digits shown in Table 1.
To identify the best way to increase accuracy in addition to feature selection, several classifiers were evaluated against the feature set and data collected. The classifiers that were used for the evaluation included multiclass SVM and K-nearest neighbors (with 5 neighbours) for Ngram model (Unigram and Bigram).

The data set used as a trial is the Item Data in Chapter 64, with 8 digits. Unlike in previous studies that only classify up to 6 digits. Based on the results of data processing, the F score measurement results are obtained as presented in Table 2. Table 2 shows that Libshort has the highest f score for the SVM classifier. The f score is obtained that is equal to 0.77 and 0.84 respectively for unigram and bigram. While the order of the lowest f-score value is in the process of categorizing text without a short text library. Representation of features is done by the n-gram model. The results show that bigram has better accuracy for all methods and classifiers. Previous studies have tested the use of n-grams in text categorization. The results show that the n-grams model will increase to n = 3, and will show a downward trend as n increases [8] Modeling topics that are unsupervised learning also gives better accuracy results on the bigram feature. These results are consistent with previous research conducted in testing the effectiveness of n-grams in modeling topics [12].

<table>
<thead>
<tr>
<th>HS-4 digit</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>6401</td>
<td>17334</td>
<td>4266</td>
</tr>
<tr>
<td>6402</td>
<td>43143</td>
<td>10785</td>
</tr>
<tr>
<td>6403</td>
<td>38790</td>
<td>9679</td>
</tr>
<tr>
<td>6404</td>
<td>15826</td>
<td>3956</td>
</tr>
<tr>
<td>6405</td>
<td>28417</td>
<td>7104</td>
</tr>
<tr>
<td>6406</td>
<td>23079</td>
<td>5769</td>
</tr>
</tbody>
</table>

**Classification**

**Table 1**

<table>
<thead>
<tr>
<th>Description Chapter 64 HS Code Dataset</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unigram</th>
<th></th>
<th></th>
<th>Bigram</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>F-score</td>
<td>R</td>
<td>P</td>
<td>F-score</td>
</tr>
<tr>
<td>Libshort</td>
<td>SVM</td>
<td>0.78</td>
<td>0.77</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.75</td>
<td>0.74</td>
<td>0.81</td>
<td>0.81</td>
<td>0.8</td>
</tr>
<tr>
<td>Text</td>
<td>SVM</td>
<td>0.48</td>
<td>0.47</td>
<td>0.64</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.76</td>
<td>0.73</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>LDA</td>
<td>SVM</td>
<td>0.43</td>
<td>0.41</td>
<td>0.47</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.65</td>
<td>0.64</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

**Topic Modelling**

Topic modeling is one of the unsupervised learning. In topic modeling, alternative hidden topics in the document are found by measuring the word distribution in hidden topics. In this study topic modeling was tested for classification based on HS codes. The topic modeling used is the Latent Dirichlect Methods. Based on the measurement of topic coherence, the unigram model produced 14 topics and 22 topics for the bigram model. The frequency distribution for the four highest topics is presented in Figure 4.
Based on Table 2 it is shown that the bigram model has better performance as a feature for product classification by topic modeling. In this study, the implementation of modeling topics on short texts has the lowest accuracy and f-scores in all classifiers. This can be due to the unique characteristics of the words that make up the description of the goods. The words used in the item description are open sentences with the frequency of each term in the description of small value. The use of topic modeling in feature extraction can be combined with the use of hyponyms or synonyms with a preliminary analysis of the dataset. The use of hyponyms or synonyms via Wordnet can improve accuracy because there are other words for certain terms that are free to use by importers. For example the description "Eva Sandal" has 2 tokens, namely: EVA and SANDAL. EVA is one type of plastic material and SANDAL can be interpreted as footwear that does not close the ankle. This definition can be translated by experts and through the use of hyponyms and synonyms then the similarity is measured. For future works, this research can also be developed through a rule-based approach and machine learning to obtain product classification algorithms that can apply throughout the chapter so that it can produce a hybrid system. Hybrid systems can be developed because in the classification of goods based on the HS Customs Code has provided a general rule is a hierarchical classification. This can be combined with the results of data processing on transaction data.

Conclusion

In this paper, experiments are carried out to classify goods in international trade. Classification is done based on the 8 digit HS system. The goods description that accompanies transaction data has an average number of words as many as 7. Classification of goods based on the HS code is a matter of categorizing short texts. The feature used is the Ngram model. The method used is Libshort, Text Categorization and topic modeling. Experiments carried out on the dataset of goods in Chapter 64 (Footwear and accessories). evaluation shows that libshort has the highest accuracy and f-score value followed by text categorization and topic modeling. SVM and KNN give two different results on the classification. In the SVM libshort, the f-score is greater than the KNN. In contrast to the other two methods SVM results in lower f-score values. The implementation of the Ngram model provides the conclusion that for short texts, the bigram feature produces a better f-score value than other methods. Based on the experimental results, it is not yet concluded whether an increase in N values on the N-Gram model will result in a better FScore value on short texts.

Acknowledgements

This research was supported by FMIPA Universitas Pakuan. We thank our colleague from Direktorat Jenderal Bea dan Cukai who provided data that greatly assisted the research.

References

Conference on Audio, Language and Image Processing, 2016, 749-753.


