Applying Machine Learning for Automatic Product Categorization

**Keywords:** Product classification, Machine Learning, Official Statistics

# Introduction

The North American Product Classification System (NAPCS) is a comprehensive, hierarchical classification system for products (goods and services) that is consistent across the three North American countries, and promotes improvements in the identification and classification of service products across international classification systems, such as the Central Product Classification System of the United Nations.

Every five years, the U.S. Census Bureau conducts an economic census, providing official benchmark measures of American business and the economy. Beginning in 2017, the economic census will use NAPCS to produce economy-wide product tabulations. Respondents are asked to report data from a long, pre-specified list of potential products in a given industry, with some lists containing more than 50 potential products. Many of the more than 1,200 NAPCS codes can be very complex and ambiguous. Businesses have expressed the desire to alternatively supply Universal Product Codes (UPC) to the U. S. Census Bureau, as this is something they are already storing in their database.

This research considers the text classification problem of predicting NAPCS classification codes, given UPC product descriptions. We present a method for automating the Economic Census by using supervised learning.

# Methods

## Dataset

To evaluate our approach to test which Machine Learning (ML) classifier would perform best in product classification, we analysed 14,000 UPC product descriptions for 44 NAPCS categories, provided to the U.S. Census Bureau by a popular drug store chain.

## Data Pre-processing

The collected data were transformed to a structured format. These steps are commonly applied for information retrieval, information extraction and data mining. This was done by applying text processing techniques on the product descriptions [1]. Punctuations and numbers were removed from the text. Next, all of the letters were converted to lowercase. All white space was removed. The final pre-processing technique was the removal of stop words.

## Feature Selection

ML algorithms require numeric feature vectors to learn the underlying representation of the dataset. For this purpose, we need to convert our text into some numeric form. We will use Term frequency-inverse document frequency or TF-IDF. It is the measure of how important the term is for a particular document in a corpus. If the term is frequent in the document and appears less frequently in the corpus, then the term is of high importance for the document [2].

The document term matrix (from TF-IDF) is usually very high dimensional and sparse. It can create issues for ML algorithms during the learning phase. Therefore, it is recommended to reduce the dimensionality of the dataset by either feature selection or dimensionality reduction methods. The former selects important features from the original feature set whereas, the latter learns new features from the original set in some other dimension. We will apply Chi-Square and Mutual information as feature selection methods, and Latent Semantic Analysis (LSA) as a dimensionality reduction technique [3].

The χ2 test is used in statistics to test the independence of two events. More precisely in feature selection, it is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent.

Mutual information is another technique to determine the relative importance of the term. It measures the number of bits required for category prediction by knowing the presence or the absence of a term in the document.

LSA is an unsupervised statistical technique used for extracting and inferring relations of expected contextual usage of words in documents. It utilizes singular value decomposition (SVD) to reduce the high dimensionality of text data. It does so by keeping the first *k* largest singular values and omitting the rest.

## Classification Model and Algorithms

Multiple classifiers were assessed against the feature set and product description data. This evaluation was conducted to identify a best choice for feature selection, and for the optimization of the accuracy rate. We use three classification algorithms to categorize products, Support Vector Machine (SVM) with linear kernel, Logistic Regression (LR), and Multinomial Naïve Bayes (MNB).

For evaluating how well the models are performing, stratified ten-fold cross-validation will be used. Generally, this is used to evaluate the performance of supervised learning algorithms and aims to ensure, each class has equal representation across each fold. Precision, Recall, F-score, and Accuracy will be calculated and be used for final model selection.

# Results

Table 1 shows the accuracy rates obtained on various feature sets, across all considered classification and feature selection models. For the LSA dimensionally reduced data, it no longer makes sense to use MNB, since the features are no longer valued in positive integers. However, we can still use SVM and LR for classification.

For the smallest feature set at size 100, LR and SVM demonstrated the best performance with all feature selection models. In the case of LSA, LR achieved an accuracy of one percentage point above SVM. The feature set of 100 derived from the LSA model, increased accuracy rates by more than 12 percentage points for SVM and LR. Performance at feature set size 1,100 showed that all models received a boost of 5 percentage points or more using the LSA model, with SVM and LR tied for best.

Table 1. Accuracy Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **100 words** | **1100 words** | **4100 words** | **7100 words** | **9100 words** |
| **Chi-Square Feature Selection with Unigrams** |
| SVM | .55 | .83 | .94 | .96 | .96 |
| Logistic Regression | .55 | .83 | .93 | .95 | .95 |
| MNB | .49 | .79 | .92 | .94 | .94 |
| **Mutual Information Feature Selection with Unigrams** |
| SVM | .64 | .87 | .95 | .95 | .96 |
| Logistic Regression | .64 | .87 | .94 | .94 | .95 |
| MNB | .57 | .83 | .93 | .94 | .94 |
| **LSA Dimensionality Reduction with Unigrams** |
| SVM | .76 | .92 | .95 | .96 | .96 |
| Logistic Regression | .77 | .92 | .95 | .95 | .95 |
| **Unigrams** |
| SVM | .62 | .87 | .95 | .95 | .96 |
| Logistic Regression | .62 | .86 | .94 | .95 | .95 |
| MNB | .60 | .83 | .93 | .93 | .93 |
| **Bigrams** |
| SVM | .47 | .62 | .71 | .75 | .76 |
| Logistic Regression | .47 | .62 | .70 | .75 | .76 |
| MNB | .47 | .62 | .71 | .75 | .77 |
| **Unigrams+Bigrams** |
| SVM | .61 | .84 | .93 | .95 | .95 |
| Logistic Regression | .61 | .85 | .93 | .94 | .95 |
| MNB | .60 | .83 | .92 | .93 | .93 |

The frequency based Unigram model, Chi-Square, and Mutual Information model achieved very good results with a feature set of 1,100, but LSA demonstrated much better results than any other model.

LR, SVM, and MNB displayed a rapid increase in accuracy at the feature set size of 4,100. With the exception of the bigram model, all our methods surpassed 90% accuracy with the feature set of 4,100. The optimal feature set size, however, seems to be at 9,100, where we continue to see very modest growth in accuracy without any model overfitting. SVM was the overall best performer as the feature set size increased, marginally beating LR and MNB by at most 2 percentage points. Table 2 summarizes precision and recall values for the frequency based unigram model.

Table 2. Classification Results using 9100 features

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Precision** | **Recall** | **F-Score** |
| Unigrams |
| SVM | .96 | .96 | .96 |
| Logistic Regression | .95 | .95 | .95 |
| MNB | .93 | .93 | .93 |

# Conclusions

We evaluated the performance of different classifiers using the 10-fold stratified cross-validation technique after applying text pre-processing, feature selection along with dimensionality reduction methods for this text classification problem. Our results show that with a small feature set LSA models give the greatest boost to the classification models. However, with a large feature set size, SVM gains the most accuracy, and the frequency based Unigram model performed at the same level as the Chi-Square, Mutual Information, and LSA models.

We believe we can achieve better consistency, and response by automating classification for U.S. businesses. We think this is a more precise strategy as indicated by our high accuracy rates obtained in this research. It was not burdensome for the drug store to supply Census with a store’s data by UPC codes. Therefore, we have created a methodology to classify products in an easier and more timely way. This work lays the foundation for not only collecting this information every 5 years, but with the timeliness that intelligent predictions provide**.**

# References

1. S. Gottipati, E-commerce product categorization, Stanford C229 Final Projects (2012).

[2] C. Chavaltada, K. Pasupa, D.R. Hardoon, A comparative study of machine learning techniques for automatic product categorisation, Advances in Neural Networks - ISNN (2017), pp. 10–17.

[3] S. Shankar and I. Lin, Applying machine learning to product categorization, Technical report, Stanford University (2011).