



# AUTOMATIC DYNAMIC PRODUCT CLASSIFICATION

Guillermo Vinué\* and Andrew Parnell

University College Dublin, School of Mathematical Sciences/The Insight Centre for Data Analytics  
Dublin, Ireland

\*guillermo.vinue@ucd.ie



## Abstract

- Online shopping is part of our everyday experience. Nowadays, more and more people buy products via the internet.
- Consumers usually read the text description of a product of interest to gain an overview of its features.
- Classification of products based on text summaries of their product description is an increasing area of activity and a very important commercial application.
- The development of a stand-alone classification method that automatically classifies products in categories is a major challenge of today in Statistics.

## Introduction

- As shown in the figure below, the product description contains highly relevant information to provide the consumer with an overview of its features and performance.

### Product Description

#### Product Description

##### De'Longhi Magnifica bean-to-cup espresso and cappuccino machine

The Magnifica ESAM4200 is a must for any budding barista taking their first step into domestic bean-to-cup machines. You can make an espresso or an Americano at the touch of a button, and using the traditional milk frother you can make the perfect lungo, latte, cappuccino, or other milky drink. It produces 15-bar restaurant and coffee house standard coffee in the comfort of your own home, and in the simplest way.



All aspects of a typical professional pump espresso machine (like the ones you see in your high street coffee bar), such as grinders, filter holders, and steam pipes, are found working inside the Magnifica ESAM4200: De'Longhi has simply redesigned, compacted, and hidden them inside the machine. Magnifica expertly performs all of these functions at the touch of one button. The traditional milk frother allows you to impress your guests by manually frothing the milk whilst making them the ideal cappuccino or latte.

#### Traditional milk frother

Café lattes, cappuccinos, and latte macchiatos can be made using the traditional milk frother to froth the milk. Using any cold milk, you can adjust the steam pressure and manually steam the milk in a barista-style fashion to create the ideal froth.

The electronic steam and coffee thermostat ensures accurate adjustable steam control. The ESAM4200 can be set to adjustable quantities of coffee and water, which help create that ideal, individually-tailored cup of coffee every time.

- A very important part of online retail store analytics is the classification of products based on text product summaries.
- Current classification is done manually, which is time-consuming and prone to errors.
- The challenge at hand is to develop a stand-alone statistical classification model, which can be used to provide brands with quick and insightful product classifications.

## Methods

- Before applying any classifier, the text of product summaries must be preprocessed. The R package **tm** is the most developed package for managing text documents ([1, 2]).
- Example of typical transformations: The initial product description is **PHILIPS HD8764/01 SAECO MINUTO. Coffee machine Espresso**
- Replace '/' with a space: **PHILIPS HD8764 01 SAECO MINUTO. Coffee machine Espresso**
- Conversion to lower case: **philips hd8764 01 saeco minuto. coffee machine espresso**
- Remove numbers: **philips hd saeco minuto. coffee machine espresso**
- Remove punctuation: **philips hd saeco minuto coffee machine espresso**
- In text mining, a *corpus* is a collection of  $n$  documents and can be represented as a *document-term matrix* (dtm).
- In a dtm, the rows are documents and the columns are words. Each entry in  $(i, j)$  is the frequency of the word  $t_j$  in document  $d_i$ :

$$X_{n \times p} = \begin{pmatrix} f(d_1, t_1) & \dots & f(d_1, t_p) \\ \vdots & \ddots & \vdots \\ f(d_n, t_1) & \dots & f(d_n, t_p) \end{pmatrix}$$

- In our case, the dtm is a matrix of  $n$  products and  $p$  words describing the products. The product categories are saved in a matrix  $Y_{n \times 1}$  consisting of  $C$  categories.
- Each one of the  $n$  products is associated with one of the  $C$  classes.
- Existing reviews about statistical classification methods for automated text categorization indicate that dominant approaches are Naive Bayes, kNN and decision trees ([4, 5]).
- We propose to establish a comparison of all these methods, also analyzing the performance of multinomial logistic regression (maxent) and discriminant analysis.

## Results in a first case study: Nestle Spain

- Extract of data for Nestle Spain.
- The texts with the products description are manipulated.
- The preprocessed texts are brought together in a dtm ( $33331 \times 726$ ).
- Every classifier is applied to a training, validation and test subsets of the dtm.
- Only those methods that have the smallest validation error rates will be used to get test error rates.
- For each category:
  1. **Training data:** Random selection of a half part of the products (the rows of the dtm) matching that category.
  2. **Validation data:** Random selection of a half part of the products not selected in the previous step.
  3. **Test data:** Remaining points.

Training dtm	Validation dtm	Test dtm
16667 (50%)	8335 (25%)	8329 (25%)

Method	Error rates	
	Validation error rate	Test error rate
kNN	0.38%	0.36%
MaxEnt.	0.35%	0.5%
LDA	2.98%	-
RF	0.65%	0.6%
NB	99.2% !!	-

- Extension of kNN to incorporate a probabilistic framework: Bayesian kNN ([3]) (new R function programmed). Its test error rate is very similar to the one returned by classical kNN.
- Taking advantage of the fact that the bayesian kNN returns probabilities for each category, just as maxent and random forests do, we can provide the customer with a classification in a number of top categories, for example, the top 3 categories, as shown in the tables below.
- Description of product 1 in the test set: *DOLCE GUSTO KRUPS MINI ME BLACK Coffee machine*. Category assigned by the three methods: *Electronics* (very suitable).

		Bayes KNN		
		C1	C2	C3
Products	Categories			
	P1	<b>Electronics (0.996)</b>	Babies (0.00038)	All sections

		Maxent		
		C1	C2	C3
Products	Categories			
	P1	<b>Electronics (0.953)</b>	All sections (0.0473)	Feeding (0.00013)

		Random Forests		
		C1	C2	C3
Products	Categories			
	P1	<b>Electronics (0.74)</b>	Feeding (0.146)	Small electr. appliance (0.114)

## Conclusions

- The main objective of this work is to develop a statistical classification method to automatically classify products based on summaries of their product descriptions.
- Further goals: An extended hierarchical model which classifies products into hierarchical categories. A dynamic extension of the classifier to identify new categories.
- This project is in collaboration with the company Clavis Insight, which is providing us with data (<http://clavisinsight.com/>).

CLAVIS INSIGHT

## References

- [1] I. Feinerer and K. Hornik. **tm: Text Mining Package**, 2014. R package version 0.6.
- [2] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25:1–54, 2014.
- [3] C.C. Holmes and N.M. Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:295–306, 2002.
- [4] A. Khan, B. Baharudin, L. Hong Lee, and K. Khan. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1:4–20, 2010.
- [5] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.