Text mining and machine learning techniques for text classification

Gianpiero Bianchi[1], Renato Bruni[2], Francesco Scalfati[1], Fabiana Bianchi[1]
1) Istat, Direzione centrale per la metodologia e disegno dei processi
statistici (DCME), Via Depretis 77, Roma, 00184 Italy
E-mail: gianbia, scalfati, fabianchi@istat.it

2) Università di Roma "Sapienza", Dip. di Ing. Informatica, Automatica e Gestionale (DIAG),
Via Ariosto 25, Roma, 00185 Italy
E-mail: bruni@dis.uniroma1.it

1. Introduction

Nowadays, the demand for timely information at a multidimensional level constantly increases, and official statistics is called to respond both efficiently and effectively. This requires a transformation of the data production model. In traditional surveys, data already held by the public administration can be integrated with innovative sources such as Big Data. This requires new strategies to tackle new data complexity and integration, but offers a great added value in terms of reduced statistical burden on the respondents while enriching the quantity, quality and timeliness of the information produced.

Text scraping and text mining techniques can be applied to substitute traditional techniques of data collection. Information from the Internet data sources can be integrated in traditional data surveys to produce higher quality estimates.

The proposed work applies the techniques described in our previous document [1]. At present, we have defined and implemented a web mining process that uses web scraping, text mining and machine learning techniques to integrate survey data and/or business register. We now propose further original enhancements to the described strategy in order to present a practically viable methodological approach to perform an automatic detection of the enterprise's characteristics by using the information contained in their websites.

The proposed procedure has allowed to produce, for the first time, experimental statistics for the yearly survey "ICT usage in Enterprises", by using Internet data to produce estimates on the following specific variables:
1. web ordering functions (e-commerce component) are available;
2. information on job vacancies;
3. links to social media (Facebook, Twitter, Instagram etc.).

ICT survey data have been used as a training set for fitting the models to predict values, and administrative data in the statistical business register have been used for handling representativeness problems [2]. Moreover, this strategy has been used in an experimental project in Istat Laboratory for Innovation (LabInn), in order to enlarge the informative content of the Statistical business register to provide concrete support for statistical production, see also [3] for details.
Internet Web pages containing the business data come without a given standard structure, so they need to be standardized and analysed first and then integrated into Statistical Business Register. Furthermore, the new information on enterprises will be used to start a more detailed statistical analysis, finding new classifications and new taxonomies to support a better interpretation of new emerging economic phenomena. In particular, they will produce a set of new enterprise-level information linked to the statistical business register, whose content allows:

- to complete the missing information in certain variables of the Business Register;
- to check some information for variables enclosed in the Business Register;

- to add new information to cover additional variables for the Business Register.

The new information covers the following types of business enterprise characteristics:

- structural characters: related to the structural features of the enterprises, such as anagraphic characteristics and personal data, business data, dimension, etc. (for instance: Tax fiscal code, VAT number, business name, company capital, telephone number, email address, certified email address, business enterprise's street address);
- qualitative characters: concern information not directly measurable but representative for the enterprise, such as, the short description of business activity, the presence of links or references to the enterprise’s social media profiles, the presence of online job application facilities, the identification of *.pdf* documents concerning financial statements or product catalogues.

The main strategy of both works is the building of a methodological approach that includes 3 main phases:

- *Web Scraping*: for web data acquisition (with different approaches depending on whether the enterprise web address is available or not in advance);
- *Text Mining*: *Text Mining* for extracting the information to integrate survey data and/or of business register.
- *Machine Learning*: for the use of algorithms that simulate a learning process for the construction of predictive models

2. The web scraping phase

In the Web Scraping phase the procedure extracts the information from each corporate website and saves it to a *NoSQL DBMS*. In this phase, there are two steps:

1. First step: The procedure identifies each enterprise on the web and creates an URLs list by means two different ways depending on URL availability:
- In case of available URL, the procedure checks URL’s validity first and if invalid, searches another mostly similar valid URL, performing a web search using search engines and computing its probability of correctness by using a machine learning approach.
- In case of non-available URL the procedure uses either the URL Retrieval techniques performing batch queries on the search engines by means of the enterprise identification characteristics, or it directly download information from some proper thematic directory sites. For each result found, probability of correctness is evaluated by using a machine learning approach, and the link whose probability exceeds a given threshold is accepted as valid.
2. Second step: The procedure extracts all the information from websites by using web scraping techniques. Besides the text appearing in the pages, it reads also some additional information: such as images, *HTML tags*, *meta-keywords*, *pdf* files, etc…

3. The text mining phase

The automatic extraction of statistical information from Internet is extremely appealing considering the huge amount of data freely available through this channel. However, the amount of data is also a big problem, and this is worsened by its completely not standardized structure, noisy and not completely homogeneous data.
Therefore, to identify the relevant parts of the extracted information a quite articulated procedure that requires the use of several steps of text mining methods and techniques has to be developed.

Text Mining is the branch of Data Mining concerning the process of deriving high-quality information from texts. References can be found for instance in [4]. Modern text mining methodologies require the integration of natural language processing techniques (see, e.g., [5]) with several advanced machine learning techniques.

Natural language processing is an approach which allows to find meaning of the free text. This is done by using several techniques such as:

- Tokenization: splits string into still useful linguistic units.
- Lemmatization: removes the inflectional ending to return the word to its basic lemma.
- Part-Of-Speech recognition: identifies a word as a particular part of speech (such as: noun, verb, etc).
- Word embedding: maps words or phrases from the vocabulary to real numbers vectors.

In general, a very high level model for text analysis includes several text processing tasks, such of these are:

- Language identification: detects the language(s) of a document.
- Information retrieval: gathers results that are relevant for the specific needs of the user.
- Information Extraction: extracts structured information from unstructured and/or semi-structured machine-readable documents.
- Summarization: provides a self-contained and internally cohesive text which serves as a selective account of the original.

In some cases, according to the specific needs, the sequential application of some of these tasks allows to obtain directly the requested information as, for instance, the structural features of the enterprises, such as anagraphic characteristics and personal data, business data, dimension, etc.

In other cases, such as predictions on specific variable (web ordering functions, information on job vacancies, etc.) needs to use machine learning methods, and the problem of automatic detection of enterprise's characteristics becomes a supervised classification problem.

4. Machine learning phase

Engineering a statistical production process which includes Big Data sources can include machine learning (ML) techniques to guarantee their automation.

The choice of which ML algorithm to apply, largely depends on the user's domain knowledge, the desired results and on the performance of computing platform. There are many different kinds of machine learning algorithms for discover patterns in Big Data that lead to actionable insights [6].

The automatic detection of enterprise's characteristics by using the information contained in their websites needs an analytical process that involves supervised classification techniques.

Supervised learning is based on the availability of a set of labeled records (training set) that constitute the source of information to learn a classifier so we need a set of websites for which we already have, or we can obtain, the class labels with respect to the considered categorization.

For this scope survey data have been used as a training set for fitting models considering both the answers provided to the survey and the texts captured on the web were available. These models can be applied to the generality of enterprises and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful. Administrative data (mainly contained in the Business Register) have been used for handling representativeness problems.

References

[1] G. BIANCHI, R. BRUNI, F. SCALFATI, F. BIANCHI Text mining and machine learning techniques for text classification, with application to the automatic categorization of websites. Advisory Committee on Statistical Methods, Rome, November 2-3, 2017

[**2]** G. BIANCHI, R. BRUNI, F. SCALFATI Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, Mathematical Problems in Engineering, vol. 2018, Article ID 7231920, 8 pages, 2018

[3] G. BIANCHI, M. CONSALVI, B. GENTILI, F. PANCELLA, F. SCALFATI, D. SUMMA New sources for the SBR: first evaluations on the feasibility of using big data in the SBR production process. 26th Meeting of the Wiesbaden Group on Business Registers - Neuchâtel, 24 – 27 September 2018

[4] R. FELDMAN, J. SANGER The Text Mining Handbook. Cambridge University Press, 2006

[5] S. BIRD, E. KLEIN, E. LOPER Natural Language Processing with Python. OReilly Media, 2009.

[6] G. BIANCHI, R. BRUNI Effective Classification using a Small Training Set based on Discretization and Statistical Analysis, IEEE Transactions on Knowledge and Data Engineering Vol. 27(9), 2349-2361, 2015.