2022 UNECE Expert Meeting on Statistical Data Collection
26-28 October 2022, Rome

A machine learning based help-desk approach for units involved in official surveys

Giampaola Bellini - Istat | Data Collection Directorate (bellini@istat.it)
Gianpiero Bianchi - Istat | Data Collection Directorate (gianbia@istat.it)
Paola Bosso - Istat | Data Collection Directorate (bosso@istat.it)
Pasquale Papa - Istat | Data Collection Directorate (papa@istat.it)

Outline

- Case study: assistance services for the survey unit requests
- Digital service management
- Automatic approach proposed
- Experimental results
- Conclusions

Case study

The work deals with the development of an optimal approach to support the assistance for the survey unit requests using automatic solutions.

- Optimize the human resources employed in the frequent and high number of manual tasks, in order to reduce costs and errors
- Improve the usability and functionality of data acquisition systems
- Simplify the request processing to minimise ticket resolution times and costs
- Provide unified outputs to various types of requests
- Coordinate and align information and requests in order to ensure consistency in communication
- Eliminate the risk of miscommunication and wrong interpretation

Digital service management

- Standardization of digital processes and digitization of the service delivery processes:
  - map the processes and needs of the people involved;
  - identify recurring activities with low added value and review workflows in order to simplify the request for services;
  - aggregate information on the people served by multiple channels in a single point of control increasing the omnichannel capabilities.
- Automation of repetitive tasks:
  - training of automatic classification models with the analysis of text;
  - automation of easy to solve tickets, in the presence of recurring problems;
  - automated opening of tickets accompanied by support data to speed up the intervention in the most complex cases.
- Improvement of the quality and perceived innovation of the service:
  - the collection of feedback;
  - satisfaction indicators;
  - proactive problem solving prior to the request.

Automatic approach proposed: data-driven procedure

The assistance requests, also known as tickets, need to be classified by topic to be answered efficiently. The effectiveness of the ticketing system would greatly increase if this semantic categorization could be carried out automatically.

- Basic features of the tickets are not meaningful for the semantic categorization, i.e.:
  - length of the message or similar measures;
  - basic presence or absence of predetermined words.

Proposed approach

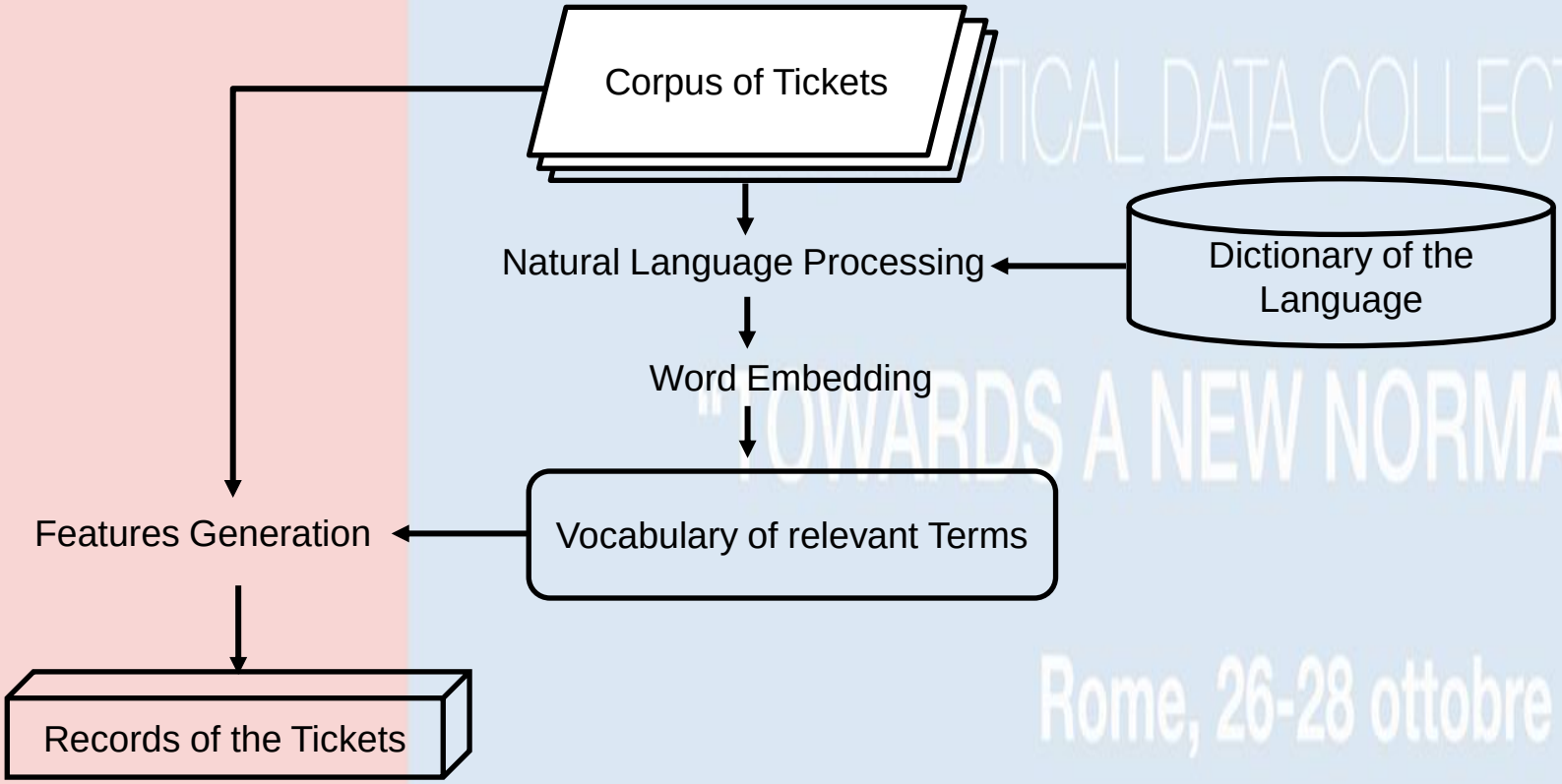
- Formal and data driven approach: all the relevant information is extracted from the data.
- Using text mining to extract the features of each ticket, and machine learning algorithms to perform the categorization.

Automatic approach proposed: main methodological contributions

- A methodology to solve the problem of the categorization of tickets written in natural language, based on text mining and machine learning:
  - automatic extraction of the meaning of the text by using natural language processing techniques;
  - followed by a classification, which is often multiclass.
- An approach to determine the hyperparameter configuration of a machine learning procedure:
  - hyperparameters are all the values needed by the classifier which are not learned by data;
  - classification algorithm needs a suitable hyperparameter configuration to produce a useful categorization.

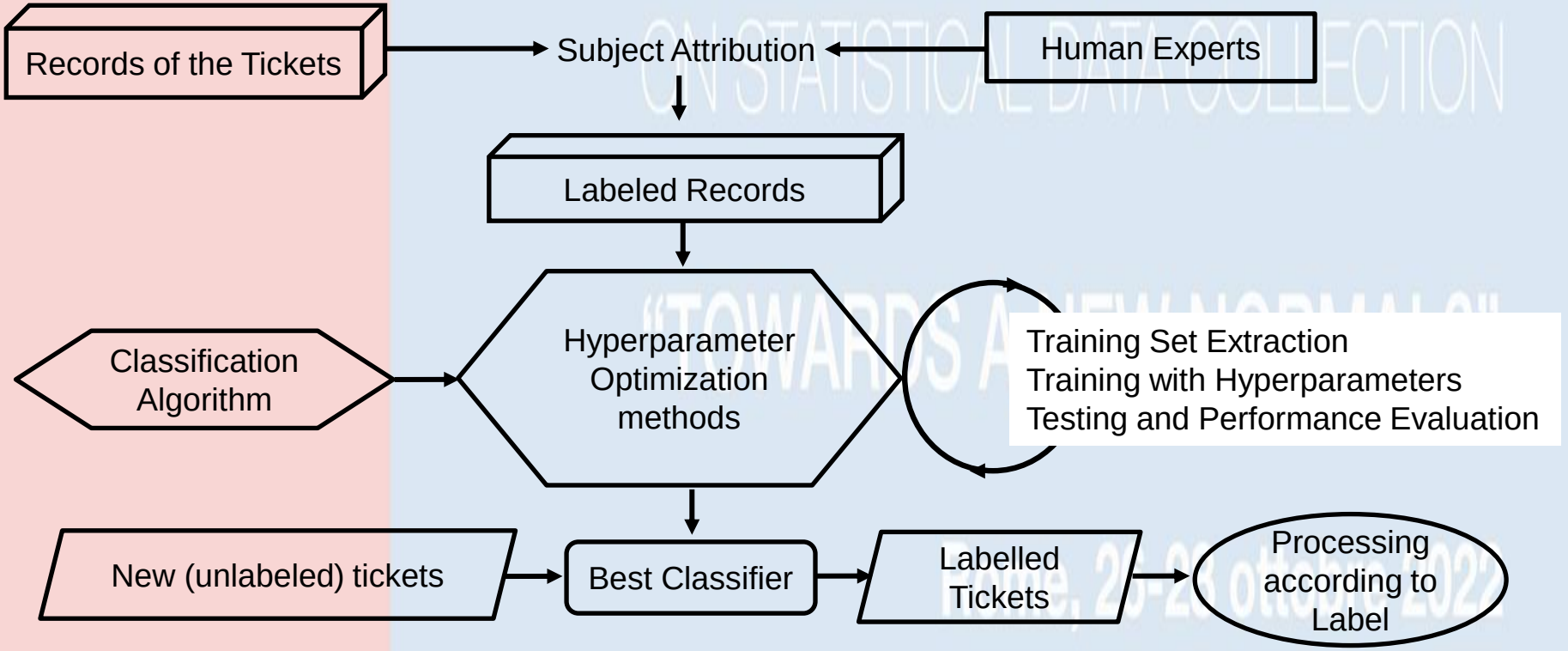
Automatic approach proposed: text mining phase

The messages' raw text is converted into a standardized form by using Natural Language Processing techniques



Automatic approach proposed: machine learning phase

Learning the classification criteria from a training set and predicting the class of unlabelled tickets.



Automatic approach proposed: Hyperparameter Optimization methods

We propose viewing the hyperparameter choice as a higher-level optimization problem:

$$\max_{H \in \mathcal{H}} P(H) \qquad (1)$$

- the hyperparameters H are the decision variables;
- the objective is the performance P of the classifier;
- the set $\mathcal{H}$ of all the feasible tuples of hyperparameter values.

- We adopt a black-box optimization approach, which does not need the explicit optimisation model, but only needs to numerically evaluate the objective function over a number of points.
- We use a recently proposed algorithm (*Liuzzi et al*, 2020) able to guarantee theoretical convergence to a local optimum of (1).

Experimental results: real-world data from the CC

- We analyzed the tickets received by Istat's Contact Center for some economic surveys:
  - 193,000 unlabeled tickets
- We considered a training set which is composed of FAQs representing various cases of assistance requests:
  - 8,000 tickets labeled by experts
  - 7 distinct classes:
    - General information
    - Usability
    - Information on questionnaire questions
    - Interaction problems with Istat
    - Eligibility
    - Indeterminable
    - Uncertain

Experimental results: Performance evaluation

- 50% as training set and 50% as test set
- Several classification approaches were applied
- Black-box approach was compared to a standard grid search
- Best results were obtained with Convolutional Neural Network

	Black box	Grid search
Hyperparameter configurations	7,680	240
Evaluated points	212	240
Time for training (sec)	318,500	375,000
Solution accuracy on test set	89.92%	86.15%
Time for prediction (sec)	170,000	200,000

- Classification procedure obtains about 90% in accuracy

Conclusions

- An automatic process of assistance to the survey units first requires the introduction of an appropriate digital service management strategy.
- Automatic solutions for the semantic categorization of tickets allow to improve both the efficiency of the ticketing system and the quality of the answers: able to perform frequent and high-volume manual tasks and extract hidden relationships and patterns.
- It is a difficult task, since it requires an automatic extraction of the meaning of the text, followed by a classification (often multiclass).
- The proposed methodology works at the formal level using a data-driven approach, and thus it could also be applied to the automatic categorization of other texts with different origins and/or languages.
- Experiments on real-world data confirm that an automatic ticket categorization is practically feasible and can reach very good accuracy in very reasonable times.

Future work

- The experimentation of technological solutions able to implement a digital platform that works as a single point of contact:
  - information on the people served by multiple channels is aggregated, increasing the omnichannel capabilities;
  - all communications are routed through a single entity like a service desk.

- Extending the analysis made for the semantic classification of the Contact Center tickets to all communication channels with respondents.

References

- Bellini G., Bianchi G., Di Paolo GG., Papa P. *Towards a selective automation process of assistance to the survey units included in business surveys*, 2022 BDCM Workshop, Oslo, 13 - 15 June 2022.
- Bianchi G., Bruni R. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. Expert Systems with Applications, vol. 142, ISSN: 0957-4174.
- Bianchi G., Bruni R., Scalfati F. Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms. Mathematical Problems in Engineering, vol. 2018, p. 1-8, ISSN: 1024-123X.
- Bianchi G., Bruni R. *Effective Classification using a small Training Set based on Discretization and Statistical Analysis*. IEEE Transactions on Knowledge and Data Engineering, vol. 27, p. 2349-2361, ISSN: 1041-4347.
- Bellini G., Monetti F., Papa P. *The impact of a centralized data collection approach on response rates of economic surveys and data quality: the Istat experience*. Statistics and Economy Journal, Vol. 100 (1) 2020 ISSN 1804-8765 (Online) ISSN 0322-788X (Print).
- Bellini G., Bosso P., Binci S., Curatolo S., Monetti F. Centralized Inbound and Outbound Contact center Service as New Strategy in Data Collection. Fifth International Workshop on Business Data Collection Methodology. Lisbon, 19-21/9/2018. https://www.ine.pt/scripts/bdcm/doc/00_BDCMLisbon2018_BP.pdf pgg 138-141.
- Liuzzi, G., Lucidi, S., Rinaldi, F. An algorithmic framework based on primitive directions and nonmonotone line searches for black-box optimization problems with integer variables. Mathematical Programming Computation 12, 673-702.
- Salemink I., Dufour S., Van Der Steen M. (2019). Vision Paper on Future Advanced Data Collection, UNECE Statistical Data Collection Workshop 'New Sources and New Technologies', Geneva 14-16/10/2019.