Citizen Generated Data and machine learning: *a way to study violence against women*

CNS15a Conferenza Nazionale di Statistica

GIANPIERO BIANCHI – ISTAT GIANBIAN@ISTAT.IT | **ALESSANDRA CAPOBIANCHI** - ISTAT ALESSANDRA.CAPOBIANCHI@ISTATI.IT
MARIA GIUSEPPINA MURATORE – ISTAT MURATORE@ISTAT.IT | **CLAUDIA.VILLANTE** – ISTAT CLAUDIA.VILLANTE@ISTAT.IT

OBJECTIVES:
Adding data-sources on VAW, by using citizen-generated data

WHICH KIND OF DATA?

1522 is a public service promoted by the Presidency of the Council of Ministers - Department for Equal Opportunities (DEO). It was activated in 2006 with the **aim of developing a broad systemic action for emerging and combating the phenomenon of intra- and extra-family violence against women.**
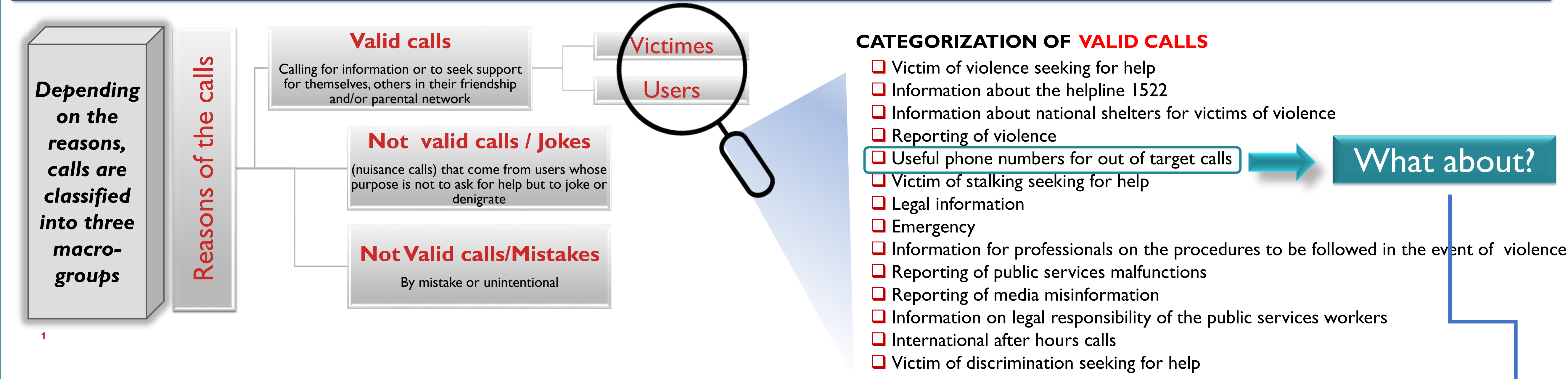
In 2009, with the entry into force of Law 38/2009 amended in 2013 on the subject of persecutory acts, it also started an action to **support stalking victims**.

1522 NUMERO ANTI VIOLENZA E STALKING

The information provided during the call is recorded on a **computerised platform** whose data has been available since January 2013. Registration takes place following questions asked by the *NGO PROFESSIONALS* of the public utility number according to **classification rules** whose filter is the **reason for the call**.

According to the cooperation agreement between DEO-ISTAT for the creation of an **'Integrated Data Collection and Processing System'**, since 2018, , the data from the computerized system in use at 1522 are sent to ISTAT, which takes care of the analysis and processing

HELPLINE 1522 - DATASET STRUCTURE

Depending on the reasons, calls are classified into three macro-groups

Reasons of the calls

Valid calls
Calling for information or to seek support for themselves, others in their friendship and/or parental network

Victimes
Users

Not valid calls / Jokes
(nuisance calls) that come from users whose purpose is not to ask for help but to joke or denigrate

Not Valid calls/Mistakes
By mistake or unintentional

CATEGORIZATION OF VALID CALLS
- Victim of violence seeking for help
- Information about the helpline 1522
- Information about national shelters for victims of violence
- Reporting of violence
- Useful phone numbers for out of target calls
- Victim of stalking seeking for help
- Legal information
- Emergency
- Information for professionals on the procedures to be followed in the event of violence
- Reporting of public services malfunctions
- Reporting of media misinformation
- Information on legal responsibility of the public services workers
- International after hours calls
- Victim of discrimination seeking for help
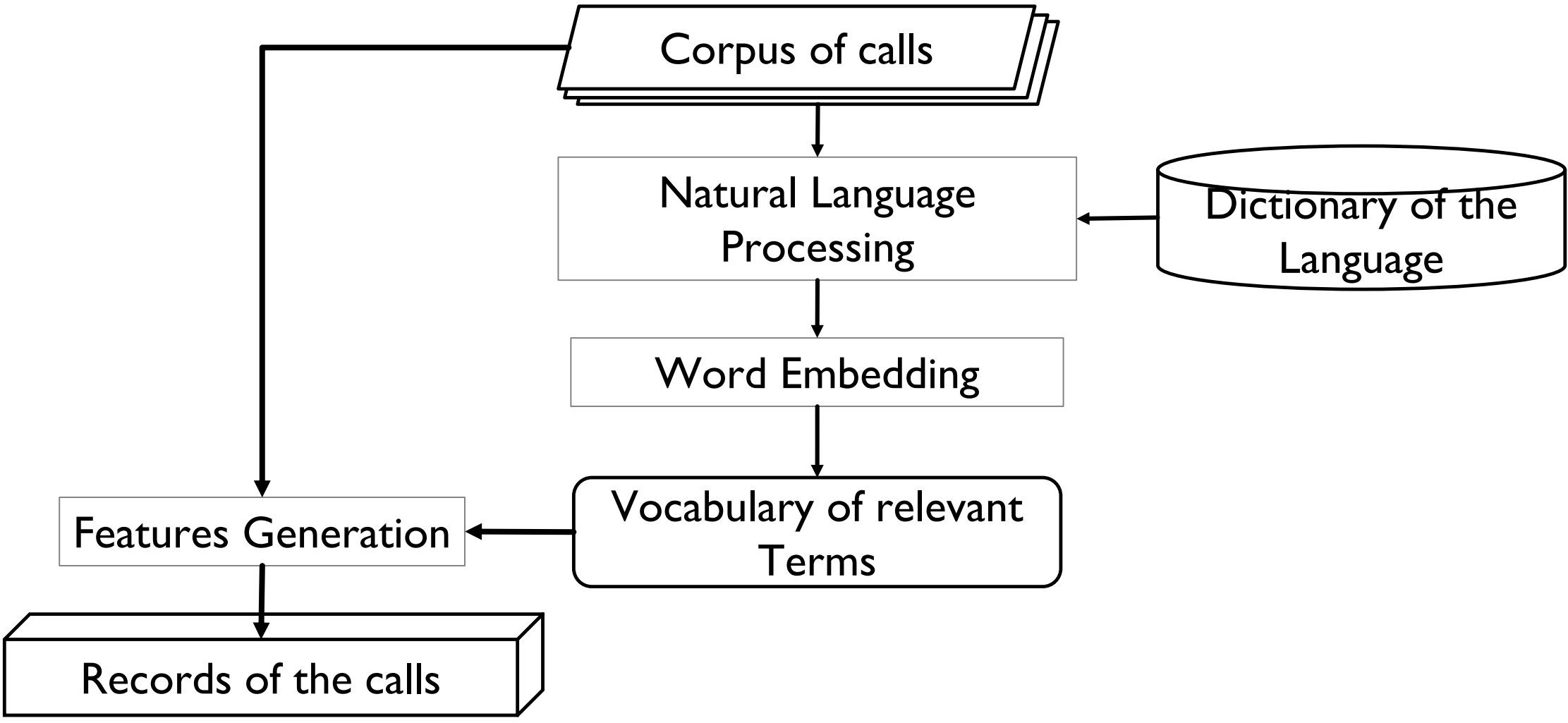
What about?

METHODOLOGICAL APPROACH

- An unsupervised text clustering and topic extraction framework has been adopted able of achieving high-quality clustering result and extracting cluster-specific topics.

- The framework includes three main components:
  - Feature Extraction: to convert the target texts into vector representations (text embeddings) by capturing the semantic information, so that it can be processed by the clustering algorithm. It also includes dimensionality reduction techniques to improve data processing.
  - Call clustering: to perform a clusterization by using the *k*-means algorithm. This algorithm aims at partitioning the calls into *k* clusters in which each observation belongs to the cluster within nearest mean, being the centroid of the cluster. The "Elbow" method is used to determine the value of *k*.
  - Topic extraction: to find main topics from each cluster using the keywords extraction methods TF-IDF (Term Frequency-Inverse Document Frequency).
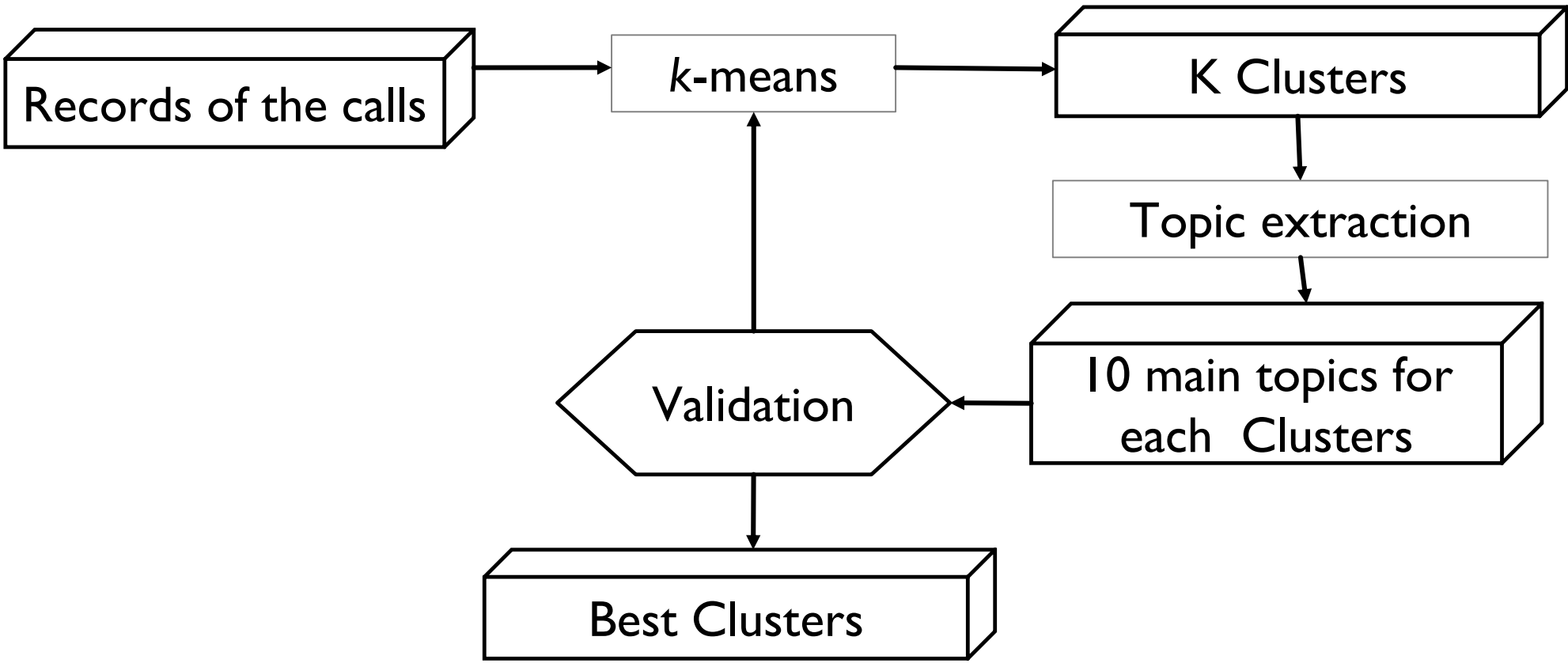
- The raw text extracted from calls is converted into a standardized form (data record) by using Natural Language Processing techniques

- Data record are categorized using the *k-means clustering algorithm*.
- A topic extraction method is applied to extract the best 5 main topics from each cluster.
- The clusters are validated by domain experts analysing the topics extracted from clusters.

Corpus of calls → Natural Language Processing ← Dictionary of the Language → Word Embedding → Vocabulary of relevant Terms → Features Generation → Records of the calls

Records of the calls → *k*-means → K Clusters → Topic extraction → 10 main topics for each Clusters → Validation → Best Clusters

RESULTS

- Analysis of the calls received by EOD's 1522 helpline:
  - 12.315 "out of the target" calls;
  - 5.761 "out of service" calls.

- 38 clusters were produced:
  - 37 clusters validated from domain experts:
  - 1 cluster with unclassifiable calls: i) 13% of "off target" calls; ii) 11% of "out of service" calls.

- Clusters have been classified **in 3 different categories**:
  1. Cluster coherent with the existing target because reporting **same contents**
  2. Cluster out of the target (not coherent with the helpline services)
  3. Cluster with new needs from the target → identifying **new classification rules**

Clusters out of the target
- Cluster 2 topics: friendly phone, helpline violence against children
- Cluster 4 topics: social service, social worker
- Cluster 6 topics: extortion victim, suffer extortion money
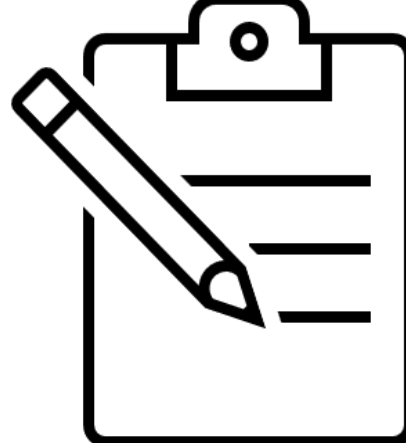
Clusters coherent with the existing target
Cluster 1 topics: victim violence, rape; Cluster 3 topics: antiviolence center, cav; Cluster 5 topics: stalking, stalker; Cluster 7 topics: sexual harrasment, sex maniac; Cluster 8 topics: free legal aid, free lawyer; Cluster 14: beating, abuse; Cluster 20: police, law enforcement; Cluster 34: report, killing; Cluster 37: 1522, info.

Clusters with new needs from the target
- Cluster 0 topics: psychiatric subject, schizophrenic
- Cluster 19: mental disorder, bipolar bipolarism

Enhancing classifications rules for the helpline, by reducing the "out of the target" calls and adding "mental desease"

NEXT STEPS

- Develop a supervised classification model that uses the validated clusters as a training set to learn the classification criteria and predict the class of unlabeled calls.
  - The predicted class corresponds to the association of the call with the reference cluster.
  - This approach would allow all unlabeled calls received by the 1522 helpline to be classified

- Improving the data gathering process of the helpline by adopting a classification model

- Generating timely and quality data from the citizen and NGOs