UNECE Expert Meeting on Statistical Data Collection and Sources 2024

Geneve 22-24 May 2024

Thematic area: *Alternative Data Sources and Process Automation*

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

Authors: *Gianpiero Bianchi - ISTAT, Alessandra Capobianchi – ISTAT, Maria Giuseppina Muratore – ISTAT, Claudia Villante - ISTAT*

Abstract

From 2018 to date ISTAT collects and releases data of 1522 helpline made by Department for Equal Opportunities of the Presidency of the Council of Ministers to support and help victims of gender-based violence and stalking, in line with what is defined within the Istanbul Convention.

Collecting and using this citizen-generated data on violence against women is of the outmost importance because complements data collected through administrative sources and dedicated population surveys. Moreover these data are timely and accurate because gathered by skilled professionals of the service, according also with the global technical guidance "Improving the collection and use of administrative data on violence against women" developed by UN Women and WHO.

The objective of the paper is to provide guidance on how to collect, process, including with machine learning techniques data from 1522, focusing on the adopted technique of using textual data, from the transcription of calls.

The paper presents the motivations, methodological choices and techniques adopted to use an essential and non- traditional data source for understanding the phenomenon of violence against women and to contribute within the multi- sources approach initiated by ISTAT for an increasingly accurate and in-depth study of the phenomenon. The impact that the adopted machine learning methodologies have had on improving quality in the data collection process is also described.

Key words: violence against women, citizen-generated data, machine learning, quality of data

1. Background framework and aims of the study

1522 is the helpline number provided by the Department of Equal Opportunities of the Presidency of the Council of Ministers (DEO) to support and help victims of gender-based violence and stalking, in line with the Istanbul Convention. 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 about persecutory acts, it also started an action to support stalking victims. The helpline number is free and active 24 hours a day and the reception is available in different languages *(Italian, English, French, Spanish, Arabic, Farsi, Albanian, Ukrainian, Russian, Portuguese, Polish).*

This helpline provides first aid information in case of emergency or useful information on the services and anti-violence centres active at territorial level to which victims of violence, or other users can turn. The database relating to the services to be addressed, is constantly updated by the Regional Administrations and the NGOs active at local level against gender-based violence from the moment of activation of a new centre or service or counter, all the indications regarding addresses and methods of service delivery are provided, allowing the 1522 operators to provide updated and timely indications. The data are available starting from January 2013. The analysis of the phenomenon of violence and stalking arising from examination of the 1522 data, therefore, provides a cross section useful in understanding the dynamics and characteristics of violence, which correspond surprisingly well with the profile already revealed by sample surveys conducted by Istat on the same subject. The recording takes place following questions posed by the operators of the toll-free number according to a standardized path whose filter is represented by the reason for the call. Depending on the different reasons for the call, the operator enters information and data, reporting what was stated by users in 1522. Based on this, calls have been classified into three macro-groups:

- ***Valid calls*** that come from interlocutors who call to get information or ask for support for themselves, for other people belonging to their friend and/or parental network

- *Invalid calls* (nuisance calls) as coming from users whose purpose is not to ask for help but to joke, denigrate the same or for unintentional mistakes.

- ***Not Valid calls***/Mistakes, due to mistake or unintentional.

Within the valid calls the professionals that gather information categorize the reason of the call are as follows:

a) Victim of violence seeking for help
b) Information about the helpline 1522
c) Information about national shelters for victims of violence
d) Reporting of violence
e) Useful phone numbers for out of target calls
f) Victim of stalking seeking for help
g) Legal information
h) Emergency
i) Information for professionals on the procedures to be followed in the event of violence
j) Reporting of public services malfunctions
k) Reporting of media misinformation
l) Information on legal responsibility of the public services workers
m) International after hours calls
n) Victim of discrimination seeking for help

Among valid calls the information reported has been further subdivided into the macro-categories “users” and “victims”. Victims are those who have suffered some form of violence and/or stalking, and whose socio economic and personal details are available. This information are collected in case of following categorization:

a) Victim of violence seeking for help
b) Reporting of violence
c) Victim of stalking seeking for help

An important consideration is the standardization process on the database. The data have been made comparable for the different years since non-homogeneous response methods have been used in the different years.

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's professionals of the public utility number according to classification rules whose filter is the reason for the call.

This standardization work has been carried out mainly to make the collected data available in the data ware house accessible through I.Stat*. See site https://www.istat.it/en/violence-against-women*. Specific attention should be paid to the number of cases: since they are calls (telephone and chat) and not people, the numbers and comments are always referred to this unit of detection and not to the user / victim who addresses the service. It is in fact possible that the same person will call the toll-free number several times, both for themselves and for others. The system to date, also for reasons of privacy, does not control this information except through a question that is addressed to the caller, asking whether or not it is the first time that the user has called the toll-free number. In the same way, since it is impossible to check the information collected during the call or the message sent, it is possible that the call is registered in the name of a possible interlocutor (other than the victim) but that it is, in fact, the same victim who does not want to report information related to himself. In this case, the database, acquiring all the information of a social and personal nature, is marked as a victim.

2. The citizen-generated data: opportunities and constraints

Citizen-generated data can significantly aid in gathering information about violence against women (VAW). This type of data collection empowers individuals and communities to report incidents, raise awareness, and drive policy changes[1].

Collecting and using this citizen-generated data on VAW, through the helpline complements data collected from administrative sources and dedicated population surveys.

Data are timely and accurate because gathered by NGOs skilled professionals of the service. As also the guideline provided by UN Woman and WHO according with "Improving the collection and use of administrative data on violence against women"[2] (UN Women and WHO.

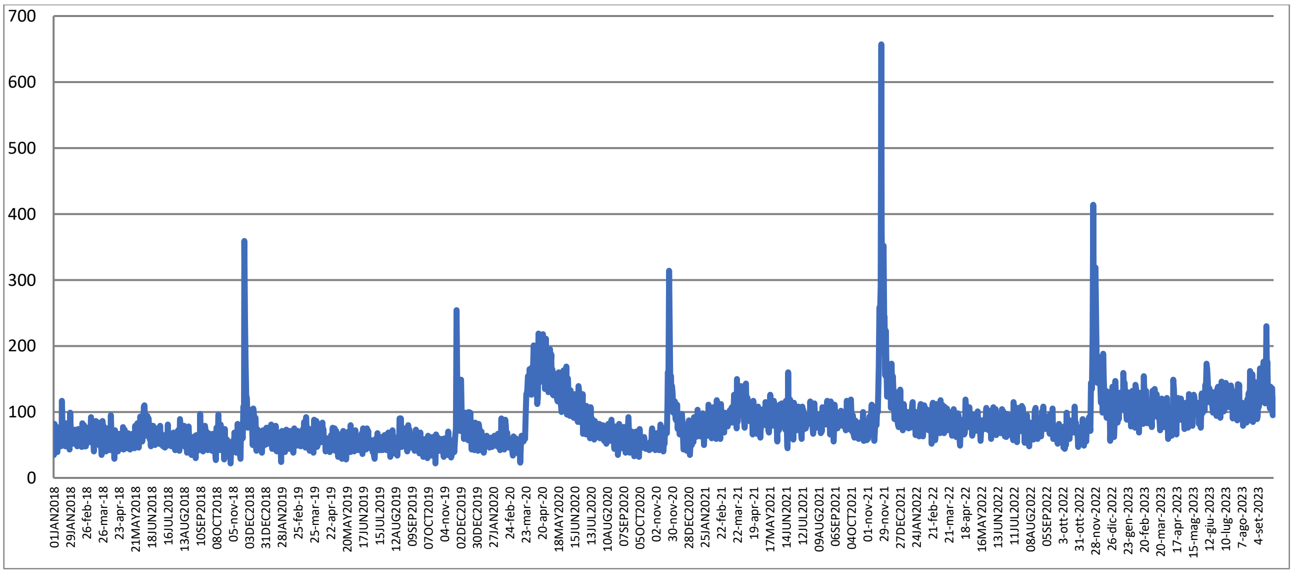
In this sense, the storage data of the 1522 hotline is an example of citizen-generated data. Citizen-driven data, often referred to as citizen-generated data (CGD), is information collected, produced or processed by individuals or groups, usually through voluntary and community activities, rather than through traditional institutional or governmental data collection methods. In fact, by using the hotline, citizens actively participate in data collection processes by reporting information on violence against women and stalking, and/or by simply asking about the hotline's services and its support to victims and users. As the following graph shows, the number of calls increases significantly during the pandemic period.

[1] For example the platform *HarassMap* in Egypt allows women to report incidents of sexual harassment via SMS or online. The data is mapped to identify hotspots, which helps in awareness and prevention efforts.

[2] UN Women, WHO, *Improving the collection and use of administrative data on violence against women*, New York: United Nations Entity for Gender Equality and the Empowerment of Women (UN Women) and World Health Organization (WHO); 2022.

Another important aspect emerging from data during the pandemic, was the particularity of users seeking for help:_ the youngest and the eldest women victims respectively from the parents and the adult children.

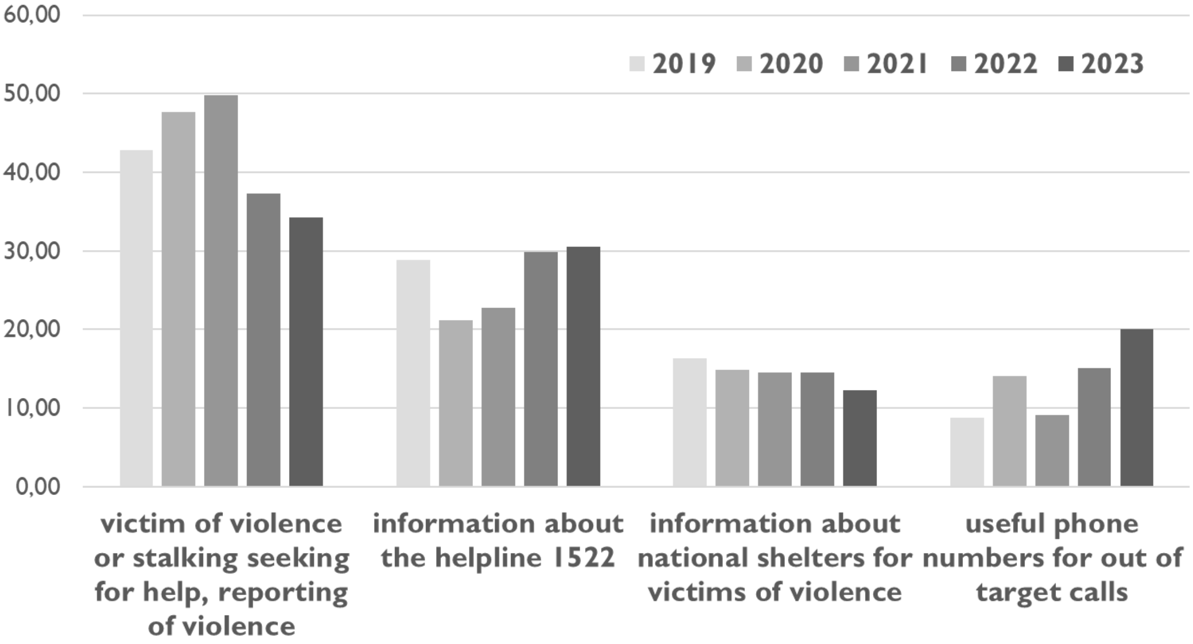
Fig 1 – Daily calls January 1, 2018 - September 30, 2023



Source: DEO - ISTAT 1522 Helpline database

Looking at the frequencies of the call's classification is very clear the main reason emerging from the users and victims. Take in consideration the same period of observation the Fig2 shows that citizens call mainly to seek help and to gather information about the 1522 service. Nevertheless the analysis of the data points out a relevant frequency of so called "out of the target". The out of target calls refer to a particular population that report socio and psychological disease.

Fig. 2 – Main reasons to call (users and victims)-First Q3th data. Years 2019-2023



Source: DEO - ISTAT 1522 Helpline database

One of the challenge of the citizen generated data is ensuring data quality due to the non-professional nature of the contributors. Thanks to the support of the NGO's professional which respond to the calls, the standardized protocols and the verification methods adopted by the 1522 helpline, ensure more reliability and quality of data.

Ensuring the accuracy and reliability of self-reported data can be challenging. Verification processes and cross-referencing with other data sources can help.

In order to enhance the classification rules for gathering information from the call, ISTAT established a strong collaborative support to the 1522's professionals aims at validating and integrating citizen driven data into official statistics by:

- Monitoring the call's categorization (classification rules) provided by the 1522 NGOs professionals according the domain rules;
- Detecting insight and new information target in order to define new classification rules;
- Improving the quality of gathering and classification rules process of data from calls;
- Producing a final classification avoiding disambiguation and more representative;
- Supporting data gathering process by using automatic categorization procedures.

Following the purpose of enhancing the quality of data storage process ISTAT decide to investigate what kind of information were classified as "out of target" reason. As the Figure 2 shows the incidence of this classification was a bit high among other reasons of calling. ISTAT decided to understand better the contents and the information storage of "ut of the target".

3. Methodology

The main purpose of the proposed approach is the semantic categorization of "out of target" calls to detect insights and new information targets.

Each call describes problems and questions as short texts in natural language. Text complexity varies from very elementary sentences to convoluted involved periods.

The requests received by telephone and transcribed by a helpline operator do not have metadata able to support the semantic categorization.

Clearly, basic features of the calls, like the length of the message or similar measures, are not meaningful for categorization. Even the basic presence or absence of predetermined words is not enough, because in many cases similar words (answer, form, compile, fill, etc.) may describe completely different types of problems. Therefore, a more in-depth analysis is needed to extract the significance of a call from its text in natural language.

In this paper, we propose an unsupervised text clustering and topic extraction framework which integrates text clustering and topic extraction into a unified framework capable of achieving high-quality clustering result and extracting cluster-specific topics.

Text clustering and topic extraction are two important tasks in text mining. In particular, topic extraction facilitates clustering.

We can first project texts into a topic space and then perform a clustering algorithm to obtain clusters that group calls based on specific characteristics.

The developed framework includes three main components:

- Feature extraction converts 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 performs 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 finds main topics from each cluster using the keywords extraction methods TF-IDF (Term Frequency-Inverse Document Frequency).

The first component includes all the functions used to convert the generic call into a data record of reasonable size, summarizing the call. After an initial Tokenization (individuation of the words within the sequence of characters) and Stop-words Elimination (elimination of useless parts, like articles, etc.), we perform Lemmatization with Part-Of-Speech recognition. This means that, for each word, we remove the inflectional ending to identify its basic lemma. This allows us to recognize together the different inflected forms of a word (e.g., plurals of nouns, tenses of the verbs, etc.). Moreover, we still keep track of which part of speech each word is (e.g., noun, verb, adjective, etc.). This Natural Language Processing (NLP) step is performed by using the "Gensim" python library from "Scikit learn" framework. Since the calls are in the Italian language, we perform the above operations by using an Italian dictionary. However, the language can easily be changed by simply switching the underlying dictionary.

Subsequently, we need to convert each text into a data record constituting a "standardized description" of the call [2]. This feature extraction is done by using the word embedding algorithm Word2vec, still in the Gensim python library. This algorithm uses a shallow neural network to learn word associations in a large corpus of text, all the calls in our case [1]. Hence, it can detect synonymous words. In particular, Word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are generated in such a way that the cosine similarity between the vectors indicates the level of semantic similarity between the corresponding words. In more detail, we have assembled a corpus composed of the text of 18,000 real calls received by the helpline 1522. After the above-described NLP steps, we identified in the word embedding operations a vocabulary with 30,000 relevant lemmas, among which we later selected the 500 most relevant lemmas at the top score of the list. By projecting on this set of 500 lemmas, each text is now converted into a vector with 500 elements, each of which is computed as the frequency of each relevant lemma in that call. Number of occurrences normalized by call-related text size

The second component performs a clusterization of data record of calls using the k-means algorithm, a well-established clustering technique. This algorithm aims at partitioning n-observations into k-clusters in which in which each observation belongs to the cluster within nearest mean, being the centroid of the cluster.

K-means algorithm requires the number of clusters k to be specified in advance. To determine the value of k representing the best compromise between distortion and number of clusters, it is often used the "elbow" method, which fits the model with a range of values for k.

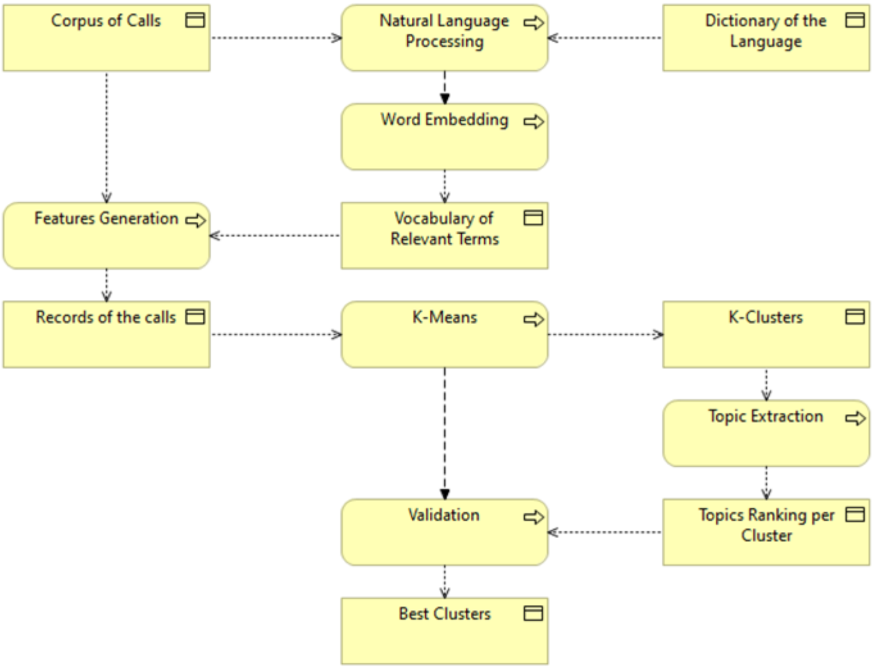
Clustering is done by using the functions k-means and k-ElbowVisualizer, from the python library scikit learn.

The third component detects the main topics for each cluster using the keywords extraction methods TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a statistical measure that evaluates how relevant a word is to a text in a collection of texts. This is done by multiplying two metrics: occurrence of a word in a given call's text, and the inverse frequency of the word across a set of call's texts. In particular, the method extracts the best five main topics for each cluster. This topic extraction is done by using the TfidfVectorizer, still in scikit learn library.

Now, domain experts validate the produced clusters by means the analysis of the topics extracted from each cluster. In the validation step, the clusters are merged, deleted or partitioned in order to produce the final classification. Validation step cycles for clustering optimization, until domain expert validate the best clusters

A scheme of the overall procedure described in this work is reported in the subsequent Figure 3.

Figure 3. Overall scheme of the proposed approach, including both the text mining phase and the machine learning phase



The result of the process showed in Figure 3 is a new classification able to avoid disambiguation and more representative of the “out of target” calls.

This classification allows us to support the data collection process using an automated procedure. Consequently, it improves the quality of the process of collecting and classifying data from calls made by the 1522 NGO professionals according to domain rules.

The proposed methodology works at the formal level using a data driven approach, and thus it could also be applied to the semantic categorization of other texts with different origins or in different languages.

4. Results

We analyzed the calls received by DEO's Contact Center: 12.315 “out of the target” calls; and 5.761 “out of service” calls.

The analysis allows to produce 38 clusters , encoded as a following:

- 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:
  - Cluster coherent with the existing target, because reporting same contents
  - Cluster out of the target (not coherent with the helpline services)
  - Cluster with new needs from the target, identifying new classification rules

More in depth:

A)

- *Cluster 2 topic: friendly phone, helpline violence against children,*
- *Cluster 4 topics: social service, social worker,*
- *Cluster 6 topics: extortion victim, suffer extortion money,*

have been classified as clusters. out of the target, because not complying with the aims of the 1522 helpline service.

B)

- *Cluster 1 topics: victim violence, rape;*
- *Cluster 3 topics: antiviolence center, cav;*
- *Cluster 5 topics: stalking, stalker;*
- *Cluster 7 topics: sexual harassment, 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.*

Have been classified as Clusters coherent with the existing target and

C)

- *Cluster 0 topics: psychiatric subject, schizophrenic,*
- *Cluster 19: mental disorder, bipolar bipolarism,*

Because relevant and coherent with the 1522 helpline service and finding new and important need of the target population have been created a new classification rule, titled “mental desease”.

The study enhanced the quality of the classifications rules for the helpline, by reducing the “out of the target” calls and adding “mental desease”.

5. Conclusion

The study allowed to improve the quality of classification’s rules of the 1522 helpline. After this study the NGO leading the service decided to better specify the “out of the target” calls, in order to reducing the big amount of invalid calls.

With this exercise on of the main goals of the citizen generated data process has been achieved, by reducing the nuisance of the calls and engaging civil society.

Citizen-generated data can significantly aid in gathering information about violence against women (VAW). This type of data collection empowers individuals and communities to report incidents, raise awareness, and drive policy changes.

Through the development of a supervised classification model using the validated clusters as a training set to learn the classification criteria and predict the class of unlabelled calls, it was possible to improve the collection and more precise classification of the data.

The tested approach would allow the classification of all unlabelled calls received by the 1522 telephone line and improve the quality of the information. The collaboration that ISTAT has established with the Department for Equal Opportunities and the NGO in charge of collecting (through its professionals) requests for help from victims and users of the 1522 service is a strategy that we can define as win-win. In one sense, it improves the quality of data collection by the NGO because a statistical point of view is adopted, and on the other, it obtains timely and useful data for policy to know the evolution of the phenomenon.

References

Bruni R., Bianchi G., & Papa P. *Hyperparameter Black-Box Optimization to Improve the Automatic Classification of Support Tickets. Algorithms*, 16(46), 1-14.

Bruni R., Bianchi G. *Website categorization: A formal approach and robustness analysis in the case of e-commerce detection*. Expert Syst. Appl. 2019, 142, 113001.

UN Women, WHO, *Improving the collection and use of administrative data on violence against women*, New York: United Nations Entity for Gender Equality and the Empowerment of Women (UN Women) and World Health Organization (WHO); 2022.

ISTAT, *The helpline 1522 during the pandemic 2021* – November 2021 https://www.istat.it/en/archivio/263909

ISTAT, Gender-based violence in the time of covid-19: calls to the 1522 helpline, Statistic Today, 13th of May, 2020. https://www.istat.it/en/archivio/245001

ISTAT, *The helpline 1522 during the pandemic (March-June 2020)*, ISTAT, August 2020, https://www.istat.it/en/archivio/246618