The Future of Internet

Methodological, technological and ethical approach to social networks analysis, data extraction and visualizations in REISearch 2017

Introduction

By including in the REISearch 2017 research process the present visualisations, we wanted to explore and communicate how European citizens express, discuss and feel about the Future of Internet populating the public sphere on social networks with their daily interactions, to:

- **create a public visual experience** of the data observed and analyzed;
- **socialize the data** and make them understandable by large audience;
- enable citizens to better understand the data they produce and their "digital" public sphere;
- **highlight the social and anthropological aspects** of the phenomena we are researching on, starting from and including **emergent expressions**.

Dealing with data – and bringing them into the public sphere – is key. Every day of our lives each of us generates progressively growing amounts of digital data: by shopping, expressing on social networks, exchanging messages, and even by traversing the spaces of the city, using our mobile phones and using our appliances and devices in our homes, offices and schools.

<< This information has started to constitute a large part of our public, private and intimate expressions >>

Publishing this set of visualisations is the starting point of **a larger data socialization process**: at the end of the research, **a full opendata set** will be released together with an **analytical report** describing the results of the whole observation.

Due to the pervasive role of technologies in contemporary societies, **openness** and **transparency** on the methodological, technological and ethical aspects of the processing and analysis of data is here considered not only a formal duty, but a need and a value.

For this reason we will fully describe in the following paragraphs:

1. the methodology and the overall harvesting and analysis process;
2. the technologies and techniques involved;
3. the critical issues;
4. the ethical issues.

[1] Process and Methodology

The visualisations are generated by using *Human Ecosystems*, a set of techniques, technologies and methodologies elaborated by HER - Human Ecosystems Relazioni.

Step One: the Harvesting Process

The first step is a harvesting process, in which major social networks are monitored in order to detect public content generated in Europe about the Future of Internet.

Sources:
public contents from Twitter, Instagram, 500 selected Facebook public groups and pages of thematic interest

Capturing content from the various sources requires different techniques.

For example, services like **Twitter and Instagram provide APIs** (Application Programming Interfaces) which allow searching for certain keywords, hashtags, geographic locations and timeframes, and, thus, to obtain the public content which was generated by users about certain topics and in relevant locations. There are limits for the usage of such APIs (for example on the number of contents which can be harvested, on the geographic area which can be searched, on the amount of time in the past for which it is possible to perform searches, and on the overall usage of the APIs themselves). Nonetheless, by combining the available data access points and modalities, it is possible to explore thoroughly the public content generated regarding specific topics and in specific locations (for example the administrative boundaries of Europe in this case).

Other services, such as **Facebook**, are much more restrictive. In effect a series of APIs and frameworks (for example the Open Graph and a few other ones) are available, but the limits for their usage are much more stringent. Content is obtained by performing an initial search for those pages and groups are relevant for the collection process, directly connecting to them (for example by using the “Join Group” function available on the social network) and, then, using the APIs, through which it is effectively possible to monitor the content which appears on them.

On top of this, all of **the services impose limits about how the harvested content can be used**. For example, it is not possible to store it directly in databases, it is necessary to provide the indication of the links from which it originates, it is necessary to provide attribution and declaration that the use is noncommercial, and similar ones. Through HER’s solutions it is possible to have all of these requirements satisfied automatically (they are specified in the various *Terms of Service agreements documents*, for users and developers, available on the respective social networks’ websites).

Step 2: Processing the Data to generate Knowledge

The **collected content is, first, anonymized and aggregated to form clusters which are suitable for analysis**, around different logics (for example by forming groups of contents by counting the mentions of a certain keyword).

<< The content is stored in our databases only in this form (anonymized and aggregated) >>

This is a very delicate stage, as it implies the verification of multiple types of conditions which not only ensure proper anonymization, but also the fact that, given an anonymized content, it is impossible (or really, really difficult) to climb back up to the original, identifiable, one. (For example: even though it is made anonymous, a geo-referenced data which is alone and isolated on a territory may make it too easy to understand to whom it refers to; this is why we remove these singularities and other similar cases from our databases).

***In general**:*
HER's systems are fully compliant with current EU regulations on privacy, personal data collection and management, and anonymization of personal data, and a dedicated team at HER monitors changes in laws and regulations to make sure that this fact remains persistent.

When this critical phase is complete, the data is processed to generate knowledge.

There are a number of processing and analysis techniques used, such as Natural Language Analysis, Emotional Analysis, Network Analysis, Geo-Referencing. They will be described in more detail in the next section.

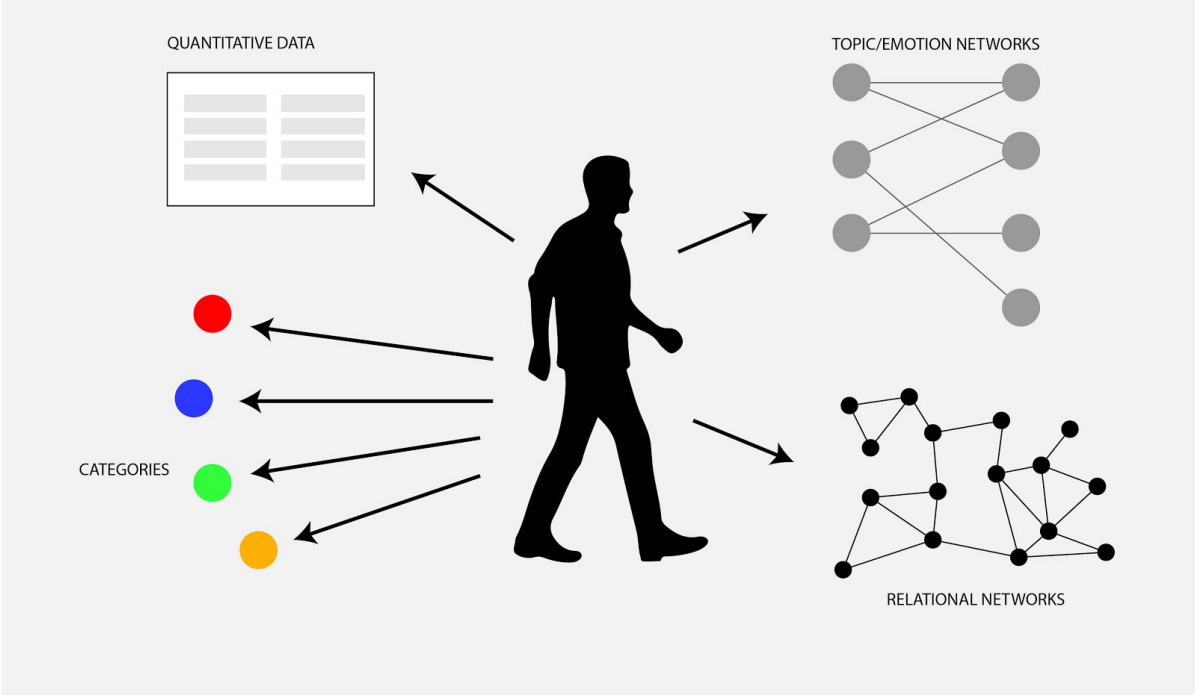
In this overview, it is important to highlight how these techniques are able to transform the unstructured data collected from social network (messages, images, comments, conversations…) and process it in order to transform it into structured data, forming the knowledge base of the research.

At this stage three typical types of knowledge are available:

- **topics**, as content is scanned for what it is talking about, and for what topics are discussed together in the same contexts;
- **emotions**, as content is scanned to gain understandings about what emotions (such as happiness, surprise, fear, anxiety, disgust, trust…) they are expressing;
- **times**, using both the content's meta-data and the phrases it actually contains to understand what time it refers to;
- **places**, where the content's meta-data (such as geographical coordinates) or sentences describe geographical locations;
- **networks**, in which the focus is to understand which people, organizations and other entities these contents put together, describing the models of relational networks and graphs (for example: do individuals talk among themselves or with organizations? or:

does information come from news items or from peer-to-peer discussions? and similar ones which allow to determine the models of communication).

All the information elements are semantically linked with each other and, thus, can be combined to infer more complex knowledge (for example, by combining knowledge about topics, places and times, we could be able to infer what the people at a certain event discussed; or by combining topics, emotions and networks we could understand which communities express which emotions about certain topics).



(Img 1: Type of Data in the Human Ecosystems knowledge base)

The content of the knowledge base is also used as a feedback process, to fine tune the data harvesting process, using a Machine Learning mechanism: here all the accumulated knowledge is used to evaluate new information to generate new knowledge about how to modify the data capture process, in terms of other words/topics to listen to, other pages, groups and communities (for example on Facebook) to include in the capturing process and other insights of similar nature.

The acquired knowledge is used in the following cycles, obtaining a system that learns and adapts to the evolving scenario (for example by understanding that at a certain time it may be interesting to include some other elements in the harvesting process, as they are particularly active and relevant).

The knowledge base is, then, used to perform some more standard analysis, such as qualitative, quantitative and community/ network analysis, to gain better understandings about the scenario that all of this information describes, such as:

- the **timelines** according to which the topics, emotions, places and communities evolve;

- the **topics**, according to which we are able to gain better understanding of how much certain topics are discussed, with which emotions, by which communities and in which places;
- the **communities**, with which we are able to understand how diverse or coherent different communities are, what they focus on, how they converge or diverge, what are their main concerns or desires;
- the **flows**, using which we are able to model how information, opinion, influence spreads;
- the **impacts**, with which we are able to gain understandings about the results of certain actions, such as how a communication campaign or even a single social networking message is able to influence people's behaviour;
- the **correlations**, with which forms a tool to help in comprehending possible cause/effect relationships;
- the **transformations**, in which it is possible to take the dimension of time into account, to study how all of the above evolve in time.

At this stage an evolving knowledge base is created.

Step 3: Socializing the Data

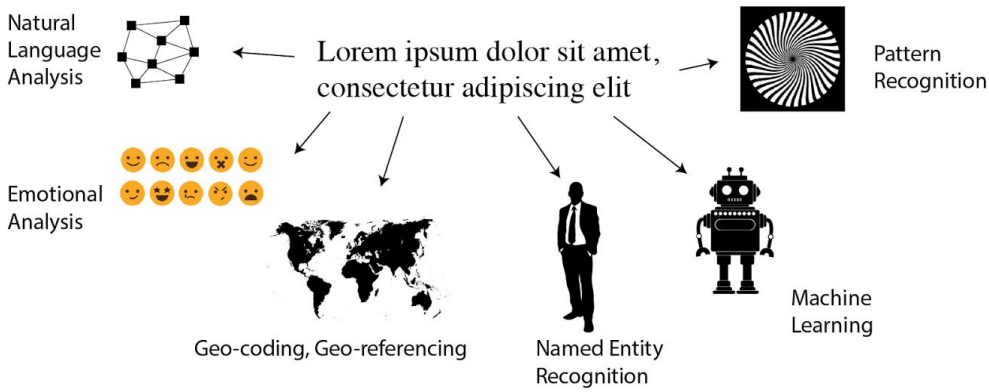
The final step is the data **socialization process**, which can take various forms.

In the context of this research, **visualizations and Open Data are the first tools** we have built to open and enable this process. Visualizations are made available through the **REIsearch website**: **each media partner involved can use it** to share its contents and diffuse them through their channels, creating a vast communication campaign at European level.

At the end of the research, together with the visualization a full **analytical report** and **a full opendata set** will be released, becoming a resource to be used by civil society, university, researchers, students, designers, artists, journalists and anyone interested in it for their own purposes.

[2] Techniques and Technologies

This section is intended to provide a short description of the main techniques and technologies used for processing the harvested data:



(Img 2.: Techniques and Technologies used in Human Ecosystems)

Some links are given to obtain further information about each technique for anyone interested to learn more on it.

Natural Language Analysis

The objective of **Natural Language Analysis** (or Natural Language Processing, **NLP**) transforms unstructured data such as text into structured data. It can be performed in multiple ways, with different objectives, such as understanding the topics which a certain text deals with, creating automatic summaries, machine translation and more.

In Human Ecosystems **NLP is performed in 29 languages**, and is used in the following ways:

- **Discourse Analysis**, which deals with understanding the structure of text and its components; for example using the way a certain sentence is written to understand if it is a question, an exclamation, a sentence providing information of some sort, an answer to a certain question, etc;
- **Semantic Analysis**, which deals with starting from text to understand its meaning, in terms of whether it assesses a certain topic and in what way, if it has a certain style for expression or if it uses a certain language;
- **Topic Discovery**, in which large numbers of sentences are observed to discover if recurring patterns may identify new topics to listen to which are relevant for the ones currently being observed; new topics come under the form of words, word patterns, sentence patterns and more;

- **Named Entity Recognition**, which uses streams of texts and their structure to identify proper names for people, places, events and more;
- **Relationship Extraction**, which uses text to identify the relationships between Named Entities (e.g.: who is married to whom; who is the employer of whom; etc.);
- **Sentiment/Emotional Analysis**, in which the words and the patterns in which words are composed are used to gain better understandings about what Sentiment the sentence is expressing (positive, negative, neutral), or, if enough information is available, what emotion it is expressing (such as joy, fear, anxiety, surprise, trust, satisfaction, etc.);
- **Information Retrieval and Information Extraction**, which, given the procedures listed above, deals with the possibility to store and extract the types of information which can be extracted from text.

More information on NLP can be found here:
https://en.wikipedia.org/wiki/Natural_language_processing

Emotional Analysis

As described in NLP, **Emotional Analysis deals with the possibility to automatically recognize emotions in text**, by recognizing how text uses word, phrase or sentence patterns.

<< In Human Ecosystems **36 main emotions** are recognized. >>

This occurs when enough evidence is present in the texts, confirming the nature of the expression being analyzed. Classification of emotions is performed using the **Circumplex Model of Emotions**.

In general, the Circumplex Model of Affect classifies emotions according to two main parameters: **Energy/Arousal** and **Comfort/Discomfort** (and, optionally, according to a series of further ones). In computationally analyzing the text, words and their combinations are accounted for according to their Energy and Comfort/Discomfort contribution to the sentence and, thus, the total can be used to infer whether a certain sentence is expressing a certain emotion. The levels of Energy and Comfort are determined through large semantic databases which Human Ecosystems has built across the years, in 29 languages, and has been built both manually, in collaboration with people and researchers from all over the world, and automatically, using machine learning to understand when certain systematic recurrencies confirm the fact that a certain expression denotes a specified level of Energy and Comfort.

More information about the Circumplex Model of emotions can be found here:
https://www2.bc.edu/~russeljm/publications/Russell1980.pdf

Network Analysis / Social Network Analysis

Network analysis studies **graphs**, **networks of relations** between discrete objects, or nodes.

In Human Ecosystems **Network Analysis is used to study the composition of the networks represented by the people, organizations, companies and not-human agents** (eg.: bots) whose expressions are captured through their public online expressions and, given these and their transformations, the flows of communication, information, knowledge take place in and through them, effectively describing how information, opinion, emotion, knowledge and influence spread across communities and cultures.

In Human Ecosystems a custom version of **Latour's ANT** (Actor Network Theory) is implemented to describe the behaviours of networks and of their participants, and to identify roles within them, such as influencers, experts, hubs, bridges among different communities.

In General, statistical models and graph-theory models are used to analyzed the networks, such as calculation of degrees, network diameters, graph densities, authorities[1], modularities [2], page rank[3], connectedness[4] and other similar ones. The results of these computations and indices are then fed to machine learning algorithms, both instantaneously and as networks evolve in time, to discover recurrencies and patterns, which may highlight the formation of interesting network/community configurations.

For more information about Network Theory:
https://en.wikipedia.org/wiki/Network_theory

For more information about Social Network Analysis:
https://en.wikipedia.org/wiki/Social_network_analysis

Geo-Referencing

This technique is the process of attributing a geographical context to a certain content.

The **geo-context** can be of multiple types:

- the **location** in which a photo has been shot;
- the **area** for which a certain content is relevant (for example a city or a state);
- the **path** along which a certain information is relevant (for example the path that takes from a certain location to another).

In Human Ecosystems geo-referencing is performed in two ways:

- **using the meta-data** included with contents, for example the geographical coordinates which social networks users can associate to their posts;
- **using the results of NLP**. In this case the **Named Entities** identified in text may be of geographical relevance (for example the name of a church, or a landmark, or the name of a restaurant); if the sentence includes enough evidence of the spatial character of the expression (for example the sentence may state that "I am going

[1] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, in Journal of the ACM 46 (5): 604–632 (1999)
[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000
[3] Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117
[4] Robert Tarjan, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972)

to…"), sufficient information may be present to identify the geographical context for the content. For this research, this modality has been used to identify whether the posts mentioned European nations, regions, cities, or universities, research centers, institutions, organizations. For this, a database of all European Nations and administrative regions, the largest 500 European cities by population, and a list of European Universities[5].

On top of that, a **GIS (Geographical Information System)** is used to establish the hierarchical characteristics of space, according to which certain coordinates are contained in certain blocks, which are contained in certain neighbourhoods, which are contained in certain zones of the city, which are contained in the city, and so on.

Machine Learning

Machine learning studies the possibility to design algorithms which are able to **recognize recurring pattern**s (in this case: patterns in texts and other parameters, such as quantities, time series etc.) and to use the fact that certain patterns have been recognized to learn, producing systems which automatically adapt themselves to changing scenarios.

In Human Ecosystems, Machine Learning is used in multiple ways:

- for the NLP techniques;
- in Topic Discovery;
- in Emotional Analysis;
- and to fine tune the data harvesting processes by adjusting what keywords, phrases and forms of sentences are monitored on social networks.

To learn more about Machine Learning:
https://en.wikipedia.org/wiki/ Machine_learning

Other Techniques

A few additional techniques have been used in this research:

- to **detect demographics**;
- to **detect the type of account** (individual, organization, bot).

To detect demographics we have used research coming primarily from health-related practices, to be able to use approaches and techniques which have been tested and evaluated in strict, formal environments, and to be able to build on medicine's ethical approaches, which are among the most stringent in regards to preserving people's rights.

For this to infer gender and age group from the harvested posts we have followed the indications emerging from the following study:

- **Cesare, N., et al** (2017) "*Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices*", arXiv:1702.01807 [cs.SI]

[5] source: https://en.wikipedia.org/wiki/Lists_of_universities_and_colleges_by_country

The study conducts a thorough review to evaluate the different modes in which researchers have documented to have extracted demographic data from social networking posts of various kind.

In this research we have selected and combined modes indicated in the paper as “scalable”, in the extraction of “gender” and “age group”:

- **M. Vicente, F. Batista, and J. P. Carvalho**, “*Twitter gender classification using user unstructured information*” in Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, 2015, pp. 1–7
- **P. A. Longley, M. Adnan, and G. Lansley**, “*The geotemporal demographics of Twitter usage*” Environ. Plan. A, vol. 47, no. 2, pp. 465–484, 2015
- **J. Chang, I. Rosenn, L. Backstrom, and C. Marlow**, “*ePluribus: Ethnicity on Social Networks.*” ICWSM, vol. 10, pp. 18–25, 2010

In our case, from a sample of 2000 random elements, the application of these algorithms has proven to be **82% accurate**.

To detect the type of account (for example, whether it is an individual, or an organization or a bot), results from these publications have been used:

- **Varol, O. et al** (2017) “*Online Human-Bot Interactions: Detection, Estimation, and Characterizatio*n” arXiv:1703.03107 [cs.SI]
- **Davis, C. A. et al** (2016) “*BotOrNot: A System to Evaluate Social Bots*” arXiv:1602.00975 [cs.SI]

In this sense, the frequencies and characteristics of the posts and of the words and tags and shares which their corresponding social network users have expressed, are used to determine a 0-100 score in which 0 means “human”, and 100 means “bot”. In the 40-80 range value, a NLP technique has been used to detect use of collective, impersonal or other “non individual” means of expression and self-representation, to attempt at detecting “organizations”, and differentiating them from both humans and bots.

From a sample of 2000 random accounts tested from the ones which have been collected, the results have proven to be **74% accurate**.

[3] Critical Issues

A series of critical issues have been identified during the research process. The following are the most relevant ones.

Privacy Laws and Regulations, and Terms of Service Documents

Most online operators provide their services under condition of accepting their **Terms of Service** (ToS) documents. These are legal documents which Internet Users acknowledge and approve when subscribing to these services.

Social networks have very complex, strict ToS documents, which are intended both to preserve people's rights while they use the online platforms, and to ensure that the business interests of the providers are protected.

Currently different ToS are provided for both regular users and for developers, specifying different levels of access, usability, availability of the data and functions made available by the different service providers.

<< Limits described in the ToS documents include the ways in which information from these platforms can be extracted and used. >>

Atthe same time, a complex set of Intellectual Property, Data Access and Usability and, even more important, Privacy and Data Protection laws and regulations exist at multiple levels, for example National and European.

<< The current european regulation on personal data has been established in the European Parliament and in the Council of 27 April 2016. >>

It has been active since May 2016, and the project is already fully compliant with it.

Establishing how the policies of Social Network providers match and are compliant to the EU policies is very difficult matter. Even though a certain amount of transparency is mandatory, the presence of proprietary technologies and the frequency at which the technological systems of these providers mutate (for maintenance, ordinary source code updates, interface changes, etc), i**t is virtually impossible to thoroughly describe how (and if) these systems comply with what the EU** has described as mandatory to preserve people's rights.

At the same time, **putting the ToS documents and the EU regulations side by side lets a plethora of grey areas emerge**, and opportunities for "interpretations" which cause this type of activity to be far from certain both in its actions and effects.

In the context of this research we consider of great importance both to preserve people's rights, according to EU laws and regulations, and to preserve the needs of businesses and organizations.

In this sense we have decided to systematically apply EU laws and regulations, first, and, then, to apply what is dictated by the ToS agreements coming from the various service providers.

This creates some uncertainties and discrepancies, most of all in the fact that it is currently not possible to understand, for organisations as well as for citizens, in what ways Social Networking providers apply the same EU laws and regulations, and in what ways these laws and regulations match what is stated in the ToD Agreements documents, and in the source code of these platforms, for lack of transparency and openness.

Considering the object and the nature of this research, inquiries, discussions, code analysis, and open sourcing of data and source code are welcome to be then analyzed in open, public, transparent processes, with all stakeholders and interested parties.

Quality of Automatic Interpretation

This issue deals with the **quality of the interpretation** of the content as processed by the automatic algorithms.

This means to try to ensure that if algorithms detect that a certain content deals with topic X, the content effectively deals with topic X. This is a very complex thing to do. While performing tasks such as NLP, algorithms are de-facto collecting bits of evidence across texts, such that at a certain point enough evidence can induce us to believe that a certain content is effectively talking about X. But these are not final determinations, they are probabilistic: for any combination of such evidence, we will be always XY% sure about this fact, and XY% will never fully be 100%, there will always be a doubt.

Human Ecosystems confronts this problem by **establishing very high thresholds**. Currently Human Ecosystems accepts a certain interpretation only if there is evidence to prove it which accounts for **more than 95%** of probability.

Irony

This issue is a peculiar version of the previous one.

Social media (and Internet in general) is a context which is characterized by high degrees of irony. This means that the situation in which someone is expressing something and really meaning its opposite will happen very often. In computational terms, this means that the situation in which an algorithm will efficiently identify topic X or emotion Y in a message and the user generating it meant the exact opposite, will happen very often. This is currently one of the most pressing issues in Natural Language Analysis: Irony.

There are a number of techniques which are currently used to mitigate these issues. All of them take into account the context in which each message is generated.

By studying the context in which a certain user communicates (her beliefs, opinions…) **we will have better tools to interpret an ironic conten**t. This is what Human Ecosystems does: if for a certain topic at least **75%** of one users' expressions is polarized in a certain way, a further expression which is polarized very differently will not be accepted immediately, but placed in a limbo, "on hold", until enough further evidence will be able to prove that the user has changed opinion.

Lack of Intentionality

With this issue we refer to the possibility that online expressions do not always reflect what online users chose to express, with intention.

This is an issue with multiple faces. For example, by understanding a certain message we could be able to collect enough evidence about a person's behaviour, or opinion. This fact is only partially related to the same person's belief system, or values, or desires. The person might have been angry, or in a hurry, or even forced to express in a certain way, for respectability, reputation, work, shyness, or multiple other reasons. Or, on the other hand, people may not consciously realize that they are debating issues in the public sphere. Or they may even not realize that they can establish such debates, and not talk at all about such issues, even if they care about them.

In general, little can be determined about the intentionality of the expressions, due to their emergent, informal character.

[4] Ethics

Many of the issues identified in the previous section can be merged together with other, more general, ones in defining the need for the composition of a comprehensive ethical code.

These are the principal elements the Code of Ethics and Conduct adopted in the present research:

- full respect and compliance for recognized laws and regulations, at regional, national, European and international levels;
- explicit and avoid conflicts of interest of any form, especially for whatever concerns the code of ethics and conduct;
- provide clear and accurate communication;
- operate with transparency and integrity;
- protect people's data and rights, by respecting current laws and regulations and by providing full access to data, information and knowledge.