III
A practical approach to exploit public data on the Internet

This chapter introduces a practical approach for using valuable social data, especially the one available in online communities of social network sites, to conduct social studies in order to increase the comprehension of important healthcare issues. This approach was designed with three stages to guide researchers to achieve their research objectives. It starts with the online search content analysis, which aims to survey how users are looking for information in the whole Web, and continues searching for a specific online community in social networking sites, where users discuss among themselves, until finding one that looks promising with regards to the research questions. It is a systematic use of free data available on the Web supported by computational tools, and therefore a guide to researchers. The results presented here have been partially published in the XII Workshop on Medical Informatics (WIM2012), a satellite event of the XXXII Congress of the Brazilian Computer Society [Car12].

III.1 The proposed approach to study healthcare based on social media

Influenced by ethnographic principles, Kozinets [Koz09] proposes a method called Netnography, where observation happens in users' natural habitat on the Internet. The content is detailed and contextualized, and can be retrieved in non-intrusive ways, enabling an opportune, effective and efficient way to process it. In this way, users are not summoned to participate in a reactive fashion (*e.g.* online surveys), which enables the analysis of freely constructed opinions and manifestations. In this research, the goal is to exploit data publicly available on the Internet that can provide insightful information to be analyzed by experts.

The proposed approach is a means by which to make opportunistic use of public data available on the Web in order to conduct research about healthcare issues. To accomplish this, the approach considers the use of several tools

already available on the Web, as well as others tailored for specific purposes. This section presents some of the tools that can be used to acquire data for analysis and describes how to integrate this data into a research report.

The proposed approach is driven by research interests (*e.g.* questions) and has three stages, each related to the source of information, which are: 1) The Internet, 2) Social Network Sites, and 3) Online Community. The Figure III.1 contains a representation of the entire approach in BPMN, showing the researcher driven by his research questions and using the appropriate data sources (Literature and The Internet) to consolidate the findings in a scientific report. As a starting point, a researcher can go to the academic literature, establishing a base upon which to begin his research quest. Notably, this approach leads to studies characterized by broad exploratory research in the early stages, which is incrementally refined into a single online community analysis. At all research stages, the researcher can have insights, which may affect or even change the original research questions, a characteristic of data-driven research [Jan91, Jup06]. Finally, the literature review is also important to anchor the research findings in established theories in order to contribute to the existing body of scientific knowledge.

The initial stage of the approach is similar to that of someone with a medical condition; it uses search engines. It is increasingly popular to search on the Web about symptoms and medication, often before looking for a doctor. Looking at the way people search for information on the Web can provide insightful data regarding population behavior [Eys09]. Moreover, search engine companies, such as Google, provide a set of tools to help service providers create better ad campaigns (*e.g.* Google Insight for Searches, Google Trends, Yahoo Clues), as this is their main source of income. The idea is to take advantage of these systems and tools, looking for general population search trends, keywords and insights.

In the first stage, the researcher should utilize the facilities of the big search engines. Such facilities provide access to systems that provide information about how people are searching on the Web. For instance, Google Insights for Search[1] allows one to compare search volume patterns across specific regions, categories, time frames and properties. Yahoo! Search Clues[2] provides an exploratory environment to find interesting patterns in what people are looking for on Yahoo! Search. Similarly, Bing formerly provided access to Bing Trends[3], but Microsoft recently shut down the service. Other services can also be used to acquire mass data information, such as those cited

[1] `http://www.google.com/insights/search`
[2] `http://clues.yahoo.com`
[3] `http://www.bing.com/trends`

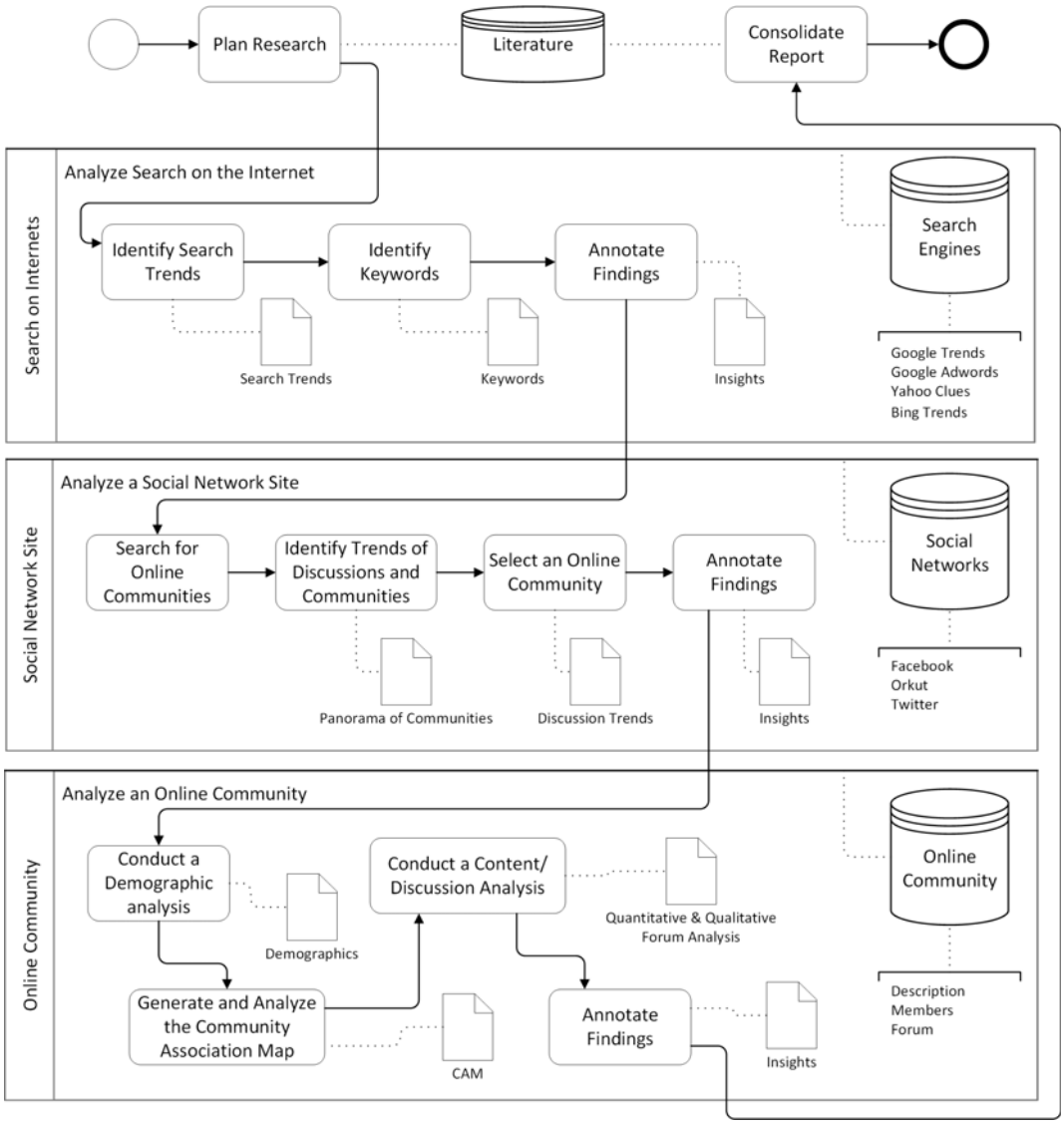

Figure III.1: The proposed research approach.

by Eysenbach [Eys09] (*e.g.* Google Ads). Despite all computational support available in these facilities, it is still up to the researchers to use these tools in order to find relevant content for the studies. For example, finding relevant keywords in Google Insights for Search is a task that starts with keywords interesting for the study, in an interactive process that explores the dataset, and relying on the researchers analytical capabilities in order to identify what is relevant to be considered for the study.

The second stage of the approach examines the world of online communities in social networking sites, looking for discussions about the topics of interest and identifying a panorama. Online communities are an important place for people on the Internet [Pre05], where users can interact with others who share a common interest, exchanging information and finding support. A more recent phenomenon is the increasing use of social networking sites [Boy07], which also

support the establishment of online communities. The approach is to use the social network sites' abilities to search for keywords related to the studying, identifying an overview of online communities.

Social networking sites, which are systems like Facebook, Twitter and Orkut, are the content source for the second stage. A researcher can use keywords identified in the previous stage to investigate what users are saying about the research topic in a social network site. The results can provide an index of user interest toward the research themes (*i.e.* trends). These results should also be considered into the decision making process about which community to study in depth. Many metrics can be used to support the research decision of which community is the most suitable to the study. These metrics are based on the information available on the social networking sites. For instance, interesting metrics for this purpose are number of members (is it big enough?) and messages (does it contain content?), or the time of last message sent (is it active?), community creation (how old is it?), and other properties as privacy settings (can I study it?). These questions and metrics are grounded on the fundamentals of virtual communities' definition. Tailored tools should be utilized in actual data processing and analysis.

The third and final stage takes an in-depth look at an online community, looking to explain the community reality as related to the research interest. A content analysis of discussion available at the community forum provides a way to understand the community. This analysis employs social science and computational techniques. For instance, a technique that might aid in performing the discussion analysis is the "Discourse of the Collective Subject", a qualitative technique with roots in the Theory of Social Representations [Lef06]. This technique has been used by by Madeira [Mad11] in analysis of online forums conducted. The aim of this technique is to identify collectives, aggregated by central ideas, and describe them through a discourse generated from a patchwork of members' speech, synthesizing the discourse as one collective subject. Other content/discussion analysis methods and techniques can also be employed for this purpose. Another example of an analytical tool is the Community Association Map [Car12], which aims to help researchers to analyze the users' interest in other communities. This tool can reveal patterns in interest and reinforce conjectures about the users' beliefs. Based on quantiative techniques, other tools can be applied to help extract information from the community, for instance, URL extractors, and demographic data descriptors.

Two examples of studies conducted as proposed by this approach are shown in Chapter V. In the first study, the goal is to identify demand patterns of carriers of hepatitis C. The other study investigates motivations for drug

abuse to start and cease, specifically with regard to the drug crack cocaine in Brazil.

(a) Discussion about the proposed approach

The proposed approach presents a practical guide to conducting studies based on data retrieved from the Internet. The main information source is based on online communities, but the search also begins with a broad view of the data available on the Web. Researchers should take care using this data, mainly because of questionable effects on the validity of the research findings. Since the social networking sites are evolving, studies which are considering using online communities available in these sites require suitable content selection. This suitable content selection can be based on information available on the online community and taking advantage of new features, like URLs, which would enable the access of content also available on other parts of the Web. Another important concern regards the ethical aspects of these studies. For that reason, this approach looks at public data available on the Web, suggesting the use of techniques that report aggregated data, therefore preserving the privacy of the individual user. In the literature, an extensive debate was found about these issues [Whi07, Koz09, Eys09]. The proposed approach should be used as a guide for the realization of a study. The several systems, services and tools support the approach execution are technologies in constant and fast evolution. Researchers may face dead-ends in data analysis due to size and time constraints; therefore, they should be creative and develop different ways to accomplish the analysis. It is best to consider the proposed approach as a practical road map to guide the data scientists through the research.

The development of new approaches and understandings of the ways in which people join forces, share and identify relevant information can foster improvement in the quality of their lives. The proposed approach provides a guide for researchers conducting studies. Although available data and tools exist to support the execution of the approach in most studies, better tools and data might improve its execution. The evolution of research requires initiatives of opportunistic and inventive use of data and technology. The proposed approach is such an initiative, showing a new way conduct those studies in practice.

III.2 A system to support a process based on the proposed approach

In order to support experts in analyzing social media such as defined in the approach presented before, a multi-agent system is proposed as a framework with tools. Figure III.2 presents a modeling view using I* notation of a system to support the proposed approach execution. This view provides a seminal requirement elicitation of all system elements, which were further developed also considering desirable key-features, in a mix of human and computation parts. The Social Media Expert Actor represents the specialist responsible for carrying out the study through the analysis of social media. The tasks are the general steps considered in the proposed approach in order to conduct studies. The Pathfinder Actor represents the task of stages 1 and 2 of the proposed approach, starting from looking for keywords and finding relevant online communities to the study. The Retriever Actor is a crawler script to scrap data out of social media sources. The Manager Actor would take care of the Efficiency soft goal by managing the number of Crawler Actors operating at the same time. The Analyst Actor has the goal to analyze the retrieved content. This analysis might be of any kind, from simple URL counting to an in-depth discourse analyze. The Assembler Actor takes care of consolidating the results and presenting the report in an appealing format.

Several desirable key-features drive the computational system design. The main design goals are as follows. I) Governance: empower the users as the "king" of the system. II) Architecture: composed by loosely-coupled and smart components. III) Strategic sourcing: find and use strategic sources of social media data. IV) Processes automation: automation of manual tasks as much as possible. V) Meaningful visualization: the process outcome must be easily understandable, seeding insights and new ideas for analysis. The multi-agent system paradigm [Woo01] provides a rich environment to address the main design goals, as shown later in this section. This work presents the seminal development of the multi-agent system to support the social media experts in their work.

(a) System key-features

The analysis of social media is a laborious process. It sets challenges in all three main stages of a study (data gathering, representation and analysis) [Cor10]. Social media experts want tools to help them with the tasks to find, organize and analyze the data. Based on the experience of building these tools [Car10], several desirable features for a system to support social

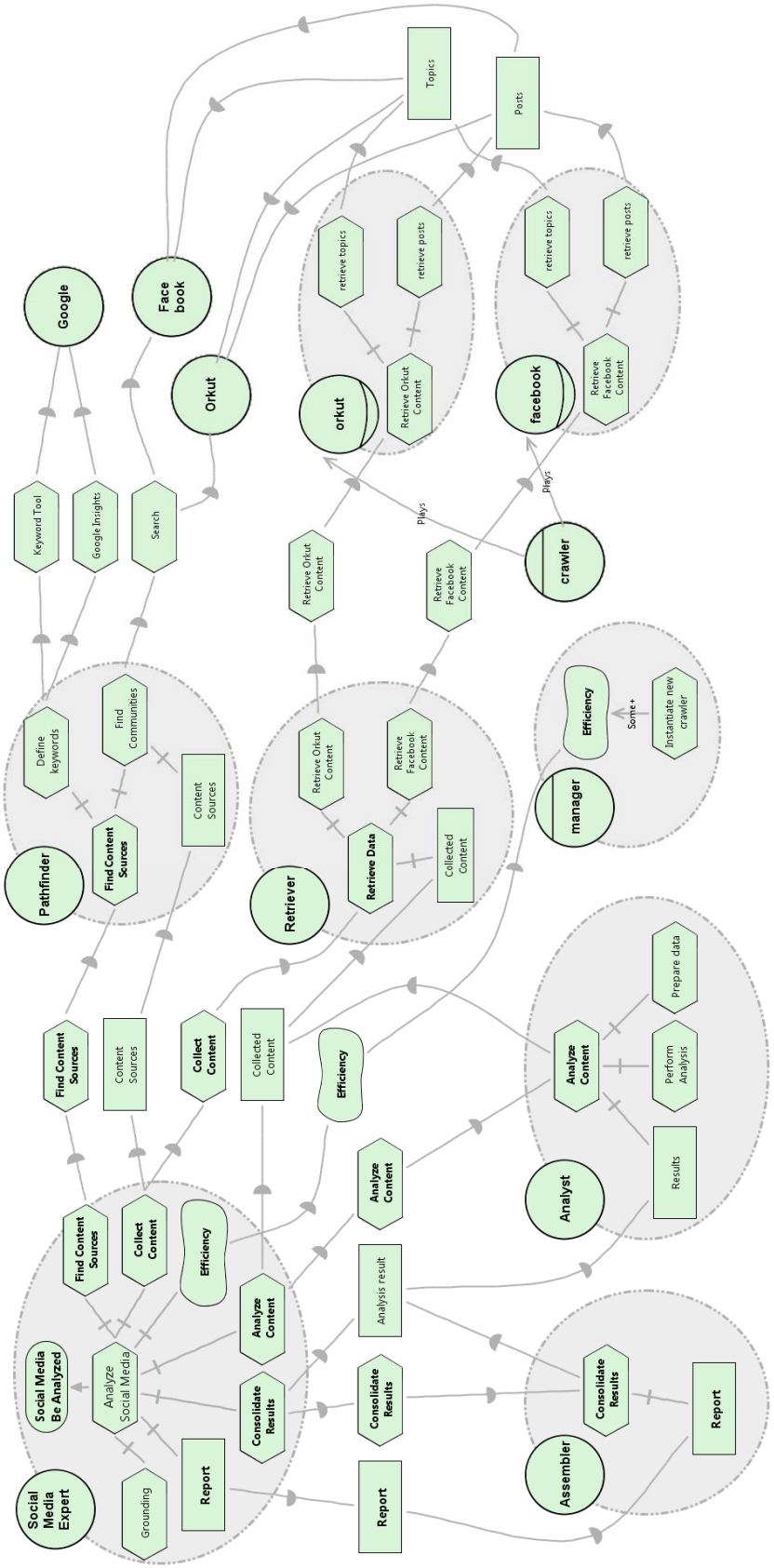

Figure III.2: I* model of a plataform to support the proposed approach execution.

media analysis have been identified. These features are grouped into five key-features, which are presented below.

Governance

Motto: empower the users as the "king" of the system.

The social media expert wants a system to support them on their duty, and not as a part of an analysis process. The users expectedly know the subject of analysis, so they want to define the data sources for the data-gathering step. They also want to get feedback from the evolution of the gathering process, and based on the status, redefine targets – broadening or shortening the boundaries of analysis. They want to try different models to represent the collected data, and experiment new measures. The system should work proactively for the users.

Architecture

Motto: be composed by loosely-coupled and smart components (*i.e.* software agents).

The methods for social media analysis are in constant evolution alongside technology. New systems and methods come out constantly improving the analysis process. To take advantage of this constant evolution, the system should have a flexible interaction between its components, preferably based on their data workflow. Users also want pro-active components with smart behaviors.

Strategic sourcing

Motto: find and use strategic sources of social media data.

The amount of social media available on the Internet is gigantic and growing. It is crucial to define clear and reasonable boundaries of analysis. The user defines primary targets for data gathering, strategically defined according to the analysis. The system should help the users to define the targets and suggest promising ones.

Processes automation

Motto: automation of manual tasks as much as possible.

Some processing steps require users' intervention. The first user interaction is to define the data sources, as mentioned before. The processing phase of the collected data also needs human expertise. The text processing tools are improving over the years, but the users still need to help to adapt the al-

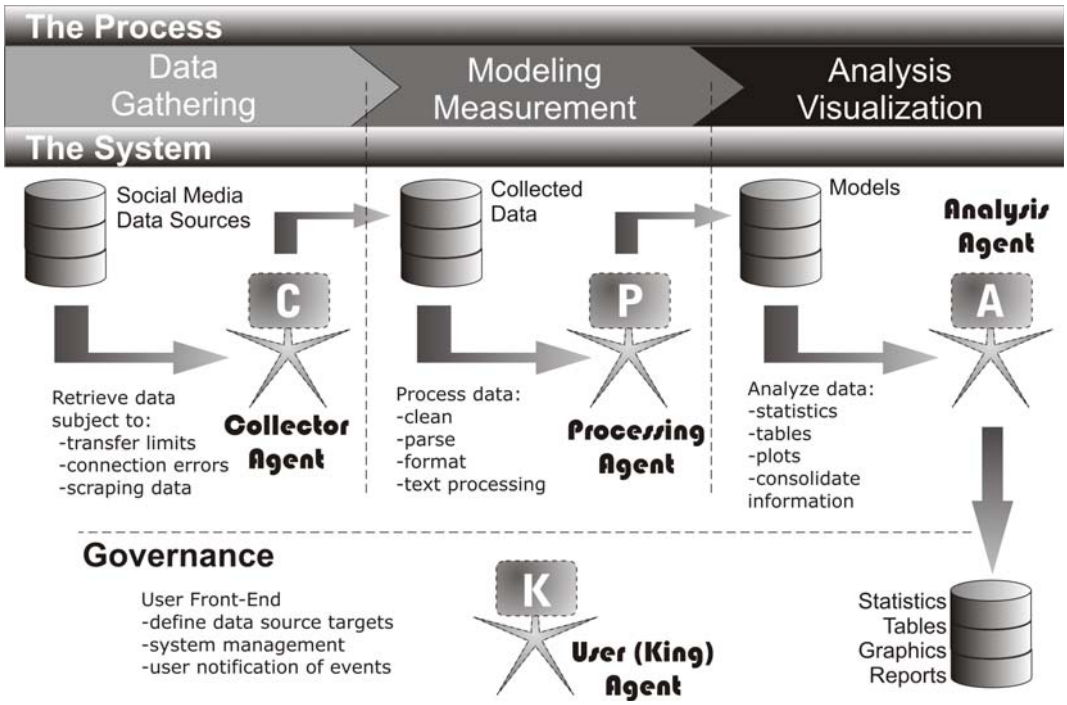

Figure III.3: The general process of social media analysis and the proposed multi-agent system to deal with its requirements.

gorithms to classify the whole data. The users should also scrutinize the data to check its quality and decide what analysis methods to use. The analytical tools may also need minor tunings, achieved by user interaction.

Meaningful visualization

Motto: the process outcome must be easily understandable, seeding insights and new ideas for analysis.

The visualization of the information plays a central role in the business. Since it is a high-level requirement, the implementation solution is to enable users to experiment visualization. The point is to leverage the system with the top-notch visualization packages to present the information in a best effort fashion.

(b) System architecture

The system architecture design has two major influences. First, the process of social media analysis defines the data workflow in the system. The other is the principles identified in desirable key-features, which were presented before. Therefore, a multi-agent system with four main agents is proposed as the system architecture. These agents are autonomous software components that cooperate to help the users to analyze social media data.

Since the system's main purpose is to support users, the autonomy and responsibility of each agent change depending on the analysis. However, the proposed architecture frames a general social media analysis process and addresses the key-features desirable by users. Even with the requirements modeled in I* for the proposed approach (see Figure III.2), the proposed architecture was designed as a solution for a general process of social media analysis, but mainly bearing in mind the proposed approach support. The choice to support a general process instead to stick with the specific requirements of the proposed solution was due the constant evolution of the technologies involved and a more specific solution would turn the architecture adaptation unfeasible in short time. Figure III.3 gives an overview of the system architecture, having on top the three steps of the social media analysis process and its support system implementation as a multi-agent system. In order to outline the agents, their main responsibilities and concerns are presented.

The Collector Agent addresses the Data Gathering process step. From social media data sources, this agent is responsible for data retrieval. Its main concerns regard connection errors (*e.g.* Internet connection, external system unavailability) and external system access limits (*i.e.* request limits). It is also desirable to help users to find new promising social media data sources. In a more sophisticated scenario, this agent can replicate its execution in different computers to overcome connections limits or problems.

The Processing Agent is responsible for the Modeling and Measurement process step. This agent can perform many tasks according to the collected data type. Generally, the first task is to clean the data from eventual noise and prepare it for further tasks. Parsing and formatting may be required tasks to transform the data into a better representation. Since social media is vastly available as text, text-processing tools can help users to characterize the data. The boundaries among the modeling, measurement and analysis are fuzzy. The idea is to prepare the data for further analysis, representing them in a more appropriate model, and to provide initial analysis as well.

The Analysis Agent has two main responsibilities. The first is to generate refined analysis and the second is to consolidate the information. This agent also interacts with the visualization packages to plot appealing information views. The process outcome is available for the users in many ways, such as statistics, tables, graphics and reports.

The User (King) Agent works in the behalf of the user, coordinating the system execution. It is the front-end with the user. This agent is the system manager, representing the users' will. The other agents have autonomy on execution, but they cooperate accordingly to what the User Agent says. This

agent can tell other agents to find ways to work faster (allocating more CPU or network resources) or even to stop working. The idea is to give the agents freedom with the allocation and use of resources, but this also means the user can control it. It also carries two other main responsibilities, one to inform the Collect Agent of the data source targets and the other to notify the users of important events.

The system agents are smart tools to help the users on their duty. They are autonomous entities, that work proactively, but are also responsive to users' wants. The system architecture has loosely coupled components, because the agents have total freedom in the way they work. The data workflow defines the agents' interactions, with an exception to address the governance requirement. The strategic sourcing and resource allocation define the pace of the system execution, also limited by the automation of the tasks involved in the process. The utmost goal is to deliver a complete final report with the social media analysis, but it is subjected to how users choose the tools that will help them to achieve it.

These smart tools can be very complex systems. A full autonomous system to support the proposed approach execution is the utmost goal. However, for the sake of this research, only few tools were developed. Most of the intelligent parts of this system are still based on human experts that perform the test. The developed tools, with their internal processes specification, are presented in the next chapter.