Social Business Intelligence in Action

Matteo Francia, Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi(✉)

DISI, University of Bologna, V.le Risorgimento 2, 40136 Bologna, Italy
{matteo.francia3,enrico.gallinucci2,matteo.golfarelli,
stefano.rizzi}@unibo.it

Abstract. Social Business Intelligence (SBI) relies on user-generated content to let decision-makers analyze their business in the light of the environmental trends. SBI projects come in a variety of shapes, with different demands. Hence, finding the right cost-benefit compromise depending on the project goals and time horizon and on the available resources may be hard for the designer. In this paper we discuss the main factors that impact this compromise aimed at providing a guideline to the design team. First we list the main architectural options and their methodological impact. Then we discuss a case study focused on an SBI project in the area of politics, aimed at assessing the effectiveness and efficiency of these options and their methodological sustainability.

Keywords: Social Business Intelligence · User-generated content · OLAP

1 Introduction

An enormous amount of *user-generated content* (UGC) related to people's tastes, opinions, and actions has been made available thanks to the omnipresent diffusion of social networks and portable devices. This huge wealth of information is raising an increasing interest from decision makers because it can give them a timely perception of the market mood and help them explain the phenomena of business and society. *Social Business Intelligence* (SBI) is the discipline that aims at combining corporate data with UGC to let decision-makers (simply called *users* from now on) analyze and improve their business based on the trends and moods perceived from the environment [4].

In the context of SBI, the most widely used category of UGC is the one coming in the form of textual *clips*. Clips can either be messages posted on social media or articles taken from on-line newspapers and magazines, or even customer comments collected on the corporate CRM. Digging information useful for users out of textual UGC requires first crawling the web to extract the clips related to a *subject area*, then enriching them in order to let as much information as possible emerge from the raw text. The subject area defines the project scope

Partially supported by the "WebPolEU: Comparing Social Media and Political Participation across EU" FIRB Project funded by MIUR.

S. Nurcan et al. (Eds.): CAiSE 2016, LNCS 9694, pp. 33–48, 2016.
DOI: 10.1007/978-3-319-39696-5_3

and extent, and can be for instance related to a brand or a specific market, or to a wider domain such as EU politics. Enrichment activities may simply identify the structured parts of a clip, such as its author, or even use NLP techniques to interpret each sentence, find the *topics* it mentions, and if possible assign a *sentiment* (also called *polarity*, i.e., positive, negative, or neutral) to it [10]. For instance, the tweet "UKIP's Essex county councillors stage protest against flying of EU flag at County Hall. Well done to them", in the subject area of EU politics, mentions topics "UKIP" and "protest" and has positive sentiment.

We call *SBI process* the one whose phases range from web crawling to users' analyses of the results. In the industrial world, the SBI process is often implemented in the so-called *social media monitoring tools* [16], i.e., commercial tools and platforms available for the analysis of UGC, such as Brandwatch, Tracx, and Clarabridge. Their main feature is the availability of a fixed set of dashboards that analyze the data from some fixed points of view (such as topic usage, topic correlation, and brand reputation) and rely on some ad-hoc KPIs (e.g., topic counting and sentiment), so they lack in providing flexible user-driven analyses.

In the academic world, the SBI "big picture" has not been deeply investigated so far. In [2] we proposed a reference architecture and an iterative methodology for designing SBI applications, and showed how its adoption can make the activities for developing and maintaining SBI processes more efficient and the SBI process itself more effective. However, we also concluded that SBI projects come in a variety of shapes, characterized by different relevance and sophistication degrees for each design task and architectural component, which results in quite different demands in terms of skills, computing infrastructure, and money. Hence, finding the right cost-benefit compromise depending on the project goals, on its time horizon, and on the available resources may be quite hard for the designer.

During the last few years we have been involved in different SBI projects. In particular, in the context of the WebPolEU project we developed an SBI platform aimed at investigating the connection between politics and social media. The project used UGC written in three languages and was focused on the 2014 European Election. This experience has motivated us in writing this paper, whose goal is to discuss the main factors that impact the above-mentioned compromise aimed at providing design guidelines to the SBI design team. To this end, first we list the main technical options for each architectural component together with their methodological implications. Then we discuss a case study focused on the WebPolEU project, aimed at assessing the effectiveness and efficiency of these options as well as the overall sustainability of the methodological approach, based on a qualitative and quantitative analysis of the critical issues related to each architectural component and design activity.

2 Architectural and Methodological Framework

The reference architecture we proposed in [2] to support the SBI process is depicted in Fig. 1. Its main highlight is the native capability of providing historical information, thus overcoming the limitations of social media monitoring tools

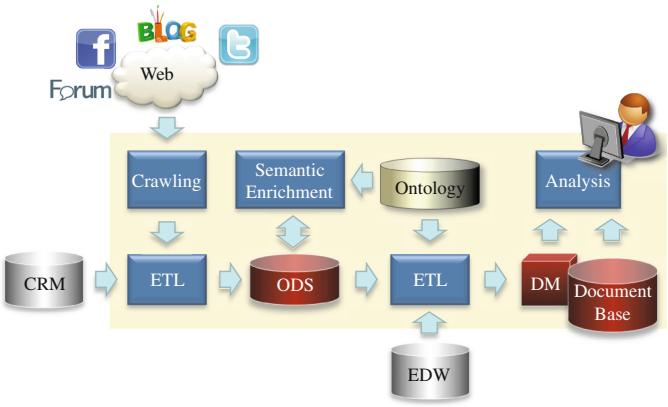

Fig. 1. A reference architecture for the SBI process

in handling the data reprocessing typically required by cleaning and semantic enrichment needs. In the following we briefly comment each component.

- The *Operational Data Store* (ODS) stores all the relevant data about clips, their authors, and their source channels; the ODS also represents all the topics within the subject area and their relationships.
- The *Data Mart* (DM) stores integrated data in the form of a set of multidimensional cubes which support the decision making process.
- The *Document-Base* stores the clips in textual form and the related meta-data to be used for text search.
- *Crawling* carries out a set of keyword-based queries aimed at retrieving the clips (and the available meta-data) that are in the scope of the subject area. The target of the crawler search could be either the whole web or a set of user-defined web sources (e.g., blogs, forums, web sites, social networks).
- *Semantic Enrichment* works on the ODS to extract the semantic information hidden in the clip texts. Such information can include its topic(s), the syntactic and semantic relationships between words, or the sentiment related to a whole sentence or to each single topic it contains.
- The *ETL* process turns the semi-structured output taken from either the crawler or the CRM into a structured form and loads it onto the ODS. Then it integrates data about clips and topics with the business data extracted from the EDW (Enterprise Data Warehouse), and loads them onto the DM.
- *Analysis* enables users to explore the UGC from different perspectives and control the overall social mood.

From the methodological point of view, we observe that the roles in charge of designing, tuning, and maintaining each component of the SBI process may vary from project to project, and so may vary the complexity of each design activity and the control the designer and the user have over it. Specifically, as claimed in [2], SBI projects can be classified into:

- *Best-of-Breed.* A best-of-breed policy is followed to acquire tools specialized in one of the parts of the SBI process. In this case, the designer has full control of the SBI process by finely tuning all its critical parameters.
- *End-to-End.* Here, an end-to-end software/service is acquired and tuned. Designers only need to carry out a limited set of tuning activities that are typically related to the subject area, while a service provider or a system integrator ensures the effectiveness of the technical phases of the SBI process.
- *Off-the-Shelf.* This type of projects consists in adopting, typically in a *as-a-service* manner, an off-the-shelf solution supporting a set of standard reports and dashboards. The designer has little or no chance of impacting on activities that are not directly related to the analysis of the final results.

Moving from level best-of-breed to off-the-shelf, projects require less technical capabilities from designers and users and ensure a shorter set-up time, but they also allow less control of the overall effectiveness and less flexibility.

3 A Case Study on EU Politics

The WebPolEU Project (http://webpoleu.altervista.org) aims at studying the connection between politics and social media. By analyzing digital literacy and online political participation, the research evaluates the inclusiveness, representativeness, and quality of online political discussion.

SBI is used in the project as an enabling technology for analyzing the UGC generated in Germany, Italy, and UK during a timespan ranging from March, 2014 to May, 2014 (the 2014 European Parliament Election was held on May 22–25, 2014). In the architecture we adopted, topics and related taxonomies are defined through Protégé; we use Brandwatch as a service for keyword-based crawling, Talend for ETL, SyN Semantic Center by SyNTHEMA for semantic enrichment (specifically, for labeling each clip with its sentiment), Oracle to store the ODS and the DM, MongoDB to store the document database for full-text search, and Mondrian as the multidimensional engine. Given the nature of the subject area, no EDW and no CRM are present in the architecture. We used the Indyco CASE tool to design the DM, and we developed an ad-hoc OLAP & dashboard interface using JavaScript, D3, and Saiku.

To enable topic-based aggregations of clips in the OLAP front-end, the classes in the domain ontology describing the subject area (that was designed together with the domain experts by classifying the topics emerged during macro-analysis) have been arranged into a *topic hierarchy* (see Fig. 2(a)). To effectively model the topic hierarchy, taking into account its specificities (it is heterogeneous, dynamic, non-onto, non-covering, and non-strict), the meta-star approach has been used [4].

4 Architectural Options

The techniques to be used to support the processes appearing in Fig. 1 may change depending on the context of each specific project, resulting in heavier or

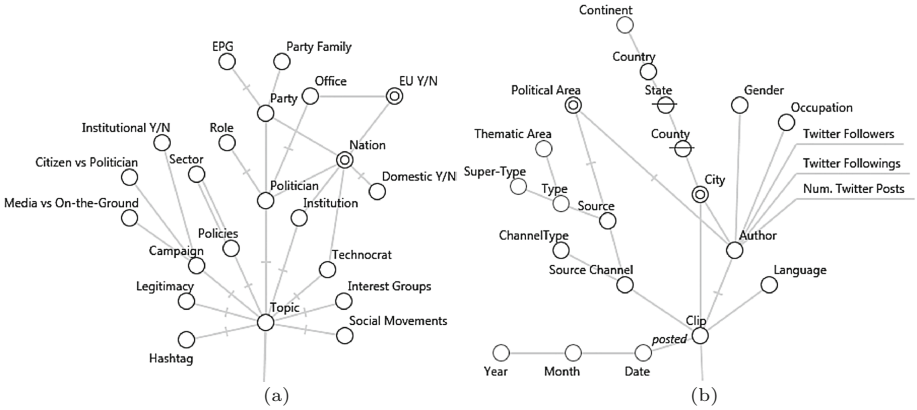

Fig. 2. A DFM representation of the topic (a) and clip (b) hierarchies for WebPolEU

lighter architectures. In the light of our experience with SBI projects of different types, in the following subsections we discuss the main options available to the design team, as well as their methodological impact.

4.1 Analysis

A component for analyzing the UGC is always present in SBI architectures, and it can take a variety of shapes characterized by quite different capabilities:

- **Dashboards** effectively summarize the trends and behaviors within the subject area, but only support a small number of predefined views and navigations (e.g., by topic or by geography).
- **Text search** enables very detailed analyses of the UGC up to the single-clip level, by supporting searches on both the clip text and its related meta-data.
- **OLAP** provides very flexible analyses based on the multidimensional metaphor, which enables users to understand in depth the market mood by slicing and drilling according to different dimensions such as time, topic, geography, UGC source, and the related hierarchies.
- **Text mining** enables advanced analyses on textual data such as clip clustering and new topic discovery [5].

Standard commercial SBI systems normally provide only dashboards and text search, and only a few of them support text mining (e.g., SAS Text Miner and Temis). Providing OLAP capabilities requires an additional layer of multidimensional data to be added to the architecture, as well as additional ETL processes that obviously increase the overall complexity. In the WebPolEU implementation, a set of cubes (see Fig. 3) are provided; noticeably, their schemata are largely project-independent, except for the topic hierarchy whose content and structure strictly depends on the domain ontology. Besides, to enable text search functionalities, the relational ODS is coupled with a document-oriented database.

4.2 ODS

In principle, the ODS component could even be dropped (in which case, the two ETL processes in Fig. 1 could be unified) since the users do not access it directly. However, the presence of the ODS—in compliance with three-tier data warehouse architectures—is warmly recommended in SBI for several reasons:

- **Buffering and early analysis.** Crawling and semantic enrichment activities have a very different timing due to the complexity of enrichment. The ODS can be seen as the *buffer* that makes the two phases independent of each other, so as to give users the possibility of timely accessing a subset of information that (i) enables some relevant early analyses; (ii) has a key methodological role for tuning the crawling and enrichment processes at the next iteration. Such information ranges from the clip meta-data returned by the crawler (e.g., source, author, and clip count) to some *quick-and-dirty* semantic enrichment.
- **Clip reprocessing.** Semantic enrichment is inherently an iterative process, due to changes in topics and in the domain ontology which may occur even months after the clips were retrieved. Storing clips in an ODS, where they can be easily queried at any time, makes reprocessing feasible.
- **Data cleaning.** It is well known that data cleaning techniques are more effective when applied to materialized data rather than when they are applied on-the-fly to a data flow. In the specific case of SBI, cleaning is necessary, for instance, to correct wrong character sequences, to repair enrichment/crawling errors which may produce wrong or incomplete results, and to filter off-topic clips based on relevance measures computed on both text and meta-data.

In our prototypical implementation, a relational ODS is used to store clips and their meta-data together with topics and their relationships. However, other alternatives could be explored. Choosing a NoSQL repository is mainly a matter of scalability, strictly related to the quantity of data to be stored and processed. In WebPolEU, about 10 millions of raw clips were retrieved and about 1.3 billions of entity occurrences were produced by semantic enrichment. Although this size is still manageable with traditional RDBMS, larger projects may make NoSQL solutions more attractive. In our experience, the main advantages of using an RDBMS are:

- The ODS plays the role of a hub for ETL data flows, and its tuples are subject to several updates to trace the process steps. This determines a transactional workload which is better handled if the ACID properties are preserved.
- The presence of a well-defined, structured, and normalized schema is very useful to process the clip meta-data.

4.3 Crawling

The crawling component is the main entry point to the SBI system for all the data that will be analyzed. From a technical point of view, the problem with crawling is to ensure that a satisfactory compromise is achieved between retrieving too much content (which adds harmful noise and leads to useless efforts

during semantic enrichment and analysis, as well as during all test activities) and retrieving too little content (which may dramatically reduce the reliability of analysis results). The two drivers that can be used to tune this compromise are *clipping* and *querying*.

Clipping is the process through which an indexed web page is parsed and every building section of the page itself is identified in order to exclude from the information extraction process all those contents that are not relevant and do not contain any useful information [19,20]. Bad clipping implies that the crawler will introduce into the system UGC filled with useless text such as hyperlinks, which will make the information almost incomprehensible for the semantic enrichment engine and often also for a human being—and also negatively affect the performance and quality of semantic enrichment activities.

Besides an accurate page clipping, the other ingredient for an effective crawling is a proper set of crawling queries. The standard way to identify relevant UGC from the web is by using Boolean keyword-based queries, where keywords considered as relevant or descriptive for the project scope are combined using different operators to instruct the crawler on the topics we are interested in and the ones that are out of scope. The operators typically provided by crawlers can be roughly classified into *Boolean* (e.g., AND, OR, NOT), *proximity* (e.g., NEAR/n), *meta* (e.g., country, site, author); wildcards are supported.

In the light of the above, it is apparent that managing and tuning the specific features of crawling to ensure its effectiveness is a burdensome and very time-consuming task. Noticeably, the roles in charge of these activity drastically depend on the project type as defined in Sect. 2: (i) in best-of-breed projects, all technical activities are in charge of the designer; (ii) in end-to-end projects, crawling templates are created and maintained by a service provider who is responsible of the clipping quality, but crawling queries are managed by the designer; (iii) in off-the-shelf projects, designers and users jointly carry out macro-analysis, but all other activities are largely in the hands of the service provider—which means that the designer can control the crawling effectiveness only to a limited extent [2]. So, from a project management point of view, the main trade-off involved in crawling is between (i) *do it yourself—but it will take a lot of time and effort* and (ii) *let the provider do it for you—but then you will have little control on the overall quality.*

4.4 Semantic Enrichment

The semantic enrichment process is maybe the one showing the widest spectrum of possible technological alternatives, with a very relevant impact on the expressiveness of the supported OLAP queries and on the accuracy of the results. Basic semantic enrichment techniques may be sufficient if users are only interested in analyzing raw data (e.g., counting the number of occurrences of each topic in the UGC); in some cases (for instance, for languages—like German—whose inherent complexity discourages automated analysis and interpretation of sentences), semantic enrichment is done by manually tagging each sentence with

its sentiment. In our WebPolEU project, semantic enrichment is achieved as the combination of different (and possibly alternative) techniques:

- **Crawler meta-data:** Each clip is equipped with several meta-data, which are mainly related to the web source (e.g., http address and web site nation), to the author (e.g., name, sex, and nationality), and to the clip itself (e.g., its language). As shown in Fig. 2(b), in WebPolEU these meta-data are used to build the clip hierarchy.
- **Information retrieval:** The content of the clips can be analyzed by searching the raw text for user-defined topics (or their aliases). Although this type of analysis is not based on an in-depth comprehension of clip semantics, it returns a quick and valuable first level of analysis of the texts. In particular it allows to count the number of occurrences of a given topic and the number of co-occurrences of a pair of topics in a clip. Figure 3(a, b) shows the IR Clip and IR Topic Occurrence cubes of the DM; each event of IR Clip represents a clip and its topics, while each event of IR Topic Occurrence represents the occurrence of a single topic within a clip.
- **Crawler sentiment:** The crawler often provides its own sentiment score. In WebPolEU we use Brandwatch, whose sentiment analysis module is based on mining rules developed for each supported language and assigns a single sentiment to each clip. In both the IR Clip and IR Topic Occurrence cubes, the crawler sentiment for each clip is modeled as a measure.
- **NLP analysis:** It is the deepest analysis raw texts undergo. As shown in Sect. 2, the commercial system SyN Semantic Center is in charge of extracting the single entities, their part-of-speech, and their semantic relationships from the raw data. Two cubes are derived through NLP analysis. The first one, NLP Entity Occurrence (Fig. 3(c)), differs from IR Topic Occurrence since it also stores all the entities (i.e., lemmas, annotated with their part-of-speech) discovered in the text. The second one, NLP Semantic CoOccurrence (Fig. 3(d)), stores semantic relationships and explicitly models couples of topics/entities in the same sentence together with an optional qualifier (e.g., Angela Merkel had lunch with Matteo Renzi).
- **Domain expert:** differently from social media monitoring solutions, SBI projects allow additional meta-data to be provided by domain experts by means of the domain ontology coded in the topic hierarchy (see Fig. 2(a)) and by additional meta-data to be added to the other hierarchies.

5 Case Study Analysis

Carrying out an SBI project requires to find the right trade-off between its effectiveness, efficiency, and sustainability, respectively expressed in terms of correctness of the results obtained, appropriateness of the response time, and time/money required to run the project. In this section we provide a quantitative evaluation of these aspects with reference to our case study.

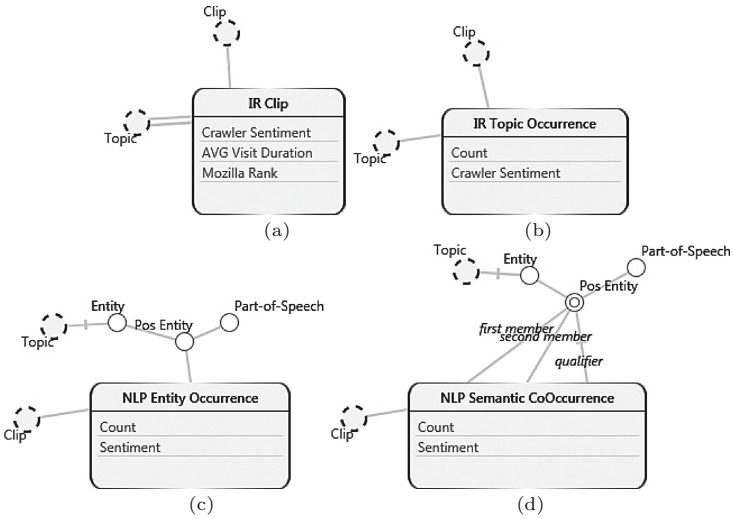

Fig. 3. A DFM representation of IR and NLP cubes. Topic and clip hierarchies have been hidden to simplify the picture

Overall, the number of collected clips in WebPolEU was around ten millions, which decreased to six millions after dropping non-relevant sources and duplicate clips. Noticeably, the quantity of information generated by the semantic enrichment process is much larger ($|\mathsf{NLP\ Entity\ Occurrence}| \approx 500\,\mathrm{M}$ for each language) and places the project on the edge of big data. The topics were provided by the team of socio-political researchers involved in WebPolEU; the number of topics is about the same (around 500) for Germany, Italy, and UK, since the same issues were discussed in the three nations. Although the number of clips collected for Germany (933 K) is quite lower than that for Italy and UK (about 3 M each), the number of occurrences generated is not so different; this is because the lower number of clips for Germany is counterbalanced by their greater average length.

5.1 Effectiveness

Our first goal is to evaluate different semantic enrichment techniques in terms of the trade-off they offer between added value on the one side, and resource demand/effort on the other. In particular, we will compare the approach based on crawler meta-data, crawler sentiment, and information retrieval (called IR in the following) against the approach based on NLP analysis (called NLP). We will focus on the Italian and English clips since they were both enriched using the same tools (Brandwatch for IR and SyN Semantic Center for NLP). As shown in Table 1(a), the two techniques find the same topic occurrences in a clip in most cases. This shows that the KPIs based on topic counting, which are widely adopted for UGC analysis, does not necessarily require the adoption of sophisticated ontology-based techniques and a full comprehension of sentence

syntax and semantic. Conversely, these techniques are required when analyzing semantic co-occurrences is one of the users' goals.

Table 1. Number of topic occurrences detected by IR and NLP (a) and number of positive, neutral, and negative clips detected by NLP, by IR, and agreed upon by NLP and IR (b)

	ITA	ENG
# Topic Occ. NLP	14 215 K	23 399 K
# Topic Occ. IR	15 401 K	25 006 K
# Shared Occ.	12 922 K	21 497 K

(a)

	ITA			ENG		
Sentiment	NLP	IR	Agreed	NLP	IR	Agreed
Positive	566 K	36 K	19 K	1090 K	142 K	107 K
Neutral	893 K	2340 K	888 K	1368 K	2973 K	1337 K
Negative	934 K	17 K	14 K	817 K	159 K	112 K

(b)

The real power of NLP comes into play when analyzing sentiment. Table 1(b) shows that Brandwatch, which adopts a rule-based technique for sentiment analysis, hardly assigns a non-neutral sentiment to a clip: most of the clips that Brandwatch labels as positive/negative are positive/negative for SyN too, while the two systems often disagree on neutral clips.

There is not much point in discussing the differences in IR and NLP sentiment without knowing which is the correct one. For this reason we evaluated the accuracy of the returned sentiment by asking five domain experts to manually tag a sample of the clips. The sample includes about 600 clips from the English corpus, equally divided by media type and NLP sentiment (as computed by Syn). Besides defining the clip sentiment as either negative, neutral, or positive, the domain experts were also asked to rate, for each clip, its *clipping quality* (i.e., the amount of non-relevant text present in the clip), which could impact on the difficulty of assigning the right sentiment, and its intrinsic *text complexity* (i.e., the effort of a human expert in assigning the sentiment due to irony, incorrect syntax, abbreviations, etc.). Table 2 shows the IR and NLP sentiment accuracy (i.e., percentage agreement with the consensus sentiment) for each sub-sample; a correct interpretation of the results requires some further explanation due to the different cardinalities of the sub-samples. It is apparent that the experts rated most of the clips as neutral—thus, a dummy classifier always stating *neutral* would most probably be very successful! Before commenting the tables, we recall that the lower bound on accuracy is 33 %, which is the percentage of success of a random classifier.

- The high accuracy achieved by IR on neutral clips is not actually due to its real capability of discerning between negative, neutral and positive clips, but rather to its inability/caution in assigning a non-neutral sentiment. Indeed, its accuracy on negative and positive clips is below that of a dummy classifier.
- When using NLP, detecting positive sentiments turns out to be much easier than identifying negative ones. This happens because positive opinions are normally characterized by enthusiastic words, while negative ones are often blurred by irony, which can hardly be detected. This is confirmed by the experts, that mostly label positive clips as having standard complexity.

– For clips whose texts complexity has been classified as hard, both IR and NLP often fail in assigning the right sentiment.
– The clipping quality impacts more on NLP than on IR accuracy. It would be interesting to investigate if this is related to the deeper level of text understanding NLP tries to achieve.

Table 2. IR and NLP sentiment accuracy for each sub-sample

Clipping Quality	Text Complexity	Negative		Neutral		Positive	
		IR	NLP	IR	NLP	IR	NLP
High	Standard	16.7%	62.7%	85.1%	39.9%	21.7%	68.3%
	Hard	15.2%	36.4%	100.0%	44.4%	0.0%	100.0%
	Overall	15.9%	49.5%	92.5%	42.2%	10.8%	84.2%
Low	Standard	20.0%	55.0%	87.8%	54.9%	28.6%	57.1%
	Hard	0.0%	0.0%	100.0%	0.0%	–	–
	Overall	10.0%	27.5%	93.9%	27.4%	28.6%	57.1%

Text Complexity	Negative		Neutral		Positive		IR	NLP
	IR	NLP	IR	NLP	IR	NLP		
Standard	18.3%	58.8%	86.4%	47.4%	25.1%	62.7%	43.3%	56.3%
Hard	7.6%	18.2%	100.0%	22.2%	0.0%	100.0%	43.0%	36.2%
Overall	13.0%	38.5%	93.2%	34.8%	16.7%	75.2%	43.2%	47.2%

As to analysis, the last phase of the SBI process, we can only give some qualitative assessment. Moving from standard dashboards to user-driven OLAP analysis has been recognized as truly valuable by the WebPolEU users since it enables them to flexibly and autonomously navigate data to get a deeper insight on the ongoing trends, leaning on hierarchies to better analyze data.

5.2 Efficiency

We start this section by mentioning how the architecture in Fig. 1 has been implemented in the WebPolEU project. ETL and analysis run on an 8-cores server with 64 GB of RAM; the text search engine runs on a 7-nodes cluster (each node equipped with a 4-cores processor and 32 GB of RAM); the semantic enrichment component runs on a 6-nodes virtual cluster (each node equipped with a 12-cores processor and 10 GB of RAM). As to the data volume, the raw crawler files take 79 GB, the ODS 481 GB, the DM 116 GB, and the documents for text search 65 GB. Noticeably, since the OLAP cubes in the DM mainly store numerical data, their required storage is lower than that of the ODS.

Table 3 shows the time required for running the main ETL flows with reference to all clips (a 20 × parallelization was adopted to maximize the throughput) and the time for the bi-directional ETL flow between the ODS and NLP semantic enrichment as a function of the clip length (here times were measured on a single-process basis). These results confirm that NLP semantic enrichment deeply impacts on the time and space required to feed the DM, so its adoption

Table 3. Average processing time in seconds for 10 000 clips; to the right, average time for NLP semantic enrichment of one clip

ETL Flow	Time per 10 K Clips
Crawling → ODS	2868
ODS ↔ IR Sem. Enrich.	180
ODS ↔ NLP Sem. Enrich.	23 035
ODS → DM (IR)	13
ODS → DM (NLP)	68
ODS → Document-Base	16

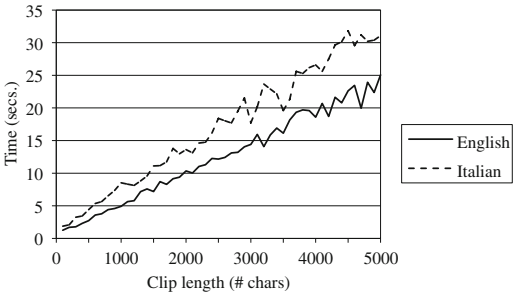

Table 4. Execution time for chart, OLAP and free-text queries

Query type	Exec. time (s)			Query example
	Min	Avg	Max	
IR charts	1.2	7.4	25.5	Daily trend of UK topic occurrences for each channel type and party
NLP charts	0.8	62.2	288.7	Top 5 entities related to the "Cameron" topic
IR OLAP	0.3	7.7	50.1	Average crawler sentiment for each party and country
NLP OLAP	0.4	14.7	79.4	Average sentiment for each topic sector and clip type
Free-text	0.2	1.1	2.9	"Europe" AND "Politics" (filter on Clip.Source = "telegraph.co.uk")

should be carefully evaluated. Interestingly, both processing time and data size are higher for Italian clips due to the greater complexity of the Italian language.

We close our efficiency analysis by showing, in Table 4, the execution time for an analysis workload including 33 queries, which can be classified into three groups corresponding to the main functions of a typical SBI platform: charts, OLAP analysis, and free-text search. The first group includes the queries whose output is used to draw the charts available in the WebPolEU interface (e.g., tag cloud, trends, etc.), while the other two groups were created by auditing and sampling the queries actually issued by WebPolEU users. Although the average query time is higher for NLP queries (because the corresponding cubes have higher cardinalities), all the groups are compatible with interactive analyses.

5.3 Sustainability

The first design iteration for WebPolEU took 84 person-days overall; of these, 18 were for designing the domain ontology (including topic definition), 21 for designing and testing semantic enrichment (in particular for tuning the dictionary), and 26 for designing and testing crawling queries. The second iteration was mostly used for tuning the ETL (20 person-days out of 30). The main critical issues related to each activity are listed below:

– Ontology design: the correctness of the results is deeply affected by the number of topics and aliases defined. For example, with reference to Fig. 2, the

number of occurrences for each topic sector depends on the topics and aliases summarizing that sector, hence, including an unbalanced number of topics for the different sectors may lead to an unfair analysis. Keeping a proper level of detail for different sectors requires a deep knowledge of the domain and related vocabulary.
- Crawling design: commercial solutions (like Brandwatch) normally limit the length of the crawling queries; this makes it harder to properly define the subject area, which is necessary to filter off-topic clips. Finding the proper formulation of queries with constraints on their length and number may become a real nightmare.
- ETL & OLAP design: although parsing a JSON file is a trivial task, handling all the possible unexpected character sequences is more tricky and requires continuous tuning along the whole project. On the other hand, unexpected character sequences often determine a failure of semantic enrichment.

6 Related Literature and Conclusion

As stated in the Introduction, only a few papers have focused on the full picture of SBI so far. Complete architectures for SBI have been proposed in [6,13]; in both cases, the basic blocks of the architecture have been identified, but still with a limited expressiveness. In particular, in [13] a comprehensive solution for the extraction of Twitter streams and the enhancement and analysis of their metadata is presented; the approach of [6] extracts sentiment data about products and their features from selected opinion websites and builds *opinion facts*. An important step towards increasing the expressiveness of SBI queries has been taken in [1], where a first advanced solution for modeling topic hierarchies has been proposed. Another step in this direction has been made in [4], where topic hierarchies are modeled by handling their dynamics and irregularity so as to enable full OLAP analyses of social data. In terms of OLAP analysis over UGC, a cube for analyzing term occurrences in documents belonging to a corpus is proposed in [9], although term categorization is very simple and does not support analyses at different levels of abstraction. In [12] the authors propose to use textual measures to summarize textual information within a cube.

As to the enabling technologies for the SBI process, a number of academic works have focused on specific issues that find application on strictly correlated fields. First of all, web crawling is a central issue in information retrieval, in whose context powerful languages to automatically and precisely capture the relevant data to be extracted were studied (e.g., [3]). In terms of semantic enrichment of raw clips and text understanding, different techniques have been studied in several areas of computer science. Whereas most of these techniques are typically tuned to perform well on a limited set of selected (web) sources, their accuracy tends to decrease when applied to a heterogeneous collection of documents extracted from multiple kinds of sources. In general, NLP approaches try to obtain a full text understanding [18], while text mining approaches rely on different techniques (e.g., n-grams) either to find interesting patterns in texts

Table 5. Summary of main architectural options

Component	Option	Pros	Cons
Analysis	Dashboard	Effective summary of trends	Low flexibility
	Text search	Detailed content analyses	Increased storage
	OLAP	High flexibility	Increased storage; extra ETL
ODS	Text mining	Enables advanced analyses	Complexity; expert analyst required
	Relational	Clip buffering, reprocessing, and cleaning; structured	Increased storage; performances
Crawling	NoSQL	Clip buffering, reprocessing, and cleaning; scalability	Low control of data transformation and quality
	Designer-managed	Good control of quality	Large effort
Sem. Enr.	Provider-managed	Small effort	Low control of quality
	Crawler meta-data	Enables clip classification and aggregation	Some complexity in collecting
	Crawler sentiment	Enables analysis of sentiment; no tuning	Unreliable for non-neutral clips
	Inf. retrieval	Enables topic occurrence analysis	Low text understanding
	NLP analysis	Enables analysis of sentiment; also reliable for non-neutral clips	Complex tuning; affected by clipping quality
	Domain expert	Enables analysis of sentiment; fully reliable	Costly; subjective

(e.g., named entities [14], relationships between topics [15], or clip sentiment [11]) or to classify/cluster them [17]. Also hybrid approaches between classical NLP and statistical techniques have been tried, either user-guided, as in [8], or automated and unsupervised, as in [6].

In this paper we have analyzed the main factors that impact on the costs and benefits of the main architectural options for SBI. A summary of the pros and cons of the different options, as emerging from our case study, is shown in Table 5. Remarkably, it turned out that crawling and semantic enrichment are the components that impact the most on the overall cost-benefit compromise. Here we summarize a few rules of thumb for making a good choice:

- The accuracy of both NLP and IR sentiment can be high on very specific sources and closed domains (such as the CRM of a bank or the movie reviews [7]), but it easily drops as soon as the domain becomes wider. Since a relevant effort is required to properly handle sentiment, the design team should carefully evaluate the use of sentiment analysis techniques by trading-off the accuracy achievable with the related costs.
- Although Twitter provides a partial analysis of the social environment, the shortness of tweets and the high percentage of non-neutral clips make it a good candidate to be the main source for an effective sentiment analysis. Indeed, experimental data show that Twitter clips yield the highest accuracy for NLP sentiment (56.6 %, vs. 51.5 % of forums and 42.4 % of news).
- Dashboards are the standard way for visualizing and analyzing data in SBI projects since they yield an immediate, easy-to-understand, and well-focused representation of results. However, as the role of SBI systems becomes more

important in companies, full-OLAP capabilities will increasingly be provided because they clearly enable more flexible and accurate analyses of the UGC.
– Off-the-shelf projects provide *quick-and-dirty* answers but preclude the possibility of carrying out in-depth analysis, tuning, reprocessing, and integration with enterprise data. They should be pursued either at an early stage of adoption of SBI solutions to assess the real value of social data for the company, or if the available resources are very limited.

References

1. Dayal, U., Gupta, C., Castellanos, M., Wang, S., Garcia-Solaco, M.: Of cubes, DAGs and hierarchical correlations: a novel conceptual model for analyzing social media data. In: Atzeni, P., Cheung, D., Ram, S. (eds.) ER 2012 Main Conference 2012. LNCS, vol. 7532, pp. 30–49. Springer, Heidelberg (2012)
2. Francia, M., Golfarelli, M., Rizzi, S.: A methodology for social BI. In: Proceedings of IDEAS, Porto, Portugal, pp. 207–216 (2014)
3. Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.J.: OXPath: a language for scalable data extraction, automation, and crawling on the deep web. VLDB J. **22**(1), 47–72 (2013)
4. Gallinucci, E., Golfarelli, M., Rizzi, S.: Advanced topic modeling for social business intelligence. Inf. Syst. **53**, 87–106 (2015)
5. Gao, W., Li, P., Darwish, K.: Joint topic modeling for event summarization across news and social media streams. In: Proceedings of CIKM, Maui, HI, pp. 1173–1182 (2012)
6. García-Moya, L., Kudama, S., Aramburu, M.J., Llavori, R.B.: Storing and analysing voice of the market data in the corporate data warehouse. Inf. Syst. Front. **15**(3), 331–349 (2013)
7. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of SIGKDD, Seattle, WA, pp. 168–177 (2004)
8. Kahan, J., Koivunen, M.R.: Annotea: an open RDF infrastructure for shared web annotations. In: Proceedings of WWW, Hong Kong, China, pp. 623–632 (2001)
9. Lee, J., Grossman, D.A., Frieder, O., McCabe, M.C.: Integrating structured data and text: a multi-dimensional approach. In: Proceedings of ITCC, Las Vegas, USA, pp. 264–271 (2000)
10. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer, New York (2012)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. CoRR cs.CL/0205070 (2002)
12. Ravat, F., Teste, O., Tournier, R., Zurfluh, G.: Top_Keyword: an aggregation function for textual document OLAP. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 55–64. Springer, Heidelberg (2008)
13. Rehman, N.U., Mansmann, S., Weiler, A., Scholl, M.H.: Building a data warehouse for Twitter stream exploration. In: Proceedings of ASONAM, Istanbul, Turkey, pp. 1341–1348 (2012)
14. Ritter, A., Clark, S., Mausam, E., O.: Named entity recognition in tweets: an experimental study. In: Proceedings of EMNLP, Edinburgh, UK, pp. 1524–1534 (2011)
15. Rosenfeld, B., Feldman, R.: Clustering for unsupervised relation identification. In: Proceedings of CIKM, Lisbon, Portugal, pp. 411–418 (2007)

16. Stavrakantonakis, I., Gagiu, A.E., Kasper, H., Toma, I., Thalhammer, A.: An approach for evaluation of social media monitoring tools. Common Value Manag. **52**, 52–64 (2012)
17. Wang, X., McCallum, A., Wei, X.: Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: Proceedings of ICDM, Washington, DC, USA, pp. 697–702 (2007)
18. Yi, J., Nasukawa, T., Bunescu, R.C., Niblack, W.: Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of ICDM, Melbourne, Florida, pp. 427–434 (2003)
19. Yi, L., Liu, B.: Web page cleaning for web mining through feature weighting. In: Proceedings of IJCAI, Acapulco, Mexico, vol. 3, pp. 43–50 (2003)
20. Yi, L., Liu, B., Li, X.: Eliminating noisy information in web pages for data mining. In: Proceedings of KDD, Washington DC, USA, pp. 296–305 (2003)