SABINE: A Multi-purpose Dataset of Semantically-Annotated Social Content

Silvana Castano[1], Alfio Ferrara[1], Enrico Gallinucci[2], Matteo Golfarelli[2](✉), Stefano Montanelli[1], Lorenzo Mosca[3], Stefano Rizzi[2], and Cristian Vaccari[4,5]

[1] DI, University of Milan, Milan, Italy
{silvana.castano,alfio.ferrara,stefano.montanelli}@unimi.it
[2] DISI, University of Bologna, Bologna, Italy
{enrico.gallinucci,matteo.golfarelli,stefano.rizzi}@unibo.it
[3] DFCS, University of Roma Tre, Rome, Italy
lorenzo.mosca@sns.it
[4] Royal Holloway University of London, Egham, UK
cristian.vaccari@rhul.ac.uk
[5] University of Bologna, Bologna, Italy

Abstract. Social Business Intelligence (SBI) is the discipline that combines corporate data with social content to let decision makers analyze the trends perceived from the environment. SBI poses research challenges in several areas, such as IR, data mining, and NLP; unfortunately, SBI research is often restrained by the lack of publicly-available, real-world data for experimenting approaches, and by the difficulties in determining a ground truth. To fill this gap we present SABINE, a modular dataset in the domain of European politics. SABINE includes 6 millions bilingual clips crawled from 50 000 web sources, each associated with metadata and sentiment scores; an ontology with 400 topics, their occurrences in the clips, and their mapping to DBpedia; two multidimensional cubes for analyzing and aggregating sentiment and semantic occurrences. We also propose a set of research challenges that can be addressed using SABINE; remarkably, the presence of an expert-validated ground truth ensures the possibility of testing approaches to the whole SBI process as well as to each single task.

Keywords: Dataset · Social technologies · Sentiment analysis
Text analysis

1 Introduction

During the last decade, an enormous amount of *user-generated content* (UGC) related to people's tastes, opinions, and actions has been made available due

Supported by the Italian Ministry of Education "Future in Research 2012" initiative for the project "Building Inclusive Societies and a Global Europe Online: Political Information and Participation on Social Media in Comparative Perspective" (www.webpoleu.net).

D. Vrandečić et al. (Eds.): ISWC 2018, LNCS 11137, pp. 70–85, 2018.
https://doi.org/10.1007/978-3-030-00668-6_5

to the omnipresent diffusion of social networks and portable devices. This huge wealth of information has raised an intense 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 analyze and improve their business based on the trends and moods perceived from the environment [9].

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 decision makers out of textual UGC requires first crawling the web to extract the clips related to a *subject area* (e.g., politics), then enriching them in order to let as much information as possible emerge from the raw text. 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* (i.e., positive, negative, or neutral) to it [12]. For instance, the tweet "UKIP's Essex county councillors stage protest against flying of EU flag", in the subject area of EU politics, mentions topics "UKIP" and "protest" and has positive sentiment. Figure 1 sketches the overall SBI process.

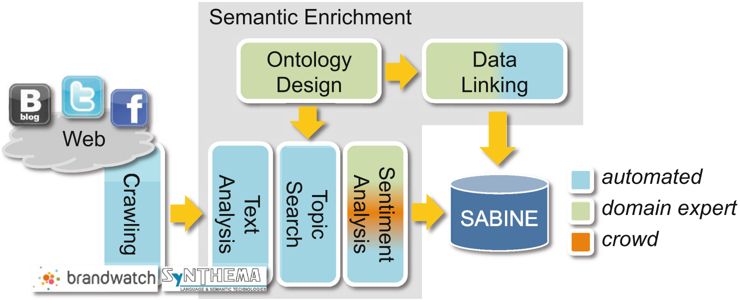

Fig. 1. The functional architecture of the SBI process which created SABINE

From a scientific point of view, SBI stands at the crossroads of several areas of Computer Science such as Database Systems, Information Retrieval, Data Mining, Natural Language Processing, and Human-Computer Interaction. Though the ongoing research in these single fields has made available a bunch of results and enabling technologies for SBI, an overall view of the related problems and solutions is still missing. Besides, the peculiarities of SBI systems open new research problems in all the previous areas. On the other hand, research developments in SBI are often restrained by the lack of publicly-available, real-world data for experimenting approaches, and by the inherent difficulties in determining a ground truth for assessing the effectiveness of an approach.

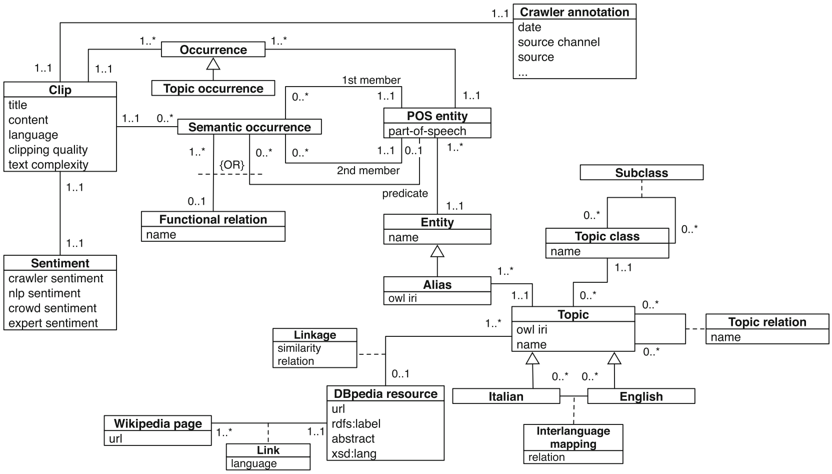

Fig. 2. UML model of SABINE

To fill this gap, in this paper we present SABINE (SociAl Business INtelligence datasEt), a modular benchmark in the domain of European politics with specific reference to the 2014 European elections. SABINE includes: 6 millions bilingual clips crawled from 50 000 web sources, each one associated with metadata and sentiment scores; an ontology with 400 topics, their occurrences in the clips, and their mapping to DBpedia; and two multidimensional cubes for analyzing and aggregating sentiment and semantic occurrences for SBI analytics purposes. Remarkably, the presence of a manually-validated ground truth for each phase of the SBI process ensures the possibility of testing approaches to the whole process as well as to each single task. In this direction, our proposal is complemented by a set of research challenges that can be addressed using SABINE; the task selection we propose is large and diverse enough to be sufficiently representative of a wide range of research tasks, ranging from content analysis to the more comprehensive SBI analytics.

The paper outline is as follows. In Sect. 3 we describe the benchmark content. In Sect. 4 we discuss the techniques adopted for building SABINE. In Sect. 5 we propose a set of SBI-related research tasks for SABINE. Finally, in Sect. 6 we draw the conclusions.

2 Related Literature

As remarked throughout the paper, the research challenges supported by SABINE span several research areas. Table 1 presents a (non-comprehensive) list of datasets available in such areas, emphasizing the novelty of SABINE as a multi-purpose dataset.

A first set of datasets comes from the information retrieval area. Probably the most popular testbed and dataset series in information retrieval was

Table 1. Functional comparison of datasets

Benchmark	Sentiment analysis	Topic search	Document classif.	Cross-language analysis	Topic discovery	Data linking	Multidim. modeling	SBI analytics
Reuters [2]	✗	✗	✓	✓	✓	✗	✗	✗
20 Newsgroups	✗	✗	✓	✗	✗	✗	✗	✗
TREC	✓	✓	✓	✓	✓	✗	✗	✗
CORA	✗	✓	✓	✗	✓	(✓)	✗	✗
CustomerReview [11]	✓	✓	✗	✗	✗	✗	✗	✗
MovieReview [17]	✓	✗	✗	✗	✗	✗	✗	✗
KDD Cup	✗	✗	✓	✓	✗	✓	✗	✗
OAEI [1]	✗	✗	✗	✓	✗	✓	✗	✗
SemEval [14]	✓	✗	✗	✓	✗	✗	✗	✗
SABINE	✓	✓	✓	✓	✓	✓	✓	✓

promoted by the Text Retrieval Conference (TREC, http://trec.nist.gov/data.html) since 1992. The datasets therein contained have been used for a variety of tasks, spanning from sentiment analysis to topic search and discovery. Similar datasets include the Reuters [2] collection (about 112,000 documents in five different languages associated with one or more thematic categories), 20 Newsgroup (http://qwone.com/~jason/20Newsgroups/), which contains 20,000 newsgroup documents, partitioned across 20 different categories, including politics and religion, and CORA (https://people.cs.umass.edu/~mccallum/data.html), a collection of different datasets supporting topic search and discovery, document classification, and information extraction.

In the fields of sentiment analysis and opinion mining, the publication of novel algorithms and techniques is often coupled with the release of datasets. CustomerReview [11] provides 4000 product reviews from Amazon and C—net.com, manually labeled as to whether an opinion is expressed and comprising the rating of the reviewer. These reviews are also grouped according to the mentioned products—whose occurrence is detected by using data mining and natural language processing techniques. Similarly, MovieReview [17] provides four datasets based on movie reviews, comprising a total of 7,000 documents and 20,000 sentences. Annotations to these reviews include the sentiment polarity (positive or negative) obtained from the original website, the rating of the user (e.g., 2.5 out of 5), and a subjectivity status (i.e., "subjective" or "objective" if it is a plot summary or a review).

Besides these datasets, we mention also international competitions that additionally provide a set of evaluation metrics and a methodology for comparing the systems that participate in the contest. KDD Cup (http://www.kdd.org/kdd-cup) is the Data Mining and Knowledge Discovery competition organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining. KDD Cup is mainly oriented towards data mining and prediction, but some editions provided data for tasks related to document classification and data linking. The Ontology Alignment Evaluation Initiative (OAEI http://oaei.ontologymatching.org/) is the main reference for ontology matching. It is specifically devoted to semistructured data linking (in particular to ontology alignment and instance matching) and supports several tasks spanning from cross-language ontology matching to instance matching. In 2016, SABINE has been used as the reference dataset for the task of inter-linguistic mapping and data linking within the OAEI Instance Matching Track [1]. Finally, the International Workshop on Semantic Evaluation (SemEval) proposes different tasks to evaluate computational semantic analysis systems, ranging from cross-lingual similarity to humor and truth detection. In particular, the competition on sentiment analysis proposed in [14] was based on a set of about 30,000 tweets, whose polarity had been manually assigned by a consensus of users via the crowdsourcing platforms Amazon's Mechanical Turk and CrowdFlower.

SABINE Contribution. With respect to these datasets, the main contribution of SABINE is the coverage of a wide range of different but related tasks and the homogeneity of the evaluation environment. The growing interest for data-intensive information systems and the coexistence of unstructured, semistructured, and structured data (not only on the web but also in the enterprise context) motivates the need for an integrated evaluation environment, where applications for search, linking, classification, analysis, and multidimensional modeling of data may be tested over the same data following a homogeneous methodological approach. This is not feasible with any of the datasets of Table 1, in that each of them provides different sets of data, different ground truth collections, and/or different evaluation methodologies. Conversely, all the research tasks supported by SABINE are built from the same collection of data by following the methodology discussed in Sects. 4 and 5 to the end of providing a comprehensive and homogeneous dataset.

3 The Content of SABINE

SABINE has been built as one of the results of the WebPolEU project (webpoleu.altervista.org), whose goal was to investigate the connection between politics and social media. SBI was used in the project as an enabling technology for analyzing the UGC generated in Italian and English during a timespan ranging from March, 2014 to May, 2014 (the 2014 European Parliament Election was held on May 22–25, 2014). By analyzing digital literacy and online political participation, the research evaluated the inclusiveness, representativeness, and quality of the online political discussion. The UML model of the SABINE content (except for the multidimensional part, whose content is described by Fig. 4) is shown in Fig. 2, while its quantitative features are summarized in Table 2 (see [7] for a more detailed profiling of the clips). The main content components of SABINE can be described as follows.

3.1 Topics and Mappings

SABINE provides about 400 relevant topics organized in a topic ontology built by domain experts (a team of five socio-political researchers). The topic ontology

Table 2. SABINE figures

Figure	ENG	ITA	*Figure*	ENG	ITA	*Figure*	ENG	ITA
# Web Sources	23K	25K	# Entities	2868K	1242K	# Entity Occurrences	511M	218M
# Topics	409	434	# Clips	3275K	2394K	# Topic Occurrences	23M	14M
# Topics Aliases	709	798	Avg Chars per Clip	2026	1677	# Semantic Occurrences	48M	35M

(modeled by classes Topic, Topic class, Subclass, and Topic relation in Fig. 2) represents the set of concepts and relationships that, on the domain experts' judgement, are relevant to the subject area; its role in the SBI process is twofold: to act as a starting point for designing effective crawling queries on the one hand, and to support analyses based on relevant concepts (e.g., how often the public debt policy is mentioned) and on their aggregations (e.g., how often the sector of economics and its policies are discussed) on the other. The class diagram for the topic ontology of the socio-political subject area of SABINE benchmark is shown in Fig. 3; for instance, topics "public debt" and "austerity " are instances of topic class Policy and are related to topic "economic policy" (Sector). To enable more accurate analyses, a large set of topic aliases (class Alias in Fig. 2) has been identified and is available for topics (e.g., "tory" is an alias for "conservative").

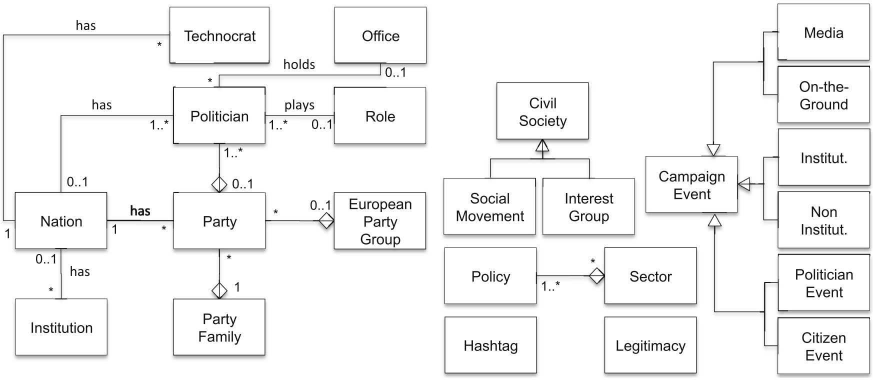

Fig. 3. The topic ontology represented as a UML class diagram

Inter-Language mappings (class Interlanguage mapping) between corresponding topics in the two benchmark languages have been manually created by the domain experts. In most cases these mappings simply express an exact translation (e.g., "immigration" is mapped onto "immigrazione" with semantics owl:sameAs), whereas they are based on weaker semantic relationships when a concept is differently expressed in the two languages (e.g., "immigration" is mapped onto "migrante irregolare" —which means illegal migrant— with semantics sabine:related). A mapping has been found only in 86% of cases since, according to the experts' judgement, some topics are specific of either UK or

Italy (e.g., "Scottish National Party" and "Quirinale"). Furthermore, topics have been linked to their corresponding DBpedia resources (classes DBpedia resource, Wikipedia page, Link, and Linkage). Linkage has been carried out automatically as described in Sect. 4.6 and then validated by domain experts.

3.2 Clips and Annotations

The benchmark provides a large corpus (around 6 millions) of raw clips (class Clip) extracted by the Brandwatch crawler from a broad set (almost 50 000) of web sources including social networks, blogs, and web sites. The most frequent clip sources are Facebook (53.8% of the clips) and Twitter (27.5% of the clips). The corpus is bilingual and *comparable*, i.e., it includes text in two languages (English and Italian) regarding similar topics [6,16]. Each clip is associated with a set of metadata (class Crawler annotation); 40 attributes overall are provided, partly returned by the crawler (e.g., title, date, source MozRank, author information, and geo-localization) and partly manually annotated by the domain experts (e.g., source type).

Clips are enriched with other relevant information resulting from clip text analysis. In particular, each clip is associated with all its occurring entities and their parts-of-speech or POSs (classes POS entity, Entity, and Occurrence). An *entity* is a concept emerging from text analysis, which is not necessarily a topic; parts-of-speech (POSs) are the roles taken by entities within a clip sentence (e.g., *noun*, *verb*, *preposition*). Among the set of entity occurrences, a relevant role is taken by the occurrences of topics and their aliases (class Topic occurrence). Finally, text analysis also led to the detection of more complex linguistic patterns involving multiple entities in the same sentence (classes Semantic occurrence and Functional relation). In particular, each semantic occurrence relates two entities by means of either a functional relation (e.g., *agent* or *qualifier*) or a predicate corresponding to an entity.

All clips are also annotated with two sentiment values (class Sentiment). The first one (*crawler sentiment*) is categorical (i.e., negative, neutral, positive); it has been determined for each clip by the Brandwatch crawler through rule-based techniques. The second one (*NLP sentiment*) is numerical; it has been determined by the SyN semantic engine for each clip sentence, then averaged for each clip (see Sect. 4.5). Finally, a subset of 2400 clips have also been labeled with a *crowd sentiment*, and half of these 2400 clips have further been labeled by domain experts (*expert sentiment*). This subset of manually-labelled clips has been created using a stratified sampling strategy based on the type of clip source (e.g., social network and blog) and on the clip sentiment.

Example 1. Here is an example of a SABINE clip: "Another compassionate conservative. Making fun of a parkinson's victim. Michael J Fox has more courage than you will ever hope to have". Some metadata for this clip are source="facebook.com", channel_type="facebook", source_type="Social network/Social media", country="US", and fb_role="audience". The only occurring topic is “conservative”; among the occurring entities we mention “compassionate”, “victim”, and “courage” (with POSs *adjective*, *noun*, and *noun* respectively). Text analysis led to find different semantic occurrences of entities with their POSs, for instance the one between “compassionate” and “conservative” (with POS *adjective* and *noun*, respectively, and functional relation *qualifier*) and the one between “have” and “courage” (with POS *verb* and *noun*, respectively, and functional relation *object*). The expert sentiment for this clip is −1 (i.e., negative), while both the crawler and the NLP sentiment are positive (because neither of the approaches was able to detect the irony in the sentence “Another compassionate conservative”). Another example of clip is “US President Barack Obama criticized Russia [...]”, which shows a semantic occurrence between entities “Barack Obama” (POS *proper noun*) and “Russia” (POS *proper noun*) involving entity “criticize” as a predicate.

3.3 Multidimensional Cubes

These are ROLAP cubes providing an easy-to-query representation of the clip content and of the outcome of the clip enrichment process. The first cube, Sentiment, is centered on clips, and it represents the set of topics appearing in each clip as well as the sentiment values computed for that clip. The second cube, Semantic Occurrence, is centered on the semantic occurrence of POS entities within clips and explicitly models couples of entities in the same sentence together with an optional predicate. The conceptual schemas of the Sentiment cube is depicted in Fig. 4 using the DFM notation [10], where cube measures are listed inside the box, dimensions are circles directly attached to the box, and hierarchies are

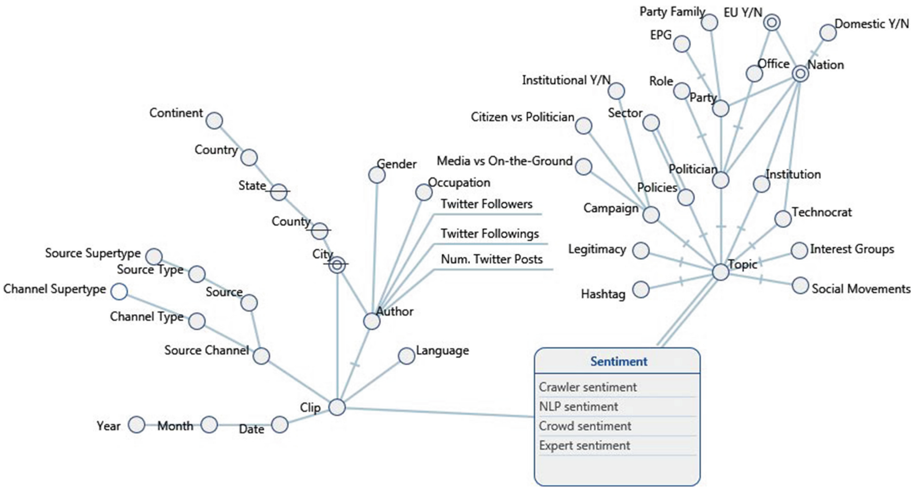

Fig. 4. A DFM representation of the Sentiment cube (for drawing simplicity, some levels of the topic hierarchy are hidden)

shown as DAGs of dimension levels. In particular, the hierarchy built on dimension Clip includes the crawler annotations, while the one built on Topic (called *topic hierarchy* from now on) derives from the topic ontology of Fig. 3 and enables topic-based aggregations of clips in the OLAP front-end. For instance, a roll-up from Politician to Party and Party Family allows to obtain the opinions about a wing as an average of the opinions about all the politicians belonging to the parties of that wing.

4 SABINE Construction Techniques

To develop SABINE we followed the methodology described in [8], which has been conceived on the one hand to support and speed up the initial design of an SBI process, on the other hand to maximize the effectiveness of the experts' analyses by continuously optimizing and refining all the design tasks. Quick tuning iterations are probably the most distinctive feature of this type of projects, and they are necessary to cope with the high fickleness of the topics covered in social conversations. The initial step consists of a *Macro Analysis* aimed at defining the project border, the main subject areas and topics. This information is the starting point for the *Ontology Design* and *Source Selection* steps, aimed at creating a topic ontology and at defining the list of web sites and social channels to be monitored, respectively. Topics and web sources are the input for *Crawling Design*, which is aimed at creating the keyword queries to be submitted to the crawling engine. To extract semantics from raw clips, a *Semantic Enrichment* phase is then triggered. At this stage, it is possible to test and tune the overall process.

In the following subsections we give further details on the techniques adopted for each task of the SBI process, using Fig. 1 as a reference.

4.1 Ontology Design

Designing the topic ontology of the European politics subject area was mainly a methodological issue. Consistently with the methodology we followed [8], we initially carried out a *macro-analysis* to identify the themes relevant to the subject area (e.g., "culture") and a first set of topics (e.g., "school"). Then, the *ontology design* phase was specifically dedicated to collecting, for each theme, a comprehensive set of topics and to arrange them within an ontology (using Protégé) by expressing inter-topic relationships (e.g., to state that "school" falls within the context of "educational policy"). Along the whole project lifetime, the topic ontology was weekly tuned and refined (in collaboration with domain experts) to accommodate new topics, topic classes, and relationships; the model of the final result is shown in Fig. 3.

4.2 Crawling

The two main design activities related to crawling are *source selection* and *crawling query design*. For crawling we adopted the Brandwatch service to ensure a satisfactory coverage of web sources along the project duration (three months). Brandwatch adopts a template-based engine, that is, it extracts only the informative UGC by detecting and discarding advertisements and banners (a process called *clipping*); it also drops duplicate clips using content aggregators. As to source selection, the Brandwatch source base has been extended with more than 100 additional domain-specific web sites suggested by our domain experts.

Keyword-based queries in Brandwatch rely on a set of 23 operators (ranging from Boolean ones to proximity ones) that allow to express filters on both textual features (e.g., the maximum distance between two words) and metadata (e.g., the author's country and the web site name). To enhance the quality of the crawling result, we weekly run a review of the set of queries initially created plus a *content relevance analysis* aimed at discarding off-topic clips.

4.3 Text Analysis

For text analysis we used the *SyN-Semantic Center* commercial engine. SyN was used for splitting clips in sentences and for extracting the single entities, their part-of-speech, and their semantic occurrences. Moreover, several techniques have been adopted to ensure homonyms and homographs identification: (i) keyword-based crawling queries have been designed to properly identify topics by explicitly excluding homonyms (e.g., to avoid occurrences of "Osborne" referring to the musician "Ozzie" rather than to the politician "George"); (ii) SyN-Semantic Center includes a module for homonyms and homographs handling based on ontological disambiguation. In detail, the linguistic and semantic text analysis made by SyN is based on morpho-syntactic, semantic, semantic role, and statistical criteria. At the heart of the lexical system lies McCord's theory of slot grammars [13]. The system analyzes each sentence, cycling through all its possible constructions and trying to assign the context-appropriate meaning to each word by establishing its context. Each slot structure can be partially or fully instantiated and it can be filled with representations from one or more statements to incrementally build the meaning of a statement. The core of the system is the SyN ontology, developed through twenty years of experiences and projects.

4.4 Topic Search

With this term we refer to the task of indexing all the occurrences of a topic (or one of its aliases) within a clip. We relied on two different techniques for searching topic occurrences in SABINE. The first one is a simple text matching technique that retrieves the exact occurrences of topics and aliases, implemented in-house as a Java algorithm. The second one is based on the results of the text analysis made by SyN, which extracts all the occurring entities in a clip; in this case, topic occurrences are obtained by linking topics and aliases to the corresponding entities.

Clearly, both techniques have pros and cons. By avoiding all kinds of text analysis, the first technique typically trades a better performance in terms of speed with a lower accuracy of the results. In particular, the results tend to include the occurrences where a topic (or an alias) is used in the clip with different semantics from the one originally meant in the topic ontology. This problem arises when topics (or aliases) in the ontology are too generic. The second technique presents the opposite challenge: by carrying out an in-depth comprehension of the clip semantics, the entities produced by SyN tend to be very specific, possibly leading to the pulverization of the same concept into a wide set of entities. Therefore, this problem arises when topics (or aliases) in the ontology are intentionally generic.

The adoption of both techniques enabled us to double-check the results and to track down the causes of conflicting results. Eventually, mismatches were manually solved in most cases, yielding a 91% agreement between the two techniques (over 14 millions occurrences).

4.5 Sentiment Analysis

Sentiment analysis is probably the hardest task in SBI; for this reason we included in SABINE both system-based and human-based sentiment scores. While system-based scores can be used as a baseline for testing other automatic techniques, human-based scores represent the ground truth.

Crawler Sentiment. This score, computed by Brandwatch, tags each clip of SABINE. The sentiment analysis component of Brandwatch is based on mining rules specifically developed for each language supported.

NLP Sentiment. SyN includes its own sentiment analysis component [15] whose score takes into account the negative or positive polarization of words and concepts, as well as the syntactical tree of the sentence being analyzed. Each clip of SABINE is tagged with this score as well.

Expert Sentiment. This score was defined for a sample of 1200 clips (600 English + 600 Italian) by asking our domain experts to manually tag them. The clips are 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 due to an inadequate template used by the crawler when clipping), 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.).

Crowd Sentiment. This score was given to a sample of about 2400 clips (1200 + 1200, including the clips tagged by experts) through a crowdsourcing process. To this end, we selected a crowd of around 900 workers within a class of bachelor-degree students in the field of humanities and political science at the University of Milano (average worker age is 21). Crowdsourcing activities were performed during one month and each worker tagged 46 clips on average. As a support we employed our Argo system (island.ricerca.di.unimi.it/projects/argo/, in Italian), which provides crowdsourcing functionalities based on *multi-worker task assignment* and *consensus evaluation* techniques [5].

4.6 Data Linking

The goal of data linking is to link ontology topics to the Linked Data Cloud. In SABINE, this has been done by coupling automated techniques with manual validation and revision by domain experts. As a first step, topic aliases were used to retrieve a set of candidate DBpedia resources for each topic t through the DBpedia Lookup Service. The degree of similarity between t and the retrieved candidates (if any) was evaluated through the HMatch matching algorithm [3]. HMatch takes into account both the linguistic information available for t (i.e., its aliases) and the ontological information (i.e., the topic class of t). Then, topic t was linked to the DBpedia resource e, among the candidates, yielding the highest degree of similarity. The link between t and e is formally defined as a 4-tuple of the form $\mathcal{L}_{t,e} = \langle t, e, \sigma_{t,e}, \rho_{t,e} \rangle$, where $\sigma_{t,e}$ is a real number in the range $[0, 1]$ representing the degree of similarity between t and e, and $\rho_{t,e}$ represents the semantics of the relation holding between t and e. Each resulting link $\mathcal{L}_{t,e}$ was submitted to domain experts to specify the most suitable semantics $\rho_{t,e}$, choosing among (i) owl:sameAs (t and e have exactly the same meaning); (ii) sabine:narrower/sabine:broader (the meaning of t is more specific/generic than the one of e); (iii) sabine:related (there is a positive association between the meanings of t and e[1]. If none of the previous options was deemed suitable by the domain experts, the link was marked as *incorrect*. The links with semantics different from owl:sameAs and all the incorrect links were submitted to a second validation round, where domain experts manually found additional DBpedia resources to be associated with the corresponding topic with owl:sameAs semantics. This procedure is crucial to ensure the quality of the resulting links. However, due to the effort required to the domain experts, this process has been feasible only for the 400 topics, but not for the entities because SABINE provides over 4 million entities, which would have resulted in an overwhelming manual activity for the experts. Table 3 shows some statistics about validation and refinement of English topics.

Example 2. As an example, we propose the link ⟨“school”, dbpedia:State_school, 0.75, owl:sameAs⟩ between topic “school” and DBpedia resource dbpedia:State_school. In the first round of validation, experts confirmed that “school”

[1] In SABINE, topics are not modeled as SKOS concepts because they include concrete entities (specific parties, institutions, people) that are better represented as OWL Named Individuals in order to keep the distinction between concepts and instances in the ontology (whereas, according to the W3C, skos:Concepts should only represent abstract entities or conceptual knowledge). Besides, not even the DBpedia resources that our topics are related to are instances of skos:Concept. Thus, we could not reuse the skos:broader and skos:narrower properties because (although their meaning is similar to the one we intend) they are used only to describe relationships between skos:Concepts.

Table 3. Results of the domain expert validation and revision for the English topics

Relation semantics	# links	Avg. similarity
owl:sameAs	252	0.812
sabine:narrower	7	0.721
sabine:broader	39	0.756
sabine:related	17	0.781
incorrect	62	0.22
sub-total	377	
expert-provided resource	135	0.433
total	512	

Most of the automatically retrieved links have been considered correct with owl:sameAs semantics (67%); 17% of the remaining links have been evaluated as correct but with semantics different from owl:sameAs. In particular, the automatic linking procedure tends to provide specific (rather than generic) DBpedia resources for topics. The automatic approach was incorrect in 16% of cases. We note also a positive correlation between the average degree of similarity associated with links and the positive evaluation provided by experts. This is important for associating a reliable degree of similarity to the links in the final dataset. Finally, for 135 topics, domain experts provided a DBpedia resource as an owl:sameAs counterpart for the topic.

can be linked to dbpedia:State_school, but with semantics sabine:broader (since school is broader than dbpedia:State_school). The resulting link was ⟨"school", dbpedia:State_school, 0.75, sabine: broader⟩. Since the semantics is different from owl:sameAs, the link was submitted to the second validation round, where we asked experts to manually find a DBpedia resource which actually has an owl:sameAs relation with "school". Experts found the DBpedia resource dbpedia:School, which leads to the addition of a second link for topic "school". The links resulting from the two validation rounds are then

$$\langle \text{``school''}, \mathsf{dbpedia{:}State\_school}, 0.75, \mathsf{sabine{:}broader} \rangle$$
$$\langle \text{``school''}, \mathsf{dbpedia{:}School}, 1.0, \mathsf{owl{:}sameAs} \rangle$$

5 Research Tasks

In this section we describe the main research tasks that are supported by the data provided in SABINE. To enable partial, ad-hoc downloads for each task, we subdivided SABINE into separate components shown as packages in Fig. 5. A package models a component of the dataset that can be downloaded separately;

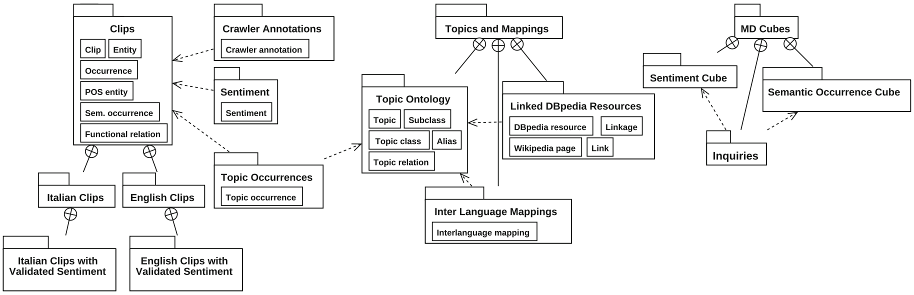

Fig. 5. Components of SABINE and their composition and dependency relationships

Table 4. Task overview

Task	Input packages	Ground truth packages	Ground truth
Content analysis tasks			
Sentiment analysis	Clips	Sentiment	Sentiment manually defined by crowd and experts
Topic search	Clips; Topic Ontology	Topic Occurrences	Topic occurrences obtained by automatic techniques and validated by experts
Document classification	Clips	Crawler Annotations	Classifications induced by the crawler annotations
Semantic analysis tasks			
Cross-language analysis	Clips; Topic ontology	Inter-Language Mappings	Mappings manually defined by experts
Topic discovery	Clips	Topic Ontology; Topic Occurrences	Topic ontology manually defined by experts and topic occurrences obtained by automatic techniques and validated by experts
Data linking	Topic Ontology	Linked DBPedia Resources	Links obtained by an automatic procedure and validated by experts
SBI analytics tasks			
Multidimensional modeling	Clips; Topic Occurrences; Topic Ontology; Sentiment; Crawler Annotations	MD Cubes	Cube schemas and instances
SBI analytics	Clips	Inquiries	Datasets resulting from the inquiries and validated by experts

one package depends on another one when an object of the former references an object of the latter. For each task, in Table 4 we summarize the SABINE component(s) to be taken in input and the ground truth we provide together with the packages where it is contained.

The main idea of SABINE research tasks is that different combinations of subsets of the main dataset (i.e., packages) can be used either as a training or as a testing set for a variety of different approaches and algorithms in the areas of *content analysis*, *semantic analysis*, and *SBI analytics*. Content analysis tasks are focused on the interpretation of the clips provided by SABINE in order to automatically assess the capability of associating them with the correct sentiment, the effectiveness of retrieval and classification of clips by topic. Semantic analysis is mainly focused on the discovery of topics in the clip collection and on the discovery of semantic relations between SABINE topics in different languages (i.e., English and Italian) and with DBpedia. Finally, the main goal of SBI analytics is to enable complex analyses of social content by integrating information obtained from a semantic enrichment process that includes techniques coming from different research fields. For this reason, the tasks in this subsection play a special role in SABINE because, unlike those described above which are related to each single phase of enrichment, they concern the capabilities of integrating the results.

6 Conclusion

In this paper we have presented SABINE, a dataset of semantically annotated social content in the domain of European politics. SABINE aims to constitute a publicly-available, real-world dataset for experimenting and comparing the most commonly performed SBI tasks, crossing the various involved research fields ranging from Database Systems, Information Retrieval, Data Mining, up to Natural Language Processing and Human-Computer Interaction. SABINE has been designed and properly packaged for modular download to enable the evaluation of a wide variety of research tasks, either separately or in combination, ranging from those more focused on content analysis, to those related to semantic analysis up to more comprehensive SBI analytics. The SABINE components related to data linking and cross-language analysis have been used for the tasks of inter-linguistic mapping and data linking within the OAEI Instance Matching Track [1]. A main technical advance of SABINE is the availability of multiple, complementary, and validated enrichments of the social content (i.e., textual clips) in form of metadata, annotations, sentiment scores, and DBpedia mappings. The availability of a user-validated ground truth, either by domain experts or by crowdsourcing or both, for each enrichment phase represents a further technical advance of SABINE. In our future work, we plan to undertake a new crowdsourcing project to manually annotate an increasingly larger amount of clips.

SABINE [4] is available for download at http://purl.org/sabine under the CC BY-NC 4.0 license; packages are made available as compressed archive files containing JSON files (the Clips package), OWL files (the Topics and Mappings package and all its sub-packages), and CSV files (all other packages).

References

1. Achichi, M., et al.: Results of the ontology alignment evaluation initiative 2016. In: CEUR Workshop Proceedings, vol. 1766, pp. 73–129. RWTH (2016)
2. Amini, M., Usunier, N., Goutte, C.: Learning from multiple partially observed views-an application to multilingual text categorization. In: Advances in Neural Information Processing Systems, pp. 28–36 (2009)
3. Castano, S., Ferrara, A., Montanelli, S.: Matching ontologies in open networked systems: techniques and applications. In: Spaccapietra, S., Atzeni, P., Chu, W.W., Catarci, T., Sycara, K.P. (eds.) Journal on Data Semantics V. LNCS, vol. 3870, pp. 25–63. Springer, Heidelberg (2006). https://doi.org/10.1007/11617808_2
4. Castano, S., et al.: SABINE: a multi-purpose dataset of semantically-annotated social content. In: Vrandečić, et al. (eds.) ISWC 2018, Part II. LNCS, vol. 11137, pp. 70–85 (2018). http://purl.org/sabine
5. Castano, S., Ferrara, A., Genta, L., Montanelli, S.: Combining crowd consensus and user trustworthiness for managing collective tasks. Futur. Gener. Comput. Syst. **54**, 378–388 (2016)
6. Chu, C., Nakazawa, T., Kurohashi, S.: Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In: Gelbukh, A. (ed.) CICLing 2014. LNCS, vol. 8404, pp. 296–309. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54903-8_25
7. Francia, M., Gallinucci, E., Golfarelli, M., Rizzi, S.: Social business intelligence in action. In: Nurcan, S., Soffer, P., Bajec, M., Eder, J. (eds.) CAiSE 2016. LNCS, vol. 9694, pp. 33–48. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39696-5_3
8. Francia, M., Golfarelli, M., Rizzi, S.: A methodology for social BI. In: Proceedings of IDEAS, pp. 207–216 (2014)
9. Gallinucci, E., Golfarelli, M., Rizzi, S.: Advanced topic modeling for social business intelligence. Inf. Syst. **53**, 87–106 (2015)
10. Golfarelli, M., Rizzi, S.: Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, New York City (2009)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the SIGKDD, Seattle, USA, pp. 168–177 (2004)
12. 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, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_13
13. McCord, M.C.: Slot grammar. In: Studer, R. (ed.) Natural Language and Logic. LNCS, vol. 459, pp. 118–145. Springer, Heidelberg (1990). https://doi.org/10.1007/3-540-53082-7_20
14. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 task 4: sentiment analysis in twitter. In: Proceedings of SemEval@NAACL-HLT, San Diego, USA, pp. 1–18 (2016)
15. Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T.: Sentiment analysis on social media. In: Proceedings of ASONAM, pp. 919–926 (2012)
16. Otero, P.G.: Learning bilingual lexicons from comparable English and Spanish corpora. In: MT Summit xI, pp. 191–198 (2007)
17. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of ACL, Michigan, USA, pp. 115–124 (2005)