SAMi2; a Semantic and Big Data Lever to Enhance Public Safety Using Social Media

Jesús del Peso, Raúl Santos, Paloma Jimeno
HI Iberia, Madrid, Spain
jdelpeso@hi-iberia.es
rsantos@hi-iberia.es
pjimeno@hi-iberia.es

Abstract: Social networking is one of the most popular online activities worldwide transforming OSNs in a mirror of society where both legal and illegal activities are present. Governments and safety and security officials have difficulties to prevent and detect these illegal actions organized through OSNs due to the sheer quantity of available data. In this paper we present the data processing approach for project SAMi2, consisting on a crawling engine, first level filtering of data based on keyword and metadata spotting followed by a semantic analysis which uses context information and models of the domain to go beyond the keyword approach. A prototype data acquisition and analysis engine is discussed and the first results for the system, which gathers and analyzes information from Twitter, are presented. This prototype is to be extended in the remainder of the project into an application for professional end-users (safety providers, government personnel) for a number of use cases which are presented.

Keywords: social network analysis, natural language understanding, security applications, computational semantics, distributed architectures.

1. Introduction, concept and high level goals

The prominence of Online Social Networks (OSNs) as a mass media of communication is part of today's daily lives for citizens worldwide. Communities and personal relations of all sorts are now inextricable from the Internet tools that have appeared over the last decade. Impacts are felt everywhere: from news spreading through personal relationships and artistic movements, OSNs have grown to cover a majority of the spectrum of human activities.

Among this, the most massive are nevertheless general-purpose ones that don't try to align to particular purposes but are rather conceived around models or paradigms of communication. For example, Facebook[1] tries to mimic the familiar concepts of traditional friendship and social neighbourhood: events, birthdays and other personal occasions are well-defined elements.

Micro-blogging sites, such as Tumblr[2] or Plurk[3], focus on short interactions, often imposing restrictions over the size of the users' inputs, where time is a crucial factor. Among these perhaps the most paradigmatic is Twitter[4], which has rapidly risen to prominence as a synonymous for timely information for emergent news. Twitter's users compose short messages, limited to 140 characters, called "tweets" to express their opinions or views. These are often composed from sensor rich mobile devices which attach metadata such as timestamps or geo-location to the messages. Twitter provides APIs that allow programmers to tap into this rich "public" stream or the dataset that is not marked by the authors as available for unrestricted publication in the website.

Parallel to technical growth, these general purpose OSNs have come to mirror the society at large. Over OSNs jobs are offered, items are bought and sold and other business information is exchanged. And in parallel to these legitimate activities, illegal activities are often discussed or planned. For example as discussed by Tonkin et al., 2012, over 600.000 tweets were deemed as relevant in the organization of the riots in London and other cities in the UK during August 2011. Government agencies and security forces have been monitoring OSNs for some years now in the search of evidence leading to early warning of organized crime such as money laundering or safety-threatening scenarios such as illegal demonstrations. This work is at the moment manual, labour intensive and thus expensive on the tax payers.
With the above technical scenario in mind, project SAMi2 was conceived as a potential helping tool for these police forces in automatizing parts of the OSN monitoring process. The main goals for the project are as follows:

[1] Facebook website: https://www.facebook.com/
[2] Tumbler website: https://www.tumblr.com/
[3] Plurk website: http://www.plurk.com/top/
[4] Twitter website: https://www.tumblr.com/

- Provide a generic technical solution to crawl the OSNs and extract relevant pieces of information leveraging the most up-to-date technologies and approaches to ensure maximum efficiency and scalability.
- Ensure that the OSN user's rights of privacy and data ownership are observed and that ethical and legal provisions in the law or the OSNs Terms of Services are upheld.
- Enforce an end-user-centric perspective to produce a professional tool for use by security stakeholders rather than other tools geared for commercial or personal use.

This paper is organized as follows: after this introduction, section 2 summarizes the application scenarios envisaged by the end-users that are involved in the project. Section 3 presents the high level architecture proposed for the analysis solution. Section 4 outlines the information flows of data in the system. Section 5 summarizes the preliminary results collected after the first implementation of the system. Finally, Section 6 discusses the conclusions of our work and future research and implementation lines.

2. Application scenarios

In accordance with our end-user-centric approach, the application scenarios were agreed with the Madrid City Police as fulfilling some useful objectives in the operative work of their social media analysis unit. The use cases were then validated by legal and ethical counsellors to ensure they did not pose any issues in those regards. Domain expertise from the Police forces remains an essential input for this system as purely data-centric analyses (e.g., statistical) can complement but not substitute years of experience by human operators.

Six application scenarios were produced for SAMi2:

1. "Escrache" or unauthorized spontaneous demonstrations at the personal residences of government officials, politicians or other public figures. The core goal of this scenario is timely detection of an upcoming "escrache" action, which are often organized via OSNs. Typical challenges include detecting the topic of conversations (the word "escrache" itself may not be mentioned).
2. Gang activity, especially when related to sports events. Often the supporters agree to meet close to the date of important matches to fight or do vandalism. Challenges in this application are detecting meeting times and dates in advance and the incorporation of intelligence from undercover agents.
3. Illegal event organization such as concerts or raves, with a similar profile of challenges to 1) and 2) above (time and location of event identification). This scenario is expanded in the example followed in sections 3 through 5 of this paper.
4. Hate speech, in which minorities are targeted and harassed via OSNs. Challenges here include the identification of a wide range of hate words and expressions under a common analysis framework.
5. Post-mortem analysis, understood as means to do analysis after the occurrence of the crime itself, e.g., after the illegal event described in 3) has happened. This present several differences with "live" analysis such as the access to data in the OSNs that might have been removed.
6. Terrorism, although falls beyond the scope of a municipal Police duties, presents challenges of data exchange and interoperability with larger organizations (national level police), which can be overcome by a precise semantic model to describe the results.

3. SAMi2 architecture

In this section an overview of the main architecture that SAMi2 is based on is depicted. This architecture has been designed with the following main goals in mind:

- *It should be general*. This architecture is intended to serve as a basis for intelligent processing systems working on quite generic information sources, ranging from structured information sources such as databases or rigidly formatted message streams to the most general unstructured natural language based textual documents and even raw multimedia sources, including images, audio and video.
- *Practical*. SAMi2, and any other system based on this architecture, tries to achieve tangible results from its very beginning, both to enable early validation and maximize interaction with final

users which therefore can help to drive the development along the right direction. In fact, user feedback is a key element in the overall conception of the information processing discussed in current paper.

- *Extensible*. The architecture makes possible to incrementally add new processing capabilities according to the final application needs –for instance, adding image analysis capabilities or multimedia sources processing when new information sources are to be injected to the system. This makes possible to follow an iterative development model. Additionally this architecture is intended to serve as a basis not only for the security focused SAMi2 system but also for many other future deep information processing based applications.

To achieve these main goals the architecture proposes the integration of a number of already existing technologies and also requires the development of a number of new processing models and tools, especially in the analysis at the semantic level of information sources.

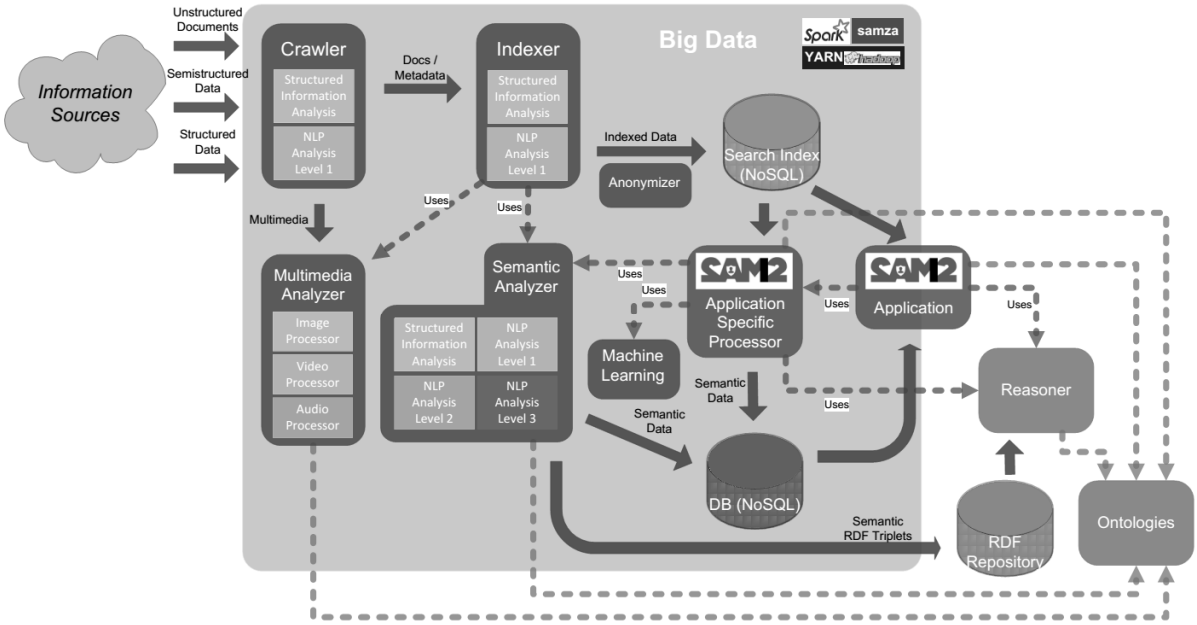

Figure 1: SAMi2 architecture

Since, as already mentioned, generic information sources are intended to be processed in current model and in order to keep consistency we will call *document* to any relevant piece of information susceptible of being processed independently. In this sense, elements such as web pages, social media posts and messages, e-mail messages or even individual images or video clips may be considered as *documents*.

The SAMi2 project in its first steps is centred on the analysis of Twitter as source of information. This social network offers a number of advantages that make it suitable as a testbed environment and starting point for a future broader coverage. Twitter is a very dynamic social network since it provides a huge amount of information related to many different topics, which is continuously generated and evolving. In addition all tweets currently published are public. This makes possible to process tweets as any other public information source. Tweets contain a lot of information, both structured and unstructured: while Twitter messages contain a large number of already formatted meta data fields, as provided by Twitter's API ("Twitter Developers," n.d.), ("The Streaming APIs," n.d.), the main informative content is still present in the form of unstructured elements, mainly as short raw texts and multimedia items. This mixed nature makes Twitter an ideal test field for the application of deep content analysis techniques and fusion of unstructured and structured data.

Figure 1 shows the main blocks composing the architecture:

- First, some quite standard processing modules, dealing with gathering of large amounts of information and performing a first level of processing are included:
  - *Crawler*. This is a generic module, in charge of gathering the different types of information pieces – documents– to be processed by the system. It is also responsible for a first level of

processing on these documents, formatting and extracting any required data field and filtering out any unnecessary element, in order to increase efficiency. For example, in the case of Twitter message processing, the crawler is responsible for the selection of the relevant JSON fields of tweets to be processed. Since this initial processing is source and domain dependent, the crawler is intended to be composed of a number of plugin-like components, each of them dealing with a specific type of information source.

- *Indexer*. It is responsible for the construction of an inverted index (Zobel and Moffat, 2006), (Zobel et al., 1998) from the contents of the analyzed documents. This is composed of different types of information elements, some of them domain dependent:
    - Structured data.
    - Information extracted from a first level of natural language processing, including identified and normalized words.
    - Some useful lexical or semantic information, such as synonyms and other lexically related terms.
    - Metadata and labels extracted from multimedia resources, including hashes for effective image indexing.

  Table 1 shows a summary of the types of elements included into the index built for Twitter messages in SAMi2 application.
  Indexing is a quite standard step in processing which allows the system to deal with huge amounts of messages, which have to be made searchable and retrievable according to a set of criteria. In SAMi2 current implementation this is achieved using an extension of the Apache Solr[5] indexer, modified to include richer lexical and semantic information.
- *Anonymizer*. Additionally, if the final application characteristics requires it, the information to be indexed may need to be anonymized, i.e. transform information in order to remove any reference to personal or private data. In fact, SAMi2 project requires the use of this module to assure that the more strict privacy policies are followed.

- A number of specific semantics based modules, providing deep information processing and knowledge modelling:
    - *Semantic analyzer*. This module is responsible for deep Natural Language Processing, including syntactic and even pragmatic analysis in order to identify main concepts and relationships between them and relevant individuals, places or temporal information. This elements can be used later as an input in more specific processing layers involving interpretation according to general or context specific ontologies. This module is also in charge of other types of high level non linguistic information, such as the generation of social models in the case of messaging information sources processing or network structures of references between documents.
      The extracted semantic elements can be of two different types: common general semantic elements, such as time and location concepts, instances or individuals mentioned in documents and domain dependent concepts and relations. The definition of these domain specific knowledge elements is done in one or more domain ontologies, which are pluggable to the semantic analyzer.
    - *Multimedia analyzer*. This module is responsible for the extraction of usable information from multimedia elements included in more complex documents or considered as documents themselves. At a first stage a set of quite simple information elements is extracted, such as image hashes (Venkatesan et al., 2000) or metadata tags already included into audio or video media. However, the general model developed here also considers more sophisticated levels of processing, such as face recognition, object or even scene recognition procedures, automatic video and audio tagging, etc. The latter functions are currently far away the scope of the SAMi2 project but, as already mentioned, the models and tools developed claims to be of more general use and serve as a basis for much more sophisticated intelligent information processing systems.
    - *Semantic framework*. This module is intended to provide semantic reasoning services to the other architecture elements –specially to the Semantic Analyzer and application level specific modules–, based on the Semantic Web model (Berners-Lee et al., 2001), (Kifer et al., 2005). A number of general ontological models are used as conceptual framework. Additionally domain level ontologies are included to add domain specific reasoning capabilities. In fact, a

[5] http://lucene.apache.org/solr/

great part of adaption to specific domains, such as different security domain profiles in the case of SAMi2, is based on domain ontologies definition and reasoning based on them.

- *Application specific processor*. This module is in charge of specific high level processing tasks for every application domain. In particular, this processor is in charge of deciding which documents need to be returned to final users as a final result. Any adaptive action to tune the system to its expected behaviour according to final user's feedback is decided here, using the features provided by the *machine learning module*.
- *Machine learning module*. It provides learning capabilities to the other modules. This module is currently based on the use of neural networks and SVNs (Support Vector Machines). Probabilistic graphical models, such as HMMs, Conditional Random Fields, are planned to be added in a near future.
- *Application.* This module provides any further adaptation to the final application domain – SAMi2 security application– and involves all Interfaces and human interaction components needed by the system including information adaptation and feedback models for users.

Table 1: Twitter information elements included into inverted index

Structured fields (from JSON message object fields)	language geoLocation information: longitude, latitude, countryCode, countryName, placeName, placeFullName, placeBoundingBox, boundingBox, placeId, placeType timestamp social information[6]: userid, profileImage, profileBackgroundImage, userScreenName, userMentions, inReplyToStatusId, statusesCount, followersCount hastags messageId
Non-structured text	Basic NLP: Text chunking, tokenization, normalization: translation from Twitter specific language, formatting, lemmatizing) Include synonyms (from WordNet[7] (Miller, 1995) or similar lexical databases like MultiWordNet[8])
Multimedia information	Image hashes, audio and video labels.[9]

The system and the processing steps described in this paper require a very high computational load, specially the semantic information processing and machine learning procedures, as they have to be performed continuously as documents arrive to the system and user feedback is provided. In addition, for the security focused applications envisioned by SAMi2, a continuous evaluation and filtering of information is required in order to provide near real time results. Therefore, a parallel processing environment must be used as a basis for the system. A solution based on the integration of a number of Big Data based open source tools, working on a cluster, has been chosen to achieve these goals. While indexing is relied to the Apache Solr engine, the main semantic and deep analysis of information and all involved machine learning is based on the Apache Hadoop[10] platform, based on the map-reduce processing model (Yang et al., 2007). All internal information managed by the system is stored in MongoDB[11] databases, allowing direct access to local data from Hadoop. This schema is scalable and allows seamlessly adding new computational resources as needed due to an increase in the volume of information to manage, including new sources or simply due to the adding of new processing features, as expected in a near future.

[6] These elements are used to build up a network model of relations between tweets and users.
[7] http://wordnet.princeton.edu/
[8] http://multiwordnet.fbk.eu/english/home.php
[9] These elements are not currently generated extracted by SAMi2 application, but are expected to be included into the next development iteration of the system.
[10] http://hadoop.apache.org/
[11] http://www.mongodb.org/

Additionally, if the fast processing of huge streams of messages or message-like information sources is needed, other specific frameworks can be used to cope with this additional load, such as Apache Samza[12].

4. Processing flows

In the SAMi2 context, overall processing is aimed to identify and return to final users a set of messages of interest, containing useful information for the Madrid municipality police, and provide these results in a near real time way in the form of alerts. In a further development step of this system, besides the identification of the aforementioned target messages a richer insight on their content is expected to be provided too.

The overall processing consists of a two phase adaptive filtering, working on different levels of processed information, from the most simple resulting from shallow processing to more complex coming from deep analysis. Figure 3 shows the steps involved in the first filter level. Since the expected total number of documents to be analyzed is far beyond the real processing capacity currently available –especially if the processing to be carried out involves deep semantic analysis– this filtering step has as main goal limiting the number of messages/documents to be analyzed in depth in the next processing steps. This first filtering is based on the construction of a tailored inverted index from the content of messages/documents coming from a first shallow processing level. It is based on quite standard information extraction techniques with a number of specific adaptations to the characteristics of the information source to be managed, in this case, Twitter. As a result, a set of information elements have been selected to be included into the inverted index, as shown in Table 1. These elements come both from the metadata provided by Twitter API and from a first natural language processing of the tweet messages text. In the next development iterations of the project a first level image analysis is expected to be also included.
Since SAMi2 project will result in a security tool for the Madrid City Council Police, the Twitter messages to be processed are written in Spanish. Therefore, a number of open source NLP tools for this language has been adapted; some other new NLP tools has also been developed in order to solve some specific problems offered by the special kind of Spanish language used in tweets. For example, a dynamic Spanish Twitter dictionary has been developed to translate some frequent Twitter specific abbreviations or simply misspellings commonly used in tweets to a normalized Spanish form and do this in a dynamic way, adapting to changes of trends of use of language in Twitter.

Table 2: High level definition for the *illegal massive parties* profile in SAMi2 application

Specific keywords	*botellón*[13], *botellona, fiesta* (party)
Types of action	*Organizar* (organize) *Acudir* (attend)
Time information	Identify if action takes place in present, past or future time.
Place information	Identify specific locations and places mentions in message text.
Social information	Size of underlying social network for processed messages. Centrality measure (Friedkin, 1991) of messages in its implicit social network.

Once tweets contents are conveniently indexed a number of queries will be performed. These queries provide the first filtering level, providing as a result a set of candidate tweets to be analyzed later by next deep analysis steps. Queries are built from a semantic model defined for the application. Therefore, for the SAMi2 application a semantic model is defined, in the form of a general security domain ontology. Final users can extend this base ontology by defining several more specific contexts of interest, called *detection profiles* in SAMi2. For example, users can specify their interest in the generation of alerts for the detection of illegal massive parties organization activities. The SAMi2

[12] http://samza.apache.org/
[13] These are specific Spanish terms used for massive informal parties organized on public streets involving alcohol consumption and frequently involving young people, which are banned in many Spanish cities.

application provides users with a set of interfaces to seamlessly define the main concepts that are involved in every detection profile. In the case of the *illegal massive parties* profile relevant concepts and simple processing rules are shown in Table 2.

Queries for the first level filtering are automatically generated by the system from the specific keywords contained in the security profile, combining them along with synonyms of the types of actions to be identified. As a result a query in the language of the Solr indexer is generated. The execution of this query provides the first level of filtering. These steps are performed by the system automatically, without other user intervention than the security profiles definition. This filtering is performed repeatedly in time slices and incrementally as new documents arrive. Note that first level filtering is designed to provide a near 100% recall –i.e. try to minimize the number of positive results filtered out, at the cost of increasing the number of negative ones–, leveraging final classification accuracy to the second filtering.

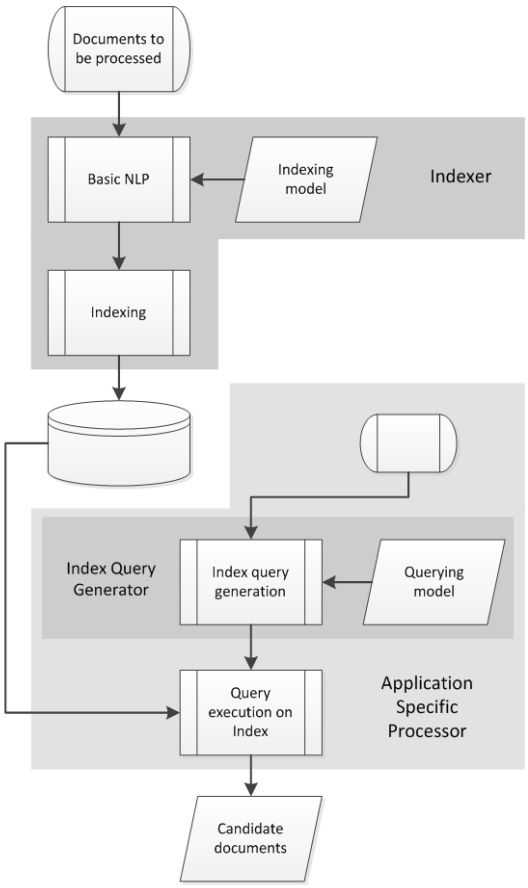

Figure 2: Document indexing and first level of filtering, based on adaptive index querying.

Figure 3 shows the next steps in the deep processing of the documents coming out from the first level of filtering. Semantic processing is the key phase which provides a real improvement from other purely statistical based methods. It has as main goal the identification and extraction of general semantic relevant elements –*semantic features*–, according to a general model, covering facets such as location and place information and the semantics of social network structures. This general semantic model takes the form of a set of general ontologies which enable complex reasoning o generic concepts. Additional semantic elements are extracted from deep content analysis according to a set of domain models specified at a high level by final users, called *security profiles* in the SAMi2 application environment.

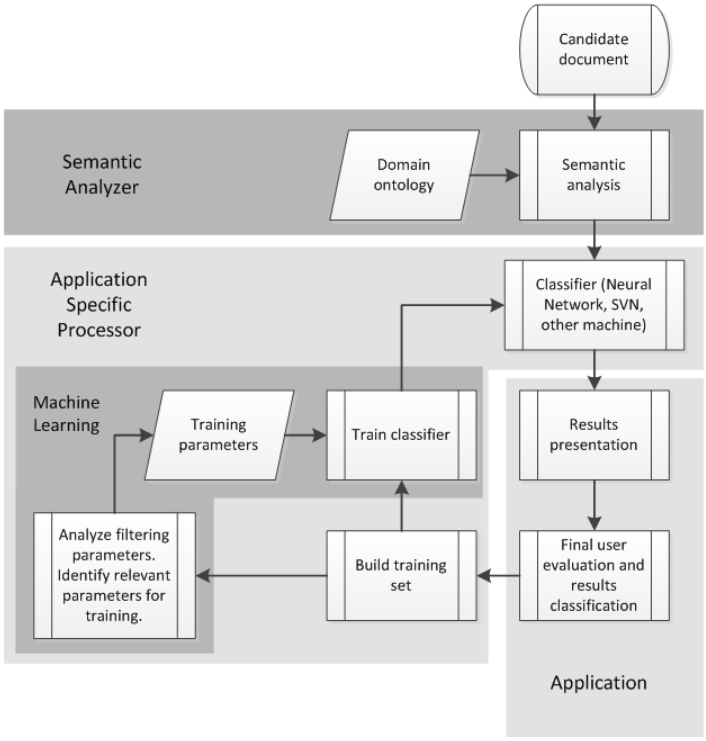

Figure 3: Semantic processing and second level filtering of documents based on semantic information. User feedback and learning is also covered.

Figure 4 shows different types of semantic processing levels, as currently considered in our proposed architecture:

- Location Analysis. At a first stage, two types of location information can be extracted from documents –specially in the case of Twitter messages: locations where the document/message where generated –very interesting in the case of Twitter–, usually provided as structured data; locations cited in texts and also in audio or video clips or even in images, which have to be extracted by language analysis and usually is application context dependent. Techniques such as named entities recognition are used for this goal.
- Time Analysis. As in location analysis, two possible levels of temporal analysis are possible: identification of the generation time of documents/messages, usually provided as structured metadata; time instants and periods mentioned in raw texts and media which typically needs to be identified using NLP techniques.
- Network/Social Analysis. Analysis of the relationships established between the persons involved in the documents/messages. This may involve, for example, relationships between current author and other messages authors –retweets in the case of Twitter– and relationships between author and persons mentioned in the message.
- Relevance Analysis. Its main goal is to rate the relative relevance of a given document with respect to other similar elements. Several relevance measures can be applied, many of which are based on measures derived from social/network analysis. Temporal and spatial information can also be used for determining relevance –for example, depending on the specific application context it may be needed to focus on certain time periods or places. In fact, in the case of SAMi2 application analysis is centered on messages coming from or related to the Madrid city area and analysis needs to be focused on the most recent messages received but not forgetting their relationships with past related messages.
- Content Analysis. This refers to the identification of relevant concepts and their relationships on raw texts and other media. Some of these elements may be dependent on the final application context and analysis is facilitated by providing a set of domain ontologies. Semantic reasoning is also usually needed in this step.
- Sentiment Analysis. This refers to the positivity or negativity of the feelings involved, implicitly –the most common case– or explicitly, in a given document.

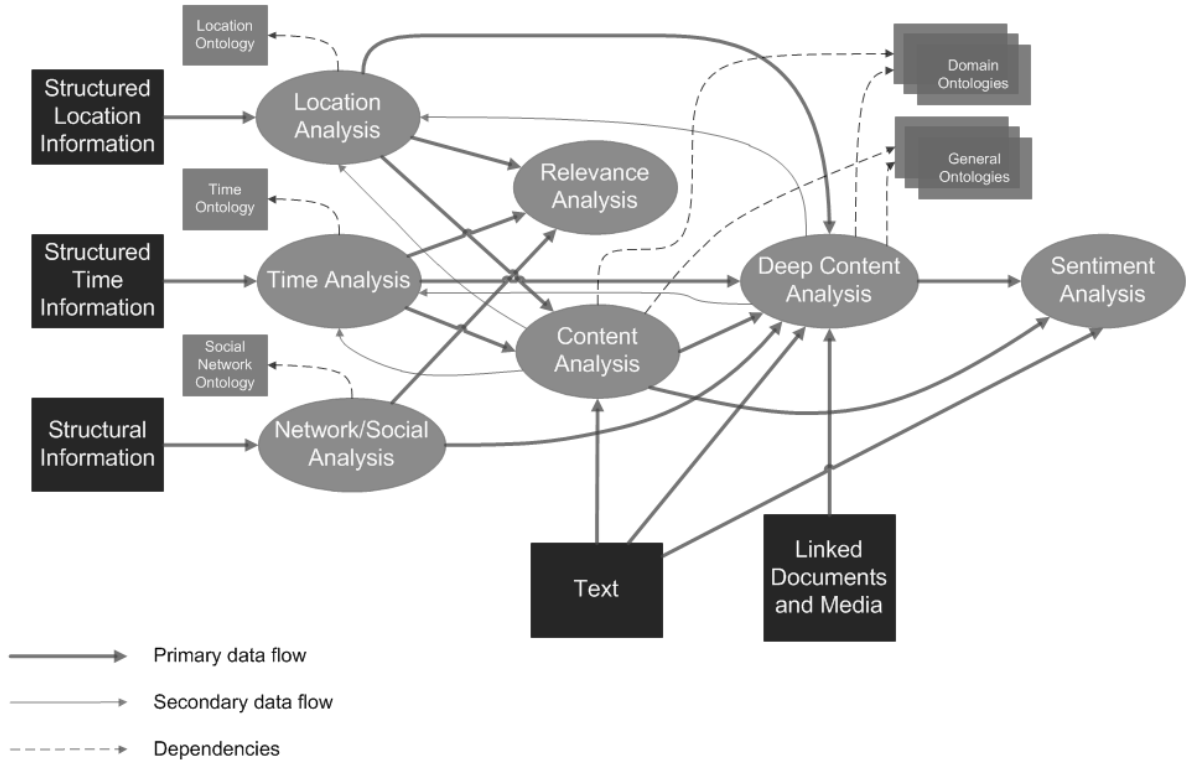

Figure 4: Semantic processing levels

The documents resulting from the first level filter are then filtered again using the *semantic features* extracted by the analyzer module. This second level filtering is made using machine learning based classifiers, such as neural networks or SVNs. The results from this second filtering are then provided to final users which are then able to evaluate if they are correct. This evaluation is fed back to the learning modules to improve next results. As it can be appreciated continued user interaction is therefore a fundamental pillar on the proposed architecture.

5. Initial results and conclusions

At the moment of writing this paper an initial version of the SAMI2 application has been implemented. Using this implementation some quite positive results has been obtained. As evaluation metrics the proportion of false negatives[14] and false positives[15] in final classification of messages has been used. It must be taken into account that the effects of these error measures are quite different in an application such as SAMi2: false positives will result in providing the final user a number of false alarms and therefore resulting in a slight overload for the human operator of the system, who should check some additional non-relevant messages; on the other hand, false positives result in hiding relevant information, which may have worse effects in overall system performance.

Table 3: Results of using semantic features for message classification in the *illegal massive parties* SAMi2 security profile.

	False positives	False negatives	Total error rate	Comments
Model trained without semantic features	16.00 %	56.96 %	44.39 %	Models are trained using only a set of relevant keywords: those with greatest TFIDF (term frequency - inverse document frequency) measure in positive examples in training set.

[14]% False negatives = (Number of incorrectly classified as negative results) / (Total number of results classified as negative) * 100

[15]% False positives = (Number of incorrectly classified as positive results) / (Total number of results classified as positive) * 100

Model trained with semantic features	7.07 %	9.04 %	7.72 %	In addition to the set of keywords used in the previous case, the following semantic features has been added to message analysis and model learning: • Time information: message is referred to the past, present or future • Concepts related to *planning* and *movement* are included in messages • Places are explicitly mentioned in messages

As shown in Table 3 a significant improvement in overall identification of relevant Twitter messages using the described semantic processing has been observed. To evaluate this improvement a comparison with similar models not using semantic features has been performed. Data from the *illegal massive parties* security profile is shown as representative results. In this test case a total of 7,260,865 tweets have been processed. The first level of filtering returned a total of 1937 messages to be semantically processed by the next components. The final number of messages identified as relevant by the whole system, after second filtering, once training is complete is 382. Single hidden layer perceptrons and SVNs has been used as classifiers. Training was based on the manual classification of the 20% of the 1937 messages resulting from first level filtering. It is worthy to be noted that the increase in filtering accuracy is specially relevant in the reduction of the proportion of false negatives. This seems to highlight the fact that some kinds of quite generic semantic features indeed help to discover implicit classification patterns otherwise hidden.
The development of the SAMi2 application and some of the modules described in the general architecture are still in process. Therefore a further improve in final results as the amount of semantic information included into the whole process is expected.

References

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. Sci. Am. 284.5, 28–37.

Friedkin, N.E., 1991. Theoretical Foundations for Centrality Measures. Am. J. Sociol. 96, 1478–1504.

Kifer, M., Bruijn, J. de, Boley, H., Fensel, D., 2005. A Realistic Architecture for the Semantic Web, in: Adi, A., Stoutenburg, S., Tabet, S. (Eds.), Rules and Rule Markup Languages for the Semantic Web, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 17–29.

Miller, G.A., 1995. WordNet: A Lexical Database for English. Commun ACM 38, 39–41. doi:10.1145/219717.219748

The Streaming APIs [WWW Document], n.d. . Twitter Dev. URL https://dev.twitter.com/streaming/overview (accessed 2.12.15).

Tonkin, E., Pfeiffer, H.D., Tourte, G., 2012. Twitter, information sharing and the London riots? Bull. Am. Soc. Inf. Sci. Technol. 38, 49–57. doi:10.1002/bult.2012.1720380212

Twitter Developers [WWW Document], n.d. . Twitter Dev. URL https://dev.twitter.com/ (accessed 2.12.15).

Venkatesan, R., Koon, S.-M., Jakubowski, M.H., Moulin, P., 2000. Robust image hashing, in: 2000 International Conference on Image Processing, 2000. Proceedings. Presented at the 2000 International Conference on Image Processing, 2000. Proceedings, pp. 664–666 vol.3. doi:10.1109/ICIP.2000.899541

Yang, H., Dasdan, A., Hsiao, R.-L., Parker, D.S., 2007. Map-reduce-merge: Simplified Relational Data Processing on Large Clusters, in: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07. ACM, New York, NY, USA, pp. 1029–1040. doi:10.1145/1247480.1247602

Zobel, J., Moffat, A., 2006. Inverted Files for Text Search Engines. ACM Comput Surv 38. doi:10.1145/1132956.1132959

Zobel, J., Moffat, A., Ramamohanarao, K., 1998. Inverted Files Versus Signature Files for Text Indexing. ACM Trans Database Syst 23, 453–490. doi:10.1145/296854.277632