TweetAlert: Semantic Analytics in Social Networks for Citizen Opinion Mining in the City of the Future

Julio Villena-Román[1,2], Adrián Luna-Cobos[1,3],
José Carlos González-Cristóbal[3,1]

[1] DAEDALUS - Data, Decisions and Language, S.A.
[2] Universidad Carlos III de Madrid
[3] Universidad Politécnica de Madrid
{jvillena,aluna}@daedalus.es, josecarlos.gonzalez@upm.es

Abstract. In this paper a highly configurable, real-time analysis system to automatically record, analyze and visualize high level aggregated information of user interventions in Twitter is described. The system is designed to provide public entities with a powerful tool to rapidly and easily understand what the citizen behavior trends are, what their opinion about city services, events, etc. is, and also may used as a primary alert system that may improve the efficiency of emergency systems. The citizen is here observed as a proactive city sensor capable of generating huge amounts of very rich, high-level and valuable data through social media platforms, which, after properly processed, summarized and annotated, allows city administrators to better understand citizen necessities. The architecture and component blocks are described and some key details of the design, implementation and scenarios of application are discussed.

Keywords: Semantic analytics, social networks, citizen, opinion, topics, classification, ontology, events, alerts, big data, city console.

1 Introduction

With the recent success and proliferation of mobile devices, the democratization of Internet accessibility and the possibility of meta-information, such as user location, user profile and demographics, etc., the vastly amount of data that is being generated has very rapidly grown. This unstructured source of data is already being used in multiple fields like sociology, advertising, etc. and may also be used to improve public administration services and functionality, as a new version of e-Gov application. User interventions in social networks often contains agreement, disagreement or comments about city services, city administrators, events in the city, etc. However, these data are not really useful unless some semantic processing or data mining technique is applied in order to automatically distinguish between relevant and not relevant information and provide a higher level of abstraction.

This work has been developed in the framework of Ciudad 2020 [1] Spanish national R&D project, which aims to achieve improvements in areas such as energetic efficiency, Internet of the Future, Internet of Things, human behaviour, environmental

sustainability and mobility and transport, in order to design the City of the Future. The project proposes a new city model designed for the citizen –*ad civitates civis*– that aims to include citizen reality into the city decisions.

Usually, the final objective of the government decisions is the citizen welfare. However, it is not always an easy task for the administration services to quickly identify the most important facts that their citizens are facing, to correctly scale them regarding their relative importance according to what citizens think about them, or just to be quick enough to recognize recent issues that may suddenly appear. In such cases, citizen opinion mining will be a key factor to identify and later solve such concerns. Therefore, the citizen is observed here from a dual point of view: on the one hand as the main user of the services that the city offers, and on the other hand, as a proactive city sensor capable of generating huge amounts of data through social media platforms. The citizen sensor is an innovative way to capture high-level heterogeneous information, very descriptive and with great value, especially when considering aggregations. If the city administrators get to properly analyze such vast amount of data coming from Social Media, they will be able to better know trends, generate hypotheses over urban behaviour models in order to improve municipal management policies, bringing them closer to the actual reality of the citizens, thus, turning them into real actors within management mechanisms of smart cities.

In such process of data understanding and mining, technologies to analyze natural language allow to semantically analyze citizen interventions in social media such as Twitter. The aim of our system is to provide city promoters with a powerful tool to rapidly and easily understand what the citizen behavior trends are, what their opinion about city services, events, etc. is, and finally to provide them a primary alert system that may improve the efficiency of emergency systems. In the same way, but applied to a smaller scale as what we propose here, the system could be used to track public services Twitter profiles' and collect user opinion about e.g., e-Gov sites or applications, allowing them to act more quickly to possible lacks of usability, services failures, etc.

The rest of the paper presents the system description and architecture, and further explores the details of each block that composes the system. Finally, a discussion and future work section with insights to improve the system currently in the development branch are presented.

2 System Architecture

We present a highly configurable, real-time analysis system to automatically record, analyze and visualize high level aggregated information of user interventions in Twitter that may be used by public entities to better understand citizen necessities. The system is composed by four main components, shown in Figure 1.

The central component is the ***datawarehouse***, the core information repository that is able to store the high volume of data that the system manages and also provides advanced search functionality to be able to exploit the information. The system is based on Elasticsearch [2], which is a flexible and powerful open source, distributed,

real-time search and analytics engine. Its distributed capabilities and the fact that it scales very good when the system grows were key factors in the selection of this architecture. Elasticsearch runs on top of Apache Lucene, so it offers quite complex search capabilities and a scalable and high-performance environment.

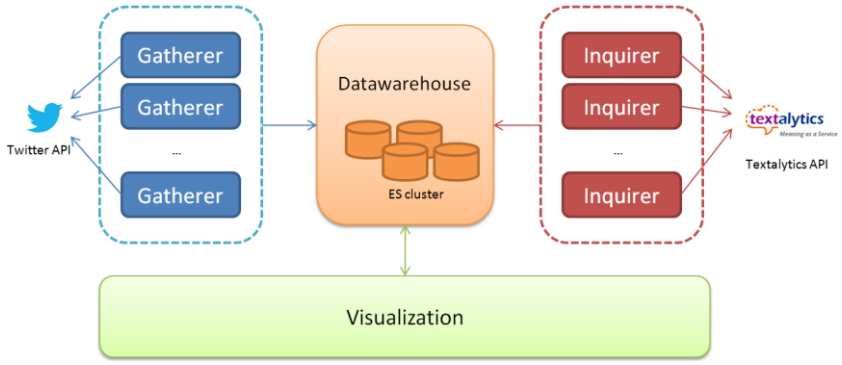

Fig. 1. System architecture

The second component is composed by a set of concurrent ***gatherer*** processes, which query the Twitter APIs [3] to collect tweets regarding to certain filters. The configuration file defines the query parameters to the Twitter streaming API, allowing to filter tweets by a list of user identifiers, a list of keywords to track (terms, hashtags) and/or a set of geographical bounding boxes to restrict the search.

The third component is composed of a set of concurrent ***inquirer*** processes, whose task is to annotate the messages using several of our Textalytics Core APIs [4]. The system is deployed to use the text classification API using two specific models specially designed for this business case (SocialMedia and CitizenSensor, described later), the topics extraction API, which extracts topics such as entities, concepts, money, URI expressions, etc., the sentiment analysis API, which extracts sentiment polarity and also subjectivity and irony indications, and finally, the user demographics API, which currently returns the gender, age and type of the author of the tweet.

Specifically, for each tweet, the system tries to identify the thematic area of the message (energy, transport, economy, politics, social interests…), concepts mentioned (city services, weather...), events to which the text refers (cultural events, soccer matches...), special alert situations (road accidents, fires, street violence, security issues…), and the specific location of the user (a building, means of transport...). This analysis is complemented by an analysis of the sentiment polarity of the message: very positive, positive, negative, very negative and neutral.

An example of an annotated tweet is shown in Figure 2, where a Twitter user alerts from a crash in a public tunnel of the city of Madrid that needed of the presence of the firemen. The system correctly detects that the issue is located in a *public road*, classifies the message in the topic of *Security* in the Citizen Sensor ontology (under *Concepts>Services>Security*) and as *Disasters and accidents* in the general Social Media ontology. Furthermore, it finds out that the entity *Calderón* (soccer stadium nearby) appears in the sentence and also several concepts: *accident*, *tunnel*, *closed*, *firefighter*,

exit, *lane*, etc. Finally, it detects that it is an objective, non ironic comment with negative polarity written out by a male aged in the range of 35 to 65 years.

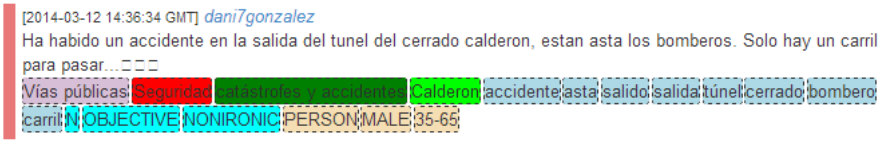

Fig. 2. Example of a tweet annotated by the system

The semantic annotation task is the highest time consuming task and constitutes most of the times the bottleneck of the system. The inquirer processes annotate the unprocessed messages in descending order of insertion time, so that the most recent information is available first to be able to react to early alerts. If the input rate of messages being indexed in the system is higher than the multi-threaded annotation rate, it is still not possible to access the high-level annotations on real time, but once this peak situation is reversed and the system manages to annotate at a higher rate than the indexing of new documents, it will start annotating the rest unprocessed documents.

Finally, the ***visualization component*** is used to exploit the annotated data. Several widgets have been developed to present the data, either just for query and reporting or also for data analytics purposes. These visualization modules can be specifically adapted to better match the city needs.

The datawarehouse and the gatherers are obviously language independent, but the inquirer components are strongly dependent on language lexicons and models. Although the text classification engine is itself language independent, classification models (consisting of training text and rules) are developed for a specific language. The topic extraction engine relies on Part-of-Speech and parsing modules specifically designed to build a sentence syntactic tree in a given language. Moreover, the sentiment analysis engine makes use of that syntactic tree and also depends on a lexicon containing polarity units and modifiers for a given language. The user demographics engine is the only module where no information in a given language is used for creating the model. Our initial business case is deployed to analyze data in Spanish, but modules exist for other languages: English, French, Italian, Portuguese and Catalan.

3 Semantic Annotation

Much effort has been invested in the semantic annotation task, specifically focusing on this scenario, tuned to properly deal with the special singularities of this kind of text snippets (tweets) that usually contain misspellings, emoticons, typographic symbols, letter/number homophones, shortenings, contractions, etc.

The inquirer provides several levels of analyses to classify the text with respect to several (customizable) categories of specifically-defined ontologies, identify topics, perform a demographics analysis to get the user age range, gender and whether he/she is a person or an organization, and sentiment analysis of polarity and subjectivity.

All modules have been exhaustively tested and successfully evaluated in various scenarios, both separately and also integrating two or more modules, in actual systems currently in production, and also in different national and international evaluation workshops such as SEPLN [5], CLEF [6] [7], NTCIR [8] and SemEval [9].

3.1 Text Classification

Another semantic annotation dimension is obtained with an automatic text classification [10] according to pre-established categories defined in a model. The algorithm used [11] [12] combines statistical classification with rule-based filtering, which allows to obtain a high degree of precision for very different environments. Two ontologies were specially designed for this system including concepts and situations that we find relevant to this particular problem; however, the system allows building particular ontologies and classification models for each scenario.

The Social Media ontology defines the general topic classification of the tweet, and contains the first-level categories shown in Figure 3a. The Citizen Sensor ontology, shown in Figure 3b, focuses on features considering the citizen as a sensor.

Category
Art and culture
Police and justice
Disasters and accidents
Economy and treasury
Education
Environment, meteorology and energy
Health
Social issues
Work
Tourism, trips and people movement
Leisure, lifestyle and free time
Politics
Religion and creed
Science and technology
Sport
Disturbances, conflicts and war
Greetings and thankfulness

Category	Subcategory
010000 Locations	010100 Government
	010200 Education
	010300 Justice
	010400 Health service
	010500 Culture
	010600 Sport
	010700 Religion
	010800 Commerce
	010900 Hotel industry
	011000 Outdoor
	011100 Means of transportation
	011200 Accommodation
	011300 Social Institution
	011400 Leisure Centre
	011500 Workplace
020000 Events	020100 Demonstration
	020200 Sport events
	020300 Conference and convention
	020400 Cultural events
	020500 Celebrations
030000 Disasters	030100 Fire
	030200 Explosion
	030300 Landslide
	030400 Avalanche
	030500 Flooding
	030600 Drought
	030700 Earthquake
	030800 Seaquake
	030900 Epidemic / Plague
	031000 Toxic discharge

Category	Subcategory
040000 Accidents	040100 Car crashes
	040200 Plane crashes
	040300 Maritime accidents
	040400 Rail accidents
	040500 Nuclear accident
050000 Criminal acts	050100 Robbery
	050200 Aggression
	050300 Harassment
	050400 Rape
	050500 Abuse / Mistreat
	050600 Kidnapping / Disappearances
	050700 Gunfire
	050800 Murder
	050900 Attempt
	051000 Drug trafficking
	051100 Intelectual property
	051200 Prostitution and Pederasty
060000 Medical emergencies	060100 Pregnancy and childbirth
	060200 Decease
	060300 Suicide
	060400 Drug addiction
	060500 Infarction
	060600 Asphyxiation
	060700 Intoxication
	060800 Injury
	060900 Burn
	061000 Fainting
	061100 Epilectic seizure
	061200 Electrocution

Category	Subcategory
070000 Weather warnings	070200 Snowfall / Frost
	070300 Heatwave
	070400 Cold snap
	070500 Storm
	070600 Hurricane / Tornado
	070700 Wind
080000 Incidents	080100 Traffic congestion
	080200 Public road damage
	080300 Interruption of supplies
	080400 Caution at the beach
090000 Concepts	090100 Lighting
	090200 Signposting
	090300 Air-conditioning
	090400 Supplies
	090500 Services
	090600 Acoustic environment
	090700 Odoriferous environment
	090800 Environment
	090900 Quality of life

Fig. 3. a) Social Media ontology; b) Citizen Sensor ontology (1st and 2nd level categories)

3.2 Topics Extraction

Topics extraction process is carried out by combining a number of complex natural language processing techniques that allow obtaining morphological, syntactic and semantic analyses of a text and using them to identify different types of significant elements. In short, the text is first divided into paragraphs, sentences and tokens, and then each token is lemmatized and tagged with its Part-Of-Speech. A rule-based parser in a series of sequential steps creates the sentence syntactic tree, detecting and tagging the existing coordinated and subordinated clauses, word groups and dependencies among them, and also recognizing named entities and concepts, based on both language resources and also language dependent heuristics (such as `[Mr.|Sir|Dr.]+NAME=>PERSON`). This process also carries out a disambiguation step for the morphosyntactic and semantic information of each token and also anaphora detection and resolution for sentence interlinking.

Currently the system is able to identify (allowing word inflections, variants and synonyms) the following topic categories: named entities (people, organizations, places, etc.), concepts (significant keywords in the text), time expressions, money expressions and URIs.

3.3 Sentiment Analysis

The system also includes functionality to perform a detailed multilingual sentiment analysis of texts from different sources. The text provided is analyzed to determine if it expresses a positive/negative/neutral sentiment polarity. First [6], the local polarity of the different sentences in the text is identified and the relationship among them is evaluated, resulting in a global polarity value for the whole text. Besides polarity at sentence and global level, natural language processing techniques also detect the polarity associated to both entities and concepts in the text (aspect-based polarity).

Moreover, although perhaps not very useful in this context, the sentiment analysis module can also detect if the text processed is subjective or objective and if it contains irony marks, both at global and sentence level, giving the user additional information about the reliability of the polarity obtained from the sentiment analysis.

3.4 User Demographics

The user demographics analysis module extracts some important demographics (type, gender, age) for a given Twitter user. State-of-the-art information extraction and text classification algorithms are used to guess those facts from his/her login, name and profile description, based on n-grams model, developed using Weka [13].

4 Visualization

The visualization is a web interface that allows to easily building complex queries in a structured way, enabling, thus, a versatile filtering of the data and high level visuali-

zation with the aim to provide the final user with a highly aggregated and condensed information at a first sight. The system is designed to provide both real time analysis and backtracking of previously stored data.

The system console is created defining several elements called widgets, in such a way that the template may be changed between different user cases (different cities and their particular needs) to adapt the system to each community.

Some of the components make use of the Highcharts JavaScript library [14] to create intuitive and interactive charts, OpenLayers [15] to display maps and geoposition information, as well as self-customized components. The user interface makes use of the capabilities of Elasticsearch, allowing the user to create their own queries by filtering on the semantic tags and aggregating information using its Facets API.

An example of an analysis dashboard using some of the built widgets is shown in next figures. Figure 4 shows filter capabilities, analysis of total number of tweets and alerts as well as last minute tracked tweets and alerts, a timeline tracking the number of tweets and alerts per minute, as well as the number of positive and negative ones.

Fig. 4. Dashboard with filters, statistics and timelines

Figure 5 presents several pie charts with user statistics (number of users by age range and gender), global sentiment polarity, and a list of the most frequent alerts, locations and events.

Fig. 5. Dashboard with user demographics, sentiment polarity, alerts and events

The console also displays a map with the locations of the alerts that contained this information and also includes the semantically annotated tweets that match the filtering criteria (Figure 6).

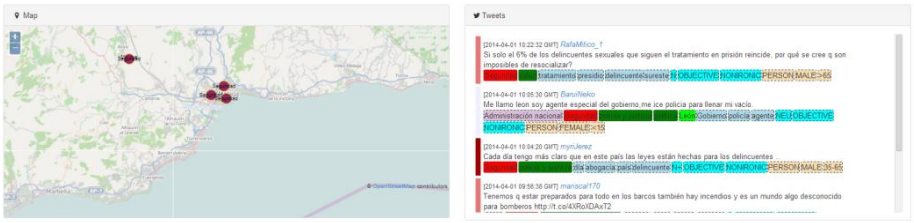

Fig. 6. Dashboard with tweets and map

Last, Figure 7 shows some widgets with tag clouds listing the most relevant (by number of appearance) topics, entities, concepts and hashtags.

Fig. 7. Dashboard with tag clouds of topics, entities, concepts and hashtags

5 Discussion and Future Work

In this work a real time semantic annotation engine for Twitter data with datawarehouse capabilities and a search engine for backtracking and later data analytics has been described. The system allows community promoters to more quickly react to specific events that may happen (catastrophes, accidents, traffic congestion, etc.), react to people feelings and detect which initiatives are more likely to be improving quality of life for their citizens, to detect the topics that are worrying the citizens... Thus, it will increase the degree of engagement of the smart cities that use the system with their citizens.

Currently the system in beta-testing process, adapting the interfaces, fine-tuning the different modules and removing noise in the annotations. The system will be deployed in different scenarios in a short or medium term. There are several business cases under negotiation. The first scenario is to build a city console for a local administration to be able to analyze in real-time the behavior and topics of interest of the citizens, with two components: a private console, internal for the city services, and a public console, a dashboard with attractive, summarized, non-confidential information to be projected or displayed at selected public locations of the city (town hall, librar-

ies, museums) or even in a LED video wall in a populous square in downtown, to engage citizens with these technologies and also promotion. The second scenario is to focus on emergencies services, providing early detection of security-related issues.

Regarding the technology, the storage capabilities of the system allow not only to analyze real time data, giving a snapshot of the current city state, but also to apply data mining algorithms to the stored data in order to better understand particularities of the population, clustering and profiling of the different groups that form the city environment, compare the singularities of the different detected clusters, etc. Currently, steps to further explore this path are being taken: city mobility analysis (how, when, why people move from one place to another), relevant topics analyzed at neighbourhood level, city reputation and brand personality, etc.

Finally, the same approach that has being used analyzing Twitter data will be used with other sources of information. The gatherer will be extended to capture data from other sources like other social network like Facebook, LinkedIn (in smart-city related groups), Tuenti, or other social sites such as YouTube, Flickr, Pinterest, etc. In addition, we want to better adapt our core models for NLP to the special features that Social Networks language introduce.

Acknowledgements. This work has been supported by several Spanish R&D projects: Ciudad2020: Hacia un nuevo modelo de ciudad inteligente sostenible (INNPRONTA IPT-20111006), MA2VICMR: Improving the access, analysis and visibility of the multilingual and multimedia information in web for the Region of Madrid (S2009/TIC-1542) and MULTIMEDICA: Multilingual Information Extraction in Health domain and application to scientific and informative documents (TIN2010-20644-C03-01).

References

1. Ciudad 2020 - Hacia un nuevo modelo de ciudad inteligente sostenible. Website. http://innprontaciudad2020.es.
2. Elasticsearch.org. Open Source Distributed Real Time Search & Analytics. http://www.elasticsearch.org
3. Twitter REST API v1.1. https://dev.twitter.com/docs/api/1.1
4. Textalytics API. http://textalytics.com
5. Díaz Esteban, A., I. Alegría, and J. Villena-Román (eds). Proceedings of the TASS workshop at SEPLN 2013. *Actas del XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural*. IV Congreso Español de Informática. 17-20 September 2013, Madrid, Spain.
6. Villena-Román, J., S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J.C. González-Cristóbal. 2012. DAEDALUS at RepLab 2012: Polarity Classification and Filtering on Twitter Data. *CLEF 2012 Labs and Workshop Notebook Papers*, Rome, Italy, September 2012.
7. Villena-Román, J., and S. Lana-Serrano. MIRACLE at VideoCLEF 2008: Topic Identification and Keyframe Extraction in Dual Language Videos. 2009. *Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language*

Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Carol Peters et al. (Eds.). Lecture Notes in Computer Science, Vol. 5706, 2009.

8. Villena-Román, J., S. Lana-Serrano, and J.C. González-Cristóbal. 2008. *MIRACLE at NTCIR-7 MOAT: First Experiments on Multilingual Opinion Analysis*. 7th NTCIR Workshop Meeting. Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. Tokio, Japón, December 2008.
9. Villena-Román, J., J. García-Morera, and J.C. González-Cristóbal. 2014. Daedalus at SemEval-2014 Task 9: Comparing Approaches for Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval'14, Dublin, Ireland, 2014 (to be published).
10. Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), pp 1–47.
11. Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Método híbrido para categorización de texto basado en aprendizaje y reglas. *Procesamiento del Lenguaje Natural*, Vol. 46, 2011, pp. 35-42.
12. Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11)*, May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press 2011.
13. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1.
14. Highcharts - Interactive JavaScript charts for your webpage. JavaScript library website. `http://www.highcharts.com`
15. Openlayers. `http://openlayers.org`