PaloAnalytics: project concept, scope and early results from the system implementation

Vassilis Poulopoulos*, Manolis Wallace*, Iraklis Varlamis*, George Caridakis†and Panagiotis Tsantilas‡

*Knowledge and Uncertainty Research Laboratory
Department of Informatics and Telecommunications, University of the Peloponnese, Tripoli, Greece
Email: {wallace, vacilos}@uop.gr, varlamis@hua.gr
†University of the Aegean, Mytilene, Greece
Email: gcari@aegean.gr
‡Palo Ltd., Kokkoni, Greece
Email: pt@paloservices.com

***Abstract*—This paper describes the national funded project entitled PaloAnalytics, which develops an innovative platform that allows companies and organizations, which operate in many countries, to monitor and analyze, in depth, the markets interest to their products and successfully plan their marketing and communication strategy, with data and insights collected from all the local media. PaloAnalytics platform will cover the need of international companies to manage their reputation and compare it with their competitors. It also allows them to investigate the impact of their products on consumers across different countries and this is achieved with the analysis of content from sites, blogs, social networks and open data. The developed services allow companies to identify both positive and negative comments and reports about their brand name and products and the individual features that formed the public opinion. The project partners explore, design and develop a range of algorithms and tools: 1) to collect and manage large amounts of data (text, multimedia and links) from online news sources, social networks and open sources, 2) to extract knowledge from textual references e.g. emotions, trends, influences, impact, results and interactions in a multilingual environment, 3) to link exported knowledge together and present it using infographics and user-friendly visualisations. In this paper, we furthermore focus on the architecture of the system that perform trending topics extraction, presenting some early results from its implementation.**

***Index Terms*—big data, data monitoring, trending topics, influencers, info graphics, data visualization, deep learning**

I. Introduction

The data that is generated daily in the world of the internet is vast. The amount of information is such that it is impossible for companies and organizations to fetch, analyze and learn from all the data produced. In this scope, PaloAnalytics is a project that aims to perform the procedures of collecting, analyzing and extracting useful information from different sources of the internet, web pages, news portals, open sources, and social media. The procedure of collecting and analyzing information from diverse sources is not something new, and has attracted research during the last 20 years [10]. It resides to the area of Data Mining [12], [11] and it focuses on Big Data analysis, which is based on multiple custom Data Warehouses [13]. In this notion, we present a project that intends to employ resources from all these sectors in order to produce its final results; in depth analysis of social media and web data in order to support organizations and companies.

Market research has proven that companies and organizations are in strong need for an holistic market monitoring and analysis service in several countries and not solely the country of their origin; or at least they are convinced that they can perform much better if they have such a tool. Besides, the competition of such companies and organizations is usually international. Furthermore, it clear that when analyzing data in an international environment each local information can easily affect the whole organization, but it is usually difficult to become a part of the organization's international policy. In general, it seems possible that data can be collected and analyzed in some extent locally but is usually not transferred as knowledge to the international level. In fact, for such organizations it would be extremely useful to utilize a unique language for all the data analyzed and it seems that the English language is acceptable and consistent. A number of tools have been developed including Mention [1] and Brandwatch [2] in order to collect and analyze data internationally but they have some major disadvantages. They focus mainly on social media and target experienced users, while in parallel they do not provide translations of reports from local languages in a universal language. Furthermore, they do not offer a homogeneous overall picture for all the countries that are of interest for a business.

In this ground, we introduce PaloAnalytics project, which intends to focus on the basic challenges that organizations face and includes the ability to have a universal monitoring tool, with links and interconnections between data collected and analyzed from a number of different sources and different

Acknowledgment: This research has been cofinanced by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE INNOVATE (project code:T1EDK-03470)

[1] https://mention.com - Mention: Scour the web, social media, and more for powerful market insights

[2] https://www.brandwatch.com/ - Brandwatch: Know what your customers think

languages. In this way the project will be the ideal solution for international companies (or companies willing to become international) and companies that their international competition affect their local business. An ideal solution, through which the organization will be able to get information out of large sets of data.

The proposed design and implementation, introduces a series of software modules that will

- analyze multilingual content posted on news sites, social networks and open data
- extract knowledge and information about products and companies, including product characteristics
- analyze sources, their influence and trends
- help in assessing the image of the business and its products as well as its competitors
- visualize the knowledge in order to easily understand the analyzed information

A number of cutting-edge technologies is employed in order to achieve all the aforementioned. Considering the unified manipulation of multilingual content, deep neural networks will be used, as they have successfully been employed in multilingual models of knowledge transfer for speech-to-text applications [14], [15]. Furthermore, deep learning algorithms have been recently used for sentiment analysis on multilingual texts with high levels of success. The selection of the English language is done in order to have a single reference language, which will enable the possibility of exploration for translation using machine learning algorithms. Due to the absence of standard text formats, or practices that are used for parallel posts in many different languages, employment of data mining techniques in a multilingual environment is not possible. This means that language agnostic techniques cannot be considered sufficient to overcome the problem of a multilingual environment. In this scope, the project utilizes translation techniques within the procedure of data mining.

Another important factor for large companies and enterprises is their influencing indicator. While examining the influence of enterprises on social networks and the web, the system uses dynamic opinion models. In this manner, it will be possible to attach opinion polarity and trust that is usually formed in data deriving from social networks. In order to record the extracted knowledge and associate it with open sources of data, techniques including Association Link Networks will be used. The latter allows definition of semantic linkages between entities extracted from texts. Finally, the architecture will be based on distributed models in order to support the large amount of data that will be generated from the data mining and analysis.

A critical part of the system is related to analysis of trending topics in social media as they reveal places with large audience. In this paper, we focus on the mechanism that handles the trending topics deriving from Twitter and we analyze its architecture and the early results from its implementation.

The rest of the paper is structured as follows: Section II presents the methodology of the project, while section III discusses the system architecture. In section IV a detailed description of each component is presented, providing more emphasis on the Trending Topics software module and its results. Section V defines the expected outcomes of the proposed system and the final section presents a discussion on the project.

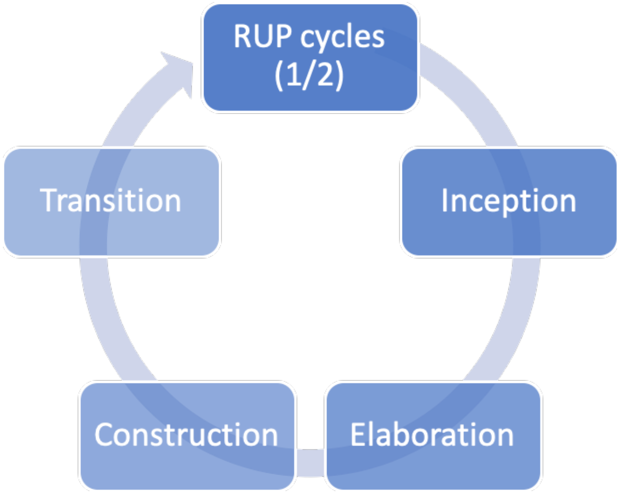

Fig. 1. Rational Unified Process

II. Methodology of the Project

Within the scope of the project it is obvious that due to the large number of different modules, the high complexity of their implementation and the importance in precision of their algorithmic procedures, an advanced methodology is employed. As such, the Rational Unified Process is used. It is a software engineering procedure that ensures producing high quality software and achieving end user needs within a specific timetable and cost. Two cycles of project evolution are followed, one that leads to the basic implementation and is longer, while a shorter one will be done in order to perform refinements. Both of the cycles will go through the same steps of development. During the first cycle the implementation will be ensured, while the second cycle will focus on the quality of the outcomes. The cycle phases include:

- Inception Phase
- Elaboration Phase
- Construction Phase
- Transition Phase

Figure 1 depicts the cycle of system design, implementation and integration.

During the inception phase a general description of the key requirements of the project is done; key points and the basic constraints are defined and the system use cases are defined in brief. An initial business case including the business framework, the success criteria and financial forecasting is the ones that lead to the project plan and to a draft business model. While analyzing the information, during the processing phase, the use case model was completed, and the final requirements were recorded. The architecture reached its final form and the project's development plan was finalized. Currently the project is under the first construction phase, where modules are implemented and starting to be integrate into the PaloAnalytics platform. Upon completion of the first phase of implementations an overall system functionality, performance and usability test will be done.

As the outcomes of the project are consumer products, at the end of the first development cycle, each individual service will be communicated to real testers (customers) for evaluation and feedback. The end of the first phase concludes with collection of feedback and system evaluation which will be the input of the second cycle of the system. The scope of the second cycle is the overall platform improvement in order to meet production requirements and high quality results. As a research project it is expected that primarily some parts of the modules will be implemented in order to meet their research scope. Nevertheless, the final outcome of the project is a production level software that should be able to produce high quality results within specific time.

The development of the platform follows a bottom-up approach, based on the proposed architecture as presented in figure 2), starting from data collection that will directly lead to data aggregation services which will be used individually. On the produced data, multilingual content analysis' services are employed, while in parallel, at this stage, business intelligence extracting solutions are applied. The availability of the proposed services will be both on Web and Mobile application enabling increased penetration into the business community. Each service is built supporting endpoint integration in order to be available for use as an individual component even for third party systems, external to PaloAnalytics platform.

This will develop a complete development stack, that will be based on multilingual content from news sites, open data sources and social media. The services of this stack are expected to attract third-party businesses companies, public bodies and researchers who will develop new management modes of business data from the sources incorporated by PaloAnalytics platform and will set up new business models on them, multiplying the benefits for the companies and organizations while maximizing the influence of the proposed solutions for the scientific and business community.

III. Architecture

According to the architecture presented in figure 2 the proposed system is divided into several components and modules enabling in this way individual design and integration. The system, though, can be separated into four major components:

- data entrance point / data storage
- deep data analysis
- semantics and metadata analysis
- point of presentation

Each of the major components consists of a number of modules in order to successfully achieve its scope. Furthermore, each component will offer services for direct data extraction and usage by third party systems.

A. Entry point

The entry point of the system is the component that is responsible for collecting and storing data from the several different sources (social media, news and open data). The data storage is built enabling several interfaces to be connected in order to fetch and store data. In general, it follows a hybrid scheme including both an SQL and a noSQL database.

The system acts as a data warehouse, including modules for data extraction, data transformation as well as data loading. The extraction of data is done from several different sources including news websites, blogging platforms, social media - focusing on text based ones - and open data sources. The data collected is transformed in order to formulate similar objects with specific unified structure. The unified structure of each unique object includes

- unique identifier
- title
- body
- source / link
- timestamp
- author

The aforementioned is the main object of the system and described the main form of data collected. A number of metadata and objects analyzing in depth each object is used including detailed information about the source, the author, accompanying multimedia and more. Figure 4 presents a generic schema of the database infrastructure that is used in order to support the essential for the system storage.

The data collected are stored on both an SQL-like storage environment as well as a noSQL environment. The hybrid scheme will help for storing elements for fast access in the noSQL nodes and collection of all the collected data in an SQL based structure for better interconnection between them and permanent storage of data with historic metadata [7]. Furthermore, a time-series database is used in order to keep track of the records that are stored in the database, including information about the source or the author. The latter is extremely useful when defining the rate of update for each source or the frequency of posting for authors and their relation to period and time.

B. System Core

The system core contains all the key elements and services of the system. It consists the basis upon which the complete system is designed and implemented. Each of the modules formulating the system core can act as an autonomous system providing endpoints for independent usage. These endpoints can also be used by the system internally in order to perform the physical interconnection between the different services.

The system core includes the following elements:

- **Named Entity Recognition (NER)**, which is a module for recognizing entities in bunches of text. A machine learning mechanism based on OpenNLP [3], a set of lan-

[3] OpenNLP: a machine learning based toolkit for the processing of natural language text. https://opennlp.apache.org/

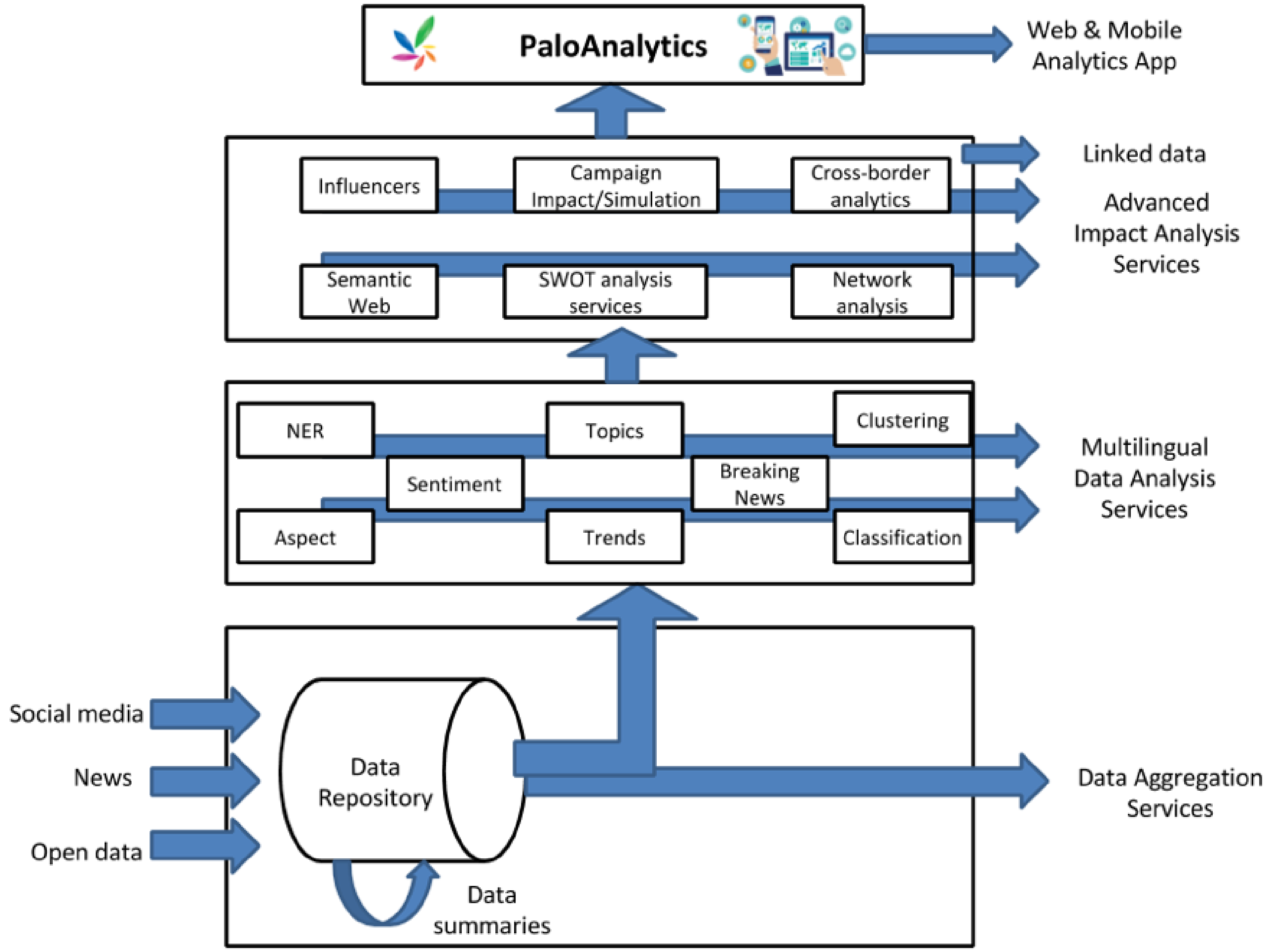

Fig. 2. Proposed architecture

guage features and a set of annotated documents for finding candidate NERs, enhanced by the use of dictionaries is used. Four models are trained for each language with annotation per person, location, product, organization. It should be considered that a larger training dataset is automatically created using frequent word sequences and Named-Entities replacement. Transfer learning and reinforcement learning techniques are tested as there is a large corpus in which we can instantly identify well-known named entities and isolate their environment [5].

- **Breaking news detection**, which is a component for recognizing important news topics. This is usually based on the number of similar articles produced in a period of time, but it should be considered that not all news topics are increased in numbers in the same manner. As such, machine learning algorithms are employed that are able to recognize breaking news based on the growth rate in time [6].
- **Clustering**, which is responsible for finding interconnections between the different entities. It should be noted that the objects collected derive from several different sources and the scope of this module is to create physical interconnections between objects having identical meaning. According to the definition of the object (without any attached metadata) the main scope of the clustering procedure is to interconnect conceptually two objects. Furthermore, as the system intends to operate regardless of the language of origin, the interconnection of the object should be language agnostic.
- **Classification**, which is a module for automatic categorization of objects to predefined categories. As the categories of the system are predefined, due to the fact that Palo is used as a news aggregation service, the categorization is done in several primal categories. The current mechanism will be enhanced in order to enable multilevel categorization including two different levels [4].
- **Sentiment Analysis**, which is responsible for extracting the polarity of the objects. A machine learning algorithm will be employed in order to replace a currently used

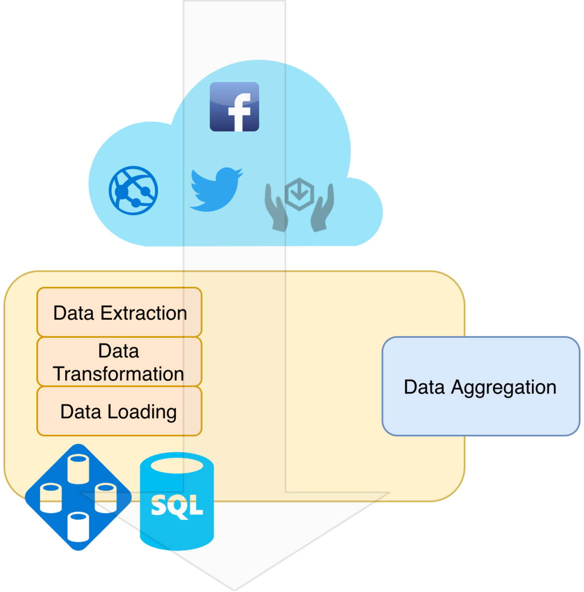

Fig. 3. Entry point component

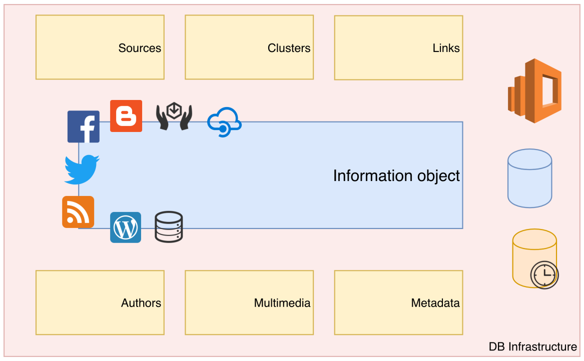

Fig. 4. Generic database scheme

algorithm based on the bag of words method [1].

- **Summarization**, which is responsible for extracting summaries out of the clusters of objects. As the clustering procedure evolves in time, the summarization procedure must adapt to changes that are done to the size of the cluster in time.
- **Trending topics detection and enrichment**, which is responsible for analyzing social media and open sources in order to detect topics that are trending and enrich them accordingly in order to detect their trends to other countries and languages.

C. *High level analysis*

The high level data analysis of the system includes a number of components that combine the outcomes of the deep data analysis and they include:

- Discovering social media influencers [2], [8]
- Applying cross-border analytics [3]
- Performing network analysis
- Exploring semantic means of the web [9]
- Simulating web and social media campaigns and measuring their impact

The aforementioned are only part of the high level analysis that can be achieved and are explored as part of the business needs and requirements. ++

D. *Frontend*

The system frontend consists of both web and mobile applications that utilize the data collected and analyzed in order to present reports, visualize data and make it easy to explore the combined information.

The web and mobile applications will have a public part that will make parts of the collected available to public. This is a news aggregation service including rich media format of data as well as interconnection of information and multilingual content. The same is for the mobile application which ca be formulated in order to enhance portability and usability of the presented content.

IV. Project Implementation and Early results

The project's implementation includes several different levels of design and implementation as it includes a large number of sub-modules that have to be integrated and work together in order to produce the final result. The central place of interaction between the components is the database that handles all the data, while each component has its own individual and independent cycle of execution. In general the project is separated into 5 different groups of implementations: the first one includes the definition of requirements, analysis of users and use cases, setting the KPIs (Key Performance Indicators and the business logic, the second includes the definition of the architecture and employment of the infrastructure that will support the system, the third part is related to core services including mainly the algorithmic part of the several systems, while the fourth part is the high level services that utilize the core of the system. Finally, the last part of the system is the execution, alpha and beta testing, definition and execution of PoC (proof of concept) scenarios as well as production environment usage. The aforementioned are necessary as the outcome of the project is a consumer software that needs to comply with the needs of organizations and businesses and should not only produce scientific results.

As the project is close to its first year of implementation some early results from services implementation are already produced. The service related to handling twitter's trending topics is presented in the next paragraphs.

A. *Handling Twitter's Trending Topics*

As the implementation for each of the subsystems advances individually and independently, we will emphasize on the component that tracks the topics in social media, specifically Twitter[4] platform, and describe how the system is able to enrich them and generate universal topics. A part, on which the system relies, is the information that is delivered back to the

[4] Twitter: https://twitter.com

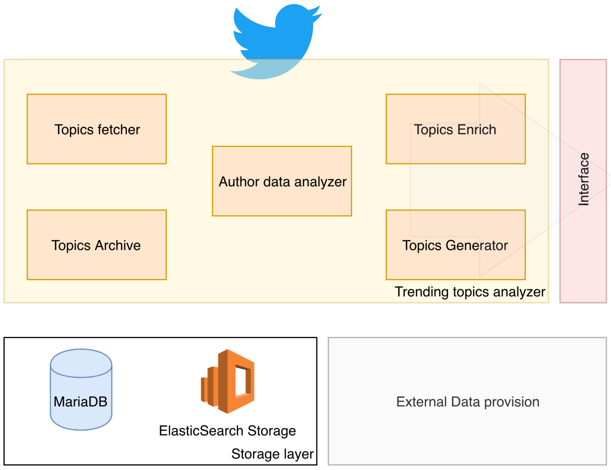

Fig. 5. Twitter trending topics component

end users. Despite the fact that twitter is able to provide its list of trending topics, it does not provide any evidence about the soundness of the topics nor the origin of the trends. Moreover, it is possible that the topics are not directly related to a company or organization, but there can certainly exist some kind of semantic interconnection that can possibly transform each trending topic to a useful marketing tool for them. The system is able to perform in depth analysis in order to extract knowledge from the list of trending topics already provided by the medium.

Figure 5 presents the component that is used in order to handle trending topics deriving from Twitter. The component is implemented by 5 main sub-modules:

- Topics fetcher
- Topics archive
- Author handler
- Topics Enrich
- Topics Generator

By using twitter's endpoint[5] it is possible to retrieve all the trending topics related to a specific place. In our case it is possible to retrieve topics for as many places as it takes in order to retrieve all information related to the countries of interest to organizations and businesses using the system. All the aforementioned is a responsibility of the **topics fetcher** module. As the topics are only updated not frequently than 5 minutes, the fetcher module is executed 5 times per hour in order to get several updates on the topics. The trending topics can be either simple words (eg. Jennifer Lopez) or hashtags (eg. BlackFriday). Hashtags are widely used in social media and are a keyword or a whole phrase used to describe a topic or a theme. It is written in one word - even it is a phrase of words - and it starts with the sign # . As the API is able to provide a set of accompanying information such as queries that lead to the trending topics as well as volume of the topic the last 24 hours, the fetcher stores all the provided data.

Topics archive is responsible for keeping a track record of all the topics, in order to monitor the rate of change to their volume in the different media. The differentiation of this module to conventional modules searching for changes in trending topics volume is the fact that our mechanism tracks the volume of the topics to all the data-sets retrieved from the system (all social media, websites and open data) in order to check the reflection of the change in volume from Twitter to all the data-sets of the system. To achieve this, the mechanism has access to "external" data, which is actually all the data that PaloAnalytics' fetcher has fetched within a period of time. The archive is a way to "cross-check", on the one hand, the soundness of high volumes of twitter trends as reflected to other sources, as well as store and monitor the duration of the "trending period" of a topic.

Another crucial factor that plays an important role for the system is the sub-module responsible for analyzing the **authors** of the data. Referred usually as "influencer", it is actually the users whose posts affect a large audience. This module tracks the records of the users, the reactions of other users towards data posted and prioritizes the users according to their affection to others. It is a means of recognizing influencers, but it is actually a method to recognize and measure the actual weight of each trending topic. The more the "influencers" related to a topic, the more importance gained from its extraction.

By using the knowledge from the aforementioned mechanisms the next one, **topics enrichment**, is responsible for enriching the lists of current extracted topics with extra data. The extra data are separated into two main categories. The first one are topics that accompany the trending ones, while the second category are topics considered to be synonyms or semantically related to the current trending topics . Finally, the last mechanism used within the trending topics module is the **topic generator**. As the system is intended to be used internationally, the system must be able to use the enriched trending topics in a specific country and language, to generate trending topics as they can occur in another country at the same time. It is not just a matter of translating current topics into several languages but recognizing the "trending" character of each term. Inevitably it is expected that for certain situations there should be common trending topics across different countries. In this case, the system is able to recognize the common topics and make interconnections between data collected in different languages.

B. *Experimental approach*

For the experimental approach of the system regarding the twitter trending topics, we constructed predefined data-sets are utilized, and the system is analyzed towards achieving interconnections between the data. By using the Twitter's endpoint that can retrieve the complete set of data posted in

[5] Twitter API: Get Trends per Place. https://developer.twitter.com/en/docs/trends/trends-for-location/api-reference/get-trends-place.html

TABLE I
COMMON TRENDING TOPICS BETWEEN GREECE AND UK

Trending Topic	Volume GR	Volume UK
#brexit	386	89376
#europe	2975	67308
#article13	504	98366
#princecharles (#$\kappa\alpha\rho o\lambda o\varsigma ingreek$)	663	103666

Twitter [6], called Twitter Stream, we are able to get all the data posted in Twitter for a specific country and language. In this way, we are able to retrieve an "offline copy" of the medium. The predefined data-sets consist of data collected for different countries and languages while the two of them were sharing common trending topics (eg. Visit of the Prince of England in Greece, leads to common trending topics in Greece and UK). In parallel, we keep an historical record of the trending topics in the two countries in the period of time that the offline copy of the data is fetched.

After collecting all the aforementioned data it is possible to examine the performance of the proposed solution. The performance is examined by making interconnections between the trending topics in both languages, and checking for semantic connection in the linked objects. The procedure was evaluated by humans using sub-sets of the linked results. In the case of an event that concerned both countries it was obvious from the results that direct linkages can be done as common Named Entities were used. In this specific scenario it was a matter of the NER sub-system to provide evidence about the common data. On the other hand it was obvious that semantic linkages were also made to data that are posted to both countries mainly because the Media and the Social media users in both countries share common concerns. The experimental procedure revealed linkages between the two countries in several topics as depicted in table I. It is obvious that the only change is the order of each topic and the volume. The order of the topic reveals a differentiation on the importance for the people, a measure that needs to be recorded, while the volume is just indicative as it derives from the amount of users in each country.

We furthermore examine the behaviour of the users posting data in both countries by checking the volume of data related to a common topic. It is expected, as already mentioned, that the difference in volume is expected due to the difference in numbers of users, but it is interesting to view if the users in both countries follow similar behavior in the change rate of the volume. At first we collect information about a trending topic that appears as trending to both languages, the one in Greek and the second in English. From figure 6 it is obvious that a topics that concerns both countries at the same time appears in both of them at the same with similar change of volume of data including the term.

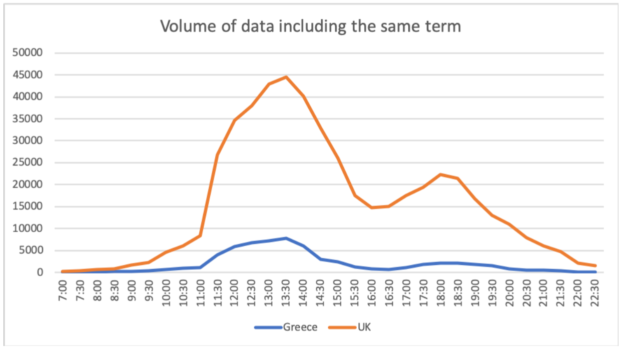

Fig. 6. Volume of data including the same topic

By recognizing the same pattern of volume change it is possible to understand that the behaviour of the users of the medium is similar. The latter leads to the assumption that the data are not only semantically connected but it is obvious that the behavior of the users in the two different countries is similar, an important factor for the marketing and knowledge acquire for businesses and organizations. Handling Twitter's trending topics is a part of the modules used in the system. We described the approach and architecture of the module as one of the modules leading its first version in the system and presented the method for evaluating its performance in different languages.

V. EXPECTED OUTCOMES

The design and development of the proposed system consists of a new and innovative product for the international market, which is expected to be the attraction for many companies and organizations primarily organizations that operate internationally. The absence of specialized competitive products in this field offers a significant advantage and allows it to be a leading player in the Greek market, which is the country of origin, and to penetrate the emerging and demanding international market of high-volume data analysis technology by providing innovative services and products.

All the aforementioned, is expected to provide a new dynamic in the field of application development in the referred emerging sectors. This is achieved by using state-of-the-art technologies and methodologies together with the extensive knowledge in the field by the partnership. At the same time, within the framework of the proposed project, the know-how acquired in the areas of large volume analysis is fully exploited, thereby enhancing the company's policy towards the increased use of cutting-edge technologies, as well as the partnerships' research background. Finally, we should consider the valuable know-how acquired by all the participating bodies during the implementation of the proposed project through the two research organizations, which will be done by the research and development in order to achieve the desired objectives. The know-how to be transferred will improve all the organizations' and especially the company's scientific potential by increasing its knowledge and expertise and consequently the company's capabilities for future support as well as developing

[6] Filter Real-time Tweets: https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html

new applications and undertaking new research projects in the context of its activities.

VI. Discussion

We presented the project PaloAnalytics, which has already reached its 9th month of undergoing. During this period the first crucial steps have been made, including the definition of the system use-cases, the formulation of the system architecture, the set-up of the system infranstructure, as well as the design and initiation of the first system components. Furthermore, the business-case is completed and the implementation of the first sub-systems is almost finalized. The infrastructure of the system is set-up and the means of integration are defined. An interesting feature of the project is the participation of two research laboratories from two different institutions in Greece, which will join their research teams to produce the results of the project. In order to achieve the objectives of the project, cutting-edge technology and algorithms are used, which means that the participants will join forces towards the research.

Despite the fact that the actual outcome of the project is minimal compared to the algorithmic procedures that lead to it, a number of related research fields will be explored during the design and implementation of the components. First of all, data mining algorithms will be researched in order to produce the optimal solution for fetching data. Furthermore, the infrastructure that stores the data is the basis of the system and as such its design and integration is part of a research and development procedure. On the other hand, a number of algorithms and techniques including deep machine learning will be investigation in order to achieve procedures listing: clustering of data (including text objects deriving from social media), summarization of clusters, named entity recognition, sentiment analysis, aspect mining and breaking news definition. Furthermore, apart from the core algorithms, a number of "high level" procedures are required in order to achieve the complete set of project scopes. These include influencers mining, semantic web, network analysis, campaign impact, swot analysis and more, which are based on the metadata that accompany the information collected and processed.

It should be noted, that all the aforementioned are not just part of a research procedure; meaning that the research should not stand on the feasibility and soundness of the results. The system is a production based environment targeting large business and organizations, which can even test and formulate the procedures and the use-case scenarios. It lies on the ground of applied research and it is expected that all the implemented solutions will be able to endure large volumes of data, users and demanding procedures.

Finally, in this paper the approach to handling Twitter's trending topics was presented as a means to examine the approach to the module's integration. Despite the fact that it seems like the module is created as an autonomous entity - which is the point of implementation - it is clear that it uses and connect data collected from the system as a whole and produces data that are stored for usage by the complete system. Moreover, the experimental procedure that the mechanism is currently undergone was presented and explained. The first early results from the module's integration were presented, including a proof of common user behaviour in two countries.

Concluding, the paper is a means of analyzing the procedures of a complex and innovative project that intends to introduce an innovative product with international usage.

References

[1] Castellano, G., Kessous, L., Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In Affect and emotion in human-computer interaction (pp. 92-103). Springer, Berlin, Heidelberg.

[2] Caridakis, G., Karpouzis, K., Wallace, M., Kessous, L., Amir, N. (2010). Multimodal users affective state analysis in naturalistic interaction. Journal on Multimodal User Interfaces, 3(1-2), 49-66.

[3] Vlachostergiou, A., Caridakis, G., Kollias, S. (2014). Investigating context awareness of affective computing systems: a critical approach. Procedia Computer Science, 39, 91-98.

[4] Varlamis, I., Tsirakis, N., Poulopoulos, V., Tsantilas, P. (2014, October). An automatic wrapper generation process for large scale crawling of news websites. In Proceedings of the 18th Panhellenic Conference on Informatics (pp. 1-6). ACM.

[5] Makrynioti, N., Grivas, A., Sardianos, C., Tsirakis, N., Varlamis, I., Vassalos, V., Poulopoulos, V. Tsantilas, P. (2017). PaloPro: a platform for knowledge extraction from big social data and the news. International Journal of Big Data Intelligence, 4(1), 3-22.

[6] Varlamis, I., Hilliard, D. F. (2017). Finding influential sources and breaking news in news media using graph analysis techniques. International Journal of Web Engineering and Technology, 12(2), 143-164.

[7] Tsirakis, N., Poulopoulos, V., Tsantilas, P., Varlamis, I. (2017). Large scale opinion mining for social, news and blog data. Journal of Systems and Software, 127, 237-248.

[8] Margaris, D., Vassilakis, C., Georgiadis, P. (2018). Query personalization using social network information and collaborative filtering techniques. Future Generation Computer Systems, 78, 440-450.

[9] Bampatzia, S., Bravo-Quezada, O. G., Antoniou, A., Nores, M. L., Wallace, M., Lepouras, G., Vassilakis, C. (2016, September). The use of semantics in the CrossCult H2020 project. In Semanitic Keyword-based Search on Structured Data Sources (pp. 190-195). Springer, Cham.

[10] Levy, A., Rajaraman, A., Ordille, J. (1996). Querying heterogeneous information sources using source descriptions. Stanford InfoLab.

[11] Hand, D. J. (2006). Data Mining. Encyclopedia of Environmetrics, 2.

[12] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

[13] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

[14] Ghoshal, A., Swietojanski, P., & Renals, S. (2013, May). Multilingual training of deep neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7319-7323). IEEE.

[15] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., ... and Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine, 29.

[16] Violos, I., Tserpes, K., Varlamis, I., Varvarigou, T. (2018). Text classification using the n-gram graph representation model over high frequency data stream. Frontiers in Applied Mathematics and Statistics, section Mathematics of Computation and Data Science Journal. doi: 10.3389/fams.2018.00041