# A platform for real-time opinion mining from social media and news streams

Nikos Tsirakis, Vasilis Poulopoulos, Panagiotis Tsantilas
Palo LTD
Kokkoni Corinthias P/C 20002, Greece
{nt,pv,pt}@paloservices.com

Iraklis Varlamis
Dept. of Informatics and Telematics,
Harokopio University of Athens
Omirou 9, Tavros, Greece
varlamis@hua.gr

***Abstract*— Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. It can best be defined by thinking of three Vs: Big data is not just about Volume, but also about Velocity and Variety. The demand for stream processing is increasing a lot these days. The reason is that often processing big volumes of data is not enough. The increasing amount of opinionated data that is published in social media, in combination with the variety of data sources has created a demanding ecosystem for stream processing. The reason is that in order to deliver high quality knowledge extraction services, several tasks of high complexity must be accomplished and the existing solutions and architectures are not sufficient for processing huge volumes of streamed data. When opinion mining is applied in social media in order to cover the needs of businesses for brand monitoring, heterogeneous data has to be processed fast, so that a firm can react to changing business conditions in real time.**

***Keywords—opinion mining; news streams; social media***

## I. INTRODUCTION

The growing enthusiasm for social media, created a new channel for companies that want to advertise their products and services, or simply want to boost and monitor their brand name. This resulted to large amounts of data, which are created daily to various social media and the news and contain mentions to products and companies.

These data can be textual or audiovisual, can be presented in a formal (e.g. product reviews) or informal way (e.g. comments), can be neutral mentions or carry an opinion about the company or product or an aspect of it [1][2].

The volume and complexity of the data that can be acquired, stored and manipulated have created a flood of data — 90 per cent of all data were generated in the last two years [3]. This creates a big challenge for companies that provide social media analytics services and cope with data from multiple data streams.

The notion of "Big Data" perfectly applies in this case, with all core big data issues to impede the task of useful knowledge extraction in real-time: scale, heterogeneity, timeliness, complexity etc. The issues that must be confronted cover the whole processing pipeline starting from data acquisition, when a decision must be made on what data to keep and what to discard, and how to store and for how long.

The value of data explodes when the can be linked, compared and analyzed in a common ground. The problem that rises here is the heterogeneity between unstructured text data from social feeds such as Twitter and Facebook, structured product reviews and ratings from sites such as Epinions and lengthy news articles that contain references to images and other multimedia. The gathering of images, videos and other multimedia content is a decision that must be taken with care, since their processing for knowledge extraction is a hard task and their requirements for storage are increased. A major challenge here is to integrate content and bring everything in a common form for analysis and presentation.

Data analysis is the next bottleneck since traditional algorithms lack of scalability and do not easily adapt to the complexity of data that needs to be analyzed. Finally, the presentation of the extracted knowledge must be carefully designed in order for the results to be self-interpreted by non-technical domain experts and assist them in getting valuable actionable knowledge.

Palo Ltd is a company specializing in information extraction from the web. It started by gathering content from news sites and blogs in Greece, about 5 years ago. The analysis was limited in clustering articles based on content similarity and presenting them in an aggregated form to the end users. PaloPro is Palo's social media analytics service, which was launched primarily in Greece but now expands to Serbia, Cyprus, Turkey and Romania. The service monitors and analyzes data from the web and social media, giving emphasis to entity extraction and sentiment analysis from text. In the same architecture, several modules for crawling, feed aggregation, text clustering, multi-document summarization, Named Entity Recognition, aspect extraction and opinion mining synthesize the ecosystem of Palo services.

PaloPro can be described as a business intelligence platform with social basis (social business intelligence) [4] that takes advantage of the knowledge of the crowd (crowd sourcing) as expressed in social media. The benefit is both for companies, which are able to monitor the popularity of their products and for buyers who receive long-term improved services and products. The interest for such a platform is increased, for example the mobile phone industry in Greece numbers 13 million active subscribers, who are active on the internet, and comment the products of the three main competitors (packages, special offers, etc.). Although

information about the popularity of each of the three partners has little value, knowledge about the course of their products in social media and the opinion formed from every new movement is valuable for any further advertisement campaign.

In the following section, we provide an overview of PaloPro service and the infrastructure that supports them. In section 3 we discuss the processing pipeline in more details and in section 4 we summarize the open issues concerning the processing of big data in a real-time environment.

## II. BACKGROUND

### A. Scientific background

The scientific interest in opinion mining and analysis is huge and reflected in numerous publications in leading scientific conferences and journals in IT and marketing. The concept of opinion mining for different aspects of an entity (aspect based sentiment mining) appeared in literature in 2009. Early research focused on multi-aspect entities such as movies [5] -and the opinions provided by the viewers' comments for the different aspects that make up the final result (actors, director, screenplay, music, etc.)- electronic devices [6] and hotels [7]. These main principles behind these works were: a) extracting opinions or emotions and b) labeling entities and their aspects (head terms) and the words that convey emotion (modifiers). To identify the sentiment (or polarity) of a comment, researchers used emotion dictionaries (comprising mostly adjectives) [8], [9] statistical techniques based on co-occurrence of head terms and modifiers, classification techniques such as SVM, Naïve Bayes, Maximum Entropy etc. [10] and in some cases, semantic and syntactic analyzers [11], [12].

Moghaddam and Ester [13] gave a new impetus to opinion mining on individual properties of commercial products from customer overviews. A typical example is the evaluation of a photo camera, where users evaluate separately the ease of use, the image quality, the shutter lag and battery duration, and behind the comments attached a positive or negative score for each aspect. This approach gives a new dimension to the problem of extracting knowledge from texts adding additional granularity levels in opinion or sentiment expressed in a text.

The fact that these opinions mining techniques were applied to commercial products has increased the interest of marketers and brand makers who want to handle the image of a product or company on the market and understand the preferences of potential customers. An interesting analysis of the economic power of the comments filed by users on product review sites has been carried by Ghose et al [14], who studied the effect of good or bad reviews to the price of a product sold on Amazon. The immediate consequence to the increase of the power of comments and opinions to the commercial products is the appearance of malicious comments (spam) with positive or negative orientation that aim to alter the real image of a product [15], [16].

### B. Competitive systems

The big interest of businesses is reflected to a number of commercial tools that provide analysis and monitoring services of the markets. The Blogmeter[1] is one such product that has been developed by the Italian company CELI and adapted for specific markets (telephony, food, fashion, etc.). It offers tools for monitoring and reporting the image of companies, products or services to social media such as facebook, twitter, google +, pinterest etc. However, the system requires that the company has a profile in the respective social media and focuses only on analyzing the information posted on the respective websites of the companies in each medium (eg likes, the followers, the retweets, etc. at pins. each company) .The SentiMeter[2] is a tool that gathers data from Twitter, Facebook, YouTube, Google+, Digg, Blogger, Tumblr and other ATOM and RSS feeds, and then allows users to create and monitor their own campaigns. It also allows creating reports and control user access to them. The Sentimarket[3] API is yet another tool for content analysis derived from social media, which allows monitoring of a market, alert creation, etc.

Other competitors in social media monitoring include: Brandwatch[4], Sysomos[5], Trackur[6], Engagor[7]. Their main characteristics are: a) they primarily collect English content or a single language content and certainly do not provide cross-country social media monitoring solutions, b) they mainly focus on social media monitoring and primarily target market analysts with good technological background, c) they provide tools for monitoring the effect of ads campaigns to the brand's image to social media but do not offer tools for interactive campaign management.

### C. Advantages of PaloPro

A major disadvantage of all the above tools is that they do not support prioritization of sources. All references to a company or product do not contribute equally to the overall feeling and by incorporating the ***importance or influence of each source*** we get better insights on how the public opinion will evolve. This prioritization of sources is automatically done in PaloPro, by employing the metadata collected from social media sources, concerning users and their buzz in the community.

An equally important lack of competitive tools is that they do not analyze the factors that lead to the increase or decrease in the popularity of an entity in social media. Although internationally there are several sites where consumers can comment on the products that interest them (eg, epinions.com, amazon.com, rateitall.com) and evaluate individual features or services (eg. tripadvisor.com), the existing social media analysis tools do not provide such detail. The aspect extraction that is performed in PaloPro, allow us to perform ***aspect-based sentiment analysis*** and provide the tools for in-depth analysis of results.

[1] http://www.blogmeter.eu
[2] https://sentimeter.com/
[3] http://www.sentimarket.com
[4] http://www.brandwatch.com
[5] http://www.sysomos.com/
[6] http://www.trackur.com/
[7] https://engagor.com/

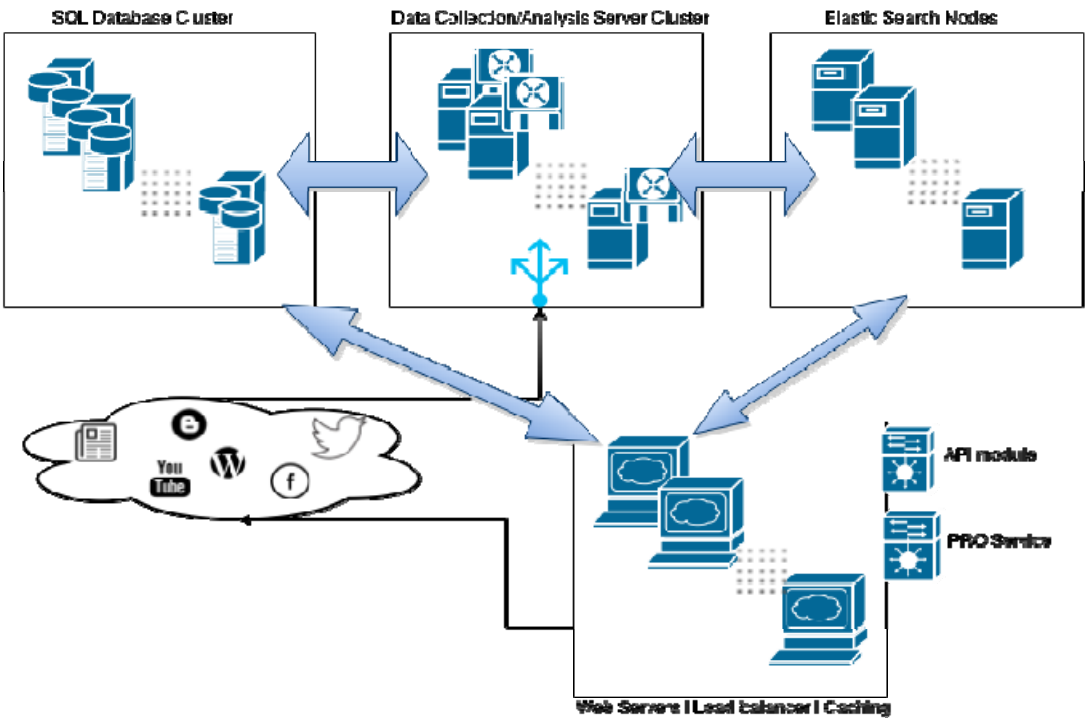


Figure 1. The architecture of PaloPro

Finally, one of the key advantages of Palo is its crawling and indexing mechanism and the efficient language agnostic techniques for Sentiment Analysis and Entity Recognition, which allow us to deploy a PaloPro services clone in a new country in a four months cycle. Shifting to a new language (or a country with multiple languages) is a problem that we already face in Palo, since we have already developed a solution for Serbia (palo.rs), Cyprus (palo.com.cy) and Greece (palo.gr) and we are deploying our services to Turkey and Romania. Our NER and Sentiment analysis tools are based on a unique knowledge building infrastructure, which exploits open multilingual resources (e.g. Wikipedia), which are available in almost any language and probabilistic (n-gram based) language agnostic techniques and can be deployed and fine-tuned with a minimum user effort.

## III. AN OVERVIEW OF PALOPRO

### *A. PaloPro media analytics services*

PaloPro service provides a tool for monitoring and analysis of opinions about user defined entities (e.g. products, persons, locations) thus creating a Reputation Management System. The user has the opportunity to view in real-time, the source of the buzz, the parameters that affect the positive, negative or neutral reputation towards an organization, brand or person and, ultimately, the overall polarity sentiment and trend on the Web. This is achieved by gathering and processing all references through natural language technologies that extract entities and opinions about these entities. Being a commercial subscription service, the requirements for accurate results are high for the underlying linguistic processing infrastructure, aiming at achieving accuracy over 87% for both the named-entity recognition and the polarity detection tasks.

The data are collected, filtered and processed by a set of crawlers, which aggregate data from different sources, including traditional news sites, blogs, forums, video comments and social media such as Twitter and Facebook posts and comments. The crawling and storage procedure is fully distributed and controlled in such a way that the system may provide a near real-time analysis to the end user. In order to achieve this, the crawling controller adjusts the frequency of visits to each source and prioritizes sources that have a higher update frequency. As a result, it providing an efficient way to instantly locate and retrieve new content. Multiple layers of spam filtering are deployed to ensure that clean data are provided to the analysis modules. The amount of documents crawled in a typical day usually exceeds 3 million documents. The lengthiest documents are collected from thousands of different websites and a huge amount of small texts comes from social media networks and specifically from Twitter. All the content that is collected is categorized on a predefined set of news domains and it is ranked for importance based on a predefined ranking of importance for sources (e.g. news portals are ranked higher than blogs).

The concept that dominates the design of PaloPro is “workspace”, a dashboard that contains visualizations of information collected for an entity of interest (e.g. a brand name and its core products). The reputation of an entity is measured on a set of user-selectable entities or user-specified keywords, which are monitored across the different news sites and social media. The user can create a new workspace and is expected to select one or more persons, companies, locations, brands, or product names from a large database of monitored

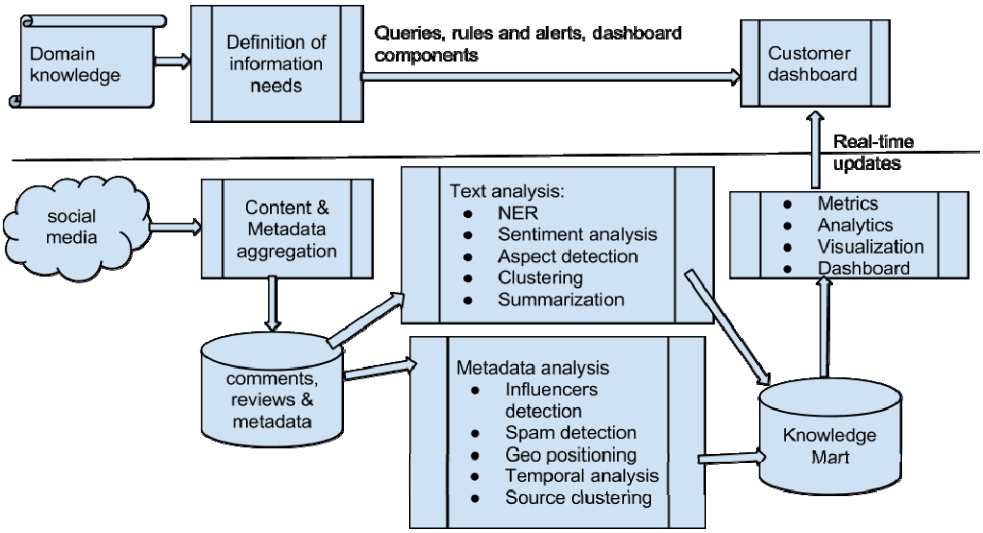


Figure 2. The information flow and the core processing tasks of PaloPro

entities, and/or define a set of keywords, in case an entity is not contained into the database of monitored objects. The user may define any number of workspaces, all of which are visible when the user logs into the system.

Through the system, a user can access information related to different entities or keywords such as persons, organizations, companies, brands, products, events etc. that are monitored by the system in the crawled corpora, along with aspect information about them. Automated alerts can be set up so that the service may deliver instant notifications whenever the data matches some predefined, user-specified criteria, as new information is extracted or when the extracted information exceeds certain user-configurable thresholds.

### B. *Infrastructure*

PaloPro follows the major requirements of real-time stream processing set out by [17] by using advanced custom coding and infrastructure software.

In order to support the data collection, storage and real-time procedures, Palo has a complex multi-level infrastructure, which consists of crawling and analysis servers, database (SQL and no-SQL) servers, web servers, caching and load balancing servers. Figure 1 depicts the generic infrastructure.

Since data sources may reside in any place in world, but end users of PaloPro come from different countries, the initial design comprises different servers per country. Recently we interconnected the data collection services for all countries and now all services use the same infrastructure. So data collection, storage, analysis and presentation are done within the same collection of servers.

In order to be able to handle the huge amounts of data collected and served to our clients, we store data in two types of databases. The first one is a *Percona MySQL database cluster*[8], while the second one is a set of nodes of *Elastic Search*[9] nodes distributed into multiple servers.

## IV. PALOPRO PROCESS FLOW

The process flow in PaloPro starts from raw data and ends up to useful business knowledge. It comprises several steps, which are depicted in Figure 2 and are explained in the following subsections. The volume, velocity and variety of data, affects the design of each step.

### A. *Data Acquisition and Recording*

PaloPro starts with the collection of content from social media, which is performed in a continuous basis (every few minutes) and results in a huge repository of textual-raw content and associated metadata that describe the source and the content itself (e.g. time information, location information, author, social medium, etc).

Social media content does not arise by itself: it is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. All these filters in Palo and PaloPro are implemented using machine learning techniques and allow new filters to be trained in such a way that they do not discard useful information. A detailed description of the news crawling mechanism of Palo is provided in [18]. This mechanism allows the administrators of Palo to quickly feed in news sources, when entering a new country and thus quickly create an initial content repository. Data from popular social media platforms is gathered using the provided APIs.

The next and most important step of PaloPro comprises the semantic analysis of texts (e.g. named entity recognition, sentiment analysis, aspect detection etc.) and the analysis of associated metadata (e.g. influential users detection, social

[8] http://www.percona.com/software/percona-xtradb-cluster
[9] https://www.elastic.co/downloads

medium impact etc.). The result of this step is a rich repository of semantically enhanced content and information concerning the social media sites and users and their influence to the social media sphere.

for querying, and the analytics tools that perform data mining tasks and statistical analyses. In PaloPro this binding is driven by the business need for information. So starting from the needs for visualization and information for the domain


Figure 3. PaloPro dashboard with real-time information on mentions' polarity, top influencers and topics of interest

### B. *Information Extraction and Cleaning*

The challenge here is to extract useful level content on-the-fly from the original content that is aggregated from the various sources. An information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis is needed. For this purpose, Palo incorporates several high parallelizable algorithms for document and sentence clustering, text summarization, named entity, aspect and opinion extraction. Using a very fast text clustering algorithm we manage to refresh our news every 3 minutes and to automatically cluster them into themes, without human intervention. Document and sentence clustering significantly reduces the load of the remaining processing pipeline since news content is highly reproduced in many sources.

For the summarization of content, we employ an efficient language-agnostic technique, which is based on n-gram graphs [19] and produces comparative results to other language dependent techniques. Finally, for entity extraction and opinion mining we implement a machine learning technique, which can be easily trained for new languages [20]. A new, highly parallelized alternative implementation, which is automatically deployed to new languages, is currently under development. The alternative takes advantage of structured collaboratively created content in order to train the respective entity extraction and opinion mining models for a new language.

### C. *Data Integration, Modelling, and Analysis*

The acquisition of data and extraction of information are the first steps towards business intelligence. However, due to the heterogeneity of information it is necessary to properly model the extracted information in order to further analyse it. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data

experts, we properly orchestrate the underlying mechanisms in order to be able to continuously feed the end-user dashboard with up-to-date knowledge about his/her company or product.

### D. *Interpretation – Visualization*

On top of the collected information and extracted knowledge, we have developed the PaloPro dashboard, which comprises sophisticated tools that allow us to depict the image of an entity (e.g. a company, a person, a product) to the social media, to measure the result of a certain action or event to an entity's image in the long run, and to drill down to the details that contributed to this result. Figure 3, provides a glance of PaloPro dashboard

## V. CHALLENGES

Having in mind the multiple phases in PaloPro stream data processing pipeline, it is interesting to consider some common challenges that underlie the different phases.

### A. *Heterogeneity*

Content in PaloPro is mainly textual. However, we also collect multimedia content which is interesting to be associated with text. In addition to this, for Greece, we collect data from Diavgeia[10], a governmental site that provides metadata and data concerning all the payments made by the public organizations to companies and individuals. Using these data, we are able to provide a detailed analysis of where public money are spend, similar to [21]. The interest for such analysis is bigger for reporters and professionals from the news industry. Although it is currently out of Palo objectives, the rich content that we continuously collect can be exploited if properly integrated.

[10] https://diavgeia.gov.gr/

### B. Scalability

A great issue that arises when we upscaled our solution was the dilemma between cloud computing and private servers. Currently PaloPro is using its own dedicated servers and the workload for them is constantly increase, thus minimizing the idle resources and making the choice of own servers more reasonable.

In terms of speed, the limits are set not by the demand for an increasing processing throughput but from the acquisition rate in conjunction with the amount of collected data. Entering a new country, such as Turkey for example, which provides 10 times the size of content that Greece provides, does not change the requirement for refresh of the news sphere every three minutes (or even less). So the scalability of algorithms and architecture must be examined accordingly.

### C. System training

The last but not least concern is the quality of the collected content and consequently of the information and services delivered. The quality standards have been defined from the 5 year presence in Greek social media analysis but the same standards must be met in a shorter period when entering a new country. A set of language agnostic methods guarantees that the quality of some modules will be the same in all countries. In the case of language specific modules, a set of tools that accelerate the training of the different models either using structured human-created content or human annotated content allows fast deployment and high quality of services.

## VI. CONCLUSIONS

PaloPro implements a holistic approach for social media analysis and monitoring of brand awareness through easy to use dashboard and support content in multiple languages. Using the business intelligence it provides, the brands and companies are able to monitor the outcomes of their campaigns using real-time analysis of the impact in social media and the news. This analysis creates a high corporate business value but on the same value creates several issues that relate to the management of big data. In this work, we presented the main challenges and summarized on the solutions that we implement.

## ACKNOWLEDGMENTS

The project is partially funded by GSRT (ICT4Growth project).

## REFERENCES

[1] Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 0165551510388123.

[2] Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S. (2014, August). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 27-35).

[3] SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years." ScienceDaily. ScienceDaily, 22 May 2013. <www.sciencedaily.com/releases/2013/05/130522085217.htm>.

[4] Dinter B., Lorenz, A. (2012). "Social Business Intelligence: a Literature Review and Research Agenda". In: Thirty Third International Conference on Information Systems (ICIS 2012). Ed. by F. George Joey. Orlando, Florida: Association for Information Systems. isbn: 978-0-615-71843-9. url: http://aisel.aisnet.org/icis2012/proceedings/ResearchInProgress/104/

[5] Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36(6), 823-848.

[6] Hu, M., Liu, B. (2004). Mining and summarizing customer reviews, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004) 168–177.

[7] Blair-Goldensohn, S., Hannan, K., McDonald, S., Neylon, T., Reis, G.A., Reynar, J. (2008). Building a sentiment summarizer for local service reviews, Proceedings of WWW 2008 Workshop: NLP Challenges in the Information Explosion Era.

[8] Hatzivassiloglou, V., McKeown, K.R. (1997) Predicting the semantic orientation of adjectives, Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL (1997).

[9] Qiu, G. Liu, B. Bu, J., Chen, C. (2009). Expanding domain sentiment lexicon through double propagation, Proceedings of the 21st International Joint Conference on Artificial Intelligence (Morgan Kaufmann, San Francisco, 2009) 1199–1204

[10] Pang, B., Lee, L. (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the Association for Computational Linguistics (2005) 115–124.

[11] Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003). Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, Proceedings of the 3rd IEEE International Conference on Data Mining (2003) 427–434.

[12] Miyoshi, T., Nakagami, Y. (2007). Sentiment classification of customer reviews on electric products, Proceedings of International Conference on Systems, Man and Cybernetics (2007) 2028–2033.

[13] Moghaddam, S., Ester, M. (2012). Aspect-based opinion mining from product reviews. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). ACM, New York, NY, USA, 1184-1184.

[14] Ghose, A., Ipeirotis, P., & Sundararajan, A. (2007, June). Opinion mining using econometrics: A case study on reputation systems. In annual meeting-association for computational linguistics (Vol. 45, No. 1, p. 416).

[15] Mukherjee, A., Liu, B., Glance, N. (2012). Spotting Fake Reviewer Groups in Consumer Reviews. International World Wide Web Conference (WWW-2012), Lyon, France, April 16-20, 2012.

[16] Jindal, N., Liu, B. (2008). Opinion Spam and Analysis. Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Feb 11-12, 2008, Stanford University, Stanford, California, USA.

[17] Stonebraker, M., Çetintemel, U., Zdonik. S. (2005). The 8 requirements of real-time stream processing. SIGMOD Rec., 34(4):42–47, 2005.

[18] Varlamis, I., Tsirakis, N., Tsantilas, P., & Poulopoulos, V. (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.

[19] Giannakopoulos, G., Kiomourtzis, G., & Karkaletsis, V. (2014). NewSum:"N-Gram Graph"-Based. Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding: Revolutionizing Knowledge Understanding, 205.

[20] Petasis, G., Spiliotopoulos, D., Tsirakis, N., & Tsantilas, P. (2014). Sentiment analysis for reputation management: Mining the greek web. In Artificial Intelligence: Methods and Applications (pp. 327-340). Springer International Publishing.

[21] Vafopoulos, M. N., Meimaris, M., Papantoniou, A., Anagnostopoulos, I., Alexiou, G., Avraam, I., ... & Loumos, V. (2012). Public Spending: Interconnecting and Visualizing Greek Public Expenditure Following Linked Open Data Directives. Available at SSRN 2064517.