The Journal of Systems and Software 127 (2017) 237–248

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Large scale opinion mining for social, news and blog data

Nikos Tsirakis[a,*], Vasilis Poulopoulos[a], Panagiotis Tsantilas[a], Iraklis Varlamis[b]

[a] Palo LTD, Kokkoni Corinthias P/C 20002, Greece
[b] Dept. of Informatics and Telematics, Harokopio University of Athens, Omirou 9, Tavros, Greece

ARTICLE INFO

Article history:
Received 24 October 2015
Revised 9 March 2016
Accepted 4 June 2016
Available online 6 June 2016

Keywords:
Opinion mining
News streams
Social media

ABSTRACT

Companies that collect and analyze data from social media, news and other data streams are faced with several challenges that concern storage and processing of huge amounts of data. When they want to serve the processed information to their customers and moreover, when they want to cover different information needs for each customer, they need solutions that process data in near real time in order to gain insights on the data in motion. The volume and volatility 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. Although, there are several solutions that can handle information of static nature and small volume quite efficiently, they usually do not scale up properly because of their high complexity. Moreover, such solutions have been designed to run once or to run in a fixed dataset and they are not sufficient for processing huge volumes of streamed data. To address this problem, a platform for real-time opinion mining is proposed. Based on prior research and real application services that have been developed, a new platform called "PaloPro" is presented to cover the needs for brand monitoring.

1. Introduction

The enormous advances in social media and their power to reflect and influence public opinion made them a domain of great interest for marketeers, communication specialists and 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. Data can be in different formats (textual, audiovisual etc), can be written in a formal (e.g. product reviews) or informal way (e.g. comments), can be objective mentions or subjective opinions about the company or product or an aspect of it (Thet et al., 2010; Pontiki et al., 2014).

The volume and complexity of the data that can be acquired, stored and manipulated, have created a flood of data 90% of all data were generated in the last two years (SINTEF, 2013). For example, in Palo, the crawler module is able to collect a large number of data which is estimated at 1000 articles per minute during the rush hours leading to an increase of 2.5Gb per month of compressed data. The data in absolute numbers are more than 10 million records per month from all the possible data sources.

The large volume and volatility create big challenges for companies that provide social media analytics services and cope with data from multiple data streams. Big players from the Web and Databases domains invest in social media analytics with generic frameworks and platforms (e.g. IBM Social Media Analytics) that emphasize on the analytics part but do not focus on text mining, or with extensions of their existing platforms (e.g. Google news lab and Google Analytics) that incorporate content from specific social media using associate data hubs and plugins (e.g. Googles social data hub). They use Twitter, Facebook and other social media APIs to collect data in streams and provide commercial archives/feeds and associated analytics.

A big challenge for social media monitoring is to link data from various sources together, compare and integrate 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.

* Corresponding author.
E-mail addresses: nt@paloservices.com (N. Tsirakis), pv@paloservices.com (V. Poulopoulos), pt@paloservices.com (P. Tsantilas), varlamis@hua.gr (I. Varlamis).

http://dx.doi.org/10.1016/j.jss.2016.06.012
0164-1212/

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. Palo Pro is Palos 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.

Palo Pro can be described as a business intelligence platform with social basis (social business intelligence) (Dinter and Lorenz, 2012) 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.

From the data stream management point of view, several research issues emerge, like: (a) approximate query processing techniques to evaluate slow and memory demanding queries, (b) sliding window query processing, (c) data sampling to handle an increased flow rate of the input stream etc. In this paper we present how the proposed platform handles such research issues, using real-time data filtering in the source, summarization of historical content and statistics computation over sliding windows. We also discuss additional technical challenges, that are not only data stream specific, are confronted, such as heterogeneity of data, multilingual content and scalability of the existing solution.

In the following section, we provide an overview of Palo Pro 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.

2. Background

2.1. Scientific background

Opinion mining is defined as the task of classifying texts into categories depending on whether they express positive or negative sentiment, or whether they enclose no emotion at all. Sentiment polarity was extracted using emotion dictionaries (comprising mostly adjectives) (Hatzivassiloglou and McKeown, 1997; Qiu et al., 2009), statistical techniques based on co-occurrence of head terms and modifiers, classification techniques such as SVM, Naïve Bayes, Maximum Entropy etc. Pang and Lee (2005) and in some cases, semantic and syntactic analyzers (Yi et al., 2003; Miyoshi and Nakagami, 2007). The topic has attracted considerable attention in recent years due to its direct applicability in real-world businesses, such as brand monitoring, marketing or prediction of election results.

The concept of ***aspect based opinion mining*** (opinion mining for different aspects of an entity) appeared in literature in 2009. Early research focused on multi-aspect entities such as movies (Thet et al., 2010) – 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 (Hu and Liu, 2004) and hotels (Blair-Goldensohn et al., 2008). The 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). However, it was until 2012, that the work of Moghaddam and Ester (2012) gave a new impetus to opinion mining on individual properties of commercial products from customer overviews. A typical example of aspect based opinion mining 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. 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 (Mukherjee et al., 2012; Jindal and Liu, 2008).

Summarization is another challenging task for social media content analysts. There are many research works that focus on the summarization (Zubiaga et al., 2012) or visualization (Hao et al., 2011) of Twitter streams. The analysis focuses on a specific entity (Hao et al., 2011) or event (Zubiaga et al., 2012) of the complete Twitter stream, so assumes a filtering step in the beginning of the pipeline. Although the volume of data for an entity or event in the unit of time in Twitter is not so impressive when compared to video streams (e.g. in Hao et al., 2011 authors report about 60,000 tweets for a movie in a five days period), the processing and visualization arise several challenges for system developers. In the case of text stream analytics it is important to summarize content and export useful features in a streamed manner, but also to keep the actual data in a separate storage for further analysis (e.g. drill through analytics, post event analytics). In addition to this, the filtering step may produce several different substreams that refer to different entities or events. In the case of PaloPro for example, social media stream monitoring is available for different countries and monitors different entities in parallel. In addition to this, since PaloPro also combines social media analytics with news analytics, it collects information from data sources around the world. An aggregation module is responsible for this, and its output is also redirected to the filtering module.

2.2. 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 (e.g. 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

[1] http://www.blogmeter.eu.
[2] https://sentimeter.com/.

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 brands image to social media but do not offer tools for interactive campaign management.

2.3. Advantages of PaloPro

All the aforementioned tools offer a uniform, yet non flexible, way of processing the different data sources, which is not always helpful. It is common that all references to a company or product do not contribute equally to the overall feeling and prioritization of sources is necessary. 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 Palo Pro, by employing the metadata collected from social media sources, concerning users and their buzz in the community.

Another advantage of the proposed system is that it performs an analysis of the factors that lead to the increase or decrease in the popularity of an entity in social media. Despite the many review sites, where consumers can comment on the products that interest them (e.g. epinions.com, amazon.com, rateitall.com) and evaluate individual features or services (e.g. tripadvisor.com), there are not yet social media analysis tools that provide such detail. This results in two separate worlds, one with fully structured data that are property of the review site owners and raw textual data that are publicly available, but are difficult to process and analyze. The aspect extraction that is performed in Palo Pro, allow us to perform *aspect-based sentiment analysis* and provide the tools for in-depth analysis of publicly available textual content.

Finally, Palo has a unique crawling and indexing mechanism and uses efficient language agnostic techniques for Sentiment Analysis and Entity Recognition, which allow the deployment of a Palo Pro services' clone in a new country in a four months cycle. This results in an explosion of the data volume that has to be analyzed. 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.

3. An overview of PaloPro

In this section, PaloPro is described in detail regarding its services, the components of the proposed architecture, the selected software and finally how all these interact.

[3] http://www.sentimarket.com.
[4] http://www.brandwatch.com.
[5] http://www.sysomos.com/.
[6] http://www.trackur.com/ .
[7] https://engagor.com/.

3.1. PaloPro media analytics services

PaloPro is an online service for monitoring user defined entities (e.g. products, persons, locations) and the sentiment about them in social media. It can be used as a Reputation Management System by companies that operate in a market or that want to enter a market. 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 fuel of PaloPro engine are the data, which are collected by domain specific web crawlers and content aggregators and are filtered and processed upon collection. Data are aggregated 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 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.

The online PaloPro dashboard allows to browse 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.

3.2. Infrastructure

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 servers, web

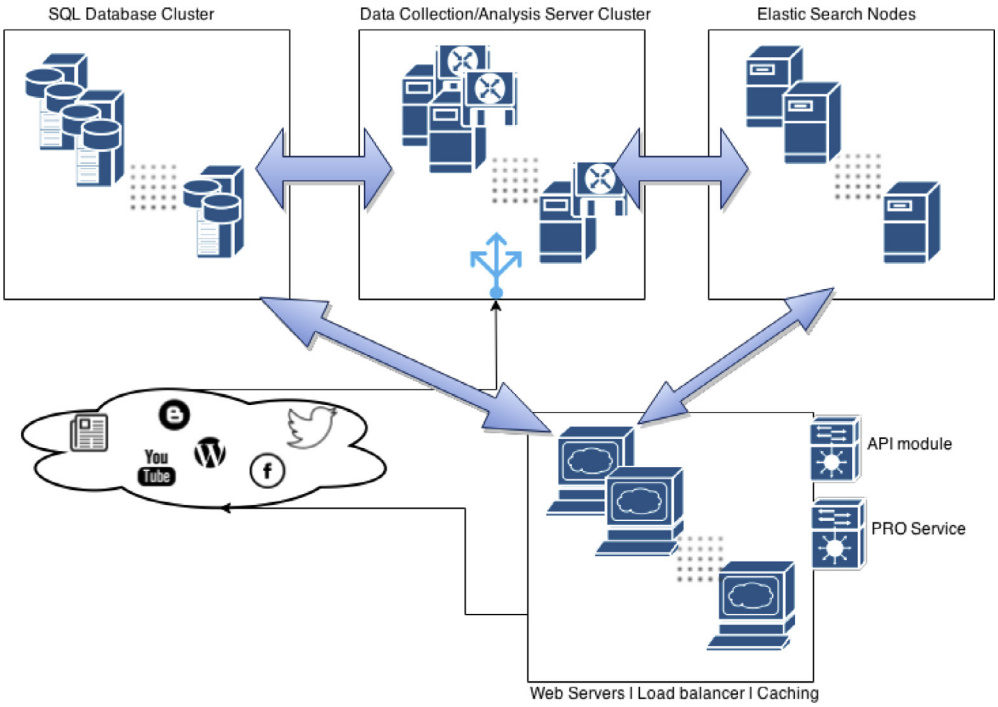

Fig. 1. The architecture of Palo Pro.

servers, caching and load balancing servers. Fig. 1 depicts the generic infrastructure.

Our system architecture consists of a set of systems in order to achieve modularity, interoperativity, scalability and the ability to handle the large sets of data that are incoming and stored to our system. This architecture was developed by a set of different software components that help us achieve having a large system able to perform all the actions needed by our infrastructure. The architecture consists of a set of clustered SQL relational databases for permanent data storage, a set of noSQL clustered nodes for fast search, indexing and quick access to real time data, a core set of application servers that perform the main system procedures which are related to collection and analysis of data and finally, a set of servers that are responsible for data presentation and include web servers, caching servers, API modules and Javascript frameworks.

The systems are divided into multiple levels in order to be able to act as a real time collecting and processing system. We selected the levels of permanent storage and indexing to be at the same level as the software running the collection and analysis software in order to be able to achieve very fast processing and analysis times. The combination of noSQL clustered infrastructure for indexing, fast search and data combination together with the SQL cluster for permanent storage of the processed data and fast access through their indexed keys provides us with a solution that can handle with ease the number of incoming data that occur even at situations where burst of data occurs. Depending on the process that is executed the system selects to utilize either the SQL or the noSQL database. Both of them are filled with data and the data collection and analysis mechanism is responsible for the data integrity between the two different clusters of databases. This means that whenever the system needs very fast access to stored data it utilizes the “fast” noSQL database to gather important information including the indexed keys to information. For more detailed information and drilling on details and meta-data the indexed keys provide easy of access to the SQL clustered database.

Selecting such an infrastructure leads to some kind of overhead to the core mechanism of our system. Despite this fact, we are able to achieve very high performance in our system. We perform the handling of the data integrity between our different data-banks and thus we are able to control the throughput of each procedure considering the real data process. All the aforementioned are not only based on hardware capabilities but also - mainly - on software solutions that can support our needs for data collection, storage, analysis and processing.

3.3. Software to achieve our infrastructure

For SQL Database Cluster, we use Percona XtraDB Cluster[8]. We chose it as an active/active high availability and high scalability open source solution for MySQL clustering. It integrates Percona Server and Percona XtraBackup with the Codership Galera library of MySQL high availability solutions in a single package that enables the creation of a cost-effective MySQL high availability cluster.

In the Data Collection and Analysis Cluster we use a set of servers where all Java Services are being executed. These services are responsible for the collection processing and analysis of the incoming data. With the aim of making these services 100% available and provide a stable environment for them in order to have consistent data we use three software packages. First, the Corosync Cluster Engine[9] which is an open source project to develop, release, and support a community-defined, open source cluster executive for use by multiple open source and commercial cluster projects or products. Second, Pacemaker[10] which is an Open Source, High Availability resource manager suitable for both small and large clusters and third, HAProxy[11] a free, very fast and reliable solution

[8] http://www.percona.com/software/percona-xtradb-cluster.
[9] https://corosync.github.io/corosync/.
[10] http://clusterlabs.org/wiki/Pacemaker.
[11] http://www.haproxy.org/.

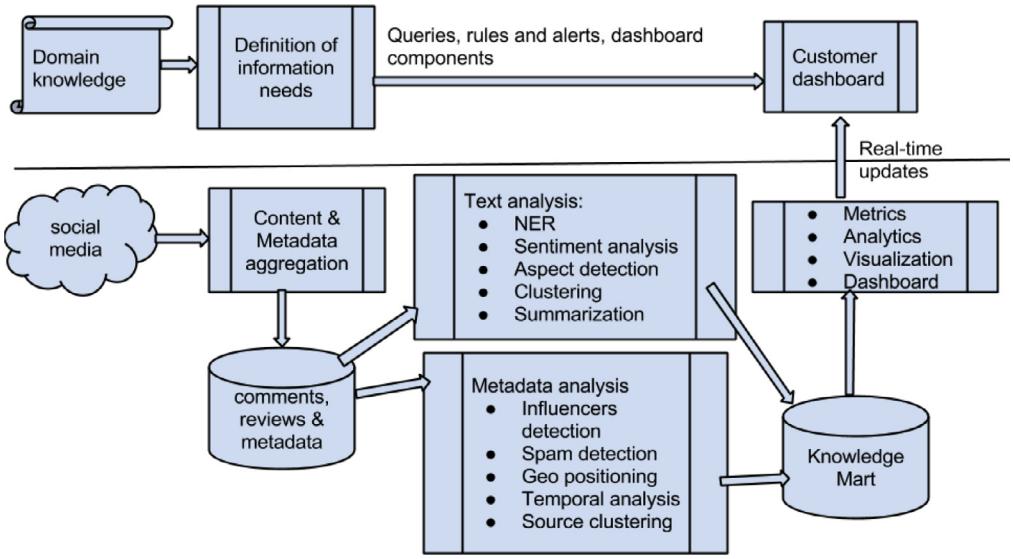

Fig. 2. The information flow and the core processing tasks of Palo Pro.

offering high availability, load balancing, and proxying for TCP and HTTP-based applications.

For Indexing, we use the most appropriate and popular engine, Elasticsearch.[12] Elasticsearch is an open-source, distributed, scalable, and highly available with real-time search and analytics capabilities providing a sophisticated RESTful API. Is built to run on NoSQL document-driven databases. Documents are stored in JSON format and fields are indexed in order to be available for search. Elasticsearch works on top of Apache Lucene to provide the most powerful full-text search capabilities. Its powerful, developer-friendly query API supports multilingual search, geo-location, contextual did-you-mean suggestions, auto-complete, and result snippets.

Finally, for the output interfaces, we use two different tools. First we use our marketing front-end tool which has been developed with Zend Framework.[13] Zend Framework 2 is an open source framework for developing web applications and services using PHP 5.3+. Zend Framework 2 uses 100% object-oriented code and utilizes most of the new features of PHP 5.3, namely namespaces, late static binding, lambda functions and closures. Second we use Kibana[14] for back-end purposes. Kibana an open-source flexible analytics and visualization platform. Provides real-time summary and charting of streaming data. It has intuitive interface for a variety of users providing instant sharing and embedding of dashboards. In other words Kibana works very easily and can be used in several ways as an out-of-box user interface platform. In Palo, Kibana helps us monitor, analyze and predict easily issues related to our data.

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 is 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,[15] while the second one is a set of nodes of Elasticsearch[16] nodes distributed into multiple servers. The main reason behind the use of double storage is the requirement for streamed processing and long term analytics in the same time. The Percona MySQL database cluster covers the latter requirement. Elasticsearch is a flexible solution for indexing, storage and retrieval of text streams (De Rooij et al., 2013) due to its document based structure (Kononenko et al., 2014).

Depending on the type of data the collection and analysis steps are taking place independently or simultaneously. In the case of crawling data, we first store the initial documents in Percona MySQL and then we analyze them, while in real time data streams, such as Twitter data, the collection and analysis steps are taking place, in memory and finally the data are uniquely stored in the Elasticsearch cluster. This provides a real time monitoring option of that data.

4. 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 Fig. 2 and are explained in the following subsections. The volume, velocity and variety of data, affects the design of each step.

4.1. Data acquisition and recording

Palo Pro starts with the collection of content from social media, which is performed in a continuous basis (every few min) 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.

[12] https://www.elastic.co/products/elasticsearch.
[13] http://framework.zend.com/.
[14] https://www.elastic.co/products/kibana.
[15] http://www.percona.com/software/percona-xtradb-cluster.
[16] https://www.elastic.co/products/elasticsearch.

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

All these filters in Palo and Palo Pro 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 Varlamis et al. (2014). 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 Palo Pro 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 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.

4.2. Information extraction and cleaning

The main challenge concerning information extraction from Palo sources, is to extract useful content on-the-fly, when the original content 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

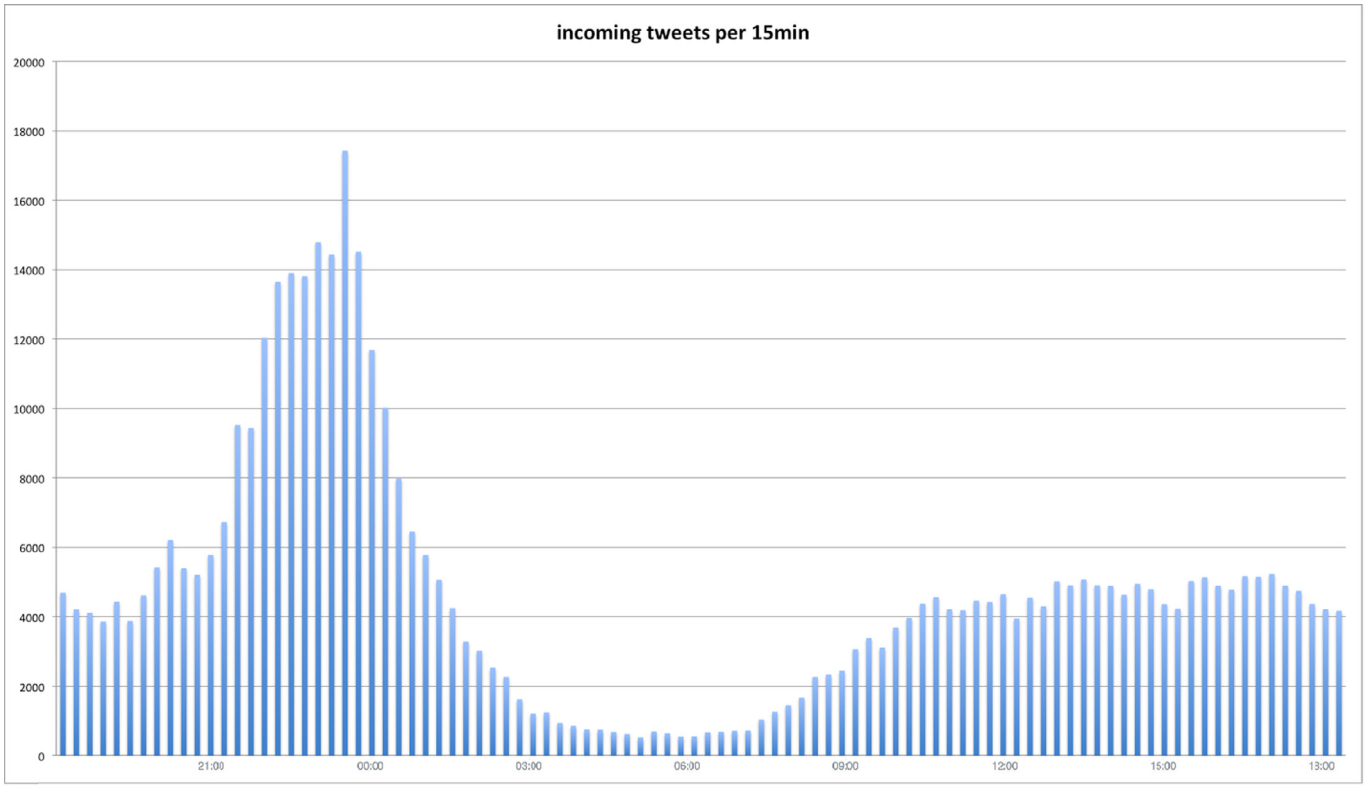

Fig. 4. The volume of tweets (the histogram bars are every 15 min) during the ministerial debate in Greece (September 2015).

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, which are explained in the following.

We use a fast text clustering algorithm (O($n \times m$), where n is the number of documents and m is the number of existing clusters), which incrementally assigns new documents to existing clusters by measuring the similarity between the cluster centroid and the new document. When the similarity of a document is below a threshold then the document forms a new cluster by itself. When a cluster stops increasing in size for a certain period, it is removed from the list of candidate clusters, thus keeping the complexity of the algorithm low. With this setup, we manage to refresh our news every 3 min and to automatically cluster them into themes, without human intervention. A similar clustering methodology is applied at sentence level, in order to create a representative summary for each cluster. An n-gram graph represention of the sentences that comprise the texts of a cluster is used as a basis for the summarization algorithm (Giannakopoulos et al., 2008). Document clustering and sentence redundancy removal significantly reduces the load of the remaining processing pipeline since news content is highly reproduced in many sources.

Content summarization employs an efficient language-agnostic graph-based technique, which represents each document or set of documents as a n-gram graph (Giannakopoulos et al., 2014). The graph is subsequently partitioned in order to highlight the document topics and the most informative content is kept as a topic representative. The method 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 (Petasis et al., 2014). 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.

4.3. Data integration, modeling, 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 analyze it. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data for querying, and the analytics tools that perform data mining tasks and statistical analyses. In Palo Pro this binding is driven by the business need for information. So starting from the needs for visualization and information for the domain 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. Orchestration comprises the connection of various content resources (a Twitter firehose, the news crawler, different social media crawlers and feeds) to different ElasticSearch nodes, for filtering and indexing content. On top of the indexing nodes, predefined queries are executed in order to aggregate data and feed the Kibana dashboard. An example of how this is done and what is the produced result is given at Section 6.

4.4. Interpretation visualization

On top of the collected information and extracted knowledge, we have developed the Palo Pro 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 entitys image in the long run, and to drill down to the details that contributed to this result. Fig. 3, provides a glance of PaloPro dashboard.

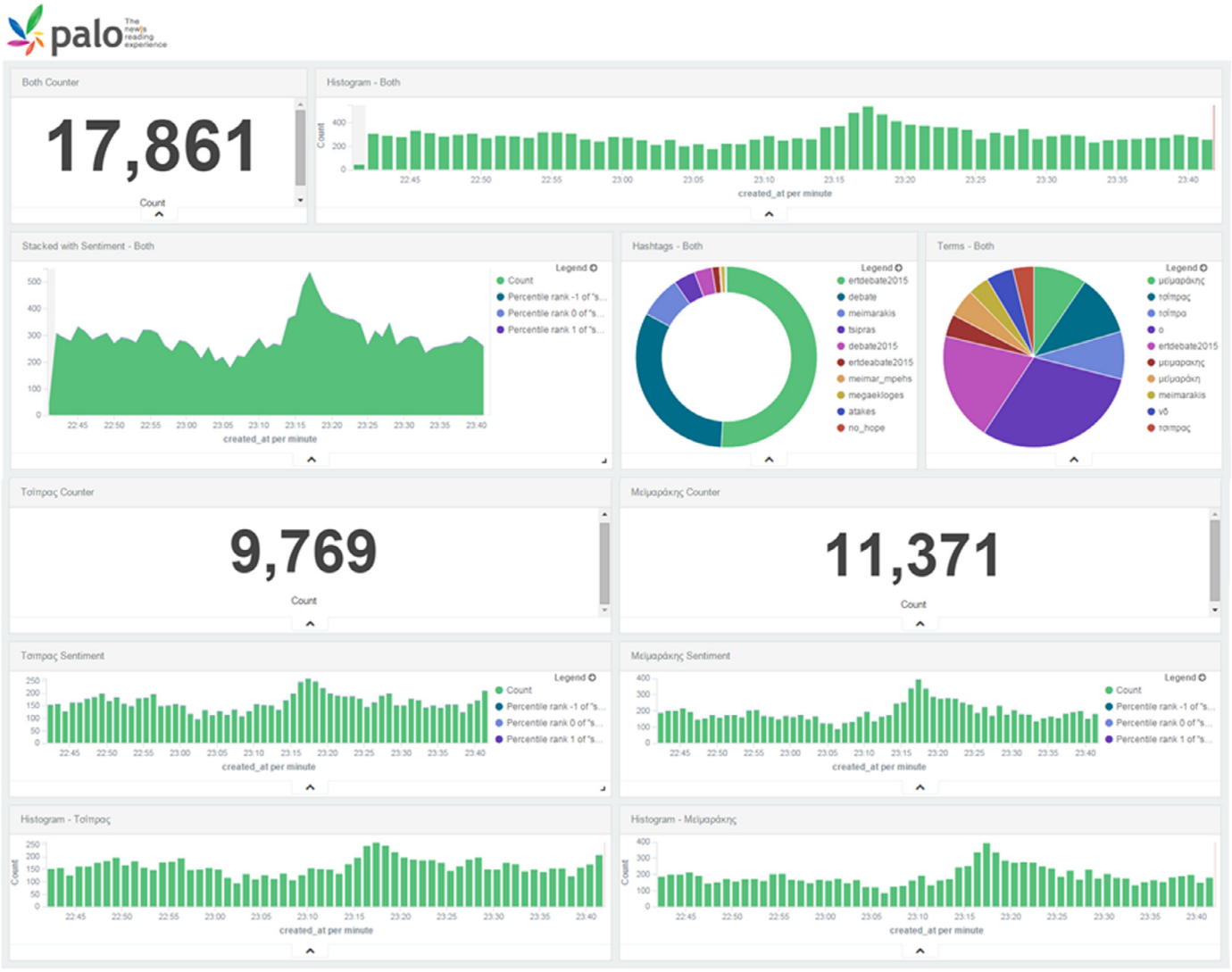

Fig. 5. A snapshot of the dashboard we launched for monitoring the evolution of the ministerial debate in Greece (September 2015).

5. Challenges

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

5.1. Scalability

An issue that arose 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.

ElasticSearch is a good solution in terms of elasticity, since it allows to adjust the number of nodes depending on the data load and since our processing is done in document or cluster (set of documents) level it is easy to parallelize processing and tackle scalability issues. As far as it concerns multi-tenancy, it mainly affects the front-end deployment, where many users may request information (the result of our processing pipeline) from the ElasticSearch node. Since all the architecture that serves user requests is on top of multiple ElasticSearch nodes for storage and indexing and a RabbitMQ cluster for request handling allows us to better balance the query load. It is on our plans to implement some load prediction algorithms in order to be able to increase or decrease the available resources in advance and guarantee high availability at minimum resource allocation.

5.2. System training

Another 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 years 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 guarantee 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.

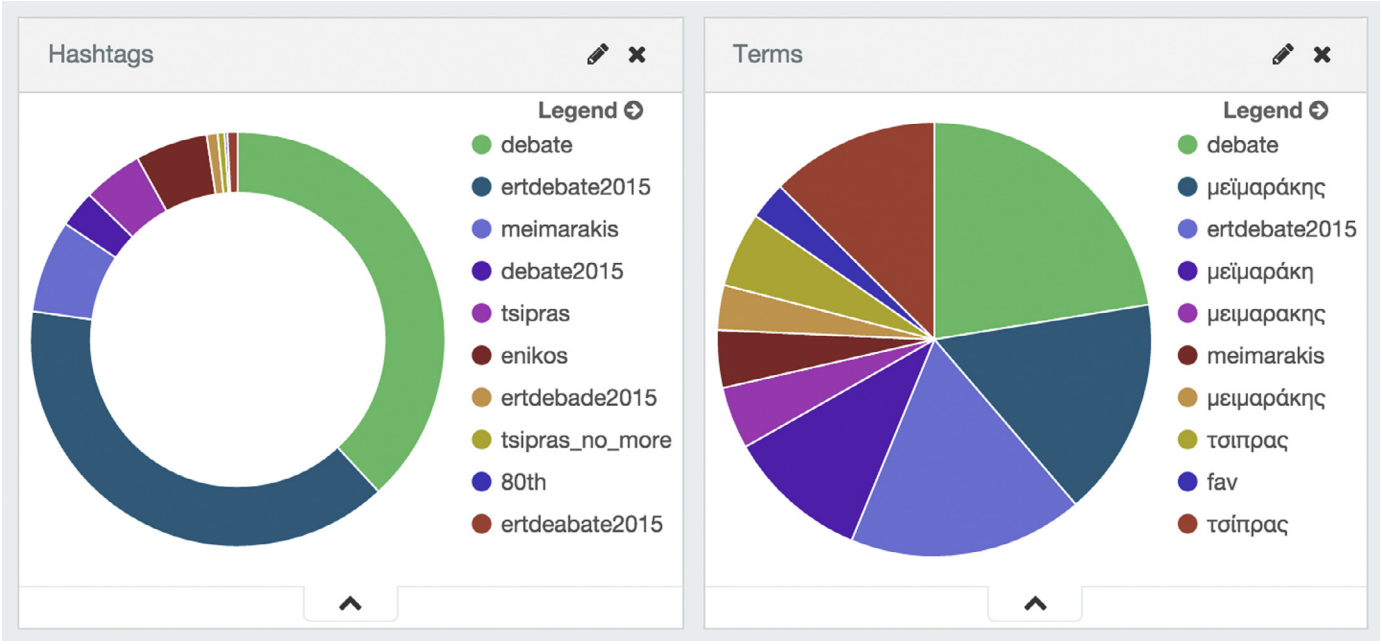

Fig. 6. Top 10 hashtags & top 10 significant terms during the Greek ministerial debate.

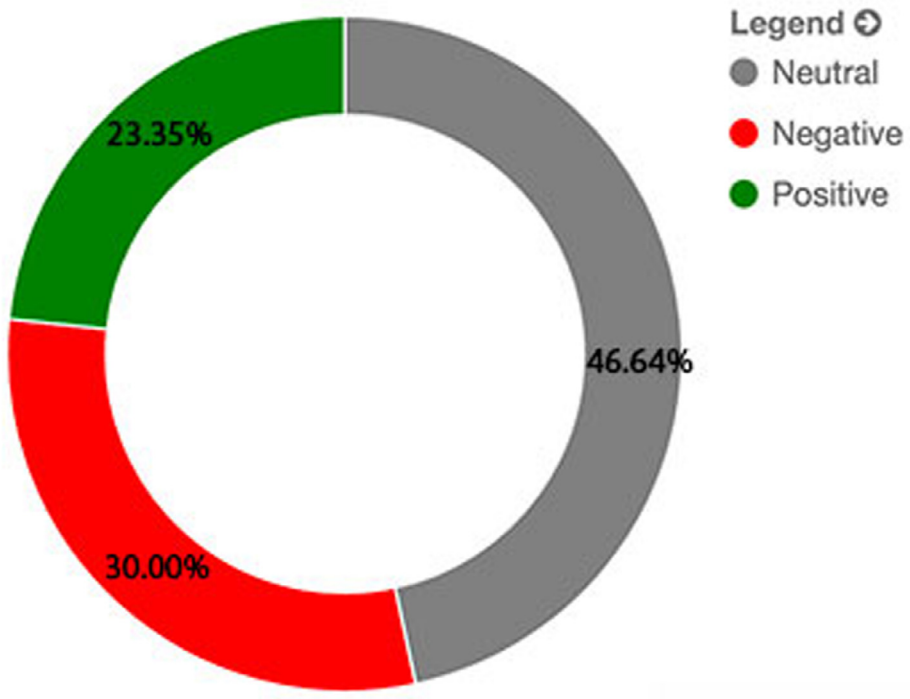

Fig. 7. A snapshot of the polarity of opinions for one of the politicians in the Greek ministerial debate.

5.3. Heterogeneity

Content in Palo Pro 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,[17] 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 Vafopoulos et al. (2012). 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.

6. Case study – Greece's presidential debate

In order to analyze the soundness of our aproach we proceeded with a case study that could easily prove our system working under heavy duty conditions. We took advantage of the fact that a debate was going to be held between the two major political parties of Greece just 6 days before the elections. During the time of

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

the debate it was expected that we would have a burst in the production of data both from online media as well as all active and politicized individuals. In fact we decided to focus on Twitter data and the period to check was decided to be 6–7 h before the beginning of the debate and the duration to be 24 h. The idea is to collect all the tweets related to the Greek language, store them into our Elastic cluster infrastructure, process the documents with our Text Analysis Module and finally visualize the results and provide a real-time monitoring dashboard of the event with the use of Kibana.[18]

6.1. Data

As already mentioned the data that was analyzed derived from Twitter. More specifically, we utilized Twitter Streaming API, customized to be used for the Greek language and for a period of 24 h. We measure the data collected between September 14, 2015, 15:00:00 until September 15, 2015, 14:59:59. What we expect is that some time before the beginning of the debate the volume of data will start to increase, we will have a huge increase during the time of the debate and we expect high volume of data the next day as well. The data was exactly as expected (see Fig. 4), as we had an interaction between 42,419 users who wrote a total of 453,130 tweets. As our focus is the debate itself and the reaction of our mechanism to the burst of incoming data we measure the incoming information during the time of the debate and we find out that 35% of the data of the 24 h was produced within the three hours that the debate lasted.

Although the numbers may sound big, they do not correspond to a big load for PaloPro processing pipeline, since at the peak time of the debate the number of tweets that we had to process hardly surpassed the 20 tweets per second. This number is 20 times bigger from the maximum number of tweets processed in Morstatter et al. (2013)[19] and comparable to the number of tweets per second available in average in the SNAP 476 million tweets dataset (Yang and Leskovec, 2011).[20] Of course, all these numbers are far from the number of tweets tweeted every second, which is around 6000.[21]

Furthermore, as the incoming data is very low during the night the production of data is more than 5 times larger than a normal incoming data rate. In a normal day, the amount of data that arrives to our processing pipeline from news, twitter and other social media streams reaches 2000 sentences per second at peak time. The current setup has been tested using synthetic load and has an average throughput of 3000 tweet-sized texts per second. The advantage of our pipeline is that it can be easily parallelized thus allowing us to add extra processing resources if required, without much effort.

All the aforementioned information imply that we discovered an ideal situation and thus a case study in order to check and analyze our system.

6.2. Results

We first used Greek stop-words in order to collect only Greek tweets from the Twitter API and then we performed text analysis in order to create any metadata for these documents. The final documents were indexed in the Elasticsearch Cluster. Then we used a collection of filters in Elasticsearch in order to create several queries and present them into the Kibana dashboard.

Using Kibana, we were able to quickly compose a dashboard for monitoring how the debate evolves (see Fig. 5), by selecting from a range of charts, which are feed by our Elasticsearch filters. The dashboard is refreshed every 30 s thus giving a complete image of the actual state of the debate and the public opinion for its competitors.

As it can be seen in the detailed Fig. 6 the dashboard is able to visualize the top scoring hashtags and top most frequently occurring terms (ignoring stopwords) and their relative frequency in a pie. It also presents the total number of tweets per politician and the respective distributions over time. It also depicts the tweet polarity against each of the two competitors using only the tweets of the last period (Fig. 7).

7. Conclusions

Palo Pro 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 651 (ICT4Growth project). Authors would like to thank A. Grivas, C. Sardianos, N. Makrinioti and V. Vasalos for their collaboration to the project.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jss.2016.06.012

References

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G.A., Reynar, J., 2008. Building a sentiment summarizer for local service reviews. WWW Workshop on NLP in the Information Explosion Era, 14.

De Rooij, O., Odijk, D., De Rijke, M., 2013. Themestreams: visualizing the stream of themes discussed in politics. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 1077–1078.

Dinter, B., Lorenz, A., 2012. Social business intelligence: a literature review and research agenda. In: Thirty Third International Conference on Information Systems (ICIS 2012).

Giannakopoulos, G., Karkaletsis, V., Vouros, G., Stamatopoulos, P., 2008. Summarization system evaluation revisited: N-gram graphs. ACM Trans. Speech Lang. Process. 5 (3), 5:1–5:39. doi:10.1145/1410358.1410359.

Giannakopoulos, G., Kiomourtzis, G., Karkaletsis, V., 2014. Newsum: "n-gram graph"-based. Innov. Doc. Summar. Tech. 205.

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D., Haug, L.-E., Hsu, M.-C., et al., 2011. Visual sentiment analysis on twitter data streams. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. IEEE, pp. 277–278.

Hatzivassiloglou, V., McKeown, K.R., 1997. Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 174–181.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 168–177.

Jindal, N., Liu, B., 2008. Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 219–230.

Kononenko, O., Baysal, O., Holmes, R., Godfrey, M.W., 2014. Mining modern repositories with elastic search. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM, pp. 328–331.

[18] https://www.elastic.co/products/kibana.

[19] Tweets for Syria have been collected from the Twitter Firehose API. Twitter Firehose is a service that provides the complete set of tweets for a given set of search terms and geolocation bounding box and is available at https://dev.twitter.com/streaming/firehose.

[20] This dataset has an average of 25 tweets per second.

[21] http://www.internetlivestats.com/twitter-statistics/.

Miyoshi, T., Nakagami, Y., 2007. Sentiment classification of customer reviews on electric products. In: Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on. IEEE, pp. 2028–2033.

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. ACM. pp. 1184–1184.

Morstatter, F., Pfeffer, J., Liu, H., Carley, K. M., 2013. Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose. arXiv preprint 1306.5204.

Mukherjee, A., Liu, B., Glance, N., 2012. Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st International Conference on World Wide Web. ACM, pp. 191–200.

Pang, B., Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 115–124.

Petasis, G., Spiliotopoulos, D., Tsirakis, N., Tsantilas, P., 2014. Sentiment analysis for reputation management: Mining the greek web. In: Artificial Intelligence: Methods and Applications. Springer, pp. 327–340.

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

Qiu, G., Liu, B., Bu, J., Chen, C., 2009. Expanding domain sentiment lexicon through double propagation.. In: IJCAI, 9, pp. 1199–1204.

SINTEF, 22 May, 2013. Big data, for better or worse: 90% of world's data generated over last two years.Sci. Daily,

Thet, T.T., Na, J.-C., Khoo, C.S., 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. J. Inform. Sci 36 (6), 823–848.

Vafopoulos, M.N., Meimaris, M., Papantoniou, A., Anagnostopoulos, I., Alexiou, G., Avraam, I., Xidias, I., Vafeiadis, G., Loumos, V., 2012. Public spending: Interconnecting and visualizing greek public expenditure following linked open data directives. Available at SSRN 2064517.

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

Yang, J., Leskovec, J., 2011. Patterns of temporal variation in online media. In: Proceedings of the fourth ACM International Conference on Web Search and Data Mining. ACM, pp. 177–186.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W., 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, pp. 427–434.

Zubiaga, A., Spina, D., Amigó, E., Gonzalo, J., 2012. Towards real-time summarization of scheduled events from twitter streams. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. ACM, pp. 319–320.

Nikos Tsirakis graduated Computer Engineering and Informatics Department, University of Patras, Greece in 2004 and then graduated Masters Program "Science and Technology of Computers" by 2006 in the same department. He received his Ph.D. in computer science from the Department of Computer Engineering & Informatics of the University of Patras, Greece, in 2010. The title of his dissertation was "Data mining techniques and their applications in data management problems and in software systems evaluation". He has published many reviewed conference publications, peer reviewed journal articles and has been invited to contribute with several book chapters. http://students.ceid.upatras.gr/~tsirakis.

Vassilis Poulopoulos obtained his diploma from Computer Engineer and Informatics Department of University of Patras in 2005 and in 2007 he completed his M.Sc. In 2010 he obtained his Ph.D. by creating an innovative platform for multi-lingual worldwide article collection including data mining, data analysis, text extraction, text categorization, text summarization and web personalization. He has more than 35 publications in International Journals, Conferences and Encyclopedias and he obtained 2 times the best paper award.

Panagiotis Tsantilas holds a degree in Physics and an M.Sc. in Computer Science and Business Administration at the University of Glasgow. His articles have been published in several Greek magazines and newspapers, and in 1995 he published his first book. He is the founder and CEO of palo LTD. Palo.gr is the leading news search engine in Greece already holding two awards distinctions in e-volution awards 2013 and in Ermis awards 2012. PaloPro awarded the evolution awards 2014 for the innovative technology used in managing the corporate reputation online.

Iraklis Varlamis is an Assistant Professor at the Department of Informatics and Telematics of Harokopio University of Athens. He obtained a Ph.D. in Computer Science from Athens University of Economics and Business of Greece in 2003. His research interests vary from data-mining and the use of semantics in web mining to graph analysis in social networks. He has published several articles in international journals and conferences in the areas of web document clustering, the use of semantics in web link analysis and web usage mining, word sense disambiguation using thesauruses, etc. More information is available at http://www.dit.hua.gr/~varlamis.