Large-scale Sentiment Analysis for Reputation Management

Conference Paper · September 2013

CITATIONS
3

READS
1,301

4 authors:

Georgios Petasis
National Center for Scientific Research Demokritos
72 PUBLICATIONS **894** CITATIONS
SEE PROFILE

Dimitris Spiliotopoulos
University of Peloponnese
112 PUBLICATIONS **942** CITATIONS
SEE PROFILE

Nikos Tsirakis
University of Patras
38 PUBLICATIONS **302** CITATIONS
SEE PROFILE

P. Tsantilas
Palo Ltd
17 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Georgios Petasis, Dimitrios Spiliotopoulos, Nikos Tsirakis, Panayiotis Tsantilas

Large-scale Sentiment Analysis for Reputation Management

Harvesting the web and social web data is a meticulous and complex task. Applying the results to a successful business case such as brand monitoring requires high precision and recall for the opinion mining and entity recognition tasks. This work reports on the integrated platform of a state of the art Named-entity Recognition and Classification (NERC) system and opinion mining methods for a Software-as-a-Service (SaaS) approach on a fully automatic service for brand monitoring for the Greek language. The service has been successfully deployed to the biggest search engine in Greece powering the large-scale linguistic and sentiment analysis of about 80.000 resources per hour.

1 Introduction

Sentiment analysis and opinion mining are relatively new areas of natural language processing that seek to capture an aspect of text beyond the purely factual. Contrary to facts, which are objective expressions about entities, events and their attributes, opinions are subjective expressions of emotions, feelings, attitudes or sentiments towards entities, events and their properties. One important aspect of opinions is the fact that they have targets: opinions are expressed for objects (i.e. entities or events) and their attributes. There are several levels of granularity regarding the detailing of the target identification in sentiment analysis. The vast majority of approaches that have been presented in the literature can be classified in the following three categories: *a*) Document level: determine whether a document expresses opinion and identify the sentiment of the document as a whole. *b*) Sentence level: identify is the sentence contains opinions and determine the sentiment of the whole sentence. *c*) Attribute level: identify object attributes and determine the sentiment towards these attributes.

Reputation management on the other hand, relates to monitoring the reputation or the public opinion of an individual, a brand or a product. Social Web is of course a valuable resource for detecting and monitoring customer feedback, in order to detect early warning signals to reputation problems and content which damages the reputation of an entity. However, the detection of this information is not an easy task, not only because of the technological challenges the identification and extraction technologies face with natural language processing, but also due to size of the social web and the amount of resources that need to be processed. As a result, the employed technologies must be both accurate in the results that they produce, and computationally efficient in order to be exploited in a commercial environment.

In this paper we present a real world application which applies natural language processing on the large scale, aiming to detect opinion polarity about a vast collection of

individuals, companies and products in the Greek Web, as harvested by the larger search engine in Greece. Commercialised under the brand name "PaloPro", this application is the first commercial automated platform for reputation management in Greece, driven by the co-operation of two companies: Intellitech[1], responsible for the linguistic analysis, and Palo[2] which harvests the Greek Web and commercialises the final product. We will try to present an overview of the "PaloPro" application and the challenges we are facing regarding the linguistic technologies employed for detecting named-entities and opinions about these entities in the context of a large-scale, real-world application for the less linguistically resourced Greek language.

The rest of the paper is organised as follows: In section 2 the application "PaloPro" is presented, while section 3 presents "OpinionBuster", which is responsible for recognising named-entities (section 3.1) and detecting polarity for each recognised entity mention (section 3.2). An empirical evaluation with the help of two manually annotated "gold" corpora is presented in section 4, along with an evaluation on two specific entities, while section 5 concludes this paper.

2 A real-world application for large-scale reputation management: "PaloPro"

PaloPro is a subscription service which aggregates all news, blog posts, discussions and videos in Greek through a simple, friendly and useful tool for monitoring and analysis, in effect 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 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 85% for both the named-entity recognition and the polarity detection tasks.

The data are collected and cleaned by a plethora of real-time crawlers, which aggregate data from different sources, including traditional news and social media such as blogs, Twitter and Facebook posts. The crawling and storage procedure is fully controlled in such a way that the system may provide a near real-time analysis to the end user. 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 2.5 million documents. The main content is collected from about 1400 different websites which are categorised based on their importance (rank) and the news domain of expertise. Most of them are news portals spanning broad domains of news articles.

Blogs are also a main source of content of journalistic orientation. The text is usually informal, sometimes with relaxed syntax. In blog posts, there is, also, a larger number of idiomatic expressions compared to traditional journalistic sources. A very large

[1] Intellitech Digital Technologies PC: http://www.intellitech.gr
[2] Palo Digital Technologies Ltd: http://www.palo.gr

portion of the data comes from social networks such as Facebook, Twitter and Youtube. The system collects and analyses the wall posts, tweets and video comments from all Greek open profiles. Forum posts is another source of content with informal text, like Blogs, but also contains continuity and related text snippets that represent the forum conversations.

PaloPro is organised around the notion of the "workspace" or "dashboard", via which the reputation of sets of user-selectable entities or user-specified keywords are monitored. The user may 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 (figure 2), 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 on to system, as shown in figure 1.

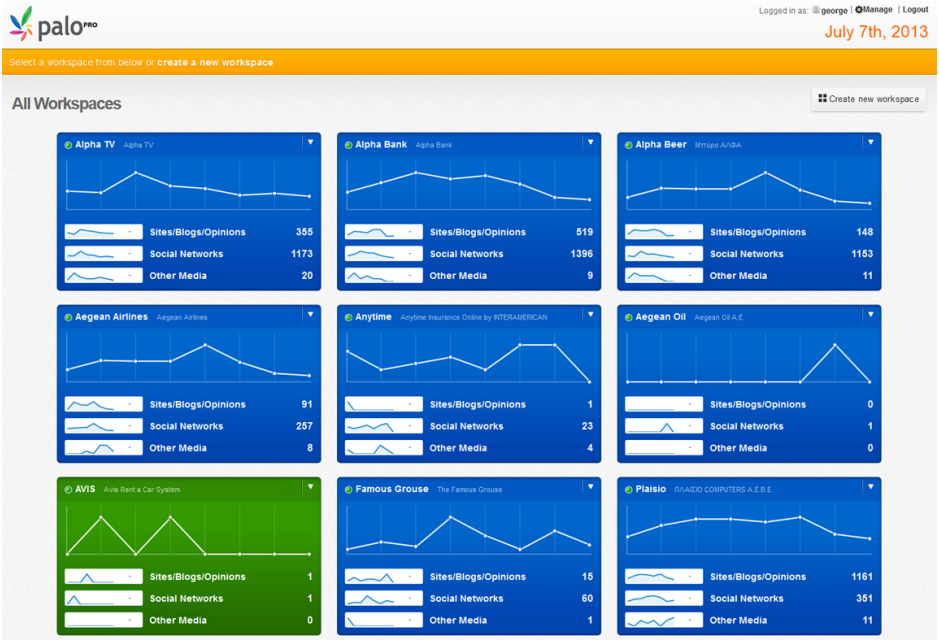

Figure 1: The initial view of PaloPro, presenting the set of user-defined workspaces.

Custom dashboards specific to business of journalistic needs may be created and edited. The auto-generated dashboard information may be accessed at any time and updated instantly with real-time data. Each dashboard, once selected, is initially presented through an overall summary, which presents activity over time (figure 4). This summary is followed by a more detailed report over the various sources, the top

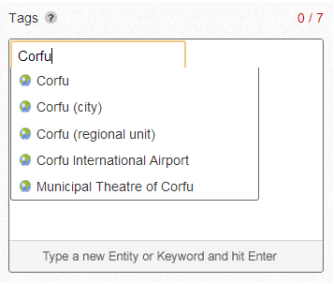

Figure 2: Selecting entities from the system's database, during workspace creation.

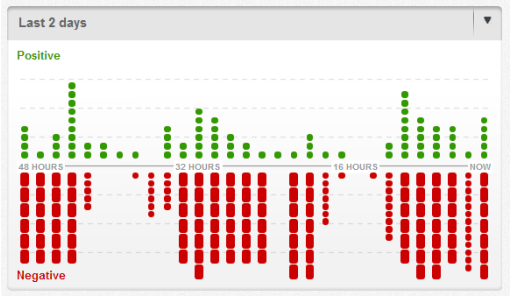

Figure 3: The "Sentiment Radar", providing an overview of the overall opinion polarity for an entity over time.

influencers, and the "sentiment radar", which present various aspects of the analytics. Various levels of abstraction may be accessed, offering schemas that range from an overview report to the detailed information extracted (as represented by the text segments that are analysed in the actual documents) about an entity. For example, figure 5 lists all the postings found in Facebook about an entity, in a specific day.

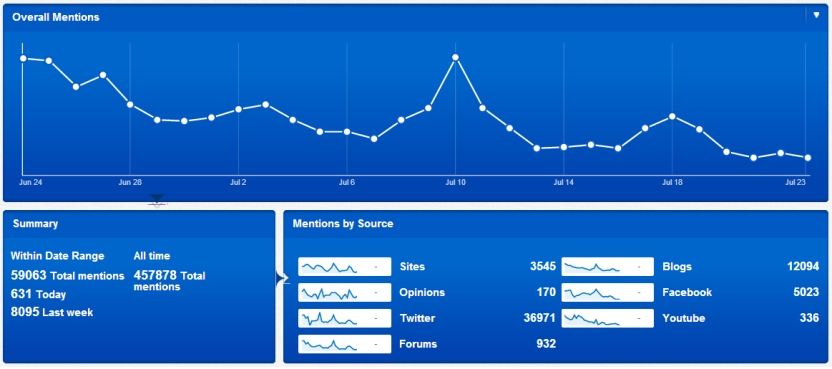

Figure 4: Overall summary of a dashboard, summarising activity over time.

Through the system, a user has access to information related to different entities or keywords such as persons, organisations, companies, brands, products, events etc. that are monitored by the system in the crawled corpora, along with sentiment (currently limited to polarity) 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

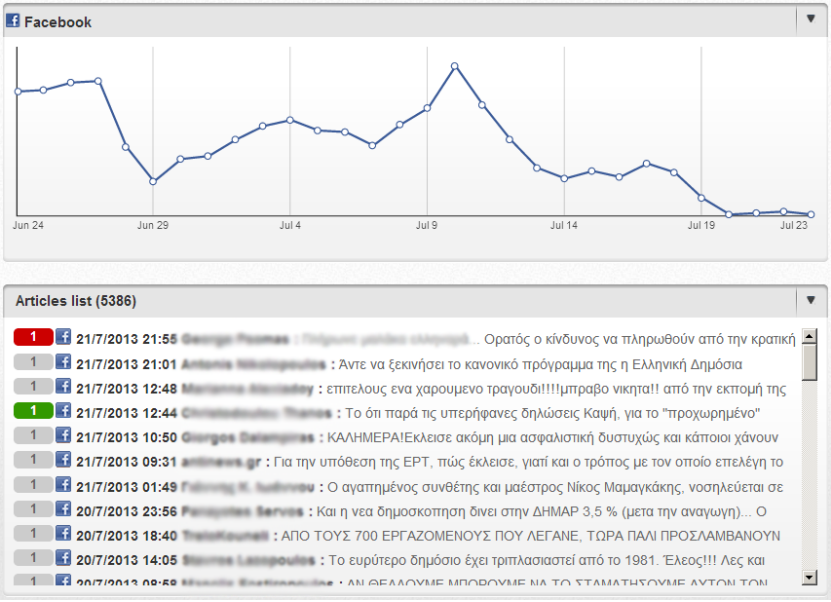

Figure 5: Facebook insights about the closing of the Greek national television organisation on June 11, 2013.

exceeds certain user-configurable thresholds. In addition, within PaloPro, a competitive reputation analysis involves identifying competitive results that rank for content and other parameters, to empower the end user to identify instances where existing content could potentially be promoted and where gaps occur in content that may be enhanced further.

3 Large-scale Polarity Detection for Entities: "OpinionBuster"

OpinionBuster is the product that powers PaloPro, as it is responsible for the extraction of named-entites and the polarity associated with their mentions in texts. Able to locate entities and polarity to a wide range of thematic domains, OpinionBuster integrates state of the art approaches for natural language processing, ranging from ontologies and rule-based systems to machine learning algorithms. OpinionBuster has been developed in the context of the Ellogon[3] (Petasis et al., 2002) language engineering platform: being coded in C/C++, Ellogon offers the required computational efficiency (both in

[3] http://www.ellogon.org

used memory and processing speed) for a commercial product, achieving a processing speed of 100 documents per second, per processing thread, on an Intel 3930K processor.

3.1 OpinionBuster: Named-entity recognition

Named Entity Recognition (NER) is the task of locating mentions of entities related to the thematic domain, and classifying these mentions into categories, with typical categories being names of persons, organisations, companies, and locations, expressions of time, monetary expressions, etc. NER is a well-established technology for English, while a significant number of approaches for other languages can be found in the literature. Furthermore, NER research has been conducted in a wide range of domains ranging from newspaper texts to highly scientific domains such as biomedicine. NER for the news domain has been promoted by a number of evaluation campaigns including the CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), the Automatic Content Extraction (ACE) program (Doddington et al., 2004) and the TAC Knowledge Base Population Evaluation task (Ji et al., 2011). The two most successful and established techniques for NER are supervised machine learning and the use of hand-written rule-sets. Both techniques require a considerable investment of manual effort, either in producing annotated training data or developing rules. As a result, more recent trends in the field tend to exploit large external knowledge sources, such as Wikipedia or DBPedia. Several approaches of NER systems for various languages are presented in (Nadeau and Sekine, 2007), while recent surveys of ontology-based information extraction systems are presented in (Karkaletsis et al., 2011; Petasis et al., 2011; Iosif et al., 2012). From a commercial point of view, NER is also an established technology supported by products such as OpenCalais[4] and AlchemyAPI[5], which support entity recognition in a small set of languages, including English, French, German, Italian, Portuguese, Russian, Spanish, and Swedish. To our knowledge, OpinionBuster is the first product that offers NER for the Greek language.

However, using any approach for reputation management requires a high level of accuracy, and usually requires entities that are not commonly found, such as political parties, or products of a specific company and its competitors. Achieving high accuracy levels without the use of domain knowledge is very difficult, if not impossible. As a result, OpinionBuster exploits a large set of knowledge sources, as well as state-of-the-art approaches for its NER component, including: *a*) Open (and linked) data based on various sources, such as the various languages of Wikipedia (as entities can appear either in Greek or in other languages, depending on their place of origin), various governmental sites that provide lists of parliament representatives and government members, the registry of companies that participate to the Greek stock market, etc. *b*) An ontology of entities maintained internally by Intellitech, which contains entities that are mainly found in the Greek market (such as products and companies), usually associated with a thematic domain which can aid in entity disambiguation. *c*) A NER extraction

[4] http://www.opencalais.com/
[5] http://www.alchemyapi.com/

grammar, specifically designed for the extraction of named-entities. This grammar is a probabilistic context-free grammar and has been automatically extracted from (manually selected) positive examples only, with the help of the eg-GRIDS+ grammatical inference algorithm (Petasis et al., 2004b,a). *d*) A thematic domain identifier, which classifies a document into one or more predefined thematic domains, aiming at providing a disambiguation context for named entity disambiguation. *e*) A machine learning based NER component based on Conditional Random Fields (Lafferty et al., 2001; Sha and Pereira, 2003; Sutton and McCallum, 2012). *f*) A rule-based co-reference resolution component. *g*) A rule-based named entity disambiguation component, which disambiguates mentions of entities that share the same surface forms according to the thematic domain(s) of the document and the context the forms appear within. *h*) A set of filtering rules, that combine the information generated by all the aforementioned components and decide upon the final classification of word forms detected as possible mentions of entities.

The motivation behind the use of so many knowledge sources and processing approaches, is of course the requirement for high accuracy. Detecting unambiguous entity names in Greek such as "Obama" may be trivial, but identifying that "Alpha" refers to "Alpha TV" station, "Alpha" beer or "Alpha Bank" requires contextual and domain information. Similarly, the word form "Aegean" may refer to even more entities, including the Aegean sea, the Aegean Airlines, 2-3 hotels named "Aegean", to "Aegean Oil", to the shipping companies "Aegean Ferries" and "Aegean Flying Dolphins", to "Aegean College", to "Aegean Power", to the newspapers "Aegean Press" and "Aegean Times", to the "Rethimno Aegean" basketball team, to the helicopter model "Aegean Hawk", etc. It is not uncommon for companies to use popular words in their names, such as the company "Πλαίσιο" ("Plaisio"), which has several meanings in Greek, including "frame" and "in the context of". Distinguishing among "στο πλαίσιο της περιοχής" (at the local "Plaisio" store) from the "στο πλαίσιο της έρευνας" (in the context of research) and the rest of the more than 3.000 expressions we have already identified the word "πλαίσιο" is used, is not an easy task.

3.2 OpinionBuster: Sentiment Analysis

Sentiment analysis and opinion mining are research areas of natural language processing that seek to capture an aspect of text beyond the purely factual. Contrary to facts, which are objective statements about entities, events and their attributes, opinions are subjective expressions of emotions, feelings, attitudes or sentiments towards entities, events and their properties. Sentiment analysis is an active and popular research field, where many new approaches are presented each year. Almost all approaches exploit a knowledge source (i.e. a lexicon) of sentiment bearing words, and try to detect sentiment or polarity of a word in its context. Sentiment Analysis within a multilingual context has to meet several challenges, especially since the majority of research on sentiment analysis has concentrated only on monolingually and mostly for English. There are mainly two approaches in multilingual sentiment analysis: statistical and

lexical. Statistical approaches need training data from different languages that are usually sparse, while lexical approaches demand lexical resources in different languages that are not always available.

In order to alleviate the problem of resources scarcity, several approaches have been presented that aim in acquisition of lexica or grammars for sentiment analysis. In (Mihalcea, 2007) an approach is presented that tries to project resources for English into Romanian. Having as a starting point a lexicon of English sentiment bearing words and annotated corpora for subjectivity (subjective/objective), both the lexicon and the annotated corpora are translated into Romanian, and subsequently used for training a statistical classifier for performing sentence level subjectivity classification. The automatic extraction of subjectivity lexica for languages with scarce resources is also studied in (Carmen Banea and Wiebe, 2008), where bootstrapping is applied on a small raw corpus initialized with a basic lexicon of a small set of seed words. The focus of this method is also on Romanian but it is applicable to any other language. Moreover, in (Devitt and Ahmad, 2007), domain specific keywords are selected by comparing the distribution of words in a domain-specific document with the distribution of words in a general domain corpus. The context of each keyword helps to produce collocation patterns. By these local grammar patterns, sentiment bearing phrases are extracted and classified. This approach applied the proposed local grammar approach for performing sentiment classification of financial news streams within a multilingual framework (English, Arabic, and Chinese).

Machine learning approaches have been broadly exploited in sentiment analysis. The cornerstone on sentiment analysis is (Pang et al., 2002) where the authors compare the effectiveness of Naive Bayes, Maximum Entropy and Support Vector Machines in order to classify opinion in movie review documents. In (Andreevskaia and Bergler, 2008) the authors combine two machine-learning classifiers through precision-based vote weighting, in order to explore the challenges of portability across domains and text genres. Their sentiment analysis system integrates a corpus-based classifier with a lexicon-based system trained on WordNet glosses and synsets, aiming at developing a system that relies on both general and domain-specific knowledge.

Besides research on sentiment classification of documents, there is also significant work on sentiment classification of sentences, mainly through supervised statistical learning. Sentiment level classification of newspaper headlines exploiting a variety of machine learning techniques has been the goal of the Affective Text Task of SemEval '07 (Strapparava and Mihalcea, 2007). A similar task, concerning the sentiment classification of newspaper headlines is addressed in (Rentoumi et al., 2009). Moreover in (Rentoumi et al., 2009), structured models such as Hidden Markov Models (HMMs) are exploited in sentiment classification of headlines. The advantage of HMMs against other machine learning approaches employed until now in sentiment analysis is that the majority of them are based on flat bag-of-features representations of sentences, without capturing the structural nature of sub-sentential interactions. In contrast, HMMs, being sequential models, encode this structural information, since sentence elements are represented as sequential features.

OpinionBuster employs a rule-based approach for performing polarity detection, based on compositional polarity classification (Klenner et al., 2009). OpinionBuster is currently restricted to detecting the positive, negative or neutral polarity for entity mentions in texts. It analyses the input texts with the aid of a polarity lexicon that specifies the prior polarity of words, which contains more than 6.000 Greek words (and more than 12.000 unique word forms, as Greek is an inflectional language). As a second step, a chunker is used to determine phrases that are the basis for a compositional treatment of phrase-level polarity assignment. Once polarity has been detected, it is distributed over the involved entity mentions with the help of subcategorization frames for verbs, which in our case are manually constructed patterns aiming at detecting the basic syntactic structures around the verbs, in order to distinguish whether the entity mentions receive or generate the polarity detected in the phrases. In case, however, a verb is encountered that cannot be handled by a rule then a simple heuristic is applied, which assigns the detected polarity to all entity mentions within the phrase.

4 Empirical evaluation

In order to evaluate our system, we have manually annotated a corpus with all mentions of entities, along with the polarity these mentions can be associated with. The corpus has been collected from two popular Greek news papers, "Real News"[6] and "Kathimerini"[7] by monitoring the RSS feeds provided by the news papers. Despite the fact that our system covers a large number of domains, we have opted to monitor only the section related to politics, from both news papers. The two newspapers were monitored for a period of about two months, from December 1^{st}, 2012 to January 31^{st}, 2013. From the collected corpus, 2,300 texts were selected and manually annotated by two annotators. Inter-annotator agreement was measured above 97% for the task of annotating mentions of entities, and above 89% for the task of polarity detection for these entity mentions.

The annotated corpus that has been used as a gold standard, contains 49,511 entity mentions. Regarding named entity recognition, OpinionBuster was able to identify 48,827 entity mentions, out of which 45,789 mentions were identical to manually annotated ones, leading to a precision of 93.78%, with a recall equal to 92.48% and an F1-Measure equal to 93.12%. Regarding polarity detection, OpinionBuster was able to identify a polarity for all entity mentions recognised (48,827), 31,754 of which were correct, leading to a precision of 65.03%, with a recall equal to 64.13% and an F1-Measure equal to 64.58%. Most of the failures related to opinion mining have to do with the absence of subcategorization frames for the involved verbs, without which our system is not able to attribute polarity to the correct entity mentions, resulting in the use of a simple heuristic that distributes the average polarity of a phrase to all entity mentions within the phrase. It should be noted that all the documents of the gold corpus used in this empirical evaluation are news items, and that the corpus does

[6] http://www.real.gr/
[7] http://www.kathimerini.gr/

not contain any comments done by users or other kind of social data, such as tweets or Facebook postings.

4.1 Evaluation on the NOMAD corpus

NOMAD[8] (Policy Formulation and Validation through non Moderated Crowd-sourcing) is an EU-funded project that aims to aid modern politicians in testing, detecting and understanding how citizens perceive their own political agendas, and also in stimulating the emergence of discussions and contributions on the informal web (e.g. forums, social networks, blogs, newsgroups and wikis), so as to gather useful feedback for immediate (re)action. In this way, politicians can create a stable feedback loop between information gathered on the Web and the definition of their political agendas based on this contribution. The ability to leverage the vast amount of user-generated content for supporting governments in their political decisions requires new ICT tools that will be able to analyze and classify the opinions expressed on the informal Web, or stimulate responses, as well as to put data from sources as diverse as blogs, online opinion polls and government reports to an effective use. NOMAD aims to introduce these different new dimensions into the experience of policy making by providing decision-makers with fully automated solutions for content search, selection, acquisition, categorization and visualization that work in a collaborative form in the policy-making arena.

One of the central elements within the NOMAD project is the identification of *arguments* in favour or against a topic, and the opinion polarity expressed on the informal Web towards these arguments. For the purposes of evaluation of the NOMAD system, a "gold" annotated corpus has been created, from 500 articles gathered from the Greek Web, relevant to the thematic domain of renewable energy sources. These 500 articles have been collected by performing queries on popular search engines using suitable terms, without restring their origin. As a result the corpus contains articles from news, sites, blogs, etc. The articles have been manually annotated with entities relevant to renewable energy sources, and arguments towards these entities, related to the advantages and disadvantages of the various energy sources (represented by the entities). In addition, each argument has been labelled with the opinion polarity of the author of each article towards the argument, using three labels "positive", "neutral" and "negative". The corpus has been annotated by two annotators, where conflicts have been resolved by a third annotator in order to create an annotated corpus of high-quality.

OpinionBuster has been applied on this corpus, with the aim to label the manually annotated arguments with opinion polarity information, exploiting both internal (the words of the arguments) and contextual information (the words of the sentence containing the argument). OpinionBuster has been applied on 120 articles (out of the 500 annotated articles), which contained 940 annotated arguments. OpinionBuster assigned a polarity label on 814 arguments, out of which 604 were correct, exhibiting a precision of 74.20%, with a recall equal to 64.25% and an F1-Measure equal to 68.87%. Most of the errors of OpinionBuster on this thematic domain can be attributed to various energy-related

[8] http://www.nomad-project.eu/

terminology that was absent from the polarity lexicon, but also on the absence of domain knowledge about specific objects, such as the negative impact nuclear power has on the public opinion, leading to negativity of arguments related to nuclear power, without this negativity being expressed linguistically in the articles. Finally, OpinionBuster failed to detect correctly situations where comparisons were made, and an energy source that pollutes the environment may be thought positively, if it is polluting less than an alternative energy source.

4.2 Case study: Empirical evaluation on two specific entities

In the previous two empirical evaluations, OpinionBuster has been evaluated with the help of manually annotated corpora, containing a wide range of entities and polarities associated to them. In this third evaluation we are going to evaluate OpinionBuster's output on real system data, as processed by PaloPro. In order to perform this evaluation, we concentrated on only two entities, the mobile telephone company "Vodafone", and the Greek bank "Τράπεζα Άλφα" ("Alpha Bank"). All documents (news articles, blogs, tweets and Facebook postings) referring to both entities, have been collected and evaluated by two annotators: For each document, the annotators have measured whether each entity has received the correct opinion polarity considering all the mentions of the entity in the document. The evaluation results are shown in tables 1 and 2, where the number of documents are displayed (both the total number of documents containing the entity, and the number of documents in which the entity has been labelled with the correct opinion polarity), along with accuracy. As we can see from the results shown in these two tables, OpinionBuster performs quite well on this task, achieving an accuracy around 80%.

	Correct	**Total**	**Accuracy**
Sites	7	9	77.78 %
Blogs	18	18	100.00 %
Facebook	166	183	90.71 %
Twitter	122	158	77.22 %
Overall	313	371	84.37 %

Table 1: Evaluation results for the entity "Vodafone".

5 Conclusions

In this paper the first large-scale real-world application for reputation management for the Greek Web has been presented. Online Reputation Management is a novel and active application area for the natural language processing research community. Sentiment analysis plays a central role in this area, as it provides the main mechanism

	Correct	Total	Accuracy
Sites	62	78	79.49 %
Blogs	55	69	79.49 %
Facebook	7	8	87.50 %
Twitter	100	129	77.52 %
Overall	223	282	79.08 %

Table 2: Evaluation results for the entity "Alpha Bank".

for keeping track of polarity, opinion, attitude, feelings on the web, etc. People use the social media to write news, blog posts, comments, reviews and tweets about all sort of different topics. Even simplistic sentiment analysis, such as polarity detection, can provide valuable incite to reputation management specialists, when tracking products, brands and individual persons, as the specialist can easily determine whether the monitored entities are viewed positively or negatively on the Web. PaloPro provides polarity analysis across the different data inputs and strives for precise and accurate results not only at the document level, but also on the attribute level by extracting opinion polarity for specific mentions of an entity in texts. Mining opinion polarity at the attribute level eliminates false results and makes the analysis more precise for the tracked entities, in comparison to polarity mining at the document or sentence level.

In PaloPro users seek to monitor their company, organization, services or products. The data related to these categories often contains valuable insights about the thoughts, needs and wants of consumers/clients. Most users do online research but most of the time it's impossible to monitor their reputation across all the data channels. PaloPro and its sentiment analysis feature ascertains how news, blogs and social media users affect reputation by applying robust sentiment analysis methods to classify polarity for this reputation. Prior to public release of this feature, potential customer survey results suggested that even opinion polarity mining can became a good starting point for creating an automated management platform for reputation.

Regarding the natural language processing infrastructure, responsible for the recognition of entity mentions and their polarity, the performance has been measured to be above 93% for the detection of entity mentions, which is an excellent result for the Greek language, on the thematic domain of news about politics. On the other hand, polarity detection did not exhibit the same performance levels, measuring a performance of about 64%, lower than the desired performance of 85%.

Acknowledgements

The authors would like to acknowledge partial support of this work from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no

288513. For more details, please see the NOMAD project's website, `http://www.nomad-project.eu`.

References

Andreevskaia, A. and Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of ACL-08: HLT*, pages 290–298, Columbus, Ohio. Association for Computational Linguistics.

Carmen Banea, R. M. and Wiebe, J. (2008). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In Nicoletta Calzolari, Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Devitt, A. and Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. In Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *ACL*. The Association for Computational Linguistics.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program - tasks, data, and evaluation. In *LREC*. European Language Resources Association.

Iosif, E., Petasis, G., and Karkaletsis, V. (2012). *Ontology-Based Information Extraction under a Bootstrapping Approach*, chapter 1, pages 1–21. IGI Global, Hershey, PA, USA.

Ji, H., Grishman, R., and Dang, H. (2011). Overview of the TAC2011 Knowledge Base Population Track. In *TAC 2011 Proceedings Papers*.

Karkaletsis, V., Fragkou, P., Petasis, G., and Iosif, E. (2011). *Ontology Based Information Extraction from Text*, volume 6050 of *Lecture Notes in Computer Science*, pages 89–109. Springer Berlin / Heidelberg.

Klenner, M., Petrakis, S., and Fahrni, A. (2009). Robust compositional polarity classification. In *Proceedings of the International Conference RANLP-2009*, pages 180–184, Borovets, Bulgaria. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In Sidner, C. L., Schultz, T., Stone, M., and Zhai, C., editors, *HLT-NAACL*, pages 196–203. The Association for Computational Linguistics.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and Spyropoulos, C. D. (2002). Ellogon: A New Text Engineering Platform. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 72–78, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E. (2011). *Ontology Population and Enrichment: State of the Art*, volume 6050 of *Lecture Notes in Computer Science*, pages 134–166. Springer Berlin / Heidelberg.

Petasis, G., Paliouras, G., Karkaletsis, V., Halatsis, C., and Spyropoulos, C. D. (2004a). E-GRIDS: Computationally Efficient Grammatical Inference from Positive Examples. *GRAMMARS*, 7:69–110. Technical Report referenced in the paper: http://www.ellogon.org/petasis/bibliography/GRAMMARS/GRAMMARS2004-SpecialIssue-Petasis-TechnicalReport.pdf.

Petasis, G., Paliouras, G., Spyropoulos, C. D., and Halatsis, C. (2004b). Eg-GRIDS: Context-Free Grammatical Inference from Positive Examples Using Genetic Search. In Paliouras, G. and Sakakibara, Y., editors, *Grammatical Inference: Algorithms and Applications, Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI 2004)*, volume 3264 of *Lecture Notes in Computer Science*, pages 223–234, Athens, Greece. Springer Berlin / Heidelberg.

Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., and Vouros, G. A. (2009). Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria. Association for Computational Linguistics.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Tjong Kim Sang, E. F. (2002). Introduction to the conll-2002 shared task: language-independent named entity recognition. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.