# EuroLoveMap: Confronting feelings from News

**Jordi Atserias**[1], **Marieke van Erp**[2], **Isa Maks**[2], **German Rigau**[3], **J. Fernando Sánchez-Rada**[4]

[1]Yahoo Labs Barcelona, [2]VU University Amsterdam, [3]The University of Basque Country,
[4]Universidad Politécnica de Madrid
jordi@yahoo-inc.com, {marieke.van.erp,e.maks}@vu.nl, german.rigau@ehu.es, jfernando@gsi.dit.upm.es

### Abstract

Opinion mining is a natural language analysis task aimed at obtaining the overall sentiment regarding a particular topic. This paper presents a prototype that presents the overall sentiment of a topic based on the geographical distribution of the sources on this topic. The prototype was developed in a single day during the hackathon organised by the OpeNER project in Amsterdam last year. The OpeNER infrastructure was used to process a large set of news articles in four different languages. Using these tools, an overall sentiment analysis was obtained for a set of topics mentioned in the news articles and presented on an interactive worldmap.

**Keywords:** Opinion Mining, Visualisation, Hackathon

## 1. Introduction

Different topics are often presented in news from different perspectives. These perspectives may differ between countries and cultures, and are brought to the fore through different communication outlets. We aim to detect these opinions from news articles from different languages to compare the polarity profiles in different countries with respect to a particular topic. Within NLP research, there is a fair body of work on opinion and sentiment analysis (Pang and Lee, 2008; Liu, 2012). Several toolkits have been developed for the detection of polarity in text, but full multilingual opinion detection which includes the holder of the opinion and the target is still lagging. The OpeNER project plans to deliver an opinion detection tool that is trained on an annotated corpus of political news and aims at a sentence-based detection of opinion expressions with their holders and targets. For this demo, however, we use the rule-based opinion tagger that was available in June 2013.

This paper presents a prototype developed in a single day during the June 2013 hackathon organised by the OpeNER project (Agerri et al., 2013)[1] in Amsterdam.[2] OpeNER aims to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts for six different languages (Maks et al., 2014). Team NAPOLEON used the OpeNER infrastructure[3] and web services[4] to obtain sentiment analyses for news articles in four different languages which were then aggregated into topics per country and presented visually on a map.

In the remainder of this contribution, we detail our system in Section 2., and present some examples in Section 3. We conclude with future work in Section 4.

## 2. Mining feelings from news using OpeNER

During the hackathon, we processed around 22,000 news articles in four different languages obtained from the RSS service of the European Media Monitor.[5] The content as well as some metadata of the newspaper articles was obtained before the hackathon. For this prototype, we decided to focus on English, Spanish, Italian and Dutch. For instance, the topic *gay marriage* was manually translated to the four languages and news articles relevant to this topic were collected and processed. An overall sentiment score was also obtained per language for each topic. Finally, the aggregated score for every topic-language pair was used for colouring a world map.

During the hackathon, we developed some software modules to process each news article through the OpeNER web services. In the remainder of this section, we detail the different steps in the workflow.

The OpeNER architecture consists of several Natural Language Processing (NLP) components. Each component is configured to take the information it requires to perform a specific analysis. KAF (Bosma et al., 2009) is used as linguistic representation. Each of the NLP processing pipelines is deployed as a Cloud Computing service using Amazon Elastic Computing Cloud[6] (Amazon EC2). Figure 1 presents an overview of the OpeNER components deployed as web services.

At the end of the different natural language processing pipelines, the extracted information is combined to obtain polarity clusters for the different topics selected.

**Language Identifier**: This component is responsible for detecting the language of an input news article and delivers it to the correct pipeline.

**Tokenizer**: This component is responsible for tokenising the text on two levels; 1) sentence level and 2) word level. This component is crucial for the rest of NLP components and is the first component in each language processing pipeline.

**Part of Speech Tagger**: This component is responsible for assigning to each token its morphological label, it also includes the lemmatisation of words. Combining the lemma and morphological label, later modules will consult a sentiment lexicon in order to assign polarity values to the words appearing in the news being processed.

**Named Entity Recognition**: This module provides Named Entity Recognition (NER) for the six languages covered by

[1] `http://www.opener-project.org`

[2] `http://opener-fp7project.rhcloud.com/2013/07/18/opener-hackathon-in-amsterdam/`

[3] `http://opener-project.github.io/`

[4] `http://opener.olery.com/`

[5] `http://emm.newsbrief.eu/overview.html`

[6] `http://aws.amazon.com/ec2`

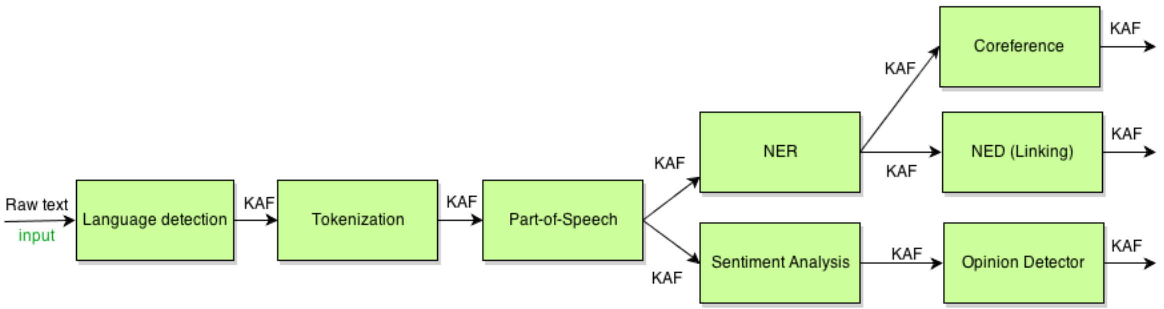

Figure 1: Overview of the components of the OpeNER pipeline

OpeNER and tries to recognize four types of named entities: persons, locations, organisations and names of miscellaneous entities that do not belong to the previous three groups.

**Named Entity Linking**: Once the named entities are recognised they can be identified or disambiguated with respect to an existing catalogue. This is required because the "surface form" of a Named Entity can actually refer to several different things in the world. Wikipedia has become the de facto standard as named entity catalogue. In OpeNER the NED component is based on the DBpedia Spotlight[7] which uses the DBpedia[8] as the resource for disambiguation entities.

**Sentiment Analysis**: The Opinion tagger we used is a rule and dictionary based tagger. It detects positive and negative polarity words (such as 'nice' and 'awful'), as well as intensifiers or weakeners (such as 'very' and 'hardly') and polarity shifters (such as 'not'). In addition, the module includes some simple rules that detect the holders and targets of the opinions related to the positive and negative polarity words.

Finally, the processed news in KAF format are stored and indexed using Solr[9] to easily query and retrieve the news articles about a selected topic. A web service was deployed to obtain json results grouping the scores detected by topic and language. The json results are then presented to the user in a world map.

## 3. Topics on EuroLoveMap

In order to test the prototype we manually selected a small number of topics in English, which were manually translated to Spanish, Italian and Dutch.[10] Table 1 presents the English topics and the corresponding translations in Spanish, Italian and Dutch used in the prototype[11].

Figure 2 presents a screenshot of the EuroLoveMap demo showing the extracted opinions on "gay marriage".

## 4. Future Work

As this is only a very first prototype built in a few hours during the previous OpeNER hackathon, there are several different avenues of research as well as engineering issues that spring from it.

To make the prototype more informative and useful for users interested in analysing trending opitions, possible extension to the prototype could be a trend line or the option to look at different snapshots of the EuroLoveMap. This could provide insights into how the opinions on the different topics evolve in different countries.

For selecting the news sources, we currently use language identification, but one preferably uses the publisher information as there may be news sources aimed at expats in languages different from the country's main language. This would not only be more precise, but also give us access to a host of background information about these sources that can be mined in order to obtain more fine-grained information. Different publishers can for example be classified as more left or right leaning. Having this information enables us to present a more fine-grained analysis of the different perspectives within a country. Information about the publisher or authors of the articles could be further mined to create authority and trust profiles using PROV-O(Moreau et al., 2012). Being able to bring up the actual text of the mined articles would make the EuroLoveMap a useful tool to for example communication scientists or anthropologists.

For this prototype, we manually selected the topics and translated them. Ideally, a system picks up on trending topics, for example by plugging into the European Media Monitor or Twitter trends and detecting which topics would be interesting to analyse. To translate these topics automatically one could imagine using DBpedia or a similar resource.

As processing the articles via the NLP pipelines is a time-consuming process, we are currently working with a static dump of processed articles. Research in for example the NewsReader[12] architecture is underway to optimise NLP pipelines further, but until then the most viable option for updating the demo would be with daily batches that are processed overnight.

[7] http://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

[8] http://dbpedia.org

[9] https://lucene.apache.org/solr/

[10] To scope the prototype, we decided to focus only on four out of the six project languages.

[11] The resulting demo can be found at http://eurolovemap.herokuapp.com/.

[12] http://www.newsreader-project.eu

| English | Spanish | Italian | Dutch |
|---|---|---|---|
| Berlusconi | Berlusconi | Berlusconi | Berlusconi |
| Boston | Boston | Boston | Boston |
| North Korea | Corea del Norte | Corea del Nord | Noord-Korea |
| Obama | Obama | Obama | Obama |
| Putin | Putin | Putin | Poetin |
| CIA | CIA | CIA | CIA |
| Snowden | Snowden | Snowden | Snowden |
| Spain | España | Spagna | Spanje |
| United States, US | Estados Unidos, E.E.U.U. | Stati Uniti | Verenigde Staten van Amerika, VS |
| Netherlands | Holanda | Olanda | Nederland, Holland |
| Italy | Italia | Italia | Italië |
| Germany | Alemania | Germania | Duitsland |
| Gay marriage, homosexual marriage | matrimonio homosexual, matrimonio gay | matrimonio gay | homohuwelijk |

Table 1: Topics and translations

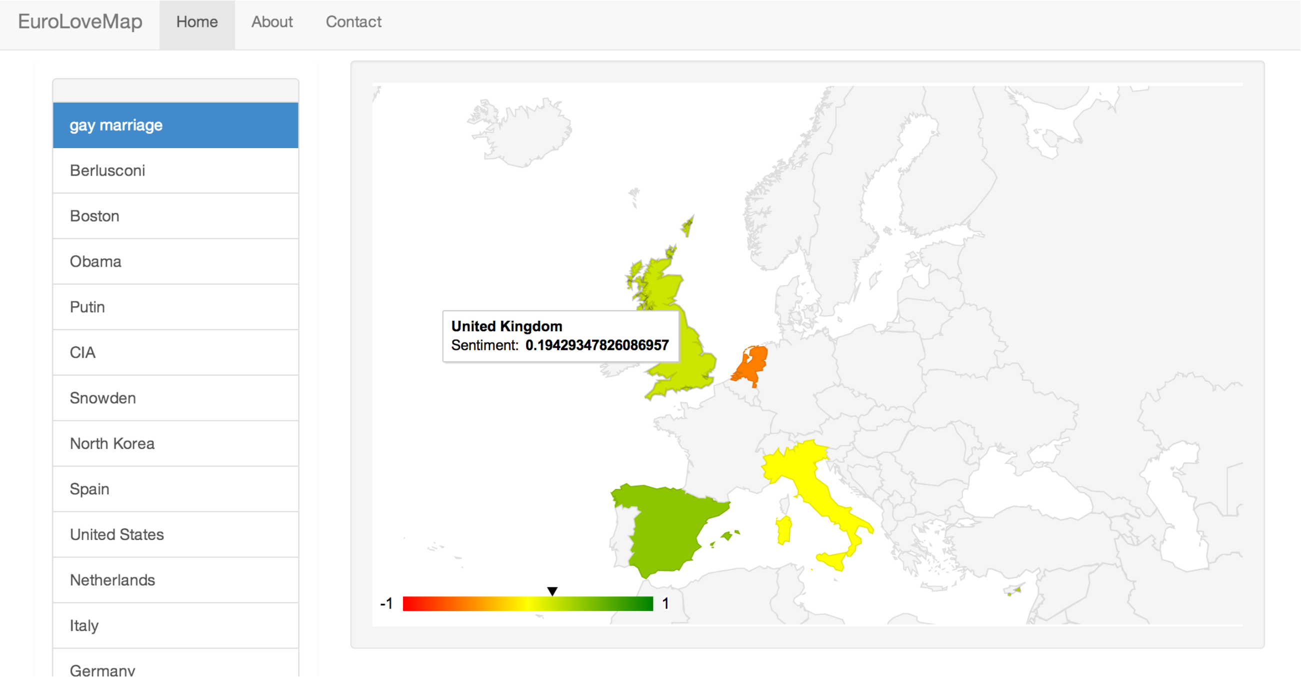

Figure 2: Screenshot of the EuroLoveMap demo showing the extracted opinions on "gay marriage"

## Acknowledgements

This research is supported by the European Union's 7th Framework Programme via the OpeNER project (ICT 296541) and the NewsReader Project (ICT-316404).

## 5. References

Agerri, R., Cuadros, M., Gaines, S., and Rigau, G. (2013). Opener: open polarity enhanced named entity recognition. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'2013).

Bosma, Wauter, Vossen, Piek, Soroa, Aitor, Rigau, German, Tesconi, Maurizio, Marchetti, Andrea, Monachini, Monica, and Aliprandi, Carlo. (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*.

Liu, Bing. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Maks, Isa, Izquierdo, Ruben, Frontini, Francesca, Azpeitia, Andoni, Agerri, Rodrigo, and Vossen, Piek. (2014). Generating polarity lexicons with wordnet propagation in 5 languages. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May.

Moreau, Luc, Missier, Paolo, Belhajjame, Khalid, B'Far, Reza, Cheney, James, Coppens, Sam, Cresswell, Stephen, Gil, Yolanda, Groth, Paul, Klyne, Graham, Lebo, Timothy, McCusker, Jim, Miles, Simon, Myers, James, Sahoo, Satya, and Tilmes, Curt. (2012). PROV-DM: The PROV Data Model. Technical report.

Pang, Bo and Lee, Lilian. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).