The Snowball effect: following opinions on controversial topics

Andoni Azpeitia
Vicomtech-IK4
Donostia-San Sebastian, Spain
aazpeitia@vicomtech.org

Alexandra Balahur
European Commission Joint Research Centre
Ispra, Italy
alexandra.balahur@jrc.ec.europa.eu

Montse Cuadros
Vicomtech-IK4
Donostia-San Sebastian, Spain
mcuadros@vicomtech.org

Antske Fokkens
VU University
Amsterdam, Netherlands
antske.fokkens@vu.nl

Ruben Izquierdo
VU University
Amsterdam, Netherlands
ruben.izquierdo@vu.nl

Abstract

This paper describes a practical application of the OpeNER[1] project technology to find trending topics in different media sources and different languages. We applied our analysis to the global scandal on leaking data involving Edward Snowden using a rule-based opinion mining tool. Results show a diversity of opinions depending on the language and the sources that were analysed. Additionally, we found an interesting division between the opinions expressed in favour of Snowden's actions and in favour of the United States' reaction towards them.

1 Introduction

Since the emergence of Social Media sources and the global need and interest to know what people think about a specific topic, the field of opinion mining has become one of the most business-interesting areas in Natural Language Processing. This field studies the sentiment that opinions generate and is popular in fields such as Brand Monitoring, Social Opinion,(Pang and Lee, 2008; Liu, 2012). For instance, a large number of companies build applications to help customers find out what the market thinks about them, what people think about their immediate competitors or simply to help them follow the news or social events that are of their concern. In order to monitor the opinion of people, there is a lack of easy-to-use resources when opinions come from multilingual sources. Most open-source and ready-to-use tools support only English. OpeNER addresses this problem, by providing a set of tools ready to integrate and use for 6 different languages. Given these tools, our goal in this paper is to present a prototype that was developed during an OpeNER hackathon (Agerri et al., 2013). The aim of this exercise was to explore what could be done in 4 hours (in terms of extracting multilingual opinions of a specific topic) of drafting and programming using the OpeNER tools. In this paper, we describe a system that gathers and analyses opinions on a trending topic (we took as example the "Snowden case"', which was at the time a very "hot" topic) in different languages from different perspectives. This paper is organized as follows: Section 2 presents the methodology performed in the project, Section 3 shows the results of the opinions and Section 4 outlines the general conclusions obtained from the project.

2 Methodology

The project presented in this paper was divided in several blocks in order to make it feasible to implement in four hours. First, three main tasks were taken into account: acquiring datasets, data processing and visualization. These tasks formed the basic preparation blocks which were carried out in parallel by different members of the team cooperating and organizing themselves.

2.1 Data acquisition

Regarding the acquisition of datasets, we employed the RSS feeds provided by the Europe Media Monitor[2] site. These output RSS and Twitter were scraped in order to find all possible news and opinions in several languages related to the "snowball effect" created by the leaking of confidential

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

[2] `http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html`

Language	Num Articles	Num Tokens
German	13	194
English	123	1903
Spanish	31	497
French	7	90
Italian	4	67
Dutch	7	122

Table 1: Number of articles per language

news by Snowden. News and Tweets containing the words Snowden in English, German, Dutch, Italian, Spanish and French were filtered. Table 2.1 shows the number of articles and total number of tokens scraped for each language. For some languages such as English, the amount of online data was bigger and so we could get a higher number of articles related to Snowden.

2.2 Data processing

The datasets were obtained and stored in raw text format and subsequently processed using the available webservices from OpeNER ([3]). OpeNER uses as standard input and output codification between the tools, an XML based format called KAF[4]. The datasets were processed through a pipeline of tools to extract the opinions using the following tools in this particular order:

- Language Detector[5]: This component detects the language that predominates in the document.
- Tokenizer[6]: This component splits the words in the document in order to segment the content units from the punctuation marks.
- POS-tagger [7]: This component detects the morphological category of each word.
- Named Entity Recognizer [8]: This component detects the different the different entities in the document and categorize them.
- Opinion mining: This component detects the opinion of the documents:
  - Ruled-based for Spanish, French, Italian and German [9]
  - Machine Learning-based for English and Dutch [10]

The opinion mining module returns single opinions found in the news, indicating what was the real opinion (the expression), what was this expression about (the opinion target), and who stated it (the opinion holder). The output of the analysis was transformed in JSON[11]. We grouped together these single opinions for each of the language to obtain an overall estimation on how many positive and negative opinions about Snowden can be found in the different languages. Furthermore, we collected opinions about the NSA and CIA. Because the NSA and CIA were the “opposing parties” in this conflict with Snowden, negative opinions about NSA and CIA contributed to positive opinions about Snowden and positive opinions on NSA and CIA increased the score of negative opinions about Snowden. Note that it is likely that most opinions about the CIA and NSA in our dataset can be seen in the context of the controversy about Snowden, since we filtered our data set to exclusively include articles on Snowden.

2.3 Visualization

Regarding the visualization of the results, all agglomerated information obtained from the Data processing was fed in a user-friendly interface which would highlight the main outcome of the analysis. In order to do this, the Django web toolkit [12] was used to generate bubbles in different colors. The bubbles show at first glance the strength of the opinions per language (big bubbles for more frequent opinions), and the polarity of these opinions (green for positive and red for negative). Figure 1 shows the opinions on Snowden we extracted from text in different languages.

3 Results

The results in the graph show that the users posting opinions in different languages have a different view on the Snowden case. Thus, the English and Italian-speaking ones seem to be more negative

[3] http://opener.olery.org

[4] http://www.opener-project.org/kaf/

[5] http://http://opener.olery.com/language-identifier

[6] http://http://opener.olery.com/tokenizer

[7] http://http://opener.olery.com/pos-tagger

[8] http://http://opener.olery.com/ner

[9] http://http://opener.olery.com/property-tagger

[10] http://http://opener.olery.com/opinion-detector

[11] http://www.opener-project.org/json/

[12] http://www.caktusgroup.com/services/custom-web-applications/

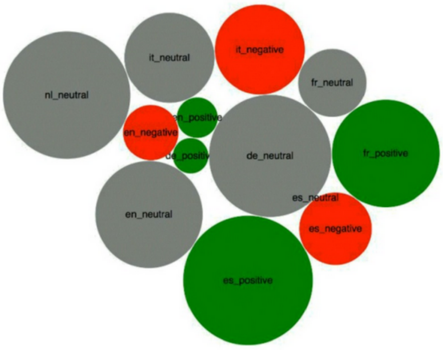

Figure 1: Overview of “Pro-Snowden” opinions in different languages and per polarity class - positive, negative, neutral

about Snowden, while the Spanish and French-speaking ones more positive. In other languages, we can see that the majority of opinions are neutral. Being able to split the results into languages (and possibly, for the future, on text sources) can thus be seen to bring more light on the perception of specific populations of controversial topics. This can be useful to many real-world applications.

4 Conclusions

This paper presents a prototype system drafted in four hours that illustrates the possibility to analyse mainstream and Social Media multilingual texts regarding a specific targeted topic and display the results of the analysis in a user-friendly way. The results in this case show that at first glance, there is a strong difference between the opinion in terms of polarity from Snowden between different languages. In this project we have also demonstrated that OpeNER webservices were ready to use in easy-to-plug-and-play way and performed the analysis fast enough to get meaningful results in a short period of time. Of course, the experiment was based in OpeNER’s organized Hackathon in Amsterdam and this exercise could only take a short period of time. Bearing this in mind, it is possible for the results to be relatively different in other conditions (e.g. by doing a more robust analysis of the datasets acquired and the results obtained).

5 Acknowledgments

This work is part of the OpeNER project funded by the European Commission 7th Framework Programme (FP7), grant agreement no 296451.

References

Rodrigo Agerri, Montse Cuadros, Seán Gaines, and German Rigau. 2013. OpeNER: Open Polarity Enhanced Named Entity Recognition. *Procesamiento del Lenguaje Natural*, 51(0).

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael.

Bob Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.