Investigating the Image of Entities in Social Media: Dataset Design and First Results

Julien Velcin†
Young-Min Kim
Stephane Bonnevay
ERIC Lab., University of Lyon 2
5 av. P. Mendès-France, Bron
f.surname@univ-lyon2.fr

Caroline Brun
Claude Roux
Xerox Research Centre Europe
6 chemin de Maupertuis, Meylan
f.surname@xrce.xerox.com

Jean-Yves Dormagen
Julien Boyadjian
Marie Neihouser
CEPEL, University of Montpellier 1
39 rue de l'Université, Montpellier
f.surname@univ-montp1.fr

Eric SanJuan
Alejandro Molina
LIA, University of Avignon
339 chemin des Meinajaries, Avignon
f.surname@univ-avignon.fr

Leila Khouas
AMI Software
46 av. Daumesnil, Paris
lkh@amisw.com

Anne Peradotto
EDF R&D
22/30 avenue Wagram, Paris
anne.peradotto@edf.fr

Abstract

The objective of this paper is to describe the design of a dataset that deals with the image (i.e., representation, web reputation) of various entities populating the Internet: politicians, celebrities, companies, brands etc. Our main contribution is to build and provide an original annotated French dataset. This dataset consists of 11 527 manually annotated tweets expressing the opinion on specific facets (e.g., ethic, communication, economic project) describing two French policitians over time. We believe that other researchers might benefit from this experience, since designing and implementing such a dataset has proven quite an interesting challenge. This design comprises different processes such as data selection, formal definition and instantiation of an image. We have set up a full open-source annotation platform. In addition to the dataset design, we present the first results that we obtained by applying clustering methods to the annotated dataset in order to extract the entity images.

Keywords: aspect-oriented opinion mining, political data, French corpus

1. Introduction

We aim at studying the image of various kinds of entities (e.g. company, politician) as it is disseminated and viewed on the Internet. "Image" here means a structured and dynamic representation that is voluntarily emitted by an entity or reflected in other people's opinions. It is today considered to be a real challenge not only to respond to specific needs in information retrieval or automatic recommendation, but also to solve important issues in political science and sociology. This research is realized under the ANR ImagiWeb project[1] involving six partners with a duration of three years and a half, ending September 2015.

This project belongs to the sentiment analysis field, with a focus on opinion mining (Pang and Lee, 2008). The main idea is to detect "what people think about a given entity" from documents content. This is implemented within more realistic task such as opinion classification, enabling a number of applications: automatic recommendation, summarization etc. Many datasets have been designed to address this issue (Pang and Lee, 2004; Hu and Liu, 2004; Stoyanov and Cardie, 2008). Each document is generally labeled with an opinion polarity ranging from two (positive and negative) to four (neutral and ambiguous in addition) polarities.

The objective of this paper is twofold. First, we present a new French political opinion dataset built for our project, together with some statistics. Even though we are interested in studying various entities, we start from political data, which gets a great attention recently. Compared to other political datasets in opinion mining (Malouf and Mullen, 2008; O'Connor et al., 2010; Wang et al., 2012) this dataset will be totally original and valuable for the community, not only for covering a new language but also for its very fine annotation granularity. Second, we experimentally show how we can automatically extract an entity image by applying clustering methods to the annotated dataset. Here, the annotated polarities constitute new features representing the document. This is completely different from the traditional opinion classification task, aiming at automatically annotating opinion polarity. At the end of the project, the extraction of entity image will be realized by a two-step approach composed by automatic polarity annotation, and clustering with time evolution.

2. Related Work and Motivation

This project is highly related to the domain of sentiment analysis, and more specifically to opinion mining, see for example (Pang and Lee, 2008) or (Esuli and Sebastiani, 2011). The main idea is to detect "what people think about a given entity" from documents content. This idea is usually implemented within a more realistic task: classifying the opinion expressed about a book, a movie… into a set of predefined polarities (e.g., positive vs. negative). The identification of opinion polarities and strength from within a

[1] http://mediamining.univ-lyon2.fr/velcin/imagiweb

† Corresponding author, julien.velcin@univ-lyon2.fr

text finds its usefulness in plenty of applications, as aforementioned. Many datasets have been designed to address such an issue. For instance, (Pang and Lee, 2004) provided several datasets about movie reviews, one annotated with positive and negative tags, and another with a rating scale. In (Ounis et al., 2009), a set of blog posts over a range of topics has been labeled either "without opinion", or with "negative opinion", "positive opinion", or a mixture of both. For French, the Deft07 competition (Grouin et al., 2009) provided various corpora (movie, book, video games user's reviews and scientific referee's reviews, but also parliamentiary debates) annotated with polarities (positive, negative and/or neutral). More recently, the SemEval contest[2], hosted by the NAACL conference, has proposed to compete on several challenges related to opinion mining. Following the work of (Spina et al., 2012), RepLab[3] is an international evaluation campaign for Online Reputation Management. The available dataset contains various entities annoted with both polarity and aspect. Here, an aspect can be a product, a key people, an event etc.

We want to contribute to these dataset experiments in at least two ways. First, our image definition is not only constituted of polarized opinions, but also of facts (e.g., polls). Moreover, these data are associated with the aspects that structure the image. We find in the literature many examples of opinion associated with aspects, also called topics or features, but these applications usually deal with product reviews (Hu and Liu, 2004). In our case, aspects are fine-grained facets that describe an entity (e.g., politician). Second, even though we are interested in studying various entities, our main case study is French politics, whereas most datasets are usually dedicated to other kind of data, e.g. movie reviews (Ghorbel and Jacot, 2011). Politics has already been addressed in previous works but mostly in English (see for instance (Malouf and Mullen, 2008), or (O'Connor et al., 2010) and (Wang et al., 2012) that deals with the US politics) and rarely with the precision we would like to reach. Furthermore, this dataset has been built with the involvement of specialists in political science. To the best of our knowledge, this dataset will be totally original and really valuable for the community if available.

3. French opinion dataset

One of our target is the image of French politicians, especially two main candidates in the last French presidential election in May 2012, Nicolas Sarkozy (former president) and François Hollande (current president). Data is constantly crawled from a representative microblog, Twitter using Twitter API[4]. In this section, we provide a brief introduction of the annotation procedure.

3.1. Image representation

Since our goal is different from other opinion mining tasks, we have carefully studied the information needed in the dataset before annotating. To this ends, we have studied existing ontology models that capture information about online communities, social networks and opinions that appear on social media. An example of such models are SIOC[5] and MARL[6]. These models rely on W3C's RDF technology which is an open Web standard. Our selected image characteristics are given in Table 1.

Table 1: List of image characteristics.

Source	source of the given text, e.g. the URL of social media.
Time	date of publication of the text.
Holder	author of the text.
Text	the text itself.
Text relation	a relation that exists between two pieces of text e.g. a retweet.
Object	entity on which the opinion is expressed (e.g. a person or a company).
Target	aspect of the entity e.g. the skills of a person.
Opinion	polarity of the opinion expressed in the text.
phrase	specific phrase where the opinion is expressed.

An "Image" is a multi-faceted representation that aggregates a set of opinions or general impressions regarding an entity. Therefore, the annotation should basically encode the different aspects on which the opinion is expressed. Designing appropriate aspects is a key element of the whole annotation process. This step has been done under the supervision of experts in political science. At the end of the process, the following 9 aspects have been selected to describe French politicians: `attribute`, `assessment`, `skills`, `ethic`, `injunction`, `communication`, `person`, `political line`, `project`, adding the `entity` itself and the case of `no aspect` to this list. The aspects are moreover decomposed into sub-aspects such as polls and supports in case of attribute, which signifies the entity's features expressed in pools and supports. 23 sub-aspects have been created for this fine-grained description.

3.2. Annotation procedure

A full system has been developed as a web application in order to easily annotate tweets and blog comments[7]. All annotations are stored in a relational database. Fig. 1 presents a snapshot of this system. The central part shows a tweet with its author and creation time. The annotation goal is to identify the point of view of the text's author by choosing an aspect/sub-aspect pair as well as an opinion polarity among 6 different ones. The annotation procedure is as follows.

Step 1 confirm the text is clearly about the given entity.

Step 2 read and understand the text to answer to the question; how does this text impact the entity's image?

Step 3 select a string that impacts the entity's image and choose a polarity among 6 modalities: very positive (++), positive (+), neutral (0), negative (-), very negative (- -), ambiguous (/). It is possible to differently tag several strings in the same text (rare for the tweets, but quite common for blog comments).

[2]http://www.cs.york.ac.uk/semeval-2013

[3]http://www.limosine-project.eu/events/replab2012

[4]https://twitter.com/twitterapi

[5]http://sioc-project.org/ontology

[6]http://www.gi2mo.org/marl/0.1/ns.html

[7]available freely on http://molina.talne.eu/sentaatool

Step 4 associate a target to the opinion. The annotator uses the closed list located on the left frame in Fig. 1 for selecting:

1. either the overall entity (e.g., the option "Entity" as in: "Sarkozy is not compatible with the French Republic"),
2. one out of the 9 aspects (e.g., "Project: overall" as in "RT @populix: Hollande's project is just hot air"),
3. a pair aspect/sub-aspect (e.g., "Project: economy" as in "the economic project of Hollande will lead France to bankruptcy").

Step 5 give a confidence to the annotation.

4. Statistics on the annotated Twitter data

A deep analysis of the data would help us to develop effective ways to extract the image of an entity. For this purpose, we provide statistics about the annotated French Twitter opinion dataset. Roughly two figures are given: first, a basic figure about opinion frequencies and polarity distributions per entity, and second, opinion disagreement among different annotators per tweet. To handle the subjectivity of annotators, we allow a tweet to be annotated at most three times by different annotators. 7283 unique tweets are annotated, of which 48% are annotated only once, 46% twice, 6% three times. At the end we obtain a total of 11 527 manually annotated tweets; 5278 tweets for François Hollande (FH) and 6241 tweets for Nicolas Sarkozy (NS).

Opinions The opinions are biased to the negative. 53% are negative while 18% are positive. We can see a slight difference between the two entities; for example, 57% of the opinions about FH are negative but only 50% for NS. However, in the period just before the election, the negativity about FH decreases as 41% while that of NS does not change. After the election the negativity about FH increases dramatically to 62%. This justifies the necessity of temporal analysis related to the image, with well-split time periods.

Aspects The 11 aspects are globally well distributed. The `entity` aspect dominates with 23%, followed by `political line` and `ethic` with 13% and 11% respectively. The evolution of frequency of each aspect according to time is interesting. In brief, some aspects are much dependent on time such as `injunction` and `communication`, obtaining very high frequencies just before the election. Both two candidates obtained positive opinions for the `injunction` because this aspect is dedicated to the clear encouragement or warning (rare) about voting for an entity. On the contrary, for the `communication` FH obtained a better score compared with his competitor. Consequently, we need to integrate aspect specificities considering temporal evolution in order to build a precise image.

Disagreement Our annotation may reflect the subjectivity of annotators more than others, because of the granularity of the labels. Analyzing the annotation disagreement among annotators per tweet would help us to understand better the opinion properties. For a reasonable analysis, we make several assumptions: (1) Very positive and positive opinions are treated as identical (same for negative). (2) Ambiguous opinions are ignored. (3) Two aspects, `no aspect` and `entity` are compatible with all other aspects since they refer to the entire entity. Table 2 shows the disagreement rates calculated on three different items, aspect, aspect/sub-aspect pair, and polarity. "Basic" and "Compatibility" rows represent the result excluding or including the compatibility assumption respectively. In brief, polarity disagreement is less than 20% whereas disagreements on aspects (sub-aspects) are more than 60% with basic analysis. However, when taking compatibility into account, disagreements on aspect and aspect/sub-aspect decrease dramatically. Since annotators may have different points of view on different aspects in the same document, we think that this rate of disagreement around 30% is reasonable.

Table 2: Disagreement rates

	aspect	aspect/sub-aspect	polarity
Basic	60%	68%	19%
Compatibility	32%	40%	19%

By taking a closer look at the annotations, we can observe that the disagreement can have various reasons. One classical example is a tweet that describes the result of a public poll. If the poll is in favour of a candidate, some annotators give a positive (resp. negative) polarity whereas others give a neutral polarity since they consider this information as a fact. Another illustration is given when selecting the aspect (sub-aspect) targeted by the opinion. For instance, a tweet related to the case Sarkozy-Kadhafi has been correctly tagged as `ethic` by the two annotators, but the chosen sub-aspect differs (`ethic:honesty` vs. `ethic:case`). This disagreement happens several times; we think that it is due to the fact that different aspects (or sub-aspects) can be selected depending on the individual point of view. Instead of misleading the automatic algorithms, this situation can reflect the diversity of interpretation. This may become a strength if the algorithms we develop take it into account accordingly.

5. Image detection via clustering

Our main objective is to use annotated data to detect the image of a given entity with machine learning techniques. Section 3.1. presented the way a global entity image could be described, we need now to define more systematically the kind of entity image on which we want to focus. A crucial point here is that an image should reflect the opinions of diverse groups, which can be detected using clustering analysis. So an image is defined by its sub-images obtained from an appropriate clustering method as follows. A timestep indicates a time interval between two adjacent times.

Definition (Image). An **Image** of an entity e is defined by $I_e = \{(t_j, c_k^e) | 1 \leq j \leq J, 1 \leq k \leq K\}$, where t_j is j^{th} timestep and c_k^e is k^{th} cluster. Each pair (t_j, c_k^e) corresponds to **sub-image** of the entity.

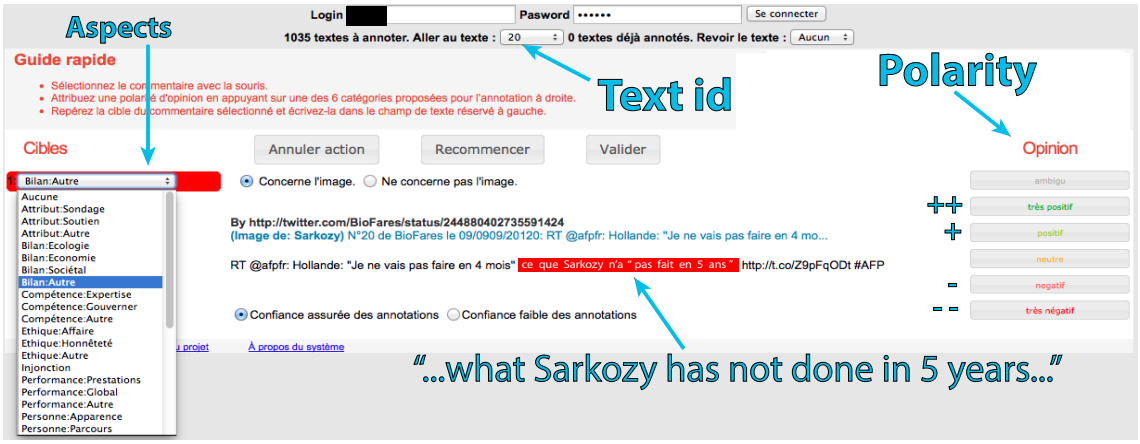

Figure 1: Snapshot of the web application used for the distributed annotation.

Clustering. We first test, in our preliminary experiments, the applicability of different popular clustering algorithms to our Twitter dataset. Timesteps start before and after the election date, however clustering is done independently for each of them. Among the three tested clustering methods, k-means, hierarchical agglomerative model, and multinomial mixture (MM) model, we found that MM yields the best results throughout several quantitative evaluations such as cluster balance, opposite polarity separation, etc.

The polarity, once detected, is represented with four different values "plus", "minus","zero", or "null", which are integrated in the numerical expression for the input instance aspects. Because tweets are very short, we used all the tweets written by a given author for a specific timestep as an input instance, in order to obtain enough co-occurrence elements to conduct a clustering analysis.

A result from a clustering process is graphically represented in Figure 2. These two different clusters have been selected from the MM clustering results, using the annotated tweets about Hollande before the election. The number of cluster is set to 9. The instance components in each cluster are grouped by aspect to show the specific interest of the cluster. The cluster on the upper figure regroups positive opinions about FH especially for `entity` and `injunction` whereas the cluster on the lower figure regroups negative ones for `entity` and `person`.

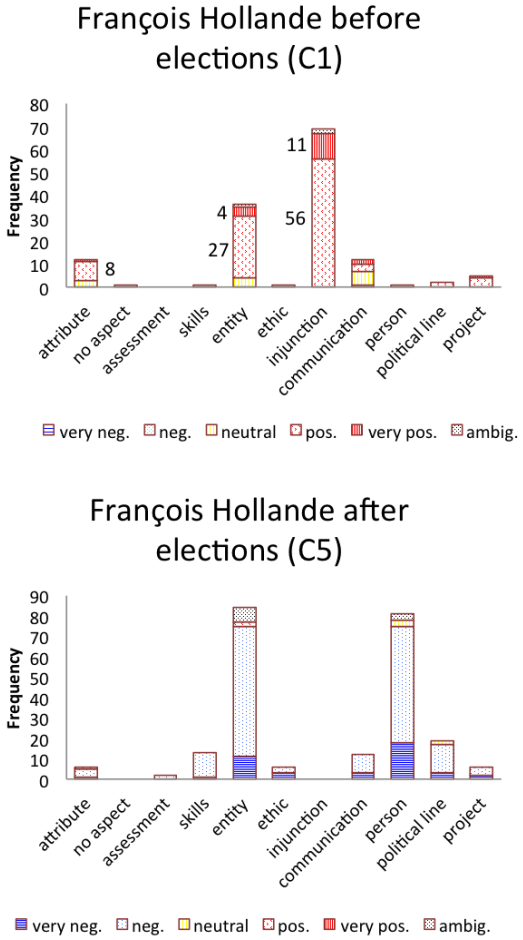

Figure 2: Graphical representation of two clusters (C1 and C2) from a clustering result with 9 clusters

6. Conclusion and perspectives

In this paper, we have presented the statistics and the annotation process of a new French political opinion dataset. We have shown how MM clustering could be applied to the dataset. However, we still need to design a proper methodology to show the evolution of an entity image throughout time, which is at the core of this project. We plan to distribute the dataset to the public in September 2014 on the ImagiWeb official website. This delay is due to the French policy on privacy that requires a clean anonymization procedure before making the dataset available to a broader audience.

At this stage of our project, we see now two main directions for our future work. First, we need to develop an opinion classification method enabling automatic aspect/polarity annotation. Our results, based on both linguistic and statistical approches, seem so far quite promising. Second, we need to develop a model, which will effectively capture the dynamics of an entity image over time. A new generative temporal-aware model, based on the MM model, has been designed and we are currently experimenting it on our dataset.

7. Acknowledgements

This work is funded by the project ImagiWeb ANR-2012-CORD-002-01. Many people have in some extent contributed to this article: J. Ah-Pine, J. V. Cossu, M. Dermouche, M. El-Bèze, C. Favre, G. Gabalda, A. Guille, A. Lauf, S. Loudcher, M. A. Rizoiu, A. Stavrianou, J.-M. Torres-Moreno.

8. References

Andrea Esuli and Fabrizio Sebastiani. 2011. Enhancing opinion extraction by automatically annotated lexical resources. In *Proceedings of the 4th conference on Human language technology: challenges for computer science and linguistics*, LTC'09, pages 500–511.

Hatem Ghorbel and David Jacot. 2011. Sentiment analysis of french movie reviews. In *Advances in Distributed Agent-Based Retrieval Tools*, volume 361 of *Studies in Computational Intelligence*, pages 97–108. Springer.

Cyril Grouin, Martine Hurault-Plantet, Patrick Paroubek, and Jean-Baptiste Berthelin. 2009. Deft'07 : une campagne d'évaluation en fouille d'opinion. In *Revue des Nouvelles Technologies de l'Information (RNTI). Fouille de données d'opinion.*, volume RNTI E-17, pages 1–24.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177. ACM.

Robert Malouf and Tony Mullen. 2008. Taking sides: User classification for informal online political discourse. In *Internet Research*, volume 18, pages 177–190.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media*.

Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. 2009. Overview of the trec-2006 blog track. In *Proceedings of the 15th Text Retrieval Conference (TREC)*. NIST.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Damiano Spina, Edgar Meij, Andrei Oghina, Minh Thuong Bui, Mathias Breuss, and Maarten de Rijke. 2012. A corpus for entity profiling in microblog posts. In *LREC workshop on language engineering for online reputation management*.

Veselin Stoyanov and Claire Cardie. 2008. Annotating topics of opinions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3213–3217. European Language Resources Association (ELRA), may.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120. Association for Computational Linguistics.