Presidential Election 2012: How French politicians tweet?

Flavien Bouillot[1] Dino Ienco[1,2] Stan Matwin[3]
Pascal Poncelet[1] Mathieu Roche[1]

[1] LIRMM – CNRS, 161 rue Ada, 34095 Montpellier, France
[2] Irstea - UMR TETIS 500 rue J.F.Breton 34093 Montpellier, France
[3] University of Ottawa. Ottawa, Ontario K1N 6N5, Canada

Abstract

Tweets exchanged over the Internet is an important source of information even if their characteristics make them difficult to analyze (e.g., a maximum of 140 characters). In this paper, we address the problem of analyzing the opinions expressed through tweets for different communities. More precisely we are interested in following, over time, what is the opinion that a community can have for a specific term or expression (e.g., is the opinion of tweets using the term "crisis" remain the same over time for a political party?). Furthermore we are also interested in shared terms or expressions between different communities. In this case our goal is to evaluate if the opinion expressed changes a lot between communities. Conducted experiments on tweets for the upcoming French Presidential election show very interesting results.

1 Introduction

In recent years, the development of social and collaborative Web 2.0 has given users more active in collaborative networks. Blogs to spread his diary, RSS news to track last information on a specific topic, tweets to publish his actions, are now extremely widespread. Easy to create and manage these tools are used by Internet users, businesses or other organizations to communicate about themselves. This data creates unexpected applications in terms of decision making. Indeed, decision maker can use these large volumes of information as new resources to automatically extract useful information.

Since its introduction in 2006, the Twitter website[1] is so developed that it is currently ranked as the 10^{th} most visited site over the world[2]. Twitter is a platform of microblogging. It means that it is a system for sharing information

[1] http://twitter.com
[2] http://www.alexa.com/siteinfo/twitter.com

where users can either follow other users who post short messages or to be followed. In January 2010, the number of exchanged tweets reached 1.2 billion and more than 40 million tweets are exchanged per day[3]. In this context, different systems can analyze this kind of data [2, 9, 6].

In this paper we briefly introduce a new approach called PoloP (*Political Opinion Mining*) which aims at following the evolution of communities over Twitter. Our main objective is to better understand the opinions that one or more communities can have for both specific terms, i.e. relevant for only one community, and for shared community terms.

Today, the tweets are also becoming an important communication medium in politics. One of the well-known example is the course of the 2008 U.S. election cycle, which resulted in the election of Senator Barack Obama, where it has been noticed how the candidates used the web and social media tools to connect to their followers and organize their campaigns. For instance, just between November 3rd and November 4th (election day), Obama gained over 10,000 new friends, while McCain only gained about 964. On Twitter, Obama gained 2865 new followers between the 3rd and 4th (for a total of 118,107), while John McCain's Twitter account only has a paltry 4942 followers in total[4]. In this paper, we evaluate, through PoloP, how French politicians use the tweets to the upcoming Presidential election in order to highligt the use of some terms or expressions.

The remainder of this paper is organized as follows. Section 2 proposes the problem statement as well as a running example. The PoloP approach is presented in Section 3. In Section 4 we present experimental results conducted on tweets for the upcoming French Presidential election. Finally, Section 5 concludes.

2 Problem Statement

In this section, we better define the problem that we address in this paper. We also propose an example that will be used all over the paper.

In the previous section, we have seen that tweets are merely reduced to 140 characters and among them meta-information can appear in the tweet. In the rest of the paper, we consider without generality that tweets are composed of terms and that, for brevity, "term" or "expression" are equivalent. Basically we assume that an expression could be obtained by any *n*-gram of several terms approach (e.g., *first lady*) in general context [4] and sentiment analysis context [7]. First of all we thus define a tweet as follows:

Definition 1 *(Tweet) Let $T = \langle U_T, \{t_1, t_2, ...t_k\}\rangle$ where U_T stands for the author id of the tweet T and t_i is a term of the tweet. Here we do not have any assumption on the term (i.e., t_i can be any meta information expressed in the*

[3] http://blog.twitter.com/2010/02/measuring-tweets.html
[4] http://www.readwriteweb.com/archives/social_media_obama_mccain_comparison.php

tweet). $user(T)$ is a function giving the author id, noted $user$, of the tweet. For the set of all tweets, we assume that $Terms$ stands for the set of all terms.

As for every tweets we are provided by a context representing several useful information, we thus assume that the following functions are supplied: *Follower(user)* will return the user id where *user* is a follower, $Following(user)$ will return the user id of its follower and $Status(user)$ gives the status of the user and finally $Tweets(user)$ will return the set of all tweets for a specific user.

In the following we consider that we are provided with a set of communities which are defined as follows:

Definition 2 *(Communities - Distribution of terms) Let $C = \{C_1, C_2, ..., C_n\}$ be a set of communities where n is the number of communities we are interested to follow. For every community C_i we assume that we are able to extract its term distribution, called D_{C_i}.*

For each community C_i, it is possible to extract the set of terms expressing opinions or not. More precisely, these sets are defined as follows.

Definition 3 *(General Terms - Specific Terms - Shared Terms) Let $\overline{TO}$ such as $\overline{TO} \subseteq Terms$ be the set of all terms expressing opinions in tweets. For a community C_i, its set of non opinion terms, i.e. specific terms, called TO_{C_i}, is such as: $TO_{C_i} = \{t | t \in D_{C_i} \wedge \nexists_{j,j \neq i} t \in D_{C_j}\}$. TO_S stands for the set of general terms without opinions but that they are very used by all communities. Basically they correspond to stop words in traditional text mining approaches. Finally, ST contains the set of all terms shared by different communities: $ST = \{t | \exists_{(i,j),i \neq j} t \in D_{C_i} \cap D_{C_j}\}$.*

The main difference between TO_S and ST is that in the first one we would like to extract terms which are very often used by all the communities. They could represent article or even tags and do not have a real interest for communities. On the contrary, ST stands for terms which are used in common by a part of communities and not by all of them and thus they can express terms of interest.

Example 1 *For instance, the tag "RT" is used by all communities and must be stored in TO_S while the term "Toulouse" even if used by all communities but but not very often should be stored in ST.*

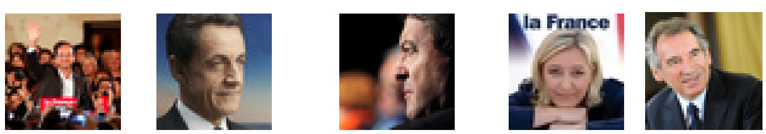
The problems we address in this paper is about the evolution over time of different categories of terms. So in the following we will mainly focus on the three following cases:

- For each term t in TO_{C_i}, we aim to automatically assign a sentiment score to t. Here we are interesting to evaluate the trend of the global opinion that a community can have for very specific terms.
- From terms in ST, our goal is to better asses how shared terms are evolving between communities.

- As PoloP is defined for analyzing in real time new tweets, we must provide a way to automatically associate the user of a tweet to a community.

In the rest of the paper we will consider the following running example. We will focus on tweets exchanged during the French Presidential election. In the beginning of April 2012, ten people are candidates. Figure 1 presents the five following politicians having more than 10% of voting intention. The main political parties are as follows: F. Hollande[5] for the *Socialist Party/PS* (center-left party), N. Sarkozy[6], the current President, for the *Union for a Popular Movement/UMP* (center-right party), J.L. Mélenchon[7] for the *Left Front/FG* (composed primarily of the French Communist Party, the Left Party and the Unitarian Left), M. Le Pen[8] for the *National Front/FN* (nationalist party) and F. Bayrou[9] for the *Democratic Movement/Modem* (center party). Some other parties are: *The Green Party/EELV* (Ecologists) with E. Joly and *New Anticapitalist Party/NPA* (Anticapitalist Party) with P. Poutou.

By analyzing the tweets expressed by politicians and followers of politicians we would like for instance extract that the term "euthanasia" was not used by any political party during the campaign till February 2012 where the socialist party candidate François Hollande gave an interview to the French magazine Marianne, claiming that he is now "not favorable" to the legalization of euthanasia. However, he added that he is "for the right to die with dignity.". Interestingly tweets expressed after this interview, by the PS community have shown that this term meanly occur with tweets having a positive sentiment, i.e. in favor of the candidate. While in the opposite party (UMP), after that Nicolas Sarkozy told to the Figaro magazine that: "Legalized euthanasia risks leading us to dangerous extremes and would be against our conception of the dignity of human beings." all the tweets expressed by the UMP community reveal that euthanasia is associated with a bad opinion. Obviously terms such as the name of the candidate will be associated in a bad opinion in the opposite party but it is interesting to evaluate when such an evolution occurs.



(a)F. Hollande (b)N. Sarkozy (c)J-L. Mélenchon (d)M. Le Pen (e)F. Bayrou

Figure 1: The main French politicians to the upcoming Presidential election

[5] http://en.wikipedia.org/wiki/Franois_Hollande
[6] http://en.wikipedia.org/wiki/Nicolas_Sarkozy
[7] http://en.wikipedia.org/wiki/Jean-Luc_Mlenchon
[8] http://en.wikipedia.org/wiki/Marine_Le_Pen
[9] http://en.wikipedia.org/wiki/Franois_Bayrou

3 The PoloP Approach

In this section we present the PoloP approach. Basically, it performs with the following steps:

1. The first step of the process aims at learning the terms used by a community. Basically, from a set of tweets from different communities C_1, ...C_n, we plan to initialize the following sets: $Tweets$, D_{C_1}, ..., D_{C_n}, $\overline{TO}$, TO_{C_1}, ... TO_{C_n}, TO_S and ST.
2. The second step addresses the problem of assigning a sentiment to all not opinion terms.
3. Finally, the third step deals with new tweets arriving and then addresses the affectation problem of these tweets to a community dynamically.

In the following subsections we present an overview of the various steps.

3.1 Step 1: Extraction of terms used by communities

Actually this step stands for the initialization process. It assumes that we are provided with a set of tweets for every communities.

3.1.1 Acquisition of relevant terms for communities

We assume that several communities are available and for each of them a set of tweets regarding these communities is also available.

Example 2 *From now, we assume that two following communities are available: C_1=centre-left/PS and C_2=center-right/UMP. Let $T_1, T_2, ...T_n$ be the tweets of the community C_1 and $T'_1, T'_2, ...T'_m$ be the tweets of the community C_2. We assume that initially tweets of communities are expressed by leaders of political parties.*

For each tweet T_i of a community, we extract the user-id and then all the associated information about followers and following people by using status, following, follower and Tweets functions. From these tweets we remove tweets from users belonging to the other community in order to keep only tweets relevant for the studied community. As there is no constraints for being followers, by removing such users we would like to minimize the number of followers that do not really belong to a party. For instance, users from the PS party can also be followers from the UMP party in order to follow the behavior of the other community.

These textual data are then gathered and cleaned (by removing tags, and so forth) to only retain relevant terms, i.e. the set $Terms$. This is performed by using any PoS tagging algorithm (e.g., Brill, TreeTagger) and by focusing only on some grammatical labels (e.g., nouns, verbs, ...). Note that at this level, we pay a particular attention to abbreviations or emoticons which will be very useful for improving the sentiment or opinion analysis phase.

3.1.2 Feature selection

The main objective here is to extract from $Tweets$ the set of TO_S (terms without opinions generally used very often by all communities) as well as SH and $\overline{TO}$ (terms of opinions). Each element of $\overline{TO}$ is a tuple: $< term, polarity, score >$. For instance $<$ "*good*"$, positive, 0.75 >$. Basically this set reveals the way that opinions are expressed into the tweets and will be used to improve the affectation of a polarity for terms of SH (Cf. Section 3.2).

With the cleaned tweets we distinguish:

1. $\overline{TO}$ **construction.** All terms expressing a clearly defined polarity as positive or negative by using the score provided by SWF[10] are kept. For instance the term *good* having a high positive score, i.e. 0.75, will be stored in $\overline{TO}$. Note that, in this phase, we do not consider the polarity according to a specific community.

   Our sentiment representation takes also into account specific lexical information such as abbreviation (e.g., *lol*) and emoticons (e.g., *:-), :-(, ...*). Actually this type of information can be very useful to get a precise emotion such as happiness, sadness, anger, sarcasm, and so forth that can be expressed in tweets [5, 8].

2. All terms without any opinion (i.e., other terms), for instance the term "employment" are now considered. They are used to build the two following sets: TO_S (words without opinions frequently used in all the communities) and TO_{C_i} (specific terms without opinions for the community C_i) and SH:

   - TO_S **construction.** Here we select common terms present in all communities. Basically this is performed by first computing for terms occurring in all communities its term frequency such as $fr_{C_i}(X) > k$ where k is a used-defined parameter (in our experiments $k = 0.025$) expressing that we would like to extract only very used terms. Then we store in TO_S each term respecting $\frac{min(fr_{C_i}(X))}{max(fr_{C_j}(X))} \sim 1$.
   - TO_{C_i} **construction.** In this step we select terms which characterizes a community, i.e. the n most frequent terms in relation to a community. In our approach two different methods are considered to select discriminant terms. Traditionally, the $TF\text{-}IDF$ measure gives greater weight to the discriminant terms [10]. As a first step, it is necessary to compute the frequency of a term (*Term Frequency*) corresponding to the number of occurrences of the term in the document[11].

[10] Francophone SentiWordnet (SWF) [1] - When such a tool is not available, all the French words are translated into English and then the English SentitWordnet can be used to get the polarity.

[11] Here *document* is used to be compliant with the original definition of the $TF\text{-}IDF$ measure and refers to a tweet in our context.

Thus, for the document d_j and the term t_i, the frequency of the term in the document is given by the following equation:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ stands for the number of occurrences of the term t_i in d_j. The denominator is the number of occurrences of all terms in the document d_j.

The IDF (*Inverse Document Frequency*) measures the importance of the term in the corpus. It is obtained by computing the logarithm of the inverse of the proportion of documents in the corpus containing the term. It is defined as follows:

$$IDF_i = log_2 \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

where $|D|$ stands for the total number of documents in the corpus and $|\{d_j : t_i \in d_j\}|$ is the number of documents having the term t_i.

Finally, the TD-IDF is obtained as follows:

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i$$

In our case, we propose a new measure [3] which does not calculate the representative terms from the number of documents but rather from the desired community. Thus, we define $IDF_{adaptive}$ as follows:

$$IDF_i^{C_k} = log_2 \frac{|E^{C_k}|}{|\{e_j^k : t_i \in e_j^{C_k}\}|} \tag{1}$$

where $|E^{C_k}|$ stands for the total number of tweets of the community C_k. $|\{e_j : t_i \in e_j^{C_k}\}|$ is relative to the number of elements of the community C_k where the term t_i appears.

To enhance the topics discussed by many users of the community regarding to a topic discussed many times by a small number of user within the community, we define TF-IDF as follows:

$$TF - IDF - NT_{i,j}^{C_k} = TF_{i,j} \times IDF_i^{C_k} \times NT_i^{C_k}$$

With :

$$NT_i^{C_k} = \frac{|\{u_j^k : t_i \in u_j^{C_k}\}|}{|U^{C_k}|} \tag{2}$$

where $|U^{C_k}|$ stands for the total number of users of the community C_k. $|\{u_j : t_i \in u_j^{C_k}\}|$ is relative to the number of users of the community C_k who use the term t_i.

Thus, we compute the value $TF\text{-}IDF\text{-}NT_{i,j}^{C_k}$, i.e. $TF\text{-}IDF_{adaptative}$ weighted with users within the community for each term t_i and can keep the n terms with the highest weights for each community.

Note that this *adaptative* approach which is $TF\text{-}IDF$-based can easily be extended to other measures (e.g., Okapi, LTU, ATC).

- SH **construction.** Finally SH is obtained from the set of all terms that appear in several communities or such as their frequency is lower than k.

3. Finally, all the terms from TO_{C_i} are combined with information coming from the status, the followers and the followings (see Section 3.1.1) to get the distribution of terms for each community D_{C_i}.

3.2 Step 2: Polarity of terms

This steps aims at providing a polarity to terms used by communities, i.e. TO_{C_i}. As tweets are a very special media, here our objective is to highlight that terms in tweets are very often associated with the sentiment expressed in the tweet. For each community, we thus score these terms as follows. We first select the tweets having the term, thanks to $\overline{TO}$, and then score the tweets as positive or negative and affect the polarity to the term. Basically here the hypothesis is based on the following assumption: *as a tweet is reduced to 140 characters and as in the tweet the term exists, the global polarity of the tweet tends to affect the term.* To improve this process we also take into account smileys that are very often used in tweets. Finally we thus affect the polarity of the tweet to the *term*.

3.3 Step 3: How to follow and evaluate?

In the following, we consider only two communities: C_1 and C_2 and we have D_{C_1} and D_{C_2}. We are also interested by following one term, *term* (from step 2).

Let us consider a timestamp of 1 day[12] and then a set of tweets $T_1, T_2, ...T_m$ having the term *term*. For each tweet T_i: we first apply $user(T_i)$ to get the user id of the tweet. From this user id (and the associated elements follower, following, ...) we are able to compute the distribution of terms for the user D_{user}. This distribution will be used to know in which class the tweet of the

[12] Actually this operation can be performed on different time granularities according to the end user.

user will be affected (thanks to D_{C_1} and D_{C_2}). For the moment, this operation is performed by using the cosine function to compare D_{user} with the available D_{C_i}. By applying the polarity of terms step, we can thus affect a new polarity to the tweet and to every terms of the tweet. By using some aggregative functions, this information can, for instance, be used to plot the evaluation of degree of polarity of terms for a community over time.

4 Experiments

4.1 Corpus

For our experiments we construct a corpus of tweets obtained via a Tweeter API by following 200 French political people from different parties cited on the Web site www.elus20.fr. Following and followers tweets of these politicians were acquired in real time. From the 12th December 2011 to the 17th April 2012, we thus obtained 1,146,617 tweets.

For each tweet, the language was automatically identified by using Textcat[13] and the recognized language is used to apply the specific Part-of-Speech Tree-Tagger tool.

4.2 Preliminary results

The used data give us preliminary conclusions. First, Table 1 presents the more retweeted users. The information regarding the often retweeted accounts gives an indication about the influence of political leaders (see Table 2).

Number of tweets	Recipients
112765	@nicolassarkozy
104944	@fhollande
56115	@nadine __morano
27486	@melenchon2012
22777	@eric_besson
21803	@bayrou
15128	@jf_cope
13801	@evajoly
12762	@vpecresse
12457	@ump

Table 1: Tweeter users

Using the method described in Section 3.1.2 (i.e., $TF\text{-}IDF_{adaptive}$), the words having higher scores are ranking (see Table 3). These results show that the current Presidential majority (i.e., UMP) cites often the candidate of the

[13] http://odur.let.rug.nl/∼vannoord/TextCat/

Number of tweets	Retweeted users
118204	rt @nicolassarkozy
91552	rt @melenchon2012
78604	rt @fhollande
28999	rt @ump
17485	rt @partisocialiste
14126	rt @nadine__morano
13479	rt @evajoly
9818	rt @cecileduflot
9814	rt @manuelvalls
9347	rt @royalsegolene

Table 2: Retweeted users' messages

main opposite party (i.e. *Hollande*).

Note that the term *Sarkozy* (i.e., the name of the current President) is not in the lists because it has not been recognized as discriminant (i.e., it is used by all the communities) but appears in the SH set.

Finally the specific communauties (i.e., Ecologists (EELV), Left Front (FdG)) returns very specific vocabulary (i.e., *pollution*, *nucléaire*, *insurrection*, *limoger*).

In order to visualize these results, a word cloud can be used (see Figures 2 and 3). The size of the words is proportional to the rank of the word from the discriminant criterion.

Figure 2: An illustration of the word cloud for the UMP.

5 Conclusion

People participating in on-line forums, microblogging or discussing on social networks leave behind them digital traces and of their opinion on a variety of

PS	UMP	EELV
hollande	hollande	nucléaire
campagne	français	écologiste
soir	soir	acter
dire	dire	écologie
pari	merci	pollution
pouvoir	concorde	projet
français	réunion	européen
changement	fort	merci
candidat	campagne	centrale
parler	bon	dire
34 537 tweets	27 539 tweets	7 916 tweets
MoDem	FdG	FN
produire	insurrection	marine
rassemblement	bastille	front
soir	front	parrainage
campagne	gauche	national
français	limoger	présidentiel
seul	révolution	communiquer
zénith	voter	var
falloir	meeting	français
dire	soir	réaction
suivre	falloir	enregistrement
5 998 tweets	5 219 tweets	283 tweets

Table 3: The Top 10 discriminant terms for 6 communities.

topics. If we knew how to aggregate and cumulatively interpret this data, we could take the pulse of the community on a given issue. For those interested in shifts of public opinion, this provides an attractive possibility of mining the voice of the people and may eventually replace public opinion polling. An additional advantage of these applications is that they deliver the pulse of the community not only to decision makers, but to the community members themselves, and will likely become one of the tools of e-democracy.

On Twitter alone, there are hundreds of millions of messages exchanged each day. While there is considerable enthusiasm being expressed for the potential pro-social contributions that Web 2.0 applications might make to optimizing human creativity, incubating innovation, informing the public and reinvigorating democracy in the process, considerable challenges remain in regard to rendering this information useful to all Internet users.

In our joint project we develop algorithms for efficient clustering, classifi-

Figure 3: An illustration of the word cloud for the PS.

cation, topic analysis and emotion analysis of social media discussions for this kind of social data.

This paper focuses on the study of tweets in the context of the Presidential French election. We plan to study the emotion we can find in this kind of data. A global process is proposed. Currently we have developed the first steps of this process in order to extract discriminant vocabulary for each community.

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 2010.

[2] J. Benhardus. Streaming trend detection in twitter. In *National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval*, University of Colarado, 2010.

[3] S. Bringay, N. Béchet, F. Bouillot, P. Poncelet, M. Roche, and M. Teisseire. Towards an on-line analysis of tweets processing. In *"Proceedings of DEXA (2), Springer Verlag, LNCS*, pages 154–161, 2011.

[4] B. Daille. Morphological rule induction for terminology acquisition. In *COLING*, pages 215–221, 2000.

[5] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of International Conference on Computational Linguistics (COLING)*, 2010.

[6] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of 2010 International Conference on Management of Data (SIGMOD 2010), Demostration*, 2010.

[7] A. Pak and P. Paroubek. Microblogging for micro sentiment analysis and opinion mining. *TAL*, 51(3):75–100, 2010.

[8] L. Ruan. Meaningful signs - emoticons. *Theory and Practice in Language Studies*, 1(1):91–94, 2011.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of 19th World Wide Web Conference (WWW 2010)*, 2010.

[10] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.