French Presidential Elections: What are the most Efficient Measures for Tweets?

Flavien Bouillot, Pascal Poncelet, Mathieu Roche, Dino Ienco, Elnaz Bigdeli, Stan Matwin

▶ **To cite this version:**

Flavien Bouillot, Pascal Poncelet, Mathieu Roche, Dino Ienco, Elnaz Bigdeli, et al.. French Presidential Elections: What are the most Efficient Measures for Tweets?. PLEAD: Proceedings of the Workshop Politics, Elections and Data, 2012, Maui, United States. pp.23-30, 10.1145/2389661.2389669 . lirmm-00801028

HAL Id: lirmm-00801028

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00801028

Submitted on 21 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

French Presidential Elections: What are the Most Efficient Measures for Tweets?

Flavien Bouillot
LIRMM
16 rue Ada, Montpellier France
bouillot@lirmm.fr

Pascal Poncelet
LIRMM
16 rue Ada, Montpellier France
poncelet@lirmm.fr

Mathieu Roche
LIRMM
16 rue Ada, Montpellier France
mroche@lirmm.fr

Dino Ienco
UMR TETIS - Irstea
361, rue J.F. Breton,
Montpellier, France
dino.ienco@teledetection.fr

Elnaz Bigdeli
University of Ottawa
800 King Edward av.
Ottawa, Ontario, Canada
elnaz@site.uottawa.ca

Stan Matwin
University of Ottawa
Ottawa, Ontario, Canada
Institute for Computer Science
Polish Academy of Sciences
stan@site.uottawa.ca

ABSTRACT

Tweets exchanged over the Internet are an important source of information even if their characteristics make them difficult to analyze (e.g., a maximum of 140 characters; noisy data). In this paper, we address the problem of extracting relevant topics through tweets coming from different communities. More precisely we are interested to address the following question: which are the most relevant terms given a community. To answer this question we define and evaluate new variants of the traditional $TF\text{-}IDF$. Furthermore we also show that our measures are well suited to recommend a community affiliation to a new user. Experiments have been conducted on tweets collected during French Presidential and Legislative elections in 2012. The results underline the quality and the usefulness of our proposal.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Experimentation

Keywords

Tweet Analysis, Information RetrievalMeasure, Community Detection, Political Data

1. INTRODUCTION

In recent years, the development of social and collaborative Web 2.0 underline the central and active role of users in collaborative networks. Blogs to spread diaries, RSS news to track last information on a specific topic, tweets to publish social 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 represents an important source of information that can be exploited in the decision making process. Indeed, decision maker can exploit these large volumes of information to automatically extract useful piece of knowledge.

Since its introduction in 2006, the Twitter website[1] has become so popular that it is currently ranked as the 10^{th} most visited site over the world[2]. Twitter is a platform for microblogging. It means that it is a system for sharing information where users can either follow other users who post short messages, or can be followed. In January 2010, the number of exchanged tweets reached 1.2 billion and more than 140 million tweets are exchanged per day[3]. In this context, different systems were proposed to analyze this flow of information [2, 9, 7]. The analysis of tweets can combine different types of information such as timeline and sentiment features [12]. In the context of blog and micro blog studies, specific features can be taken into account such as hashtags [1]. The normalization of these features by using specific heuristics [4] and speech recognition devices can improve the quality of the final result [5].

The growing use of this technology starts to influence many aspect of the real life. One example of impact in real life is the use of social media in politics campaign [8]. Well-known high-impact use of Twitter happened in the course of the 2008 U.S. election cycle, which resulted in the election of Senator Barack Obama. In this episode, 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

[1]http://twitter.com
[2]http://www.alexa.com/siteinfo/twitter.com
[3]http://blog.twitter.com/2011/03/numbers.html

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].

The work presented in this paper is a part of the POLOP Project[5] (*Political Opinion Mining*) which aims to cope with the analysis of the evolution of French political communities over Twitter during 2012 both in terms of relevant terms, opinions, behaviors. 2012 is particularly important for French political communities dues the two main elections: Presidential and Legislative. Figure 1 presents the timeline with the main events related to this period. As we can notice the official campaign started in April even if the main candidates were known in December. The 6th of May was the final Presidential election where F. Hollande has been elected and the legislative elections were finished one month after.

In this paper we address the problem to select specific keywords for different communities over Tweet social media. In particular we develop our approach in the context of Political opinions analysis. We are particularly interested in the best measures to evaluate the most relevant terms for a specific community. Actually lots of efficient measures (e.g. $TF\text{-}IDF$, $Okapi\text{-}BM25$) which are statistics have been proposed by the Information Retrieval or the Text Mining fields to extract the most representative words in documents. Our main contributions are the following:

- We propose two new measures and compare them with the well-known $TF\text{-}IDF$. These measures were specifically designed to better highlight the importance of terms for communities;
- We show how our measures are also very useful for assigning a new user interested in the domain shared by communities, i.e. political domain in our concern, to the appropriate community;
- We conduct experiments on more than 2,122,012 tweets from 213,005 users and particularly on six political parties in order to evaluate both the measures and the classification.

The remainder of this paper is organized as follows. Section 2 proposes the problem statement as well as a running example. New $TF\text{-}IDF$-based measures are defined to extract relevant terms used by communities are presented in Section 3. In Section 4 we present experimental results conducted to compare measures. Section 5 reports the results obtained when affecting a new user to a specific community. Finally, Section 6 concludes and presents future work.

2. PROBLEM STATEMENT

In this section we define our problem and we supply practical examples to illustrate our approach

First of all we recall the main characteristics of tweets. Tweets are merely reduced to 140 characters. When a user follows a person, it receives all messages from this person, and conversely, when that user tweets all his followers will receive its messages. Tweets are associated with meta-information that cannot be included in messages (e.g., date, location ...) or included in the message in the form of tags having a special meaning: for example the tag @username means that you are sending a message to a particular user, the # topic assigns a specific topic, RT means that the message was retweeted, i.e. sent to all the followers. In the rest of the paper, we consider without loss of generality that a tweet is composed of terms and we use the word *term* and *expression* interchangeably. More formally a tweet is defined as follows: 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 make any assumption on the term (i.e., t_i can be any meta information expressed in the tweet).

In this paper we focus on different communities. We consider a community as a set of twitter accounts belonging to people or organizations who share political opinions. There may be accounts of political figures, official accounts of political parties or accounts created during the campaign to support a candidate. But this may also be people who do not belong to a political party but that interact with account mentioned above.
Let $C = \{C_1, C_2, ..., C_m\}$ be a set of communities where m is the number of communities we are interested to follow. For every community C_i we assume that we are able to extract its user distribution, called D_{U_i} and its term distribution, called D_{C_i}. In other words, D_{U_i} stands for the users of a community and D_{C_i} stands for the set of terms used by users of a community.

The problem we address in this paper is the following. For a set of communities, $C = \{C_1, C_2, ..., C_m\}$, how to efficiently build $\text{TOP}_{C_i}\text{-}K$, the subset of D_{C_i} which corresponds to the K most specific terms for a specific community C_i with $1 \leq i \leq m$.

In the rest of the paper we consider the following running example. We focus on tweets exchanged during the French Presidential election. In the beginning of April 2012, ten people were candidates. Figure 2 presents six politicians with more than 5% of voting intent. The main political parties and leaders are as follows: F. Hollande[6], the current president, for the *Socialist Party/PS* (centre-left party), N. Sarkozy[7] for the *Union for a Popular Movement/UMP* (centre-right party), J.L. Mélenchon[8] for the *Left Front/FG* (composed primarily of the French Communist Party, the Left Party and the Unitarian Left), M. Le Pen[9] for the *National Front/FN* (nationalist party), F. Bayrou[10] for the *Democratic Movement/Modem* (center party) and E. Joly[11] for the *green/EELV* (green party).

By analyzing the tweets expressed by politicians and followers of politicians, for instance, we have found out that the terms *Change*, *Accommodations* or *Health* were frequently used by the socialist party candidate Francois Hollande while *President* or *Hollande*, i.e. the name of the opposite party leader were extensively used by the Union of Popular Move-

[4]http://www.readwriteweb.com/archives/social_media_obama_mccain_comparison.php
[5]http://www.lirmm.fr/~bouillot/polop/polop.html
[6]http://en.wikipedia.org/wiki/François_Hollande
[7]http://en.wikipedia.org/wiki/Nicolas_Sarkozy
[8]http://en.wikipedia.org/wiki/Jean-Luc_Mélenchon
[9]http://en.wikipedia.org/wiki/Marine_Le_Pen
[10]http://en.wikipedia.org/wiki/François_Bayrou
[11]http://en.wikipedia.org/wiki/Eva_Joly

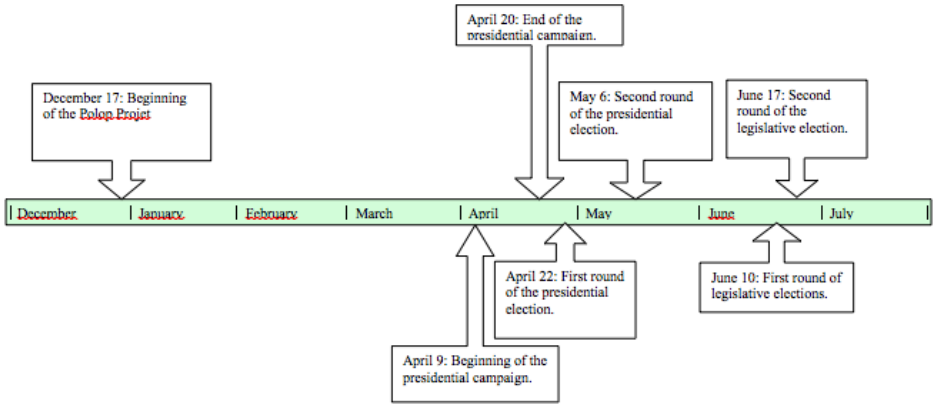

Figure 1: French elections on 2012.

ment. Interestingly we would like to highlight some specific terms used by extreme parties. For instance, the term *Revolution* was used only by the Left Party. Furthermore by considering different time granularities we would like to highlight specific terms used for instance for a week preceding important events (e.g. political meeting, ...).

F. Hollande *N. Sarkozy* *J-L. Mélenchon*

M. Le Pen *F. Bayrou* *E. Joly*

Figure 2: The main French politicians of the Presidential election

More precisely, let us consider the following distribution of terms, denoted by t_i for brevity, for the PS and the UMP: $D_{PS} = \{t_1, t_{12}, t_{13}, t_{14}, t_{15}, t_{16}, t_{17}, t_{18}\}$ and $D_{UMP} = \{t_1, t_{22}, t_{23}, t_{24}, t_{25}, t_{26}, t_{27}, t_{28}\}$. We can notice that t_1 belongs to both D_{PS} and D_{UMP}. The three most relevant terms in D_{C_i} can be:

- TOP_{PS}-3 $= \{t_1, t_{12}, t_{17}\}$
- TOP_{UMP}-3 $= \{t_1, t_{24}, t_{25}\}$

Here, we can notice that t_1 can appear both in TOP_{PS}-3 and TOP_{UMP}-3. This is because this term is not used by all communities and then is becoming relevant for the PS and UMP among others.

3. EXTRACTING RELEVANT TERMS USED BY COMMUNITIES

In this section, we propose two measures to extract discriminant terms for a community, i.e. TOP_{C_i}-K. First of all, as these measures are TF-IDF-based we recall its definition. Second we introduce and justify the two new measures.

Traditionally, the TF-IDF measure (Term Frequency - Inverse Document Frequency), introduced by [10], gives greater weight to the discriminant terms. 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[12]. 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 [3], in a very different context, we proposed a new measure called TF-$IDF_{adaptative}$. This measure has been defined in order not to focus on the number of documents but rather to the number of documents for a specific class. So in our case, this measure seems well adapted to the concept of communities as it does not calculate the representative

[12]Here *document* is used to be compliant with the original definition of the TF-IDF measure and refers to a tweet in our context.

terms from the number of documents but rather from the desired community. Thus, we define $IDF_{adaptive}$ as follows:

$$IDF_i^{C_l} = log_2 \frac{m}{|\{C_l : t_i \in C_l\}|} \quad (1)$$

where m stands for the total number of communities. $|\{C_l : t_i \in C_l\}|$ is the number of communities C_l where the term t_i appears.

Basically, usage TF-IDF on tweets can be very different than on other kinds of documents. As tweets are easy to write, it is clear that some users can overwhelm a community with a not shared topic or even by re-tweeting in an abusive way. To enhance the topics discussed by many users of the community, as opposed to a topic discussed many times by a small number of user within the community and to avoid this problem, we propose the $TF\text{-}IDF_{adaptative\_normalize}$ defined as follows:

$$TF - IDF - NT_{i,j}^{C_l} = TF_{i,j} \times IDF_i^{C_l} \times NT_i^{C_l} \quad (2)$$

With:

$$NT_i^{C_l} = \frac{|\{u_j : t_i \in u_j^{C_l}\}|}{|U^{C_l}|} \quad (3)$$

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

In the rest of the document we adopt the following notations:

- $TF\text{-}IDF$: the traditional measure of the $TF\text{-}IDF$ which is independent of the communities;
- $TF\text{-}IDF_a^{C_i}$: the $TF\text{-}IDF_{adaptative}$ measure for a specific community C_i;
- $TF\text{-}IDF_{a-n}^{C_i}$: the $TF\text{-}IDF_{adaptative\_normalize}$ measure for a specific community C_i.

4. EXPERIMENTS

In this section, we report experiments conducted on Tweets from the French Presidential and legislative elections. Our corpus has been constructed by using the Twitter API (Stream Api[13]). It has been built by following 200 French political leaders from different parties cited in www.elus20.fr. The limit of 200 is due to the limitation of the API.

The streaming API retrieves in real time all tweets from the 200 accounts and all tweets which cite, retweet, mention or answer these leaders. In addition to the text message, a number of meta-informations are recovered by the streaming API (time and date, user's informations, location's informations, number of answers, number of retweet, ...).

The following preprocessing is performed on the tweets. Firstly, we keep the text, the user name (normally first name and last name), the user account name, and the date of sending. The different tags (i.e. RT, #, @) and links are then extracted from the message text and the language is determined by using TextCat[14]. On French tweet, we annotate the text with part-of-speech and lemma information via TreeTagger[15]. With the lemmatized terms we also kept proper nouns (e.g. Hollande, Sarkozy[16]), acronyms (e.g. PS, UMP) and hashtag (e.g. #FH2012 for F. Holllande, #NS2012 for N. Sarkozy).

From the 12th December 2011 to the 19th June 2012, we thus obtained 2,122,012 tweets from 213,005 users. For 130,618 tweets, 232 users can unambiguously be assigned to a political party (i.e. user is a politician or an official political community account). Even if lots of parties were involved in the election, we mainly focused on the most important ones and then kept only tweets for six parties: PS, UMP, FG, FN, Modem, EELV. This pruned the data to 118,572 tweets and 185 users. Note that all these tweets and users have been obtained after the processing step. Note that in the following all the French terms have been translated in English.

For experiments, we have considered different time granularities: days, weeks and months. Actually the distribution of tweets over time can differ significantly, i.e. for some events lots of tweets can be sent in one day while sometimes the number of exchanged tweets is not significant for a long time. In our dataset, conducted experiments showed that evaluating results with the day granularity is useless since very often the number of tweets is not sufficient to expose a trend. Interestingly we noticed that there is no significant difference between weeks and months. So in the following reported results will focus on weeks. Obviously displaying all values for each week has a poor interest so results will be presented as an average of values for each week.

In the next sections we evaluate our measures both for intra and inter-communities. Our goal is to evaluate if our measures are well adapted to really extract significant terms.

4.1 Intra-community analysis

In this section, we aim at evaluating the most significant measure when addressing only users in a community. More precisely, here we do not compare terms among several communities. First of all, for each community and for every week, we extract the TOP-K terms varying K from 1, 10, 50 and 100 by using different measures. We then compute for each week percentage of overlap between pairs of the top-K measures and report the average overlap.

First of all we describe the results when comparing the two new measures with $TF\text{-}IDF$. Table 1 presents the comparison with $TF\text{-}IDF_a^{C_i}$ while Table 2 reports results with $TF\text{-}IDF_{a-n}^{C_i}$. For example, in Table 1, the first value *78%* means that in the whole period there is on average of 78% of common terms, but that does not necessary imply that the overlapping is the same for every week. Interestingly we can note that even with a high value of K (i.e. 50 or 100), the proposed measures extract quite different terms. For instance the overlapping for the PS is 45%. That result is interesting since it shows that even if the two new approaches are based on $TF\text{-}IDF$ they identify new terms. Table 3 reports the comparison between $TF\text{-}IDF_{a-n}^{C_i}$ and

[13]https://dev.twitter.com/docs/streaming-apis

[14]http://www.let.rug.nl/~vannoord/TextCat/

[15]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[16]Current and past Presidents

Table 1: Percent of common words in TOP-K according to $TF\text{-}IDF$ and $TF\text{-}IDF_a^{C_i}$

Community	TOP-1	TOP-10	TOP-50	TOP-100
PS	78%	34%	37%	45%
UMP	30%	35%	42%	47%
EELV	65%	47%	51%	48%
MoDem	65%	53%	49%	54%
FG	91%	69%	57%	58%
FN	35%	55%	92%	97%

Table 2: Percent of common words in TOP-K according to $TF\text{-}IDF$ and $TF\text{-}IDF_{a-n}^{C_i}$

Community	TOP-1	TOP-10	TOP-50	TOP-100
PS	83%	53%	61%	66%
UMP	39%	54%	62%	68%
EELV	65%	57%	60%	64%
MoDem	52%	42%	59%	64%
FG	70%	66%	61%	65%
FN	35%	55%	92%	97%

$TF\text{-}IDF_a^{C_i}$. Here also we can notice that the overlapping is not so high for $K = 100$. The high percent of community FN found for $K = 50$ or and $K = 100$ is not significative in our experiments. This is explained by the fact that a very small number of terms were exchanged for this community every week. Actually in our dataset, we obtained only 474 tweets for this party over the period, an average of 18 tweets per week, representing no more than 4987 terms. Presented results also show that the normalization proposed by the $TF\text{-}IDF_{a-n}^{C_i}$ is useful since it emphasizes terms used by a small number of people, as opposed to to those used by the entire community.

Here we give an example of the ten most relevant terms for week 11 (March 12th to March 18th) for community UMP (Table 4).

We can notice some terms present in our analysis:

- nouns and verbs: french, capital, wish
- Proper nouns: city (Lyon, Recy, Meaux), people (Hollande, Sarkozy)
- Hashtag: campaign slogan (#Lafranceforte), french media (#Lgj, #Pdc, #Dpda), official campaign hashtag (#Ns2012)

In this section we have shown that the three $TF\text{-}IDF$ give different information, in the next section we investigate

Table 3: Percent of common words in TOP-K according to $TF\text{-}IDF_a^{C_i}$ and $TF\text{-}IDF_{a-n}^{C_i}$

Community	TOP-1	TOP-10	TOP-50	TOP-100
PS	74%	50%	56%	62%
UMP	48%	52%	60%	65%
EELV	91%	70%	70%	67%
MoDem	65%	73%	69%	76%
FG	83%	80%	80%	82%
FN	100%	100%	100%	100%

Table 4: Example of the TOP-10 most relevant terms according to the different measures for the community UMP for the week 11

$TF\text{-}IDF$	$TF\text{-}IDF_a^{C_i}$	$TF\text{-}IDF_{a-n}^{C_i}$
#Ns2012	#Ns2012	#Ns2012
#Pdc	#Lafranceforte	#Lafranceforte
France	Lyon	#Franceforte
#Dpda	#Franceforte	Lyon
Sarkozy	#Pdc	#Pdc
Hollande	#Dpda	#Dpda
capital	#Lgj	France
Lyon	Meaux	Sarkozy
french	capital	Hollande
wish	France	Recy

Table 5: Percent of common words in TOP-k between two communities according to $TF\text{-}IDF$

Communities	TOP-1	TOP-10	TOP-50	TOP-100
PS-UMP	0%	30%	43%	45%
PS-EELV	0%	3%	23%	27%
PS-MoDem	0%	10%	24%	26%
PS-FG	0%	2%	15%	21%
PS-FN	0%	1%	7%	9%
UMP-EELV	0%	4%	19%	24%
UMP-MoDem	0%	10%	24%	24%
UMP-FG	0%	2%	14%	18%
UMP-FN	0%	3%	8%	8%
EELV-MoDem	0%	2%	14%	16%
EELV-FG	0%	1%	10%	15%
EELV-FN	0%	3%	5%	6%
MoDem-FG	0%	1%	10%	13%
MoDem-FN	0%	1%	5%	6%
FG-FN	0%	5%	8%	12%

if there is the best $TF\text{-}IDF$ to determine relevant terms of communities.

4.2 Inter-community analysis

In this section, we compare the proximity between communities according to different $TF\text{-}IDF$ measures. We want to know which $TF\text{-}IDF$ is the most discriminating of a community. A discriminating $TF\text{-}IDF$ is a $TF\text{-}IDF$ which returns for a given community some relevant terms not present among terms relevant for other communities.
For each week we compare, for the different TOP-K, the percent of common words between communities according to $TF\text{-}IDF_c$ (Table 5), then the percent of common words according to $TF\text{-}IDF_a^{C_i}$ (Table 7) and finally the percent of common words according to $TF\text{-}IDF_{a-n}^{C_i}$ (Table 8).
We give an example of the most ten relevant terms according to $TF\text{-}IDF_c$ for the week 11 (Mars 12th to Mars 18th) for the six communities studied (Table 6).
According to $TF\text{-}IDF_c$ (Table 5), we see that communities PS and UMP share almost half of there TOP-50 (and TOP-100) most relevant terms. PS (and UMP) share almost a quarter of their terms with communities EELV and Modem. There are no proximity between other communities.
We give an example of TOP-10 most relevant terms accord-

Table 6: Example of TOP-10 most relevant terms according to $TF\text{-}IDF$ for the week 11

PS	UMP	EELV
#Fh2012	#Ns2012	Joly
#Hollande2012	#Pdc	#Eelv
#Dpda	France	project
Hollande	#Dpda	Eva
Sarkozy	Sarkozy	nuclear
Marseille	Hollande	Secteeelv
Europa	capital	Strasbourg
France	Lyon	record
owe	French	thanks
can	wish	european
MoDem	FG	FN
Bayrou	Bastille	#Mlp
#Lafrancesolidaire	#Placeaupeuple	election
association	insurrection	presidential
François	Clermont	official
Puteaux	MéLenchon	candidature
#Lgj	#Rfi	prevent
France	Jean-Luc	battle
have to	#Fdg	declaration
Hirsch	Pierrelaurent	Marine
can	left	tuesday

ing to $TF\text{-}IDF$ for the week 11.

According to $TF\text{-}IDF_a$ (Table 7), we see that communities PS and UMP share the most of their TOP-50 (and TOP-100) relevant terms but $TF\text{-}IDF_a$ is more discriminating than $TF\text{-}IDF$ because the percent of sharing is only 15%-18% instead of 43%-45% and there are no proximity with other communities.
Finally we focus on $TF\text{-}IDF^{Ci}_{a-n}$ (Table 8), we see again a proximity between communities PS and UMP. The results are worse than those obtain with $TF\text{-}IDF^{Ci}_{a}$.

Based on these findings, we believe that $TF\text{-}IDF^{Ci}_{a}$ seems to be the best candidate for create $\text{TOP}_{C_i}\text{-}K$. In the next section we interest on an other dataset used for calculate $\text{TOP}_{C_i}\text{-}K$.

4.3 Datasets analysis

In previous experiments we considered that terms are: nouns, adjectives and verbs in their lemmatized forms as well as proper nouns (e.g. Hollande, Sarkozy), acronyms (e.g. PS, UMP) and hashtag (e.g. #FH2012, #NS2012). In the following we will call this dataset as "Dataset 1". Now we would like to evaluate if hashtags, proper nouns and acronyms are really useful. We then build a new dataset, called "Dataset 2", by keeping only terms (nouns, verbs and adjectives) in their lemmatized forms. Table 9 summarizes the difference and similarity of the two datasets.

We show in Table 10 the most TOP-10 relevant terms according to $TF\text{-}IDF^{Ci}_{a-n}$ for community PS for week 12 (March 19th to March 25th). These dates have been chosen because at the same moment a dramatic event happened in

Table 7: Percent of common words in TOP-K between two communities according to $TF\text{-}IDF^{Ci}_{a}$

Communities	TOP-1	TOP-10	TOP-50	TOP-100
PS-UMP	4%	8%	15%	18%
PS-EELV	0%	1%	3%	4%
PS-MoDem	2%	0%	4%	6%
PS-FG	0%	0%	1%	3%
PS-FN	0%	0%	0%	1%
UMP-EELV	0%	0%	2%	3%
UMP-MoDem	0%	1%	4%	6%
UMP-FG	0%	0%	2%	3%
UMP-FN	0%	0%	0%	1%
EELV-MoDem	0%	1%	1%	1%
EELV-FG	0%	0%	10%	1%
EELV-FN	0%	0%	0%	0%
MoDem-FG	0%	1%	1%	1%
MoDem-FN	0%	0%	0%	0%
FG-FN	0%	1%	2%	2%

Table 8: Percent of common words in TOP-K between two communities according to $TF\text{-}IDF^{Ci}_{a-n}$

Communities	TOP-1	TOP-10	TOP-50	TOP-100
PS-UMP	4%	27%	37%	38%
PS-EELV	0%	4%	13%	14%
PS-MoDem	0%	6%	12%	16%
PS-FG	0%	1%	5%	8%
PS-FN	0%	0%	1%	2%
UMP-EELV	0%	3%	10%	14%
UMP-MoDem	0%	5%	11%	13%
UMP-FG	0%	1%	5%	8%
UMP-FN	0%	0%	1%	1%
EELV-MoDem	0%	2%	4%	7%
EELV-FG	0%	0%	3%	6%
EELV-FN	0%	1%	0%	1%
MoDem-FG	0%	1%	2%	4%
MoDem-FN	0%	0%	0%	1%
FG-FN	0%	1%	2%	3%

Table 9: Summary of Dataset 1 and Dataset 2

	Dataset 1	Dataset 2
Common nouns	X	X
Verbs	X	X
Adjectives	X	X
Proper nouns	X	
Hashtag	X	
Acronyms	X	

Table 10: TOP-10 most relevant terms according to $TF\text{-}IDF_a^{Ci}$ for community PS for the week 12

PS	
Dataset 1	Dataset 2
#Fh2012	victim
Aurillac	republic
Toulouse	evening
Hollande	inauguration
victim	moment
republic	family
#Hollande2012	racism
Lyon	commend
Florencecassez	change
France	respect

Toulouse on of March 19th children were killed at school[17].

Despite the fact that TOP-K issue from Dataset 1 and Dataset 2 are very different, we do not observe significant difference in intra-community analysis and inter-community analysis when comparing results from both dataset.

5. HOW TO ASSIGN A NEW USER TO A COMMUNITY?

In our experiments we address several issues. For example, we show that different $TF\text{-}IDF$-based measures give different results (Section 4.1) so there can be useful depending on the context. In our context of political community detection, we ask if there is a better measure to extract relevant terms of a community (Section 4.2). We conclude that an adaptive approach is very useful to extract relevant and specific terms for a community.

Contrary to our intuitions, we show in Section 4.3 that hashtag, proper nouns and acronyms does not help us to better determine community. They can naturally determine a community but not more nor less than common nouns, verbs and adjectives. There are few proper nouns, hashtag and acronyms which are specific to a single community.

With these findings we show how *learning from tweet data* we can attempt to determine a political community for users who are not necessary politicians, as follows:

1. For each week and each community we calculate TOP-K relevant terms according to $TF\text{-}IDF_a^{Ci}$.
2. We determine which is the most frequent community for this user for a given week by comparing the number of terms shared between user terms and different TOP-K.
3. Then we compute all the candidates for each week of this user and we assign the community that occurs most frequently.

We check 1,052 politically marked twitter accounts from the six main political communities. This set will be used as *test data*. Selected users is considering as unambiguous. They all meet four requirements:

- Users are not used in the construction of TOP-K most relevant terms.
- They stated in their description a political preference (e. g. *"support Nicolas Sarkozy"*, *"activist PS"*, *"Regional Councillor EELV"*)
- The description contains no semantic inconsistencies (e. g. *"I support PS but I vote for Sarkozy"*, *"Fan of NS2012 and FH2012"*)
- They sent more than 5 tweets

To avoid introducing bias, the user's description is used for manual classification where content of tweets is used for automatic classification.

First results seem satisfactory.

- If we focus on TOP-10, we classify 98.8% of users for a total of 93.7% assigned correctly.
- If we focus on TOP-50, we classify 99.9% of users for a total of 93.9% assigned correctly.
- If we focus on TOP-100, we classify 99.8% of users for a total of 94.3% assigned correctly.

We compute recall, precision and F-Measure (Table 11 and Table 12).

Table 11: Macro-average: Recall, Precision and F-Measure

TOP-K	Recall	Precision	F-Measure
TOP-10	0.85	0.95	0.90
TOP-50	0.87	0.89	0.88
TOP-100	0.86	0.91	0.88

Table 12: Micro-average: Recall, Precision and F-Measure

TOP-K	Recall	Precision	F-Measure
TOP-10	0.94	0.95	0.94
TOP-50	0.94	0.94	0.94
TOP-100	0.94	0.94	0.94

It is interesting to note that results are improved if we select only users with more than 20 tweets (there are 678 users in our datasets).

- If we focus on TOP-10, we classify 99.9% of users for a total of 96.6% assigned correctly.
- If we focus on TOP-50, we classify 100% of users for a total of 96.5% assigned correctly.
- If we focus on TOP-100, we classify 100% of users for a total of 97.1% assigned correctly.

We compute recall, precision and F-Measure (Table 13 and Table 14).

[17]http://www.bbc.co.uk/news/world-us-canada-17426313

Table 13: Macro-average: Recall, Precision and F-Measure

TOP-K	Recall	Precision	F-Measure
TOP-10	0.88	0.97	0.92
TOP-50	0.89	0.98	0.93
TOP-100	0.88	0.98	0.93

Table 14: Micro-average: Recall, Precision and F-Measure

TOP-K	Recall	Precision	F-Measure
TOP-10	0.97	0.97	0.97
TOP-50	0.96	0.96	0.96
TOP-100	0.97	0.97	0.97

6. CONCLUSIONS

People participating in on-line forums, microblogging or discussing on social networks leave behind them digital traces of their opinion on a variety of topics. If we knew how to aggregate and cumulatively interpret this data, we could se them to determine communities of users. 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.

This paper focused on the study of tweets in the context of the Presidential and legislative French elections and proposed several measures that have been proven as efficient for extracting discriminant terms for communities. We also illustrated that extracted terms are very useful for automatically assign a user to a specific community.

The study discussed in this paper focuses on the TF-IDF weight. In future work, we can integrate this type of weight to enhance the matrix representation of tweet data. After applying this kind of weight for tweet features, a process based on Latent Semantic Analysis (LSA) can be performed [6, 13]. The result will be a compressed version of the original matrix of textual corpus.
Eventually, the weights proposed in this paper can be combined with topic-modeling approaches [11] in order to predict the community of a tweet.

7. REFERENCES

[1] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of COLING*, pages 36–44, 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] A. Joshi, A. Balamurali, P. Bhattacharyya, and R. K. Mohanty. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of ACL (System Demonstrations)*, pages 127–132, 2011.

[5] C. Kobus, F. Yvon, and G. Damnati. Normalizing sms: are two metaphors better than one ? In *COLING*, pages 441–448, 2008.

[6] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[7] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of 2010 International Conference on Management of Data (SIGMOD 2010), Demonstration*, pages 1155–1158, 2010.

[8] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of KDD*, pages 430–438, 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)*, pages 851–860, 2010.

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

[11] J. Tang, R. Jin, and J. Zhang. A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 1055–1060, 2008.

[12] B. Tsolmon, A. Kwon, and K.-S. Lee. Extracting social events based on timeline and sentiment analysis in twitter corpus. In *Proceedings of NLDB, Springer Verlag, LNCS*, 2012.

[13] P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *12th European Conference on Machine Learning*, pages 491–502, 2001.