Debate on Political Reforms in Twitter: A Hashtag-driven Analysis of Political Polarization

Mirko Lai, Cristina Bosco and Viviana Patti
Dipartimento di Informatica
University of Turin, Italy
Email: {lai,bosco,patti}@di.unito.it

Daniela Virone
Scuola di Dottorato in Studi Umanistici
University of Turin, Italy
Email: daniela.virone@unito.it

***Abstract*—Political debates about a reform may sparkle national controversies, by leading members of the community to polarize their opinions and sentiment about the topic addressed. With the rise of social media like Twitter users are encouraged to voice and share their strong and polarized views and in general people are exposed to broader viewpoints than they were before. The large amount of user-generated social data available is a great opportunity to investigate the communicative behaviors emerging in the context of such political debates and to shed some light on the way communities of users with different roles in the society and different political sentiment interact. In this paper we focussed on communications in Twitter around the reform of marriage in France in 2012 and 2013 – "Le Mariage Pour Tous" – which had been the subject of debate and controversy. We collected a corpus of tweets tagged by the hashtag #mariagepourtous, created to mark the messages about the reform. We applied different kinds of analysis on our dataset based on linguistic and non linguistic features of the observed data in order to investigate the communicative behavior in using subjective and evaluative language on a political topic. The analysis leaded also to reflect on the impact of different typologies of users involved in the virtual debate which included both political messages created by media organizations and by other individual users, from ordinary citizens to politicians or celebrities.**

I. Introduction

The analysis of user generated texts from social media can involve, in particular, the analysis of subjective opinions and sentiments [1], [2]. Several recent works deal with this task mainly following approaches based on lexical resources developed according to cognitive models, or on machine learning algorithms applied to annotated corpora of data from social media. But both the approaches are prone to various kinds of problems. On the one hand, the limited availability of adequate lexical databases, in particular for less resourced languages, causes hardships in the application of lexicon-based approaches. On the other hand, the a priori definition of the words describing private affective states according to known cognitive models, as usually done in affective lexica, can produce affective notions which can be useful in general but not adequate for some specific kind of context or discourse domain [3].
For what concerns instead corpus-based approaches, they are featured by the prevalence of statistics on linguistics. Moreover data sparseness make them reliable only when based on very large annotated corpora, whose development is a very time-consuming task. Finally, several different kinds of annotation have been proposed, but often they are broad grained and only include a very basic set of tags [4].
This motivates the recent trends towards hybrid approaches or to computational semantics oriented frameworks where a global notion of communication is involved, which includes e.g. context, themes, dialogical dynamics in order to detect the affective content even if it is not directly expressed by words, like for instance when the user exploits figurative language (irony or metaphors) or in general when the communicated content does not correspond to words meaning but depends on other communicative behavior.

On this perspective, a particular interesting domain is related to the political debates. In the last years social media, and in particular Twitter, have been used in electoral campaigns by different actors involved in the process: by campaign staff in order to disseminate information, organize events; by the news media in order to inform and promote news content; and by voters to express and share political opinions. Therefore recently many studies focussed on understanding the phenomenon, by studying the effect of this technology on the election outcomes [5], its possible use to gauge the political sentiment [6], or by studying the networks of communication in order to investigate the political polarization issue [7].

This study contributes to this area, by examining a different kind of political debate on Twitter: a debate around a reform. Communications dynamics and phenomena related to political sentiment polarization are present but can be interestingly different here. In absence of a candidate, the arguments in favor or against the reform being discussed move to the foreground, and users are more prone to communicate not only a general appreciation or criticism towards a proposed reform, but also a vision about the object of the reform, which is expressed in messages and it is an interesting aspect to analyze. In this paper we present a detailed analysis of a dataset from Twitter driven by a hashtag, #mariagepourtous, created to mark the messages about the reform "Le Mariage Pour Tous" (Marriage for all), discussed in France in 2012 and 2013. This analysis is mainly motivated by the need to set a framework for discovering and investigating various aspects of communication in social media that can be formalized in the annotation of corpora for sentiment analysis, and then exploited for improving the detection of sentiment by systems. We applied different kinds of lexical and pragmatic analysis on our dataset, based on linguistic and non linguistic features

of the observed data, in order to shed some light also on the impact on the virtual debate of different typologies of users involved: media organizations, but also individual users, from ordinary citizens to politicians or celebrities. In particular by applying a diachronic analysis we attempt to answer the question: "How the hashtag #mariagepourtous spread, how and where is born, who are the major disseminators?" in order to get some first insight about how the debate around "Marriage pour tous" spread in Twitter. Moreover, by applying a synchronic analysis on the dataset we try to give a first answer to the question: "What are the main semantic areas of the vision accompanying the opinions for and against the reform?".
This work is collocated in the wider context of an ongoing project about the study of communication in different media. It represents therefore the first step of a study which involves the application of similar analysis to datasets collected from journals and parliamentary debates, and finally the comparison of the results achieved in social media, journals and debates. This also motivated the choice of the selected topic –the debate about homosexual couple wedding in France– which allows for the collection of large datasets from all the media involved in the research where the topic has been extensively discussed during a wide time slot. The hypothesis we are going to investigate during the project as future work is if there are similarities of differences between social media and other media that can be automatically detected.

The paper is organized as follows. The next two sections respectively describe related works and the data set, showing the criteria and methodologies applied for the selection of data. Fourth section is devoted to the analysis of collected data, while the last one describes future development of this research.

II. Related work

Several works rely on sentiment analysis techniques [1] to analyze politics [6], [8], [9], a domain where the problems related to the exploitation of figurative language devices described in [10], [4], [11], [12], [13], [14] and in the Semeval15-11 shared task [15] have been detected as frequent. Moreover, some research focussed on aspects concerning the political polarization in Twitter [7], [5] which are very interesting also in the dataset we are analyzing in the current work. Other works instead addressed the issues related to the arguments accompanying the political messages, like [16] where an analysis devoted to discover in the tweets the argumentation related to evaluative discourse is presented and applied to the case of the racism anti-Rom in the Web; it is shown that a discourse where a form of evaluation is expressed does not necessarily exploits semantic and linguistic markers traditionally linked to the evaluation, but it can be also based on dialogical and dialectical components. While the analysis of Twitter political debates related to election campaigns became quite popular in the last years, since there is a lot of interest in developing tools to automatically gauge the political sentiment in order to predict the election outcome, the idea to focus the analysis on the debate around a reform can lead to get some new insights on the communicative behavior in using subjective and evaluative language in politics.

Moreover, most of the works carried on so far in this area focus their analysis on English datasets and rely on the use of lexical and affective resources which are available only for English, e.g. [17], [18]. With respect to the availability of affective lexica and resources for sentiment analysis, French can be currently considered among the under-resourced languages. Nevertheless, in the last few years some effort has been devoted to the development of new annotated data to be exploited in this area, see e.g. [19], [20], [21].

III. Collection and composition of the data set

As introduced above, this work is collocated in the context of an ongoing project about communication in different media and is focussed on the debate about homosexual couple wedding in France. The project includes the collection of the following data sets from different media and sources:

- TW-MariagePourTous corpus: texts collected from Twitter by filtering the tweets posted in the time-lapse 16th December 2010 - 20th July 2013 for French language and for the presence of the hashtag *#mariagepourtous* and without retweets
- NEWS-MariagePourTous corpus: texts collected from French newspapers, in particular from LeMonde online and from the sources made available by the Factiva search engine[1], published in the time-lapse 7th June 2011 - 4th February 2013 and filtered by the keyword *#mariagepourtous*
- NEWSTITLE-MariagePourTous corpus: texts collected as the NEWS-MariagePourTous corpus and covering the same period, but including only the titles of the articles
- DEBAT-MariagePourTous corpus: texts collected from parliamentary debates about the first discussion of the bill on homosexual wedding (meetings of the National Assembly and Senate of the French Parliament from 27th January 2013 to 12th February 2013) and the following meetings (from 4th to 12 April 2013 and from 15th to 23th April 2013) where the bill has been approved[2].

The larger corpus is that from newspapers, i.e. NEWS-MariagePourTous, which includes around 24,000 articles, while the smaller is NEWSTITLE-MariagePourTous.
For what concerns the TW-MariagePourTous corpus (henceforth TW-MPT), the corpus on which is based the current analysis, it includes 254,366 original messages, 88,157 of which have been re-tweeted by one or more user during the time of the corpus collection. Each tweet is associated with the metadata related to the posting time and the user that posted it, information that we exploited in the analysis presented in section IV-A and IV-B.

Hashtags are single words or expressions (with words not separated by spaces) preceded by the symbol '#' well known in Twitter. Hashtags allows users to create communities of people,

[1] See http://new.dowjones.com/products/factiva/.

[2] See http://www.assemblee-nationale.fr/14/debats/ for the transcription of debates of the National Assembly, and http://www.senat.fr/seances/comptes-rendus.html for the debates in Senate made available by the French Government.

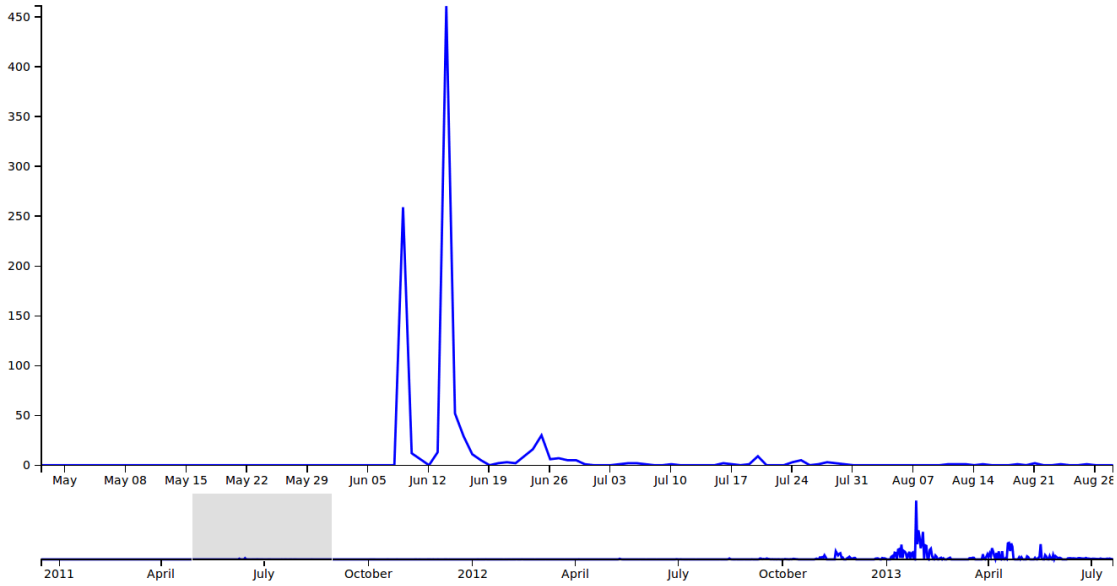

Fig. 1. The first peak of *#mariagepourtous* in June 2011.

interested in the same topic by making it easier for them to find and share information related to it [22]. The creation of a new hashtag, as a new linguistic device, can be motivated by the need of making widely known information that can vary from social to politics, or drawing attention to a commercial product. We can see for instance the hashtags/slogans created during election campaigns. Even if not all the hashtags generated are destined to be exploited and shared by large people groups, in general, when a user exploits an existing hashtag, he/she wants to be recognized as belonging to the group using it, to be accepted within the dialogical and social context growing around the topic [23], but not necessarily in order to assume the same opinion about the content of the hashtag. For instance, *#mariagepourtous* has been used by people expressing both positive and negative opinions about homosexual wedding in France.

By selecting a hashtag as our main filtering criterion, we easily collected several arguments and different opinions expressed by the persons interested in the web debate about the topic, circumscribing the collection to posts which are really related to the debate and therefore of interest for our research. But we can also observe the "life" of the hashtag: how it propagates among Twitter users, the time when it is newly proposed, then when it is accepted by the community and finally not more used (see IV). For this purpose we analyzed *#mariagepourtous* starting from the first tweet where it occurred and observing it during the following three years. The specific socio-politic topic linked to the hashtag, together with its large diffusion, makes it especially adequate for the study of the dynamics of communication in social media and for the comparison with what happens in other media about the topic represented in Twitter by using *#mariagepourtous*.

It should be furthermore observed that the selection of *#mariagepourtous*, among the other hashtags featuring the discussion about the marriage in France, is motivated by its exploitation as a *linguistic formula* which can be analyzed as a sort of locution providing a focus for the debate and richly attested in all the four collected corpora, and not only in TW-MPT. The analysis applied in this study on TW-MPT for detecting co-occurrences of the hashtag with a selection of themes (see section IV) can be, therefore, applied also to other corpora e.g. for detecting the relationships between *#mariagepourtous* and particular morpho-syntactic structures in NEWS-MariagePourTous corpus, as it has been done in [24].

IV. Data analysis

We applied to our dataset different kinds of analysis based on linguistic and non linguistic features of the observed data, in order to detect the pragmatic nature of communicative behavior of users in exploiting subjective and evaluative language. We followed a unsupervised approach in all the experiments we performed. As previously reported, French is in fact currently featured by a very limited amount of annotated data and lexical resources which can be used for sentiment analysis and similar tasks. This limited our application of sentiment analysis techniques only to preliminary tests and motivated our approach and plans for future development, which include the manual annotation of a portion of the TW-MPT corpus.

The analysis we applied to the data set are organized according to two main directions:

- the *diachronic analysis* concerns data features that are strongly related to the passage of time, such as the frequency of the hashtag in defined time slots, and the relationships between this frequency and the events happened during the same time slots and spread by other media, like tv, newspapers and parliamentary debates; it is allowed by the association to the tweets of metadata showing time and user

- the *synchronic analysis* concerns instead the features of the discourse centered on *#mariagepourtous* which can be studied without the involvement of considerations about the passage of time.

Putting together these directions we can obtain hints about the communicative dynamics acting around the hashtag and in general in Twitter on the topic we are investigating.

A. Diachronic analysis

We performed according to the diachronic perspective mainly analyses devoted, on the one hand, to the observation of the birth and life of the hashtag, and, on the other hand, of the users' behavior. The first kind of analysis consists in the detection of the variation of the frequency of *#mariagepourtous* in Twitter during time and in the observation of possible relationships between the variation of the hashtag frequency and the socio-political events.
Starting from the first exploitation of the hashtag, in 16th December 2010, *#mariagepourtous* has never been used until June 2011. (see the first six tweets including the hashtag in Fig. 3 [3]). From June 2011 the frequency of the hashtag slowly and progressively augmented until September 2012 and several events influenced its frequency causing various peaks of usage. They are mainly related to politics, e.g. parliamentary debates, the election campaign, manifestations about homosexual wedding, which can be seen in table I. The

TABLE I. THE MORE RELEVANT EVENTS IN THE FIRST TIME SLOT OF THE COLLECTION OF THE TW-MPT

time	event
June 2011	the bill on homosexual wedding is rejected by the parliament
January 2012	the parliament debate about the bill starts again
March 2012	electoral campaign and election of the president of the Republic
September 2012	interview to Christiane Taubira

image in fig. 1 shows the details of the peak related to the first of the relevant events cited in table I, i.e. the rejection of the bill on homosexual wedding by the parliament in June 2011. The following and also most strongly influencing event that we can observe is the interview to the deputy Christiane Taubira at the beginning of September 2012, whose peak is shown in fig. 2.
According to the frequency of the hashtag it can be drawn a clear distinction between the time slot preceding this event and that following it. In fact, in the first time slot (16th December 2010 - 11th September 2012) only 1,130 users did exploit the hashtag, producing 3,528 tweets, while in the second period (11th September 2012 - 20th July 2013) 50,513 users exploited *#mariagepourtous* and posted 250,838 messages. It can be interesting to notice that referring to the corpus NEWS-MariagePourTous, we can see that the linguistic expression

[3]English translation of the posts follows: Stéphane Pillet @ alexisgirszonas: "If #marriagegay is more sensational and so it makes more sense for medias, #mariagepourtous is more comprehensible and right"; Alexis Girszonas @stephanepillet: "#marriageforall the couples of human being; this poor Mme Barèges likes it +1 for you ;) "; Stéphane Pillet @GillesBonMaury: "On #marriage it's all talk and no trousers, it do not serves people going straits to the point. ;) #mariagepourtous"; Stéphane Pillet: "#marriageforall the couples of human being, this poor Mme Barèges likes it"; Stéphane Pillet: " #mariagegay What's this? Marriage is not hetero or homo (not at all gay). The PS (socialist party) defends #marriageforall the couples"; et-aloes dot net: "[SOS HOMOPHOBIE news] #QPC Open #marriage is fight against #homophobia URL #mariagepourtous".

Fig. 3. The first six tweets including the hashtag *#mariagepourtous*

"mariage pour tous" has been used in newspapers from June 2011, but it is only from March 2012 that the hashtag can be found in media other than Twitter. This shows that from the end of the first time slot, *#mariagepourtous* became a linguistic device used by a community that spans beyond Twitter.
This clear cut among two different time slots allowed by a relevant event, i.e. the interview to Taubira, motivated the data subdivision we applied in the rest of the analysis: the Twitter-Pre-Taubira part of the corpus (henceforth TWPre-corpus) and the Twitter-Post-Taubira (henceforth TWPost-corpus). We separately interrogated the data of the two sub–corpora in order to find information about the negotiation of the hashtag and the progressive increasing exploitation of it, until it achieved the status of expression shared by a community. Moreover, this distinction has been useful to describe the behavior of users according to their contribution to the debate around *#mariagepourtous* in the different time slots. The users' behavior is in fact among the aspects we investigated in our dataset, both in the perspective of the single user and in that of the community, by observing how the opinion of them can be relevant within the debate and, therefore, identifying among them the most influential.
Focusing first on the TWPre-corpus and then on the TWPost-corpus, we detected the presence of users that can be considered as more influential for the debate under various respects. For what concerns the TWPre-corpus, we found that among the 1,130 active users producing original posts (i.e. excluding the re-tweeted messages) with the hashtag *#mariagepourtous* only 84 generated the larger amount of posts (more than 5 each), while the others 1,046 posted 1 (750 users) or between 2 and 5 tweets (296 users) each. We call those 84 users hashtag pioneers. We can classify the most active pioneers, i.e. those posting larger amount of messages and listed in table II, in different groups according to their social role: parliamentary

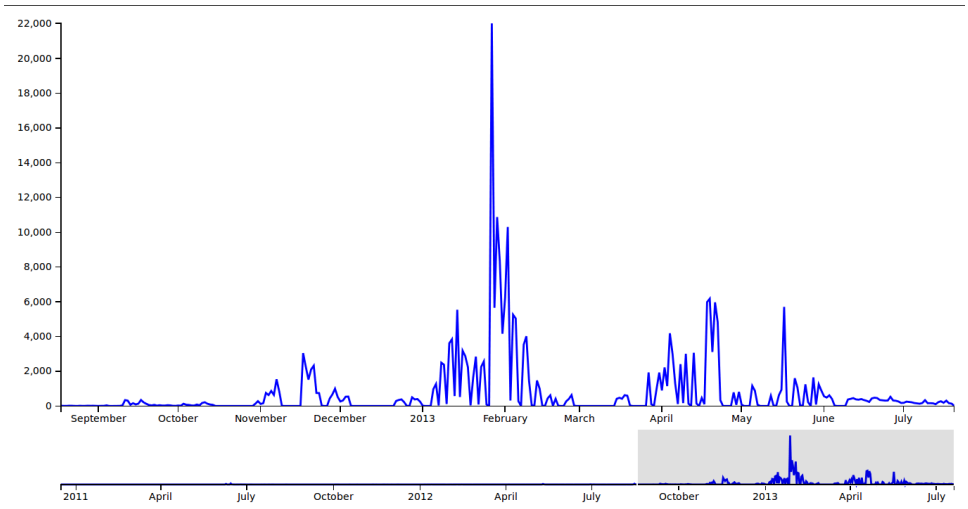

Fig. 2. The peak of *#mariagepourtous* in February 2013, in the time slot following the interview of Taubira.

TABLE II. THE USERS THAT USED MORE FREQUENTLY THE HASHTAG *#mariagepourtous* IN TW-PRE-CORPUS, I.E. THE PIONEERS OF THE HASHTAG.

user	tweets
JeanLucRomero	135
GekkoHopman	112
ProjetEntourage	101
Engagement31	96
JeromePasanau	88
Yagg	79
Funny_Fog	76
Pascal_Lelievre	75
unevisionautre	73
GillesBonMaury	72

deputies, i.e. Gilles Bon Maury and Jean Luc Romero, affiliate with the party that in the following months presented the bill; an online journal (Yagg) and an association (Projectentourage) and several single users.

For what concerns the TWPost-corpus, a similar trend in the distribution of tweets and users than in TWPre-corpus, can be observed, see the representation of the distribution in fig. 4. In this corpus we observed not only the most active users, i.e. those posting the most of tweets, but also how much each tweet influenced the debate, i.e. how many times it has been retweeted by other users. Table III shows the list of the more active users, while table IV shows the authors whose posts have been more often retweeted together with the average of how many times these messages have been retweeted. A sample of the most retweeted messages can be seen in Figure 5[4]. We can observe that no one of the users detected as most active in the TWPre-corpus, i.e. the pioneers, maintained a relevant role in the TWPost-corpus and some of the most active users of the TWPost-corpus activated his/her account after the

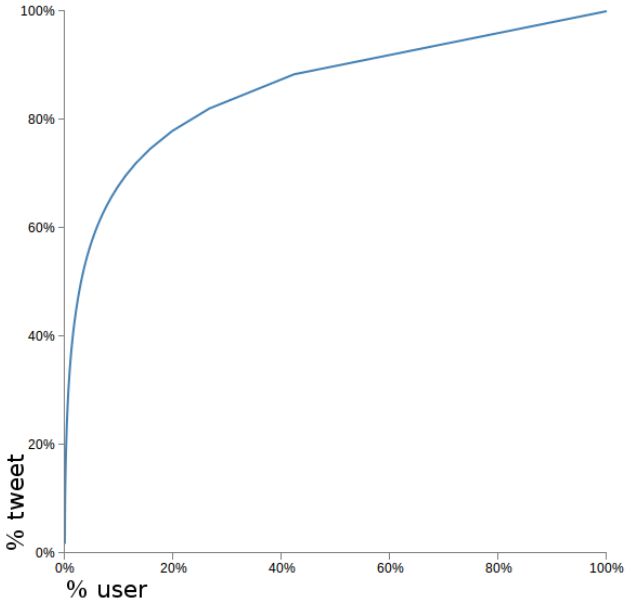

Fig. 4. The cumulative frequency distribution of tweets and users in TW-Post-corpus.

Taubira interview. Nevertheless, the more influential users, i.e. the opinion leaders which posted the most influential tweets, do not coincide with the most active users since they posted a very limited amount of messages.

Another observation about context, is that it can include also other hashtags. Table V reports the more frequent of them aggregating the cases of variations when they occur in more than one form[5] (see hashtags marked in the table with *). The hashtags used together with *#mariagepourtous* play the

[4]English translation of the posts follows: Michael Youn: "For the day of the march, a lonely slogan: A Gay marriage is better than a sad marriage #mariagepourtous"; Gunther Love: "I really want to know how many divorced people will be on the street to defend the holy values of marriage? #mariagepourtous"; Elio di Rupo: "I'm proud of our country modernity where every couples have the right of getting married. #mariagepourtous #Belgium"; Conseil constist: "Decision n 2013-669 DC, Bill opening marriage to same-sex-couples #mariagepourtous Compliant URL".

[5]We considered as variation of a hashtag the hashtag that begins with the same word, but includes a different final part.

TABLE III. THE USERS THAT USED MORE FREQUENTLY THE HASHTAG *#mariagepourtous* IN TW-POST-CORPUS, I.E. THE OPINION LEADERS.

user	tweets
fandetv	4277
cutesmilingcat	2194
jrossignol	1718
Hirschfeld_J	1710
LeMariagePrTous	1129
Yasmilady	1084
JackyMAJDA	953
Pridemap	913
jsherpin	893

TABLE IV. THE USERS WHOSE POSTS HAS BEEN MOSTLY RETWEETED.

user	tweets	retweets	retweets/tweets
MichaelYoun	2	4354	2177
eliodirupo	1	1100	1100
Conseil_constit	1	917	917
kavanaghanthony	1	612	612
ChTaubira	1	479	479
farrugiadom	1	417	417
AmandineDu38_	1	399	399
lebonlebon	1	327	327
youssouphamusik	3	976	325

Fig. 5. The most retweeted posts

role of indicators of sub-topics. See for instance *DirectAN* and *DIrectSenat* for indicating parliamentary debates related to homosexual wedding, or PMA (Procreation Medicalement Assistee (*Assisted reproductive technology*)) and GPA (Gestation Pour Autrui (*surrogate motherhood*)) which are themes discussed in the same context.

TABLE V. THE FREQUENCY OF THE HASHTAGS EXPLOITED TOGETHER WITH *#mariagepourtous*

mariagepourtous	250,235
DirectAN	25,100
manifpourtous	14,031
DirectSenat	5,055
UMP	4,831
MariageGay	4,115
PMA*	3,361
homophobie*	3,101
LGBT*	2,861
GPA	2,851
Hollande*	2,626
Taubira*	2,391
PS *	2,098

B. *Synchronic analysis*

The framework of analysis presented in this section aims at shedding some light on the main semantic areas of the vision accompanying the opinions for and against the reform. This will be a useful basis for further analysis related to detect the political sentiment in the messages, with the final aim to detect not a generic sentiment on the reform, but the polarity of the political sentiment at a finer-grain of granularity, distinguishing the different aspects discussed by the community in the public debate. All the analyses performed here are carried on according to a synchronic perspective and refer to the data of TW-Post-corpus since this period offers a larger amount of data to be analyzed and represents the time frame where *#mariagepourtous* is shared by the community involved in the debate.

We organized our analysis in two steps. The first is the detection of the main topics related to the debate and their classification according to a few of semantic areas: family, socio-political debate, legal aspects, public manifestations. The tag cloud can be seen in Figure 6. While the detection of topics is based on tag cloud-based techniques, as described below, the labeling of each area has been done by using the more frequent and representative word of the cloud area itself. In particular, in the upper left side of the cloud, that labeled as family, we can see *mariage* (marriage), *enfants* (children), *parents* (parents) and *pére* (father); *loi* (law), *égalité* (equality) and *droits* (rights) feature instead the lower left areas, which has been labeled as legal aspects. Among the more frequent words for the area labeled as public manifestation, i.e. the lower right one, there are *manif* and *manifestation* (manifestation); and finally the area labeled as socio-political debate includes words as *debat* (debate), *deputés* (deputies) and *opposants* (opponents).

In order to build the cloud we created a directed network of tokens such that an edge exists between two tokens if they co-occur in the same tweet (excluding self-loops). The width of the edges matches the number of tweets in which the two linked tokens co-occur. Then, we excluded from tokens:

- urls, hashtags, mentions and numbers;
- tokens without lemma[6];
- tokens with POS different from ['NOM', 'VER','ADJ'][7];

[6]We exploited TreeTagger as lemmatizer: http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/.

[7]We refer to http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/data/french-tagset.htmlforthemeaningofthetags.

In order to reduce the complexity of the network, we removed the edges having width < 4, by obtaining a network composed by 35,146 nodes and 34,4425 edges. Therefore, by exploiting the community detection methods in [25], we detected 15,828 communities. Among such communities only four big communities emerged (the others have less than 1% of nodes). The sub-network of the biggest four communities is composed by 18,581 nodes and 328607 edges: 52,87% and 95,41% of the total network, respectively. Finally, we plotted in the word clouds the twenty tokens with higher degree for each community; the font size is proportional to the node degree.
In order to validate this strategy, following the same methodology and using the same parameters, we are creating a tag cloud for the DEBAT-MariagePourTous corpus and a comparison of the results is matter of future work.

In a second phase we focussed on extracting some information on the *context* where the hashtag *#mariagepourtous* occurs, observing the set of words before and after the hashtag in posts. We investigated this notion in the data set by tokenizing each tweet and collecting and counting the expressions that can be found close to the hashtag *#mariagepourtous* in posts. In order to mainly focus on words carrying semantic content, we applied a morphological analysis for excluding from this count determiners, the prepositions with the exception of "pour" (*for*) and "contre" (*against*), punctuation marks, numerals, screen names, hashtags (already considered in the previous analysis) and urls. We limited our observation to the remaining four words which precede or follow the hashtag *#mariagepourtous*, thus distinguishing the left and right contexts (left-MPT and right-MPT henceforth). By restricting the range of considered words to the four words preceding or following the hashtag we have significantly increased the possibilities to capture expressions semantically relevant and linked to it, for instance, in a simple sentence, usually composed by not less than three tokens.

The most interesting observation emerging from this experiment are two: the frequent occurrence of the expressions "pour" and "contre" in the left-MPT context, and the frequent occurrence of the verb "etre" (*to be*) in the right-MPT context. Such observations suggest that users tend to express a kind of rough polarization about the reform in the left-MPT context, while expressions and arguments related to a more articulated evaluation or definition of the hashtag (and therefore of the reform at issue) seem to be concentrated in the right-MPT context. Therefore, in order to get some deeper insights on the main semantic areas of the vision accompanying the opinions for and against the reform, a strategy is to focus on the words following the copula (i.e. *#mariagepourtous* is ...), where we can find the evaluation or definition given by the user for the hashtag. A qualitative analysis of the expressions following the copula shows that most of users prefer to express their own definition or evaluation (e.g. "[...] est le fruit d'une culture libérale ; c'est à dire la liberté de choisir et d'entreprendre." ("[...] it's the result of a liberal culture; that means our freedom to choose and built."), or "[...] est un combat laïque et rèpublicain face fanatisme politique et religieux" ("[] it is the laic and republican fight against politic and religious fanatism")), while very few users shared the same definition (e.g. " est une loi pour les gays et non pour les homosexuels URL" ("it is a bill for gay and homosexuals URL")).
Therefore, in order to collect a set of tweets where users express their evaluations, we focused our analysis on posts where the hashtag goes with the copular verb. This sub–corpus of TW-MPT, including 1,322 occurrences of *c'est* (it is), 1,859 of *est* (is) and 301 of *n'est* (it isn't), is the dataset we selected for the application of a currently on going sentiment analysis oriented annotation.

The natural subsequent step of analysis would be to apply some sentiment analysis on the right-MPT context in order to detect the polarity of the opinions that contributed to the debate.
We tried, therefore, some first experiment with sentiment analysis techniques exploiting an available resource, which is currently among the only affective lexical resources for French[8]. The experiment confirms the limits known in literature about the exploitation of resources not developed for the domain, see e.g. [21]: a qualitative evaluation of the polarities detected shows that they can be considered only partially reliable and in particular when determined by the occurrence of words referring to words which assume the same polarity in almost all contexts, like e.g. "egalité" (*equality*) or "liberté" (*freedom*), but not in the opposite case, like e.g. "manifestation" (*manifestation*) that in our context is mainly considered as assuming a negative polarity. A further investigation of the sentiment analysis issue (resources and approach) is matter of future work.

V. Conclusions and future work

The paper presents a first analysis of a corpus from Twitter on the political debate around the reform "Le mariage pour tous" addressing the topic of homosexual wedding in France. The development and analysis of the Twitter dataset is collocated in a larger project devoted to the study of communication dynamics in social and other media.

The analyses performed raises several questions to be answered by future work.

The main drawback to develop a sentiment analysis system for French is the lack of linguistic resources such as sentiment lexicons and also reliable corpora manually annotated with sentiment polarity. In order to overcome this limitation, we plan to develop a preliminary sentiment analysis system by automatically translating the lexical resources available for English to French and by developing a reduced corpus of manually annotated tweets to train the system, by adapting the methodologies applied in [26] in the contest of the Sentiment POLarity Classification shared task proposed at the Evalita evaluation campaign on of natural language processing and speech tools for Italian [27], [28].
For what concerns the sentiment analysis issue, we should investigate also the use of a linguistic knowledge derived by a more structured and wider notion of context which includes syntactic chunks or spans over a large word space, or the exploitation of domain-independent expressive signals such as emoticons and emojis.

The irony issue will be another challenge to address (consider that some of the most influential users in the debate resulted to be comedians). On this perspective, a manual annotation of a portion of the TW-MPT corpus with sentiment

[8] http://duckpond.wesleyan.edu/twitter-project/data/

Fig. 6. A cloud-style representation of words distribution in the dataset. It includes four sections respectively showing (from left to right and from top to bottom) the more used words for the following semantic areas: family, socio-political debate, legal aspects, public manifestations.

and irony labels (similarly as done in [4]) will be of great help. We are currently working on the development of a sub–corpus, mentioned in section IV-B, where sentiment polarity and irony will be annotated with respect to specific topics and semantic areas, selected by relying on the techniques described in the previous section.

Moreover, while in this paper we have considered only the relationship between Twitter and events happened in politics, in our future plans there is the idea of evaluating the relationships also with other media and public discussions, which strongly influenced the debate, e.g. online journals and newspapers, but also the parliamentary debates. Another aspect that can be relevant in a social perspective, it is also considering the network of re-tweets as well as modeling the list of the users of the hashtag as a network.

References

[1] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis (Foundations and Trends(R) in Information Retrieval)*. Now Publishers Inc., 2008.

[2] B. Liu, *Sentiment analysis and subjectivity*. Taylor and Francis Group, Boca, 2010.

[3] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[4] C. Bosco, V. Patti, and A. Bolioli, "Developing corpora for sentiment analysis: The case of irony and Senti–TUT," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 55–63, 2013.

[5] J. Skilters, M. Kreile, U. Bojars, I. Brikse, J. Pencis, and L. Uzule, "The pragmatics of political messages in twitter communication," in *ESWC Workshops*, ser. Lecture Notes in Computer Science, R. Garcia-Castro, D. Fensel, and G. Antoniou, Eds., vol. 7117. Springer, 2011, pp. 100–111.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proceedings of the ICWSM-11*, Barcelona, Spain, 2011, pp. 178–185.

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, "Political polarization on twitter," in *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847

[8] H. Li, X. Cheng, K. Adson, T. Kirshboim, and F. Xu, "Annotating opinions in German political news," in *Proceedings of the LREC'12*, Istanbul, Turkey, 2012, pp. 1183–1188.

[9] Y. He, H. Saif, Z. Wei, and K.-F. Wong, "Quantising opinions for political tweets analysis," in *Proceedings of the LREC'12*, Istanbul, Turkey, 2012, pp. 3901–3906.

[10] D. Maynard and M. Greenwood, "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis." in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: ELRA, may 2014.

[11] A. Reyes, P. Rosso, and D. Buscaldi, "From humor recognition to irony detection: The figurative language of social media," *Data Knowledge Engineering*, vol. 74, pp. 1–12, 2012.

[12] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[13] A. Gianti, C. Bosco, V. Patti, A. Bolioli, and L. D. Caro, "Annotating irony in a novel italian corpus for sentiment analysis," in *Proceedings of the 4th Workshop ES3*, Istanbul, Turkey, 2012, pp. 1–7.

[14] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in Twitter and Amazon," in *Proceedings of the CONLL'11*, Portland, Oregon (USA), 2011, pp. 107–116.

[15] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, A. Reyes, and J. Barnden, "Semeval-2015 task 11: Sentiment analysis of figurative language in twitter," in *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM*, 2015.

[16] E. Eensoo and M. Valette, "Approche textuelle pour le traitement

automatique du discours evaluatif," *Langue francaise*, no. 4, pp. 109–124, 2014.

[17] A. Esuli, S. Baccianella, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. of LREC'10*. ELRA, 2010.

[18] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[19] A. Fraisse and P. Paroubek, "Toward a unifying model for opinion, sentiment and emotion information extraction," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3881–3886.

[20] ——, "Twitter as a comparable corpus to build multilingual affective lexicons," in *Proceedings of the LREC'14 Workshop on Building and Using Comparable Corpora*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 17–21.

[21] Y. Bestgen, "Building affective lexicons from specific corpora for automatic sentiment analysis," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008, pp. 496–500.

[22] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto, "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 58–65.

[23] F. Chiusaroli, "Scritture brevi oggi. tra convenzione e sistema," in *Scritture brevi di oggi*, F. Chiusaroli and F. M. Zanzotto, Eds. Università Orientale di Napoli, 2012, pp. 4–44.

[24] D. Virone and M. Lai, "Dans un corpus hybride: les messages twittés, l'hypertextualité et la formule 2.0," in *Actes du Colloque international ICODOC 2015: Icar COlloque DOCtorants/DOCteurs*, 2015, pp. 66–67.

[25] V. Blondel, J. Guillaume, R. Lambiotte, and E. Mech, "Fast unfolding of communities in large networks," *J. Stat. Mech*, 2008.

[26] I. Hernandez-Farias, D. Buscaldi, and B. Priego-Sánchez, "IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task," in *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA2014*, Pisa, Italy, 2014, pp. 75–81.

[27] V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso, "Overview of the Evalita 2014 SENTIment POLarity Classification Task," in *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*. Pisa, Italy: Pisa University Press, 2014, pp. 50–57.

[28] G. Attardi, V. Basile, C. Bosco, T. Caselli, F. Dell'Orletta, S. Montemagni, V. Patti, M. Simi, and R. Sprugnoli, "State of the art language technologies for italian: The EVALITA 2014 perspective," *Journal of Intelligenza Artificiale*, vol. 9, no. 1, pp. 43–61, 2015.