Building a Corpus on a Debate on Political Reform in Twitter

Mirko Lai[1], Daniela Virone[2], Cristina Bosco[1], Viviana Patti[1]
[1]Dipartimento di Informatica, Università degli Studi di Torino
[2]Dipartimento di studi umanistici, Università degli Studi di Torino
{lai,bosco,patti}@di.unito.it, dvirone@unito.it

Abstract

English. The paper describes a project for the development of a French corpus for sentiment analysis focused on the texts generated by the participants to a debate about a political reform, i.e. the bill on homosexual wedding in France. Beyond the description of the data set, the paper shows the methodologies applied in the collection and annotation of data. The collection has been driven by the detection of the hashtag mainly used by the participants to the debate, while the annotation has been based on polarity but also extended with the target semantic areas involved in the debate.

Italiano. *L'articolo descrive un progetto per lo sviluppo di un corpus per la sentiment analysis composto di testi in francese prodotti dai partecipanti ad un dibattito su una riforma politica, i.e. la legge sul matrimonio gay in Francia. Oltre a descrivere il dataset, l'articolo mostra le metodologie applicate nella raccolta e nell'annotazione dei dati. La raccolta dei dati è stata guidata dalla presenza dell'hashtag maggiormente utilizzato dai partecipanti al dibattito, mentre l'annotazione è basata oltre che sulla polarità anche sulle aree semantiche toccate dai partecipanti nel dibattito.*

1 Introduction

The recent trends in sentiment analysis are towards hybrid approaches or to computational semantics oriented frameworks where linguistic and pragmatic knowledge are encompassed for describing a global notion of communication. This notion 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 (e.g. irony, metaphor or hyperbole) or, in general, when the communicated content does not correspond to words meaning but depends also on other communicative behaviors.

On this perspective, a particular interesting domain is related to the political debates and, in particular, in the specific form that such debates assumes in social media, which strongly differentiate them from other kinds of classical conversational contexts (Rajadesingan and Liu, 2014). 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 staffs 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 focused on understanding the phenomenon, by studying the effect of this technology on the election outcomes (Skilters et al., 2011), its possible use to gauge the political sentiment (Tumasjan et al., 2011) and the users' stance on controversial topics (Rajadesingan and Liu, 2014), or by studying the networks of communication in order to investigate the political polarization issue (Conover et al., 2011).

This study contributes to this area by showing a methodology for the collection and annotation of a data set composed by texts from different media where a political debate has been developed. As a starting point of the project in this paper we will present a dataset of Twitter messages in French language about the reform "Le Mariage Pour Tous" (Marriage for everyone, i.e. marriage for all), discussed in France in 2012 and 2013. The collection of th dataset has been driven by a hashtag, i.e. #mariagepourtous, created to mark the messages about the debate on the reform, while the

selection of tags to be annotated has been based on the detection and analysis of the semantic areas involved in users posts. The detection of these areas is the result of a set of analysis we applied on the corpus described in more details in (Lai et al., 2015).

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 in instead devoted to the annotation of collected data.

2 Related work

Several works rely on sentiment analysis techniques (Pang and Lee, 2008) to analyze politics (Tumasjan et al., 2011; Li et al., 2012; He et al., 2012), a domain where the problems related to the exploitation of figurative language devices described in (Maynard and Greenwood, 2014; Bosco et al., 2013; Reyes et al., 2012; Reyes et al., 2013; Gianti et al., 2012; Davidov et al., 2011) and in the Semeval15-11 shared task (Ghosh et al., 2015) have been detected as frequent. Moreover, some research focused on aspects concerning the political polarization in Twitter (Conover et al., 2011; Skilters et al., 2011), or on the detection of the stance of Twitter users from their tweets debating a controversial topic, such as abortion, gun reforms and so on (Rajadesingan and Liu, 2014)[1]. All such perspectives are very interesting also in the dataset we are describing in the current work.

Other works, instead, addressed the issues related to the arguments accompanying the political messages, like (Eensoo and Valette, 2014) 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. This is a strong motivation for the development of annotated corpora where this kind of knowledge can be reliably described. 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.

Finally let us to notice that most of the works carried on so far in this area focus their analysis on English datasets only, while under this respect several languages, like French or Italian, are currently under-resourced, with some exception (Stranisci et al., 2015).

3 Collection and composition of the data set

This work is collocated in the context of an ongoing project about communication in different media and is focused on the debate about homosexual couple wedding in France. The project includes the collection of the following datasets from different media and sources:

- TW-MariagePourTous: texts from Twitter selected 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*.
- NEWS-MariagePourTous: texts from French newspapers, i.e. LeMonde online and sources retrieved by using the Factiva search engine[2], published in the time-lapse 7th June 2011 - 4th February 2013 and filtered by the keyword *#mariagepourtous*.
- NEWSTITLE-MariagePourTous: titles only of the texts collected in the NEWS-MariagePourTous corpus.
- DEBAT-MariagePourTous: texts 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[3].

The largest corpus is NEWS-MariagePourTous, which includes around 24,000 articles, while

[1] A new task on *Detecting Stance in Tweets* has been proposed in Semeval-2016 (Task 6) as part of the Sentiment Analysis Track: `http://alt.qcri.org/semeval2016/task6/`

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

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

Figure 1: A cloud-style representation of words distribution in the TW-MPT dataset.

NEWSTITLE-MariagePourTous is the smallest.

The current study focus on the MariagePourTous dataset (henceforth TW-MPT), which includes 254,366 messages. 88,157 of them have been retweetted by one or more users during the time of the corpus collection[4]. Each tweet is associated with the metadata related to the posting time and the user that posted it, information that can be exploited in the analysis of data.

The collection of this corpus is based on the detection of the hashtag *#mariagepourtous*. Hashtags are single words or expressions (with words not separated by spaces) preceded by the symbol '#', well known in Twitter and exploited by users to create communities of people interested in the same topic (Cunha et al., 2011), by making it easier for them to find and share information related to it (think, for instance, of the hashtags/slogans created during election campaigns). 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 (Chiusaroli, 2012), 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.

[4]We didn't include in the annotated corpus the retweetted messages but we have this information available for further processing and statistics (Lai et al., 2015).

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. Furthermore, we could observe the "life" of the hashtag during its first propagation among Twitter users, and then diffusion within the community (see (Lai et al., 2015)).

4 Data analysis and data-driven annotation

As previously reported, the limited amount of resources available for French sentiment analysis makes the development of a sentiment annotation for the TW-MPT corpus an especially valuable effort. Nevertheless, our main goal was to test a methodology for the definition of a data-driven annotation scheme, which can be applied also in other cases, and, in particular, in socio-political debates for making explicit the features of this kind of conversational context. Therefore, our annotation scheme extends the standard annotation for marking the polarity of opinions and sentiments, usually applied in corpora annotated for sentiment analysis, by including both tags for marking figurative language devices and a set of semantically oriented labels. The analyses based on linguistic and non linguistic features described in (Lai et al., 2015), which we applied for detecting the dynamics of communicative behavior of

users in exploiting subjective and evaluative language, meaningfully helped us in designing this annotation scheme.

For what concerns polarity, we applied in this project the same approach applied in (Gianti et al., 2012; Bosco et al., 2013; Bosco et al., 2014; Bosco et al., 2015) for the annotation of Italian corpora for sentiment analysis, which includes the tags of table 1.

The annotation of figurative devices is based on three labels: HUM POS for marking the presence of irony featured by positive polarity, HUM NEG for negative irony, and a yes/no feature for METAPHOR. Also other figurative devices (e.g. hyperbole) can be of interest for sentiment analysis, but the extension of the schema in this direction will be object of future work.

label	polarity
POS	positive
NEG	negative
NONE	neutral
MIXED	both positive and negative
UN	unintelligible content
RP	repetition of a post

Table 1: Polarity tags annotated in the TW-MPT corpus.

For what concerns, instead, the set of semantically oriented tags, we defined them according to an analysis of the dataset. In fact, observing the corpus and the other collected data, we hypothesized that the debate developed around some particular topic, and the experiments performed validated our hypothesis. We classified the most frequently occurring words by the application of the cloud extraction techniques described in (Lai et al., 2015) to the full TW-MPT corpus tag. The result is that represented in Figure 1, showing that user tweets focused on few quite sharply distinguishable semantic areas encompassing several other relevant discussed themes: *family* (we labeled as FAMILLE), *legal aspects* (we labeled as LOI), *public manifestations* (we labeled as MANIF), *socio-political debate* (we labeled as DEBAT).

The annotation scheme has been applied on a first portion of 2,872 tweets of the TW-MPT corpus, i.e. all the posts where the hashtag occurs immediately after or before the verb "etre" (*to be*), namely the messages where the hashtag is in some way evaluated or defined by users. After the discussion and definition of a set of the guidelines to be shared, the annotation has been done by two independent skilled annotators, and the disagreement has been calculated and analyzed. Before the final release of the corpus, which will be available soon, a third annotation will be applied in order to improve the reliability of data, but some preliminary hints can be derived from the analysis of the currently available data.

The disagreement on polarity appears in 861 of the 2,872 annotated tweets, but it is mainly focused on cases where irony is involved. In 184 tweets only one annotator detected irony when the other doesn't, but both detected the same polarity. For instance, annotator-1 used POS and annotator-2 used HUM-POS, or viceversa or the same with the labels HUM-NEG and NEG. This confirms the hypothesis of a variable perception of irony among humans (González-Ibáñez et al., 2011). Only a limited amount of cases (177) have been found where the annotators disagreed annotating opposed polarities (i.e. POS and NEG). In a few remaining cases the disagreement depends on the annotation of a neutral polarity versus a defined polarity (173) or mixed polarity with respect to a sharp one (86). For what concerns the annotation of the semantic areas, it is featured by a very high agreement (the annotators selected the same label in 1958 cases). However, further investigations are needed in order to find areas where they mainly disagree. Finally, for what concerns the detection of metaphors, the related annotation is still in progress, as it has been applied to the corpus in a second stage.

5 Conclusions and future work

The paper presents a data-driven methodology for collecting and annotating corpora for sentiment analysis, which has been applied to a French corpus of a Twitter debate about a political reform. The collection is driven by a hashtag expoited by users expressing opinions of a controversial topic. The annotation is based on a set classical polarity labels, extended with tags for figurative language devices (i.e. irony) and for a few semantic areas detected in posts, intended as *aspects* of the reform on which users express their opinions.

The investigation of further aspects and information sources that can be found in data, e.g. emojis, links and images, is matter of future work.

Acknowledgements

The authors thank all the persons who supported the work, and in particular Federica Ramires that meaningfully contributed to the annotation and analysis of the corpus as part of her Bachelor's degree thesis.

References

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti–TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti, and Emilio Sulis. 2014. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. ELRA.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2015. Developing corpora for sentiment analysis: The case of irony and senti-tut (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4158–4162. AAAI Press / International Joint Conferences on Artificial Intelligence.

Francesca Chiusaroli. 2012. Scritture brevi oggi. tra convenzione e sistema. In Francesca Chiusaroli and Fabio Massimo Zanzotto, editors, *Scritture brevi di oggi*, pages 4–44. Università Orientale di Napoli.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos Andre Goncalves, and Fabricio Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, Portland, Oregon. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2011. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the CONLL'11*, pages 107–116, Portland, Oregon (USA).

Egle Eensoo and Mathieu Valette. 2014. Approche textuelle pour le traitement automatique du discours evaluatif. *Langue francaise*, 184(4):109–124.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. 2015. 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*.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2013)*, pages 1–7, Istanbul, Turkey. ELRA.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586. Association for Computational Linguistics.

Yulan He, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong. 2012. Quantising opinions for political tweets analysis. In *Proceedings of the LREC'12*, pages 3901–3906, Istanbul, Turkey.

Mirko Lai, Daniela Virone, Cristina Bosco, and Viviana Patti. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015)*. IEEE. In press.

Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim, and Feiyu Xu. 2012. Annotating opinions in German political news. In *Proceedings of the LREC'12*, pages 1183–1188, Istanbul, Turkey.

Diana Maynard and Mark Greenwood. 2014. 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, may. ELRA.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis (Foundations and Trends(R) in Information Retrieval)*. Now Publishers Inc.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in twitter debates. In William G. Kennedy, Nitin Agarwal, and Shanchieh Jay Yang, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 8393 of *Lecture Notes in Computer Science*, pages 153–160. Springer International Publishing.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowledge Engineering*, 74:1–12.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Jurgis Skilters, Monika Kreile, Uldis Bojars, Inta Brikse, Janis Pencis, and Laura Uzule. 2011. The pragmatics of political messages in twitter communication. In Raul Garcia-Castro, Dieter Fensel, and Grigoris Antoniou, editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Marco Stranisci, Cristina Bosco, Patti Viviana, and Delia Irazú Hernández Farias. 2015. Analyzing and annotating for sentiment analysis the socio-political debate on "La Buona Scuola". In *Proceedings of the 2th Italian Conference on Computational Linguistics (CLiC-IT 2015)*, Trento, Italy. In Press.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2011. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the ICWSM-11*, pages 178–185, Barcelona, Spain.