

Analyzing and annotating for sentiment analysis the socio-political debate on #labuonascuola

Marco Stranisci^{1,2}, Cristina Bosco¹, Viviana Patti¹, Delia Irazú Hernández Farías^{1,3}

¹Dipartimento di Informatica, Università di Torino

²Cooperativa weLaika, Torino

³Universitat Politècnica de Valencia

{bosco,patti}@di.unito.it, marco.stranisci@welaika.com,
dhernandez1@dsic.upv.es

Abstract

English. The paper describes a research about the socio-political debate on the reform of the education sector in Italy. It includes the development of an Italian dataset for sentiment analysis from two different comparable sources: Twitter and the online institutional platform implemented for supporting the debate. We describe the collection methodology, which is based on theoretical hypotheses about the communicative behavior of actors in the debate, the annotation scheme and the results of its application to the collected dataset. Finally, a comparative analysis of data is presented.

Italiano. *L'articolo descrive un progetto di ricerca sul dibattito socio-politico sulla riforma della scuola in Italia, che include lo sviluppo di un dataset per la sentiment analysis della lingua italiana estratto da due differenti fonti tra loro confrontabili: Twitter e la piattaforma istituzionale online implementata per supportare il dibattito. Viene evidenziata la metodologia utilizzata per la raccolta dei dati, basata su ipotesi teoriche circa le modalità di comunicazione in atto nel dibattito. Si descrive lo schema di annotazione, la sua applicazione ai dati raccolti, per concludere con un'analisi comparativa.*

1 Introduction

The widespread diffusion of social media in the last years led to a significant growth of interest in the field of opinion and sentiment analysis of user generated contents (Bing, 2012; Cambria et al., 2013). The first applications of these techniques

were focusing on the users' reviews for commercial products and services (e.g. books, shoes, hotels and restaurants), but they quickly extended their scope to other interesting topics, like politics. Applications of sentiment analysis to politics can be mainly investigated under two perspectives: on one hand, many works focus on the possibility of predicting the election results through the analysis of the sentiment conveyed by data extracted from social media (Ceron et al., 2014; Tumasjan et al., 2011; Sang and Bos, 2012; Wang et al., 2012); on the other hand, the power of social media as "a trigger that can lead to administrative, political and societal changes" (Maynard and Funk, 2011) is also an interesting subject to investigate (Lai et al., 2015). This paper mainly focuses on the last perspective. Our aim is indeed the creation of a manual annotated corpus for sentiment analysis to investigate the dynamics of communication between politics and civil society as structured in Twitter and social media. We focused from the beginning our attention mainly on Twitter because of the relevance explicitly given to this media in the communication dynamics of the current government. In order to describe and model this communicative behavior of the government, we assume the theoretical framework known in literature as *framing*, which consists in making especially salient in communication some selected aspect of a perceived reality (Entman, 1993).

The data selected to create the corpus have been chosen by analyzing in Twitter and other contexts the diffusion of three hashtags, i.e. #labuonascuola, #italicum, #jobsact. In particular, we focus on #labuonascuola (*the good school*), which was coined to communicate the school reform proposed by the actual government.

A side effect of our work is the development of a new lexical resource for sentiment analysis in Italian, a currently under-resourced language. Among the existing resources let us mention Senti-TUT

(Bosco et al., 2013), which has been exploited together with the TWITA corpus (Basile and Nissim, 2013) for building the training and testing datasets in the SENTiment POLarity Classification shared task (Basile et al., 2014) recently proposed during the last edition of the evaluation campaign for Italian NLP tools and resources (Attardi et al., 2015). The Sentipolc's dataset includes tweets collected during the alternation between Berlusconi and Monti on the chair of Prime Minister of the Italian government. The current proposal aims at expanding the available Italian Twitter data annotated with sentiment labels on the topic of politics, and it is compatible with the existing datasets w.r.t. the annotation scheme and other features.

The paper is organized as follows. The next section describes the dataset mainly focusing on collection. In the section 3 we describe the annotation applied to the collected data and the annotation process. Section 4 concludes the paper with a discussion of the analysis applied to the dataset.

2 Data collection: criteria and subcorpora

In this section, we describe the methodology applied in collection, which depends on some assumption about the dynamics of the debate, and the features of the resulting dataset, which is organized in two different subcorpora: the Twitter dataset (TW-BS) and the dataset including the textual comments extracted from the online consultation about the reform (WEB-BS).

In order to describe the communicative behavior of the government, we assume, as a theoretical hypothesis, that the communication strategy acting in the debate can be usefully modeled by exploiting *frames*. In political communication, this cognitive strategy led to impose a narration to opponents (Conoscenti, 2011).

Following this hypothesis we can see that the Prime Minister and his staff coined two categories of frames by hashtagging in order to impose a narration to the public opinion: the first one aimed at legitimating the new born government and its novelty in the political arena (#lavoltabuona; #passodopopasso); the other one in order to create a general agreement on some proposal (#labuonascuola, #italicum, #jobsact). Each of these hashtags could be considered as an indicator of a frame created for elaborating a storytelling on the three most important reforms pro-

posed by the government respectively on school, job and elections.

The observation of Twitter in this perspective led us to focus on messages featured by the presence of the three keywords #labuonascuola, #jobsact, #italicum, and posted from February 22th, 2014 (establishment of the new government) to December 31st, 2014. First, we collected all Italian tweets in this time slot (218,938,438 posts), then we filtered out them using the three hashtags. With 28,363 occurrences #labuonascuola, even if attested later than the others, is featured by the higher frequency, which occur respectively 27,320 (#jobsact) and 3,974 (#italicum) times. This prevalence is due not only to the general interest for the topic, but in particular to the activation by the government of an online consultation on school reform through the website <https://labuonascuola.gov.it>.

The first corpus we collected, WEB-BS henceforth, includes therefore texts from this online consultation¹. We collected 4,129 messages composed by short texts posted in the consultation platform. All contents were manually tagged by authors with one among the 53 sub-topics labels made available, and organized by themselves in four categories: 'what I liked' (642), 'what I didn't like' (892), 'what is missing' (675) and 'new integration' (1,920). So, the label which conveys a positive opinion represented the 15.55% of the total. Otherwise, the negative label has been used the 21.60% of times. This manual classification in sub-topic and polarity categories of the messages, makes the WEB-BS dataset especially interesting, since the explicit tagging applied by the users can be in principle compared with the results of some automatic sentiment or topic detection engine. Moreover, let us observe that even if the WEB-BS corpus shares linguistic features with the corpus extracted from Twitter described below, it represents a different global context (Sperber and Wilson, 1986) (Yus, 2001) that orients, at the pragmatic level (Bazzanella, 2010), users in the expression of their opinions.

The second corpus we collected is composed of texts from Twitter focused on the debate on school (TW-BS henceforth), selected by filtering Twitter data exploiting the previously cited "fram-

¹Users could participate to the consultation in different ways: as single users, filling out a survey, or as a group taking part to a debate about a particular topic or aspect of the reform.

ing” hashtags. We focused our attention on tweets posted from September 3rd, 2014 (when the consultation was launched by the government with a press conference) to November 15th, 2014. In addition to #labuonascuola, we used also keywords like ‘la buona scuola’, ‘buona scuola’, ‘riforma scuola’, ‘riforma istruzione’. The resulting dataset is composed of 35,148 tweets, which was first reduced to 11,818 after removing retweets, and then to 8,594 after a manual revision devoted to further deletion of duplicates and partial duplicates. A quantitative analysis of the collected data shows us that 4,244 users contributed to the debate on Twitter. Among them, only 1,238 (29,2%) posted at least 2 messages and produced 5,588 tweets, 65% of the total. If we consider the hashtags’ occurrences, #labuonascuola appears 5,346 times, while its parodic reprise is very infrequent: 108 total occurrences for three hashtags #lacattivascuola - #thebadschool, #lascuolaingiusta - #theunfairschool, and #labuonasola - #thegoodswindle.

3 Annotation and disagreement analysis

The annotation process involved 8 people with different background and skills, three males and five women. The task was marking each post with a polarity and one or more topic according to the set of tags described below.

For what concerns polarity, we assumed the same labels exploited in the Senti-TUT annotation schema: NEG for negative polarity, POS for positive, MIXED for positive and negative polarity both, NONE in the case of neutral polarity. Finally, we annotated irony, whose recognition is a very challenging task for the automatic detection of sentiment because the inferring process goes beyond syntax or semantics (Reyes et al., 2013; Reyes and Rosso, 2014; Maynard and Greenwood, 2014; Ghosh et al., 2015). As in Sentipolc (Basile et al., 2014), we were interested in annotating manually the polarity of the ironic tweets, where the presence of ironic devices can work as an unexpected “polarity reverser” (e.g. one says something “good to mean something “bad). So, we coined two labels: HUM NEG for tagging tweets ironic and negative, and conversely HUM POS for tagging the ones that were both positive and ironic. The set of labels was completed by a tag for marking unintelligible tweets (UN), one for duplicates (RT), and NP for texts about not related topic.

As far as topics are concerned, among the 53

categories used in the WEB-BS corpus, we selected the 13 most frequent, which occur 2,182 times in the consultation website: docenti - *teachers*, valutazione - *evaluation*, formazione - *training*, alternanza scuola/lavoro - *school-work*, investimenti - *investments*, reclutamento - *recruitment*, curriculum - *curriculum*, innovazione - *innovation*, lingue - *languages*, merito - *merit*, presidi - *headmasters*, studenti - *students*, and retribuzione - *remuneration*. Furthermore, we coined two more general labels for tweets addressing a sub-topic not present in categories, and for tweets just indirectly targeted to school reform.

In order to limit biases among annotators and to make well shared the meaning of all the labels to be annotated, we produced a document including guidelines for annotations, several examples of polarity-labeled tweets, three glossaries about the meaning of the topics- and some recurrent terms on the school reform.

The final dataset, manually annotated by two independent human annotators and cleaned from duplicates, not related, and unintelligible tweets, consists of 7,049 posts. 4,813 out of the total amount of annotated tweets, were tagged with the same label by both annotators. This is the current result for TW-BS; the label distribution is shown in 1. The inter-annotator agreement at this stage was $\kappa = 0.492$ (a moderate agreement). A qualitative analysis of disagreement (the 31.8% of the data) shows that the discrepancies very often depend on the presence of irony which has been detected only by one of the annotators even if both the humans performing the task detected the same polarity. This confirms the fact known in literature that irony is perceived in different ways and frequency by humans, as in the following example which showed a disagreement between annotators:

‘Ho letto le 136 pagine della riforma della scuola, finisce che i giovani si diplomano e vanno all’estero.
#labuonascuola’
‘I read the 136 pages about the school reform, it ends with youngs who graduate and go abroad.
#labuonascuola’

The remaining part of disagreeing annotations can be reported mainly as cases where one annotator detected a polarity and the other annotated the post as neutral. In order to extend the dataset, we are planning to apply a third independent annotation on the posts with disagreeing annotations.

4 Analysis of corpora

The analysis is centered on two main aspects of the annotation, i.e. polarities and topics, in the perspective of label frequency and relationships between labels and disagreement.

Table 1 shows the frequency of the labels exploited for polarity and a high frequency of the neutral label can be observed in this graphic. When discussing the guidelines for the application of labels, we decided to use the NONE label for marking all the cases where textual features that explicitly refer to a polarized opinion couldn't be detected. A further investigation would be necessary in order to make a distinction between neutral subjectivity (e.g. expressions of hope, without a positive or negative valence) and pure objectivity (Wilson, 2008; Liu, 2010).

For what concerns tweets marked as positive and negative, if we hold together the ironic-polarized tweets with their corresponding labels, we have 924 negatives (37.09% of the total) against 263 positive posts (10.7%). The disparity is amplified when we take into account just ironic tweets. The use of irony for conveying a positive opinion is very rare (18 occurrences only).

label	occurrences
<i>NONE</i>	2,469
<i>NEG</i>	1381
<i>POS</i>	497
<i>HUM NEG</i>	404
<i>HUM POS</i>	18
<i>MIXED</i>	44

Table 1: Labels distribution in TW-BS.

A comparative analysis of polarity distribution in the TW-BS and the WEB-BS corpora has shown further important differences. The distribution of polarity is more balanced in the latter than in the former, where negative polarity prevails, while irony, frequently occurring in the Twitter corpus is almost absent in the other one. This confirms our theoretical hypothesis that the global contexts underlying these datasets are different, but also raises issues about the higher politeness and the cooperativeness applied by users in the consultation with respect to what is expressed in a social media context like Twitter. Furthermore, the nature of ironic posts on Twitter deserves further and deeper investigations, e.g. about the relation between the presence or the absence of ironic tweets and the

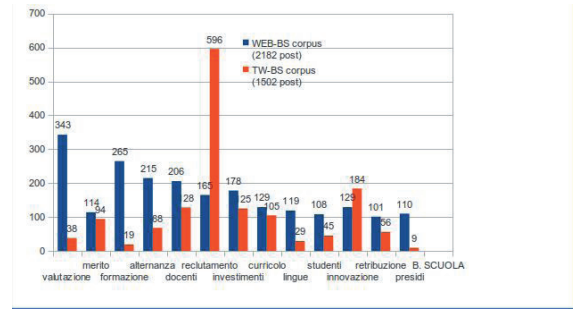


Figure 1: the use of topic labels in WEB-BS corpus, and TW-BS corpus

occurrence of particular events, like the press conference that launched the reform. For what concerns instead the analysis of topics, we observed that, even if the disagreement has not been high (31.4%), the annotators mostly did agree on the generic label BUONA SCUOLA, which occurs 4,071 times with the agreement of two annotators. This is confirmed by the limited exploitation of the more specific labels for the annotation of topic: the total amount of all the specific labels is 1,502. Moreover, it emerges a difference between the topics selected by users in WEB-BS corpus, and the ones annotated in the TW-BS corpus. This difference between contents proposed by the government and the topics spread out from the micro-blogging platform can be observed by looking at the different distribution of the labels in the two contexts. If we consider just the 13 label used both for TW-BS corpus, and WEB-BS corpus, we can notice important differences. For instance, VALUTAZIONE was the mainly used during the debate (15.72%), but attested few times in Twitter (2.52%). Otherwise, the label RECLUTAMENTO, which was used only the 7.56% of the times in the WEB-BS corpus, is the most frequent in the TW-corpus (39.68% of the occurrences).

5 Conclusions

The paper describes a project for the analysis of a socio-political debate in a sentiment analysis perspective. A novel resource is presented by describing the collection and the annotation of the dataset organized in two subcorpora according to the source the texts have been extracted from: one from Twitter and one from the institutional online consultation platform. A first analysis of the resulting dataset is presented, which takes into account also a comparative perspective.

Acknowledgements

The authors thank all the persons who supported the work. We are grateful to our annotators, in particular to Valentina Azer, Marta Benenti, Martina Brignolo, Enrico Grosso and Maurizio Stranisci. This research is supported in part by Fondazione Giovanni Gorla e Fondazione CRT (grant Master dei Talenti 2014; Marco Stranisci) and in part by the National Council for Science and Technology, CONACyT Mexico (Grant No. 218109, CVU-369616; D.I. Hernández Farías).

References

- G. Attardi, V. Basile, C. Bosco, T. Caselli, F. Dell'Orletta, S. Montemagni, V. Patti, M. Simi, and R. Sprugnoli. 2015. State of the art language technologies for Italian: The Evalita 2014 perspective. *Journal of Intelligenza Artificiale*, 9(1):43–61.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia. Association for Computational Linguistics.
- Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. 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)*, pages 50–57, Pisa, Italy. Pisa University Press.
- Carla Bazzanella. 2010. Contextual constraints in cmc narrative. In Christian Hoffmann, editor, *Narrative Revisited*, pages 19–38. John Benjamins Publishing Company.
- Liu Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- 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.
- E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.
- Andrea Ceron, Luigi Curini, and Iacus M. Stefano. 2014. *Social Media e Sentiment Analysis: l'evoluzione dei fenomeni sociali attraverso la rete*. Springer.
- Michelangelo Conoscenti. 2011. *The Reframer: An Analysis of Barack Obama Political Discourse (2004-2010)*. Bulzoni Editore.
- Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.
- A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, A. Reyes, and J. 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*.
- Mirko Lai, Daniela Virone, Cristina Bosco, and Viviana Patti. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *Proc. of 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), Special Track on Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis.*, Paros, France. IEEE. In press.
- Bing Liu. 2010. *Sentiment analysis and subjectivity*. Taylor and Francis Group, Boca.
- Diana Maynard and Adam Funk. 2011. Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference Workshop*, pages 88–99.
- 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.
- Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614.
- Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.
- Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dan Sperber and Deirdre Wilson. 1986. *Relevance: communication and cognition*. Basil Blackwell.
- Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Weppe. 2011. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the ICWSM-11*, pages 178–185, Barcelona, Spain.
- Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations, ACL '12*, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states*. Ph.D. thesis, University of Pittsburgh.

Francisco Yus. 2001. *Ciberpragmatica : el uso del lenguaje en Internet*. Ariel.