Tweeting in the Debate about Catalan Elections

Cristina Bosco[1], Mirko Lai[1,2], Viviana Patti[1], Francisco M. Rangel Pardo[2], Paolo Rosso[2]
[1]Università degli Studi di Torino, [2]Universitat Politècnica de València
{bosco,lai,patti}@di.unito.it,
francisco.rangel@autoritas.es, prosso@dsic.upv.es

Abstract

The paper introduces a new annotated Spanish and Catalan data set for Sentiment Analysis about the Catalan separatism and the related debate held in social media at the end of 2015. It focuses on the collection of data, where we dealt with the exploitation in the debate of two languages, i.e. Spanish and Catalan, and on the design of the annotation scheme, previously applied in the development of other corpora about political debates, which extends a polarity label set by making available tags for irony and semantic oriented labels. The annotation process is presented and the detected disagreement discussed.

Keywords: annotation, sentiment, figurative language, Spanish, politics, Twitter,

1. Introduction

Texts generated by users within the context of social media can be a great opportunity for moving onward the development of corpus-based techniques for Sentiment Analysis and Opinion Mining (SA&OM). In this paper, we present the preliminary findings of an ongoing project for the development of a new annotated corpus for the application on Spanish and Catalan of SA&OM techniques, called TWitter-CatalanSeparatism (henceforth TW-CaSe). It collects texts from Twitter about the debate in Catalonia (Spain) on the elections and on the separation of the region from Spain.

The development of this resource is collocated within the wider context of a research about communication in socio-political debates which is featured by a semantically oriented methodology for the annotation of data sets for SA&OM. We adopted an approach based on a global notion of communication oriented towards a holistic comprehension of all the parts of the message, which includes e.g. context, themes, and 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.

The approach has been tested until now on texts from two different socio-political debates, namely the debate on the homosexual wedding in France (Bosco et al., 2015; Lai et al., 2015; Bosco et al., 2016) and that on the reform of the school and education sector in Italy (Stranisci et al., 2015; Stranisci et al., 2016). The new corpus described in the present paper will spread out the multilingual perspective by adding to the data for Italian and French those for Spanish and Catalan. Because of the differences in topics and languages, these corpora considered together will allow us to test the relative independence of the approach from topic and language, but also to prepare the ground for future cross-linguistic comparisons. These resources can indeed shed some light on the way communities of users with different roles in the society and different political sentiment interact one another. Moreover, the novelty of this work consists in both developing currently missing resources and extending the treatment of political texts for SA&OM (Conover et al., 2011a; Li et al., 2012; Conover et al., 2011b; Skilters et al., 2011) towards the field of discussions about controversial topics.

Let us notice that French, Spanish and Catalan are currently under resourced languages w.r.t. English[1], even if, in the last few years several efforts have been devoted to the development of new annotated data and affective lexicons, see e.g. the recent attempts to automatically build such resources (Bestgen, 2008; Fraisse and Paroubek, 2014a; Fraisse and Paroubek, 2014b). Similarly, a very limited amount of resources for SA&OM are available for Italian, except for some recent efforts such as the Senti-TUT corpus[2], where a set of Twitter posts have been manually annotated with affective polarity and irony, and the Sentix affective lexicon (Basile and Nissim, 2013), developed in the context of the TWITA project by the alignment of several resources, including SentiWordNet (Esuli et al., 2010), a well-known sentiment lexicon for English.

Furthermore, the resources can be also of some interest for training systems in stance detection, i.e. the task of automatically determining from text whether the author is in favor, against or neutral with respect to a given target when the topic is controversial, which is currently considered as a crucial issue for SA systems (see e.g. the Semeval 2016's Task about *Detecting Stance in Twitter* within the Sentiment analysis Track[3]).

The paper is organized as follows. First, it describes the collection and the annotation of the data set, then, it shows the preliminary analysis done on the data together with the analysis of the disagreement detected on the first portion of the data set which has been annotated.

[1]See results and cross-language comparison published at http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison in the context of META–NET, the Network of Excellence forging the multilingual Europe technology alliance, where Spanish, French and Catalan are among the languages discussed.

[2]http://www.di.unito.it/~tutreeb/corpora.html

[3]http://alt.qcri.org/semeval2016/task6/

2. The debate about Catalonia separatism and the collected corpus

The Catalonia region, located in northeastern Spain, was in the ancient past independent of the Iberian Peninsula government, featured by its own language (i.e. Catalan), laws and customs. More recently, the region was granted a degree of autonomy once more in 1977, but calls for complete independence grew steadily until July 2010, when the Constitutional Court in Madrid overruled part of the 2006 autonomy statute, stating that there is no legal basis for recognizing Catalonia as a nation within Spain. The economic crisis in Spain has only served to magnify calls for Catalan independence and Catalan nationalists held an unofficial poll in November 2014 achieving a large majority of votes for independence. The vote was non-binding as the Constitutional Court had ruled it illegal. But the secessionists viewed it as a defining moment and the declared regional elections in September 2015 have been a de facto referendum on independence. Catalan nationalist parties won an absolute majority in the 135-seat regional assembly and on 9 November pushed through a motion to start the process towards independence. The Spanish government has hit back, declaring the secessionist step unconstitutional.

As usual in the last few years in the debates about social and political topics, the debate on Catalan separatism involved a massive exploitation of social media by users interested in the discussion. For drawing attention to the related issues, as happens for commercial products or political elections (Sang and Bos, 2012), they created some new hashtag for making widely known information and their opinions. Among them *#Independencia* is one of the hashtags which has been accepted within the dialogical and social context growing around the topic, and largely exploited within the debate.

At the current stage of the development of our project we exploited the hashtag *#Independencia* as the first keyword for filtering data to be included in the TW-CaSe corpus. Nevertheless, because of the complexity of the debate and of the various social and political involved entities, this corpus will be extended in the next future by exploiting other filtering keywords and hashtags for collecting new data both for Spanish and Catalan, with the main aim to adequately represent the scenario. For the present time *#Independencia* allowed us the selection of about 3,500 original messages collected between the end of September and December 2015, and which have been also largely retweeted[4].

3. Annotation and analysis

The posts collected are featured by Spanish or Catalan language (almost equally represented), which cannot be automatically distinguished during collection. Only the posts in Spanish have been annotated until now, while the others will be annotated as a second step. In order to identify the two languages in the collected posts and to select a Spanish section of the corpus for accomplishing the annotation of irony and polarity, we involved two human annotators, both skilled in Spanish and Catalan language, that annotated both all the tweets of our collection, thus producing a pair of annotation for each tweet. The result is shown in Fig. 1. Overall, 2,274 tweets were identified as written in Catalan (CA) and 1,045 as written in Spanish (SP). Remaining tweets, such as tweets including just a list of hashtags consisting of both words in Spanish and Catalan, were labelled as UN.

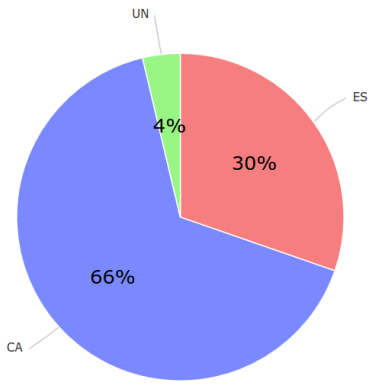

Figure 1: The distribution of languages in the 3,500 collected posts.

The 1,045 tweets identified as written in Spanish are included in the the Spanish section of our corpus (TW-CaSeSP henceforth) and has been annotated for polarity and irony according to annotation scheme and process described below. Let us observe that, not surprisingly, the data is unbalanced in terms of languages. The Catalan independence debate concerns a region with a very high percentage of people understanding and speaking the Catalan language. The distribution suggests that users posting about this issue using the selected hashtag are mostly Catalan speakers. A related issue concerns the possible bias of the dataset in a particular political viewpoint (e.g. more independence-wishers). Our plan to extend TW-CaSe by exploiting other keywords and hashtags for collecting new data both for Spanish and Catalan is also aimed to address this issue, having a wider coverage of debate in Twitter.

3.1. Annotation scheme: polarity and irony

As far as the design of the annotation scheme is involved, we applied the annotation exploited in (Bosco et al., 2013; Bosco et al., 2014; Basile et al., 2014) for marking the polarity of opinions and sentiments, extended with the labels UN and RP for marking unintelligible and repeated content respectively (see table 1).

As observed above, the data set includes texts both in Spanish and Catalan, but the annotation has been applied until now only to the tweets in Spanish, while the annotation for Catalan is undergoing. Moreover, we included additional tags for figurative language devices, i.e. irony and metaphor. In particular, HUMNEG, HUMPOS and HUMNONE are the labels we used for marking the presence of

[4]The dataset was collected with the Cosmos tool by Autoritas (`http://www.autoritas.net`) in the framework of ECOPORTUNITY IPT-2012-1220-430000 project funded by Spanish Ministry of Economics.

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-CaSe corpus.

irony (together with the information about the intended polarity), as summarized in Table 2.

label	figurative device
HUMPOS	positive irony
HUMNEG	negative irony
HUMNONE	neutral irony

Table 2: Tags annotated in the TW-CaSeSP corpus for figurative language uses.

The following are examples of application of the labels in the annotation of our corpus.

HUMNEG: *@junqueras Pues por la pinta, debes tener más cruces que la Carretera de Vicálvaro #IndependenciaCataluña.* (@junqueras Well, from the looks of it, you must have more intersections[5] than the Carretera de Vicálvaro #IndependenciaCataluña.).
HUMNONE: *ERC dice que si los catalanes votan #independencia no lo parará "ni Dios ni Rajoy". http://t.co/o7oU2JFbeC http://t.co/KAfchlWg8V* (ERC says that if the Catalans vote #independencia "neither God nor Rajoy" will stop him.)
HUMPOS: *Esto es tambien culpa de Mas? #JuntsPelSi #27S2015 #27S #independencia https://t.co/9wNRe7kmrN* (Is this also the fault of Mas? #JuntsPelSi (united for yes).)

As future work, it is also planned the annotation of metaphorical expressions that will be done by using the label METAPHOR, applied yet in the French corpus (Bosco et al., 2016).

3.2. Annotation process and analysis

We collected until now the annotation of two skilled humans for each Spanish post of TW-CaSe. The detected inter-annotator agreement at this stage was $\kappa = 0.662$[6]. Polarity and irony labels were distributed as follows: NONE 56,9%, POS 5,6%, NEG 25,9%, MIXED 5,9%, HUMPOS 0,2%, HUMNEG 4,6%, HUMNONE 1%, as shown in Fig 2.
A third annotation via the Crowdflower platform[7], a crowdsourcing platform for manual annotation often used in the community (Ghosh et al., 2015), is under development in order to reduce the detected disagreement and to improve the reliability of the data set for the release of the corpus.
Let us observe that most of the tweets in TW-CaSeSP with a polarity valence are negative, both in the absence and presence of irony. This is not surprising when we interpret this result as an indicator of the stance of the users on the Catalan independence. Indeed, we can hypothesize that most of the users in favor of independency will tweet in Catalan, and they are under represented in this Spanish section of TW-CaSe. The extension of the corpus with a Catalan section will be essential in order to have an overall picture of the debate in terms of sentiment expressed in social media.
Thanks to the association of each message with the metadata related to the author and posting time, and in order to better understand the conversational context growing around the debate, we are also currently performing a set of analysis according to the model described in (Lai et al., 2015) and in (Bosco et al., 2015). In particular, we are collecting the list of users that more frequently posted messages about the debate and the presence among them of opinion leaders, the frequency of tweets in different days and weeks in order to see the possible relationships between events and communication in Twitter.
Moreover, the presence of two different languages in the corpus gave us the possibility of a new perspective analysis, seeing how different opinions are distributed in the two different groups participating to the debate.
Finally, the couple of corpora on the socio-political debates held in France and Italy, will be used for the development of comparisons and for the investigation of communicative dynamics.

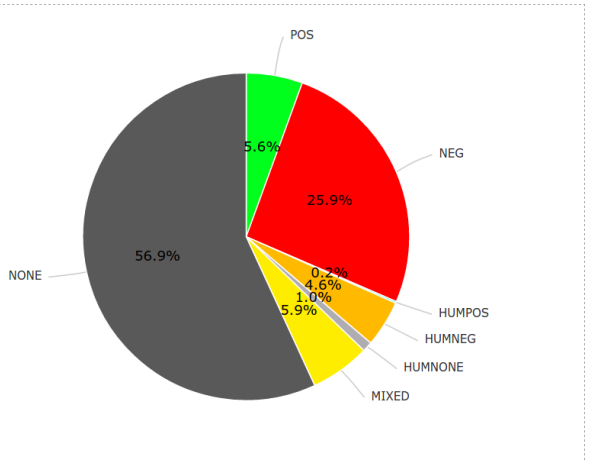

Figure 2: The distribution of polarity tags in the 822 agreed annotated posts of TW-CaSeSP.

4. Conclusions

The paper presents the ongoing development of a novel Spanish and Catalan corpus annotated for Sentiment Analysis and Opinion Mining. The corpus is part of a project for the study of communication in political debates oriented to a multilingual and holistic perspective. The annotation scheme is the same applied in other corpora developed for Italian and French within the context of the same project,

[5]The word 'cruces' can also be translated with 'crosses'.
[6]The value is calculated by considering the labels related to polarity and irony: POS, NEG, NONE, MIXED, HUMPOS, HUMNEG, HUMNONE.
[7]http://www.crowdflower.com/

and includes, beyond the polarity labels, also tags for marking figurative uses of language, in particular irony.
The contribute of the project consists both in making available data sets for currently under resourced languages and in preparing the ground for investigate communication dynamics in political debates and to do that also in a multilingual perspective.

5. Acknowledgement

We would like to acknowledge for their contribute in the annotation of the Spanish portion of the corpus Valeria Petta, that did that in accomplishment of her master's degree thesis, and Delia Irazú Hernández Farias, PhD student under the joint supervision of the Universitat Politècnica de València and the Università degli Studi di Torino. The work of Viviana Patti was partially carried out at the Universitat Politècnica de València in the framework of a three-month fellowship of the University of Turin co-funded by Fondazione CRT (WWS2 Program). Paolo Rosso has been partially funded by SomEMBED MINECO TIN2015-71147-C2-1-P research project and by the Generalitat Valenciana under the grant ALMAPATER (PrometeoII/2014/030).

Basile, V. and Nissim, M. (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.

Basile, V., Bolioli, A., Nissim, M., Patti, V., and Rosso, P. (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.

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

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

Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G., Sanguinetti, M., and Sulis, E. (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 Open Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. ELRA.

Bosco, C., Patti, V., Lai, M., and Virone, D. (2015). Building a corpus on a debate on political reform in Twitter. In *Proceedings of CLIC-2015*, pages 171–176.

Bosco, C., Lai, M., Patti, V., and Virone, D. (2016). Tweeting and being ironic in the debate about a political reform: the French annotated corpus TWitter-MariagePourTous. In *Proceedings of LREC 2016*. To appear.

Conover, M., Gonçalves, B., and Ratkiewicz, J. (2011a). Predicting the political alignment of Twitter users. In *Proceeding of the IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199, Los Angeles, CA, USA. Academy of Science and Engineering.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011b). Political polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Esuli, A., Baccianella, S., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC'10*. ELRA.

Fraisse, A. and Paroubek, P. (2014a). 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)*, pages 3881–3886, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fraisse, A. and Paroubek, P. (2014b). Twitter as a comparable corpus to build multilingual affective lexicons. In *Proceedings of the LREC'14 Workshop on Building and Using Comparable Corpora*, pages 17–21, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado, June. Association for Computational Linguistics.

Lai, M., Virone, D., Bosco, C., and Patti, V. (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*, Paris, France. IEEE.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in German political news. In *Proceedings of the LREC'12*, pages 1183–1188, Istanbul, Turkey.

Sang, E. T. K. and Bos, J. (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.

Skilters, J., Kreile, M., Bojars, U., Brikse, I., Pencis, J., and Uzule, L. (2011). The pragmatics of political messages in Twitter communication. In Raul Garcia-Castro, et al., editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Stranisci, M., Bosco, C., Patti, V., and Hernández-Farias, I. (2015). Analyzing and annotating for sentiment analysis the socio-political debate on #labuonascuola. In *Proceedings of CLIC.it 2015*, pages 274–279.

Stranisci, M., Bosco, C., Hernández-Farias, I., and Patti, V. (2016). Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *Proceedings of LREC 2016*. To appear.