Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT (Extended Abstract) *

Cristina Bosco[1] and **Viviana Patti**[1] and **Andrea Bolioli**[2]
[1]Dipartimento di Informatica, Università di Torino, Italy
{bosco,patti}@di.unito.it
[2]CELI srl, Torino, Italy
abolioli@celi.it

Abstract

This paper focusses on the main issues related to the development of a corpus for opinion and sentiment analysis, with a special attention to irony, and presents as a case study Senti–TUT, a project for Italian aimed at investigating sentiment and irony in social media. We present the Senti–TUT corpus, a collection of texts from Twitter annotated with sentiment polarity. We describe the dataset, the annotation, the methodologies applied and our investigations on two important features of irony: polarity reversing and emotion expressions.

1 Introduction

Mining opinions and sentiments from natural language is an extremely difficult task. It involves a deep understanding of explicit and implicit information conveyed by language structures – whether in a single word or an entire document. Recently proposed approaches, which rely on a structured notion of text [Johansson and Moschitti, 2013], are oriented to capture information going beyond the word level to outperform social media search tools in terms of portability and performance. Among them, several are based on statistical and machine learning NLP and assume as prerequisite human annotation of texts, both as ground truth data for measuring the accuracy of classification algorithms and as training data for supervised machine learning. In this article, we discuss the problems underlying the development of written-text corpora for Opinion Mining and Sentiment Analysis (OM&SA). We briefly survey the research area and refer to the specific case of irony, a linguistic device that's especially challenging for NLP and is common in social media [Ghosh *et al.*, 2015]. As a case study, we present the Senti–TUT Twitter corpus that was designed to study sentiment and irony for Italian, a language currently under-resourced for OM&SA.

2 Developing Corpora for Opinion and Sentiment Analysis

The development of a corpus consists in three main steps: collection, annotation and analysis. Each of them is strongly influenced by the others. For instance, the analysis and exploitation of a corpus can reveal limits of the annotation or data sampling, which can be respectively addressed by improving annotation and collecting more adequate data.

2.1 Collection

Most of the corpora designed for OM&SA are collected from web services which provide comments on commercial products, like reviews posted on Amazon [Davidov *et al.*, 2011; Filatova, 2012] blogs and micro-blogs like Facebook and Twitter, in order to provide insights about people's sentiments about celebrities or politics, see e.g. USA [Tumasjan *et al.*, 2011], German [Li *et al.*, 2012] or UK elections [He *et al.*, 2012]. Often the OM&SA corpora are the result of sampling and filtering oriented to a particular target or source. Data selection and filtering are usually based on keywords and hashtags. Moreover, metadata on time and geolocations, users' age, gender, background and social environment, or communicative goals, enable the detection of sentiment variation or trends. Text selection should be driven by considerations about text genres, which are featured by the exploitation and frequency of different linguistic structures and devices, subjectiveness vs objectiveness, message length.

The most frequently used collection methodologies are Web crawling and scraping, or calling the Web APIs exposed by the service (Google Reader's API, Twitter's API, and so on) and the Really Simple Syndication (RSS) feeds, especially for the collection of data from blogs and social media. Another recent methodology for building OM&SA corpora, as well as resources for other tasks, is crowdsourcing [Wang *et al.*, 2013; Filatova, 2012].

2.2 Annotation

The annotation step includes a scheme's definition and its application to the collected data. The scheme's design is an effort in the perspective of data classification that leads to theoretical assumptions about the concepts to be annotated. It defines what kind of information must be annotated, the inventory of markers to be used, and the annotation's granularity. In OM&SA, this is especially challenging because we lack an agreed model or theory about these massively complex phenomena. Research in psychology outlines three main approaches to modeling emotions and sentiments: the categorical, the dimensional, and the appraisal-based approach.

*This paper is an extended abstract of the IEEE Intelligent System Journal [Bosco *et al.*, 2013]

The most widespread are the categorical and the dimensional ones, which describe emotions by marking a small set of discrete categories and scoring properties like polarity or valence (positive/negative) and arousal (active/passive) in a continuous range of values [Cowie *et al.*, 2011]. Accordingly, the kinds of knowledge usually annotated are the sentiment's category (hate vs love), polarity (positive vs negative), the source and target toward which the sentiment is directed. Annotations can be based on polarity labels, possibly equipped with intensity ratings, which also helps us classify texts where mixed sentiments are expressed. They can also be based on labels representing different emotions [Roberts *et al.*, 2012]. When complex knowledge is involved, it can be helpful to rely on structured knowledge of affective information, such as categorization models expressed by ontologies, which can work as a shared guideline for the annotators.

Most social data are made up of unstructured texts containing all of the ambiguities found in spoken communications. Thus, annotations at both the document and subdocument levels can provide relevant contributions. At the document level, the annotated units' length varies from posts composed of one or two sentences to much longer documents. Considering whole documents provides a broader knowledge about context, which is a precious element, especially in irony and sarcasm detection. Analysis at the subdocument level, instead, is concerned with distinguishing the portions of text containing sentiment expressions. It presupposes that texts have been tokenized with the parts of speech (PoS) tagged and syntactically analyzed. However, the results are often limited by the text's ungrammaticality. The two annotation levels can offer complementary information. For instance, resolving anaphora and prepositional phrase attachments can be a prerequisite for identifying the target or source of an emotion; detecting emotional adjectives by PoS tagging can improve classifications based on document-level annotation.

Applying the annotation scheme to the data necessarily involves more than one annotator to release reliable and unbiased data within the limits of a task inherently affected by subjectivity. The resulting inter-annotator disagreement is measured [Wiebe *et al.*, 2005; Momtazi, 2012] and sometimes solved. The most commonly applied measures are those inspired by the Cohen's κ coefficient [Artstein and Poesio, 2008]. Best practices to limit and solve the disagreement consist of setting up guidelines shared among the annotators.

2.3 Analysis and Exploitation

Annotated corpora for OM&SA are useful in the training and testing of machine learning statistical tools for the classification of emotions and sentiments. Results are strongly influenced by both the quantity and quality of data. Error detection and quality control techniques have been developed, and often the exploitation itself of the data discloses possible errors. A strategy that can give very useful hints about the reliability of the annotated data is the comparison between the results of automated classification and human annotation.

Labeling schemes are always the outcome of a tension between simplicity and complexity, but instead of investing efforts in a minimal labeling, it is recommended to construct a richer labeling supporting different uses of the annotated material, see Cowie et al. in [Cowie *et al.*, 2011]. Re-usability and portability are indeed important measures for datasets that strive for being suitable to the development of integrated emotion-oriented computing systems. This motivates the efforts devoted to the definition and dissemination of standards for the annotation of data also with respect to OM&SA, see Schröder et al. in [Cowie *et al.*, 2011].

3 The Senti-TUT Project

We present the Senti–TUT project, as a case study for the issues raised in the previous section (http://www.di.unito.it/~tutreeb/sentiTUT.html). The major aims of the project are the development of a resource currently missing for Italian, and the study of a particular linguistic device: irony. This motivated the selection of data domain and source, i.e. politics and Twitter: tweets expressing political opinions contain extensive use of irony. Irony is recognized in literature as a specific phenomenon which can harm sentiment analysis and opinion mining systems [Davidov *et al.*, 2011]. To deal with this issue, we extended a traditional polarity-based framework with a new dimension which explicitly accounts for irony.

3.1 Irony, Sarcasm and the Like

Among the different perspectives and computational approaches for identifying irony, some researchers focus on machine-learning algorithms for automatic recognition, while others focus on corpus generation or on the identification of linguistic and metalinguistic features useful for automatic detection [Filatova, 2012; Davidov *et al.*, 2011; Reyes *et al.*, 2013; Maynard and Greenwood, 2014].

Theoretical accounts suggest different ways of explaining the meaning of irony as the assumption of an opposite or different meaning from what is literally said, that is irony can play the role of *polarity reverser*, a very interesting aspect to be checked in a social media corpus for OM&SA. Other factors to be considered are text context and common ground [Gibbs and Colston, 2007], often preconditions for understanding if a text utterance is ironic. Another issue concerns boundaries among irony and other figurative devices, such as sarcasm, satire or humor. According to literature, boundaries in meaning between different types of irony are fuzzy [Gibbs and Colston, 2007], and this makes more suitable annotations where different types of irony are not distinguished, as the one adopted in Senti–TUT. However, as results in [Reyes *et al.*, 2012] suggest, also in case of figurative languages the choice among coarse or finer-grained annotation could lead to different outcomes in the analysis.

Even if there is no agreement on a formal definition of irony psychological experiments, have delivered evidence that humans can reliably identify ironic text utterances from an early age in life. These findings provide grounds for developing manually annotated corpora for irony detection.

3.2 Data Collection

Senti–TUT includes two corpora, namely TWNews and TWSpino, composed by tweets (shorter than less than 140 characters) with a focus on politics, a domain where irony is

frequently exploited by humans. We collected Italian Twitter messages posted during the weeks that have seen the change of government in Italy, after Mario Monti was nominated to replace Silvio Berlusconi as prime minister (from October 16th, 2011 to February 3rd, 2012). Applying a filtering with keywords and/or hashtags, like "mario monti/#monti", "governo monti/#monti", "professor monti/#monti", etc., and then removing retweets and incomprehensible posts, we defined a corpus of 3,288 Tweets. For what concerns TWSpino, it is composed of 1,159 messages from the Twitter section of Spinoza (http://www.spinoza.it), a very popular Italian blog of posts with sharp satire on politics. We extracted posts published from July 2009 to February 2012 and removed advertising (1.5%). Since there is a collective agreement about the fact that these posts include irony mostly about politics, they represent a natural way to extend the sampling of ironic expressions, also without filtering.

3.3 Annotation

We considered as document the single tweet and we annotated the sentiment towards Monti and the new government exploiting the following tags:

POS (positive)

NEG (negative)

HUM (ironic)

MIXED (POS and NEG both)

NONE (objective, none of the above)

Let us see some examples:

TWNews-24 (tagged as POS)
'Marc Lazar: "Napolitano? L'Europa lo ammira. Mario Monti? Puo' salvare l'Italia"'
(Marc Lazar: "Napolitano? Europe admires him. Mario Monti? He can save Italy")

TWNews-124 (tagged as NEG)
'Monti e' un uomo dei poteri che stanno affondando il nostro paese.'
(Monti is a man of the powers that are sinking our country.)

TWNews-440 (tagged as HUM)
'Siamo sull'orlo del precipizio, ma con me faremo un passo avanti (Mario Monti)'
(We're on the cliff's edge, but with me we will make a great leap forward (Mario Monti))

The annotation, manually performed, begins with a phase where five human annotators (two males and three woman, varying ages) collectively annotated a small set of data (200 tweets), attaining a general agreement on the exploitation of the labels. Then, we annotated all the data producing for each tweet not less than two independent annotations. The agreement calculated at this stage, according to the Cohen's κ score, was satisfactory: $\kappa = 0.65$. In order to extend our dataset, we applied a third independent annotation on the cases where the disagreement has been detected (about 25% of the data). After that, the cases where the disagreement persists, i.e. all annotators selected different tags, have been discarded as too ambiguous to be classified. 3,288 tweets are the final result for TWNews.

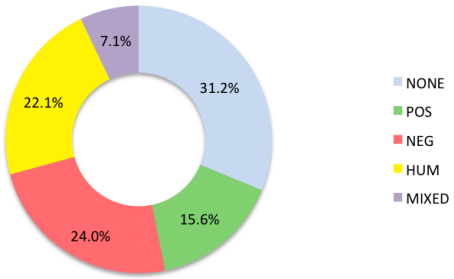

Figure 1: Distribution of the Senti–TUT tags in TWNews

3.4 Corpus Analysis and Exploitation

To get a better sense of how we might use Senti–TUT for future classification tasks, we analyzed the manual annotations. Fig. 1 shows a sample of the distribution of tags referring to the TWNews corpus. Among the features expressed in our corpora, we focus on polarity reversing and emotional expressions.

Polarity Reversing in Ironic Tweets

The first test we tried concerns the hypothesis that ironic expressions play the role of polarity reversers. As we can observe, for instance, in tweet TWNews–440, the explicit meaning of an ironic expression can be the opposite of the real intended one; therefore, irony can undermine the accuracy of a sentiment classifier that isn't irony-aware. To validate such a hypothesis and offer hints about the frequency of this phenomenon, we compared the classification expressed by humans (naturally irony-aware) and that of an automatic (not irony-aware) classifier, such as Blogmeter. We focused on 723 ironic tweets from TWNews, henceforth denoted as TWNews-Hum. The task for a couple of human annotators (H) and Blogmeter classifier (BC) was to apply the tags Pos, Neg, None, or Mixed to TWNews-Hum. The BC implements a pipeline of NLP processes within the Apache UIMA framework. It doesn't use machine-learning techniques, but similar to Diana Maynard and her colleagues' work [Maynard *et al.*, 2012] it adopts a rule-based approach to sentiment analysis, which relies primarily on sentiment lexicons (almost 8,450 words and expressions) and sentiment grammar expressed by compositional rules. Assuming that polarity reversing is a phenomenon that we can observe when an expression is clearly identified as positive, and the opposite makes it negative (or vice versa), let's focus on tweets classified by BC as positive (143) or negative (208). Excluding the 30 tweets where human annotators disagreed, we obtained a set of 321 posts. On those data, we detected a variation between BC and H classification, taken as an indicator of polarity reversal. We observed this variation in most of the selected tweets (68.5%). In some cases, there was a full reversal (varying from a polarity to its opposite), which is almost always from positive (BC) to negative polarity (H). In other cases there was an attenuation of the polarity, mainly from negative (BC) to neutral (H). We summarize the results in Table 1, where Btag $\rightarrow$ Htag denotes the direction of the polarity variation from the

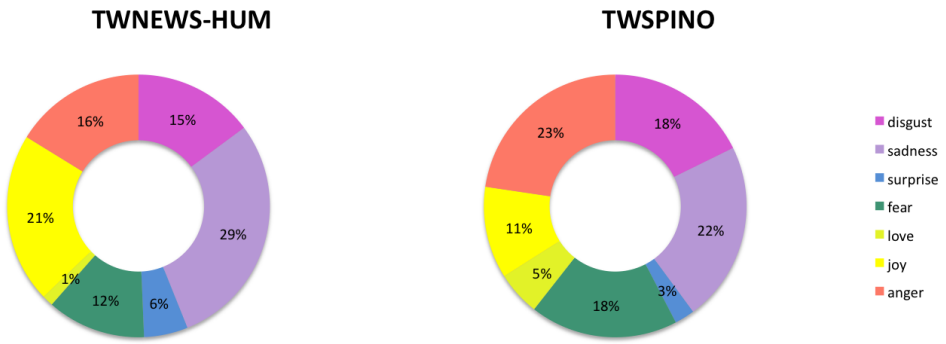

Figure 2: Emotion distribution in the ironic emotional tweets of TWNews (left) and TWSpino (right).

Blogmeter to the human classification. Although the dataset limited size and its particular domain and text genre makes our results preliminary, the theoretical accounts seem to be confirmed.

full reversal 37.3%:	33.6%	POS	→	NEG
	3.7%	NEG	→	POS
attenuation 62.7%:	40.5%	NEG	→	NONE
	22.2%	POS	→	NONE

Table 1: Polarity variations in ironic tweets showing the reversal phenomena

Emotions in Ironic Tweets

Another interesting challenge is to apply to our dataset emotion detection techniques (beyond positive or negative valence), like in [Reyes *et al.*, 2012], and to reflect on relationships between irony and emotions. We have applied rule-based automatic classification techniques provided by Blogmeter to annotate our ironic tweets (723 of TWNews-Hum and 1,159 of TWSpino) according to six ontology categories (based on Ekman's six basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *love*) [Roberts *et al.*, 2012]. These emotions are expressed only in the 20% of our dataset and differently distributed in the corpora, as shown in Fig. 2. In TWNews-Hum, the most common emotions were *sadness* (29.1%) and *joy* (20.9%), followed by *anger*, *disgust*, and *fear*. *Surprise* was rare, and *love* was almost nonexistent. TWSpino contains instead more negative emotions: *anger* (22.7%) and *sadness* (22.2%), followed by *fear* and *disgust*. Positive emotions, such as *joy* and *love*, have fewer occurrences, and *surprise* is rare. The first observation that emerges from these results concerns the emotions detected and typology of irony. For instance, it's interesting that in TWNews-Hum, the most common emotions are joy and sadness – human emotions conceptualized in terms of polar opposites. Accordingly, we observe a wider variety of typologies of irony in those tweets, which range from sarcastic posts aimed at wounding their target to facetious tweets expressing a kind of "genteel irony". By contrast, in TWSpino, the detected emotions have mostly a negative connotation, and the typologies of irony expressed are more homogeneous and are mainly restricted to sarcasm and political satire. This could be related to the fact that Spinoza's posts are selected and revised by an editorial staff, which explicitly characterize the blog as satiric. In contrast, TWNews collects tweets spontaneously posted by Italian Twitter users on Monti's government; it then presents multiple voices of a virtual political chat space, where irony is used not only to work off the anger, but also to ease the strain.

4 Lessons Learned and Future Challenges

Beyond developing a missing resource for Italian – extended in [Bosco *et al.*, 2014] and exploited in the Evalita 2014 Sentipolc shared task on sentiment analysis on Italian tweets [Basile *et al.*, 2014] – the primary purpose of the Senti–TUT Twitter corpus was to study irony. Interestingly, we found that irony is often used in conjunction with a seemingly positive statement to reflect a negative one, but rarely is it the other way around. This is in accordance with theoretical accounts, which note that expressing a positive attitude in a negative mode is rare and harder for humans to process, as compared to expressing a negative attitude in a positive mode [Gibbs and Colston, 2007]. Other features we detected about irony are incongruity and contextual imbalance, the use of adult slang, echoic irony, language jokes (which often exploit ambiguities involving the politicians' proper nouns), and references to television series. Our analysis shows also that the Senti–TUT corpus can be representative for a wide range of ironic phenomena, from bitter sarcasm to genteel irony. Therefore, an interesting direction to investigate is to define a finer-grained annotation scheme for irony, where different ways of expressing irony are distinguished. However, this requires reflection on the relationships between irony and sarcasm; on the differences between irony, parody, and satire [Gibbs and Colston, 2007]; and on the representative textual features that distinguish these phenomena.

For what concerns emotions, we proposed a measure that relies on Blogmeter's techniques applied to the ironic tweets of Senti–TUT. An interesting step forward would be to refer to richer semantic models [Cambria and Hussain, 2012], to enable reasoning about semantic relations among emotions.

References

[Artstein and Poesio, 2008] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December 2008.

[Basile *et al.*, 2014] Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo 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)*, pages 50–57, Pisa, Italy, 2014. Pisa University Press.

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

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

[Cambria and Hussain, 2012] Erik Cambria and Amir Hussain. *Sentic Computing: Techniques, Tools, and Applications*. Springer Briefs in Cognitive Computation Series. Springer-Verlag GmbH, 2012.

[Cowie *et al.*, 2011] Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors. *Emotion-Oriented Systems*. Springer Berlin Heidelberg, 2011.

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

[Filatova, 2012] Elena Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the LREC'12*, pages 392–398, Istanbul, Turkey, 2012.

[Ghosh *et al.*, 2015] 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.

[Gibbs and Colston, 2007] Raymond W Gibbs and Herbert L. Colston, editors. *Irony in Language and Thought*. Routledge (Taylor and Francis), New York, 2007.

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

[Johansson and Moschitti, 2013] Richard Johansson and Alessandro Moschitti. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3), 2013.

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

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

[Maynard *et al.*, 2012] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of tNLP can u tag #usergeneratedcontent?! Workshop at LREC'12*, pages 15–22, Istanbul, Turkey, 2012.

[Momtazi, 2012] Saeedeh Momtazi. Fine-grained German sentiment analysis on social media. In *Proceedings of the LREC'12*, pages 1215–1220, Istanbul, Turkey, 2012.

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

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

[Roberts *et al.*, 2012] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. Empatweet: Annotating and detecting emotions on Twitter. In *Proceedings of the LREC'12*, pages 3806–3813, Istanbul, Turkey, 2012.

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

[Wang *et al.*, 2013] Aobo Wang, CongDuyVu Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31, 2013.

[Wiebe *et al.*, 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.