Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT

Developing corpora for sentiment analysis and opinion mining: the case of irony and Senti-TUT

Cristina Bosco, *Member, IEEE*, Viviana Patti, *Member, IEEE* and Andrea Bolioli

***Abstract*—In recent years several efforts were devoted to automatically mining opinions and sentiments from natural language in social media messages, news and commercial product reviews. Since this task involves a deep understanding of the explicit and implicit information conveyed by the language, most of the approaches refer to annotated corpora. However, the development of this kind of resource raises several new challenges due both to the specificity of the data from such domains and text genres, and to the knowledge to be annotated.**

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, an ongoing project for Italian aimed at investigating sentiment and irony about politics in social media. We introduce and analyze the Senti-TUT corpus, a collection of texts from Twitter annotated morpho-syntactically and 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.

***Index Terms*— Corpora for sentiment analysis, Social media, Irony, Italian.**

1 Introduction

MINING opinions and sentiments from natural language is an extremely difficult task, which involves a deep understanding of most of the explicit and implicit information conveyed by language structures, from single words to entire document. The growth of the Social Web and the availability of a dynamic corpus of user-generated contents, such as product reviews or statistical polling data, makes necessary to deal with the cognitive and affective information conveyed by expressive texts reflecting spontaneous user responses.

For this task, rudimentary approaches, mainly based on single words or flat structures, are followed by social media search tools, e.g. Social Mention, Twitter Sentiment, Twendz, Twitrratr[1], where users enter a term and get back the negative and positive posts that contain it. However, recent approaches are oriented to capture information going beyond the word level to outperform social media search tools in terms of portability and performance, by relying on a more structured [1], multi-faceted and semantic notion of text [14]. 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.

The development of annotated corpora for Opinion Mining and Sentiment Analysis (OM&SA), on the one hand, benefits from the know-how gained during the last twenty years in corpus-based NLP, where linguistic databases are crucial. On the other hand, since OM&SA involves particular linguistic and non linguistic knowledge, new languages, text styles and domains, several new challenges must be faced, and new concept-level approaches, which foresee the use of semantic and affective resources for annotation must be explored.

In this paper we discuss the problems underlying the development of corpora of written text for OM&SA. We briefly survey the research area and we refer to the specific case of irony, a linguistic device which is especially challenging for NLP and very common in texts from social media. As a case study, we present the Senti-TUT Twitter corpus, designed to study irony for Italian, a language currently less-resourced for OM&SA.

The next section describes the issues related to collecting, annotating and analyzing corpora for OM&SA. Section 3 presents the theoretical accounts and the applicative challenges related to irony raised during the Senti-TUT development. A discussion of lessons learned and challenges ends the paper.

2 Developing corpora for Opinion and Sentiment Analysis

The development of a corpus consists in three main steps: collection, annotation and analysis (Fig. 1). 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: what, from where, how?

The issues related to collection mainly refer to the selection

This work has been partially funded by the PARLI Project (MIUR PRIN 2008). We are grateful to our annotators and to CELI Torino for providing the facilities offered by the Blogmeter platform.

C. Bosco and V. Patti are with the University of Torino, Dipartimento di Informatica, Corso Svizzera, 185, 10149, Torino, Italy, phone: +39-0116706711, fax +39-011751603. E-mail: {bosco, patti}@di.unito.it

A. Bolioli is with CELI srl, Blogmeter, Torino, Via San Quintino 31, 10121, Torino, Italy, phone: +39-0115627115, fax +39-0115064086. E-mail: abolioli@celi.it

[1] http://socialmention.com; http://www.sentiment140.com; \\http://twendz.waggeneredstrom.com; http://twitrratr.com.

of data and composition of the corpus (i.e. what), the choice of the data source (i.e. from where), but also to the collection methodologies applied (i.e. how).

Figure 1: Steps in the development of a corpus.

It is the task for which the resource is developed that usually drives the decisions about *what* data to collect and *from where*. Most of the corpora designed for OM&SA are collected from web services which provide comments on commercial products, like reviews posted on Amazon [2], [3]. Others are extracted from 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 [4], German [5] or UK elections [6]. Less frequently also other kinds of text are collected, see e.g. [7], [8] for corpora and tasks about OM&SA in emails and suicide notes respectively.

Often the OM&SA corpora are indeed the result of sampling and filtering oriented to a particular target or source, in contrast to resources for other tasks, like parsing, where the focus is mainly on building larger and balanced collections of texts as spontaneously occurring (unrestricted). Data selection and filtering are usually based on keywords like named entities or metadata released by the authors of posts in micro-blogs, like the hashtags exploited for irony and sarcasm in [9], [2]. Moreover, metadata on time and geolocations, users' age, gender, background and social environment, or communicative goals, enable the detection of sentiment variation or trends.

Also text genre has to be taken into account during collection, since each genre is featured by a different manner of expressing opinions and sentiments [10], and by the exploitation of different linguistic structures and devices. For instance, texts from blogs are highly subjective while those from newspapers want to give to the reader an impression of objectivity. Limits imposed by social media on the message length influence instead the morphological and syntactic structure of posts, while the frequency of figurative devices can significantly vary in different domains such as Twitter and product reviews.

For what concerns collection methodologies (i.e. how to collect data) the most used is web crawling and scraping, or calling the web Application Programming Interfaces (APIs) exposed by the service (Google Reader's API, Twitter's API, etc.) and the Really Simple Syndication (RSS) feeds, especially for the collection of data from blogs and social media. Crowdsourcing also has been recently applied for building OM&SA corpora as well as resources for other tasks [11], [3].

2.2 Annotation: what kind of information and annotated in which way?

The annotation step includes the definition of a scheme and its application to the collected data, but also the assessment of the material by the evaluation of inter-annotator agreement.

The design of the scheme is an effort in the perspective of data classification binding us to make theoretical assumptions about the concepts to be annotated. It defines what kind of information has to be annotated, the inventory of markers to be used, and the granularity of the annotation. In the case of OM&SA, this is especially challenging because an agreed model or theory about these massively complex phenomena is missing. Research in psychology outlines three main approaches to modeling emotions and sentiments: the categorical, the dimensional, and the appraisal-based approach. The most widespread representations are the categorical and the dimensional ones, which describe, respectively, emotions by marking a small set of discrete categories, and by scoring properties like polarity or valence (positive/negative) and arousal (active/passive) in a continuous range of values, see Schröder in [12]. Accordingly, the kinds of knowledge usually annotated are the category of the sentiment (hate vs love), the polarity of the sentiment (positive vs negative), the source and the target toward which the sentiment is directed, the intensity. Annotations can be based on simple broad polarity labels, possibly equipped with intensity ratings allowing to also deal with the classification of texts where mixed sentiments are expressed [13], or based on labels representing different emotions, e.g. the Ekman's basic emotions [10]. When complex knowledge is involved, as in case of emotional categories, it can be very helpful to rely on structured knowledge of affective information, like affective categorization models expressed by ontologies, better still if psychologically motivated, such as the *Hourglass of Emotions* in [14] [22], inspired by the Plutchik's studies on human emotions, where 24 emotional categories are organized into four affective dimensions, and can blend to form compound emotions. An ontology which encodes knowledge about emotions can work as a guideline to be shared by the annotators in order to develop a common understanding about emotions and their relationships [10]. Moreover, it can support comparison and aggregation among results of the emotional analysis, as in case of the Hourglass model [22].

For what concerns the annotation granularity, since opinions and sentiments are often expressed implicitly through context and domain dependent concepts, it is important to rely on approaches going beyond the syntactic level, as the hoped for in the sentic computing approach to OM&SA [14]. Most of data are unstructured text containing all of the ambiguities found in spoken communications. For these reasons annotation at both *document* and *subdocument* levels can give relevant contributes. At the document level, the length of the

annotated units varies from that of posts composed of one or two sentences, to that of wider documents. Considering whole documents makes available broader knowledge about context, a precious element, especially in irony and sarcasm detection [3], [5]. Different annotations for context-dependent and context-independent opinions can also be useful [5]. Analysis at subdocument level is instead concerned with distinguishing the portions of a text (words, phrases or more complex structures) containing sentiment expressions. It presupposes that texts were tokenized, Part of Speech (PoS) tagged and often also syntactically analyzed. But the result of such kind of analyses is often limited by text ungrammaticality.

Online social data, remain hardly accessible to classical NLP techniques. They are specifically meant for human consumption and their automatic analysis involves a deep understanding of natural language text by machines, from which we are still very far. To support NLP, a promising approach is to apply new paradigms of semantic annotation, relying on resources such as SenticNet (http://sentic.net), an affective common-sense knowledge resource which enables the inference of both the conceptual and emotional information associated with natural language opinions, and, hence, a easier extraction of the concept-level sentiment conveyed by word-level natural language texts [21].

Moreover, features that vary from one language to another also dramatically decrease the portability of tools and the suitability of the annotations, like word order and morphological richness, see [15], [13] for German. On this perspective, let us notice that most of the available resources are in English with a few exceptions, such as the multilingual dataset automatically annotated for subjectivity in [16]. The two annotation levels can offer complementary information. For instance, resolution of anaphora and prepositional phrase attachments can be a prerequisite for the identification of target or source of an emotion. While the detection of emotional adjectives by PoS tagging, can improve classifications based on document level annotation.

For what concerns the application of the annotation scheme to the data (i.e. annotated in which way), it is usually supported by semi-automatic tools and necessarily involves more than one annotator, in order to release reliable and unbiased data, within the limits of a task inherently affected by subjectivity. The proper number of annotators depends also on the difficulty of the task [5]. The resulting inter-annotator disagreement is measured [15], [13], and possibly solved. The most commonly applied measures are those inspired by the Cohen's κ coefficient [17]. Best practices to limit and solve the disagreement consist in setting up guidelines shared among the annotators, or annotating and discussing collectively portions of data [10].

2.3 Analysis and exploitation of a corpus

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 [10].

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 [12]. 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 for several NLP tasks[2], in the past, and, more recently, discussed also for OM&SA, see Schröder et al. in [12].

3 THE SENTI-TUT PROJECT

We present the Senti-TUT project [18], 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 [2]. 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 focus on machine learning algorithms for automatic recognition, others put the accent on corpus generation or on the identification of linguistic and meta-linguistic features useful for automatic detection [3], [2], [9], [19], Strapparava et al. in [12]. In the following, we will briefly recall theoretical issues and key aspects to be considered in developing a corpus for irony detection.

Relevant contributions on irony can be found in a wide range of disciplines ranging from linguistic to psychology [20]. The rhetorical tradition treated irony as the figure of speech in which the meaning is the opposite of the literal meaning. Modern Gricean pragmatic theory has not departed radically from this view. Another interesting account within the relevance theory by Sperber and Wilson (Chapter 3, in [20]) suggests that irony is a variety of echoic use of language, where the communicator dissociates himself from the echoed opinion.

Theoretical accounts suggest different ways of explaining the meaning of irony as the assumption of an opposite or different

[2] See TEI (Text Encoding Initiative, http://www.tei-c.org/index.html), EAGLES (Expert Advisory Group on Language Engineering Standards) and CES (Corpus Encoding Standard, http://www.cs.vassar.edu/CES/).

meaning from what is literally said. Under this perspective, it is clear that irony can play the role of *polarity reverser* with respect to the words used in the text unit. This is one of the most interesting aspects to check in a social media corpus for sentiment analysis, as we will see in Sec. 3.4.1.

Other factors to be considered are text context and common ground [20], which according to psychological models of language use, are often preconditions for understanding if a text utterance is ironic. Consider, for instance, Facebook comment threads. Here the dialogical context can be essential to detect irony, since often threads implicitly refer to a common ground restricted to a group of friends, thus making the irony recognition harder for the others. Instead, in case of Twitter, posts do not follow a conversation thread showing, on this respect, a *contextless nature* [2]. Furthermore, even if identifying irony in tweets often require world knowledge, post authors usually refer to a *broader* common ground (i.e. knowledge about news or VIPs), by expressing irony differently than in conversations among friends.

Another issue concerns boundaries among irony and other figurative devices, such us sarcasm, satire or humor. According to literature, boundaries in meaning between different types of irony are fuzzy [20]. This could be an argument in favor of annotation approaches where different types of irony are not distinguished, as the one adopted in Senti-TUT. However, as results in [9] suggest, also in case of figurative languages the choice among coarse or finer-grained annotation could lead to different outcomes in the analysis.

Psychological studies underline also the subjectivity of irony perception, regardless of the different world knowledge or limitedness of a shared context: different people could consider a given post ironic or sarcastic “to some degree”. Annotation schemes can deal with this aspect, by allowing to assign intensity ratings to ironic annotations [2], but also by implementing careful disagreement evaluation. Moreover, even if, as in case of most figurative devices, there is no agreement on a formal definition of irony, psychological experiments brought some evidence that humans can reliably identify ironic text utterances, also in early ages of their life. These findings provide a grounding to the development of manually annotated corpora for irony detection.

3.2 Data collection

Senti-TUT includes two Twitter corpora namely TWNEWS and TWSPINO, with a focus on politics, a domain where irony is frequently exploited by humans. Tweets are composed by less than 140 characters distributed in one or more short sentences.

TWNEWS corpus has been extracted by applying filters based on time and metadata, aimed at selecting posts where a variety of opinions about politics is represented. For collection and filtering we relied on Blogmeter social media monitoring platform (http://www.blogmeter.eu), which exploits Twitter API to extract the tweets. 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). We used the list of keywords and/or hashtags ”mario monti/#monti”, ”governo monti/#monti”, ”professor monti/#monti” (lowercase or capitalized) for selecting about 19,000 tweets on Monti government. Retweets were, then, removed as not relevant with respect to our task of irony and sentiment analysis, and this resulted in a collection of about 11,000 tweets. 70% of those tweets were further discarded by annotators as ungrammatical, not well-written, duplicated (but not marked as RT) or incomprehensible without their context: even if tweets do not follow a conversation thread, a notion of context is spreading in the data by means of repetitions and reprises of previous posts. The final result are the 3,288 posts of TWNEWS, annotated as reported in Section 3.3.

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

In order to make the collected data adequate for studying irony, we designed and applied them an annotation both at document and subdocument level. The former is oriented to the description of tweet polarity, while the latter is based on an existing schema representing the morphology and syntax of the reference language.

The annotation at the document level is suitable for high-level tasks, such as classifying the polarity of a given text, in line with the general idea that very little can be gained by complex linguistic processing for tasks such as text categorization and search. The annotation at subdocument level benefits from the experience gained in corpus-based NLP tasks such as e.g. PoS tagging and parsing, and it is instead in line with more recent works [1] where the task is not only to find a piece of opinionated text, but also to extract a structured representation of the opinion (e.g. determining the holder and the target), inspired by experience in information extraction, semantic role labeling and structured machine learning.

3.3.1 Morphological and syntactic annotation

The morphological and syntactic annotation of Senti-TUT is done according to the format developed and applied in the Turin University Treebank (TUT), http://www.di.unito.it/~tutreeb. This is a freely available resource developed by the NLP group at University of Turin, by applying the Turin University Linguistic Environment (TULE, http://www.tule.di.unito.it/), whose pipeline includes tokenization, morphological and syntactic analysis. It has been successfully exploited as testbed for parsing in the evaluation campaigns for Italian parsing (http://www.evalita.it/).

Below a post from TWSPINO[3] represented according to TUT format, which includes a very detailed morphological tag set, essential feature for describing a language with a rich inflection, and a large inventory of grammatical relations labeling the edges of the dependency trees, in order to describe the argument structure of the sentence:

```
1 La (IL ART DEF F SING) [7;VERB-SUBJ]
2 spazzatura (SPAZZATURA NOUN COMMON F SING) [1;DET+DEF-ARG]
3 di (DI PREP MONO) [2;PREP-RMOD]
4 Napoli (NAPOLI NOUN PROPER F SING CITY) [3;PREP-ARG]
5 si (SI PRON REFL-IMPERS ALLVAL ALLVAL 3 CLITIC) [7;VERB-OBJ]
6 sta (STARE VERBAUX IND PRES 3 SING) [7;AUX]
7 decomponendo (DECOMPORRE VERBMAIN GER PRES) [0;TOP-VERB]
8 . (. PUNCT) [7;END]
1 Concorrera` (CONCORRERE VERBMAIN IND FUT 3 SING) [0;TOP-VERB]
1.10 t [] (T PRON PERS ALLVAL ALLVAL ALLVAL) [1;VERB-SUBJ]
2 al (A PREP MONO) [1;VERB-INDCOMPL]
2.1 al (IL ART DEF M SING) [2;PREP-ARG]
3 Nobel (NOBEL NOUN PROPER) [2.1;DET+DEF-ARG]
4 per (PER PREP MONO) [3;PREP-RMOD]
5 la (IL ART DEF F SING) [4;PREP-ARG]
6 chimica (CHIMICA NOUN COMMON F SING) [5;DET+DEF-ARG]
7 . (. PUNCT) [1;END]
```

3.3.2 Tweet-level sentiment and irony annotation

We considered as document the single tweet and we annotated therefore one of the following sentiment tags for each tweet, by evaluating basically the sentiment towards Monti and the new government:

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? Può salvare l'Italia"'
(Marc Lazar: "Napolitano? Europe admires him. Mario Monti? He can save Italy")

TWNEWS-124 (tagged as NEG)
'Monti è 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))

TWNEWS-3198 (tagged as MIXED)
'Brindo alle dimissioni di Berlusconi ma sul governo Monti non mi faccio illusioni'
(I drink a toast to the Berlusconi's resignation, but I have no illusion about the Monti's government)

TWNEWS-123 (tagged as NONE)
'Mario Monti premier? Tutte le indiscrezioni.'
(Mario Monti premier? All the gossips.)

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 (around 2%, an interesting sample to analyze for future work). 3,288 tweets are the final result for TWNEWS.

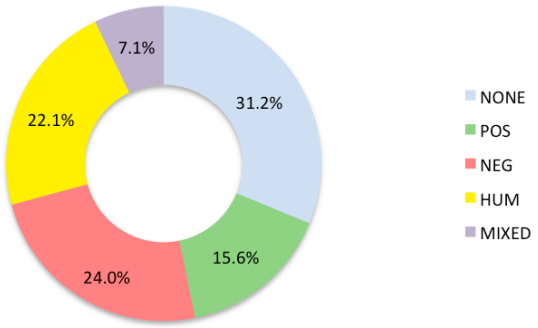


Figure 2: Distribution of the Senti–TUT tags in TWNEWS.

3.4 Corpus analysis and exploitation

In order to have hints about the future use of Senti-TUT for classification tasks, we performed an analysis of the manual annotation. For what concerns the distribution of tags, a sample is shown in Fig. 2 referring to the TWNEWS corpus. Among the features expressed in our corpora, we have chosen to focus on polarity reversing and emotional expressions.

3.4.1 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, Sec. 3.3, 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 not irony-aware. In order to validate such hypothesis and have some hints about the frequency of this phenomenon, we developed a comparison between the classification expressed by humans, naturally irony-aware, and that of an automatic not irony-aware classifier, i.e. Blogmeter. We focussed on the 723 ironic tweets of TWNEWS, henceforth denoted as TWNEWS-HUM. The task for both a couple of human annotators (H) and Blogmeter classifier (BC) consisted in applying the tags POS, NEG, NONE or MIXED to TWNEWS-HUM. BC implements a pipeline of NLP processes within the Apache UIMA

[3] TWSPINO-216: '*La spazzatura di Napoli si sta decomponendo. Concorrerà al Nobel per la chimica.*' (The garbage of Naples is becoming rotten. It will apply for the Nobel prize in Chemistry).

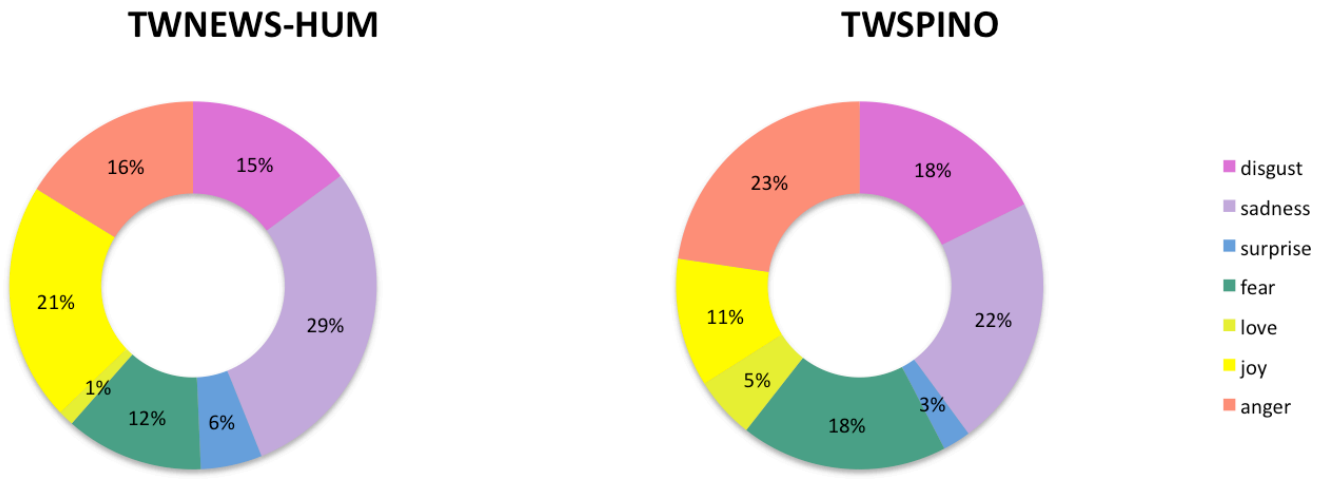


Figure 3: Emotion distribution in the ironic emotional tweets of TWNEWS (left) and TWSPINO (right).

framework. It does not uses machine learning techniques, but, similarly to [19], adopts a rule-based approach to sentiment analysis, which relies primarily on sentiment lexicons (almost 8,450 words and expressions) and sentiment grammars expressed by compositional rules. Assuming that the polarity reversing is a phenomenon which can be observed when an expression is clearly identified as positive and the reversing makes it negative (or vice versa), let us focus on tweets classified by BC as positive (143) or negative (208). Excluding the 30 tweets where the human annotators disagree, we obtained a set of 321 tweets. On those data we detected the variation between BC and H classification, taken as an indicator of polarity reversing. We observed a variation in most of the selected tweets (68.5%): in some cases a full reversing (variation from a polarity to its opposite), which is almost always from positive (BC) to negative polarity (H), in the remaining cases an attenuation of the polarity, mainly from negative (BC) to neutral (H). Results are summarized hereafter, where Btag → Htag denotes a shift from the Blogmeter to the human classification:

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

Although the limited size of the dataset and its particular domain and text genre make our results preliminary, the theoretical accounts seem to be confirmed.

3.4.2 Emotions in ironic tweets

Another interesting challenge is to apply to our dataset emotion detection techniques (beyond positive or negative valence), like in [9], and to reflect on relationships between irony and emotions. We have applied rule-based automatic classification techniques provided by Blogmeter in order to annotate our ironic tweets (723 of TWNEWS-HUM and 1,159 of TWSPINO) according to the six categories of the ontology in [10]. It includes the Ekman's six basic emotions, ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE, plus LOVE. These emotions are expressed only in the 20% of our dataset and differently distributed in the corpora, as shown in Fig. 3. In TWNEWS-HUM the most common emotions were SADNESS (29.1%) and JOY (20.9%), followed by ANGER, DISGUST and FEAR. SURPRISE was rare. LOVE almost inexistent. TWSPINO contains instead more negative emotions: ANGER (22.7%), SADNESS (22.2%), followed by FEAR and DISGUST. Positive emotions, like JOY and LOVE, have fewer occurrences, and SURPRISE is rare.

The first observation that emerges from the above results concerns emotion detected and typology of irony. In TWNEWS the most common emotions are JOY and SADNESS, that well-known studies on human emotions conceptualize 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", that does not involve necessarily a negative attitude, but can be playful and aimed at producing a comic or parodic effect, or at strengthening ties with the virtual interlocutors. In contrast, in TWSPINO detected emotions have mostly a negative connotation and the typologies of irony expressed are more homogeneous and mainly restricted to sarcasm and political satire. This can be related to the fact that Spinoza's posts are selected and revised by an editorial staff. Moreover, Spinoza's editors explicitly characterize the blog as satiric, then the post selection is oriented to publish the sharpest wits, often with the goal to "hit and sunk" the target. In contrast, TWNEWS collects tweets spontaneously posted by Italian Twitter users on Monti's government, then it expresses multiple voices of a virtual political agora, where irony is used not only to work off the anger, but also to ease the strain.

4 LESSONS LEARNED AND FUTURE CHALLENGES

Beyond the development of a missing resource for Italian,

the major aim of the creation of the Senti-TUT Twitter corpus is studying irony, rather than Twitter as a whole. This motivated the filtering by hashtags and keywords towards politics we applied for collecting the corpus. Nevertheless, only a small portion of those data resulted to be suitable for annotation, due to the specificity of Twitter-language features, that make often posts unintelligible for humans and machines: very high frequency of ungrammaticality, repetitions, SPAM and context dependency. This is a first lesson learned and an interesting issue for future work.

We applied to the dataset both sentiment and irony annotation at the tweet-level, and morpho-syntactic annotation at subdocument level, then, we studied the polarity reversing phenomenon and the distribution of emotions in the ironic tweets. We have found that irony is very often used in conjunction with a seemingly positive statement, to reflect a negative one, but rarely the other way around. This is in accordance with theoretical accounts, where it is noticed that expressing positive attitudes in a negative mode is rare and harder to process for humans, than expressions of negative attitude in a positive mode (see Attardo on *asteism* in [20], Ch. 6). 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, that confirm the importance of shared knowledge in irony detection. A formal account and a measure of these phenomena is a matter of future work. It will require a finer granularity in text analysis, in line with [1], and the use of common-sense knowledge bases to extract the latent semantics from text, as hoped for in concept-level approaches to OM&SA [14], especially to measure incongruity and contextual imbalance in terms of semantic relatedness of concepts expressed in ironic texts [9]. For this purpose we are devoting our efforts to the application of a semantic annotation based on the major semantic resources currently available for Italian (BabelNet,WordNet).

Our analysis shows also that the Senti-TUT corpus can be representative for a quite wide range of ironic phenomena, from bitter sarcasm to forms of 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 would require to reflect on the relationships between irony and sarcasm, on the differences between irony, parody and satire [20], and on textual features representative for the phenomena to distinguish: challenging but not trivial issues.

For what concerns the emotional ground, we proposed a measure, relying on emotion annotation techniques provided by Blogmeter applied to the ironic tweets of the Senti-TUT dataset. Blogmeter adopts a rule-based approach to sentiment analysis, and have been recently tested in an experiment of automatic emotion annotation on a corpus of 31 million Italian tweets, with the set of emotions used in [10]. An interesting step forward could be to refer to a richer semantic model, e.g. the Hourglass of emotions [22], in order to enable reasoning about semantic relations among emotions, i.e. similarities, opposites, intensities. Even if we can currently report on a limited exploitation of our data in automatic classification tasks, see experiments in [18], the lessons learned from the data analysis give useful hints about future directions.

REFERENCES

[1] R. Johansson and A. Moschitti, "Relational features in fine-grained opinion analysis," Computational Linguistics, 2012, in press.

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

[3] E. Filatova, "Irony and sarcasm: Corpus generation and analysis using crowdsourcing," in Proceedings of the LREC'12, Istanbul, Turkey, 2012, pp. 392–398.

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

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

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

[7] S. M. Mohammad and T. Yang, "Tracking sentiment in mail: How genders differ on emotional axes," in Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), Portland, Oregon (USA), 2011, pp. 70–79.

[8] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew, "Sentiment analysis of suicide notes: A shared task," Biomedical Informatics Insights, vol. 5, no. Suppl.1, pp. 3–16, 2012.

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

[10] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, "Empatweet: Annotating and detecting emotions on Twitter," in Proceedings of the LREC'12, Istanbul, Turkey, 2012, pp. 3806–3813.

[11] A. Wang, C. Hoang, and M. Y. Kan, "Perspectives on crowd-sourcing annotations for natural language processing," Language Resources and Evaluation, vol. in press, pp. 1–23, 2012.

[12] R. Cowie, C. Pelachaud, and P. Petta, Eds., Emotion-Oriented Systems. Springer Berlin Heidelberg, 2011.

[13] S. Momtazi, "Fine-grained German sentiment analysis on social media," in Proceedings of the LREC'12, Istanbul, Turkey, 2012, pp. 1215–1220.

[14] E. Cambria and A. Hussain, Sentic Computing: Techniques, Tools, and Applications. Dordrecht, Netherlands: Springer, 2012.

[15] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," Language Resources and Evaluation, vol. 39, no. 2-3, pp. 165–210, 2005.

[16] C.Banea, R.Mihalcea, and J. Wiebe,"Multilingual subjectivity: are more languages better?" in Proceedings of the COLING'10, Beijing, China, 2010, pp. 28–36.

[17] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," Computational Linguistics, vol. 34, no. 4, pp. 555–596, December 2008.

[18] A. Gianti, C. Bosco, V. Patti, A. Bolioli, and L. Di Caro, "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, Istanbul, Turkey, 2012, pp. 1–7.

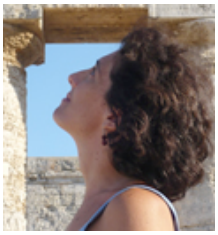
[19] D. Maynard, K. Bontcheva, and D. Rout, "Challenges in developing opinion mining tools for social media," in Proceedings of tNLP can u tag #usergeneratedcontent?! Workshop at LREC'12, Istanbul, Turkey, 2012, pp. 15–22.

[20] R.W. Gibbs and H.L. Colston, Eds., Irony in Language and Thought. New York: Routledge (Taylor and Francis), 2007.

[21] E. Cambria, C. Havasi and A. Hussain, "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis", in Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, Marco Island Florida, 2012, pp. 202-207.

[22] E. Cambria, A. Livingstone, and A. Hussain. The Hourglass of Emotions. In: LNCS, vol. 7403, pp. 144-157, Springer, 2012.

Cristina Bosco PhD in Computer Science, is an assistant professor at the Department of Computer Science of the University of Torino (Italy). Her interests include: dependency and constituency parsing, linguistic resources with morphological and syntactic annotation, evaluation, sentiment analysis (see her publications at http://www.di.unito.it/~bosco).
She is responsible for the TUT project (http://www.di.unito.it/~tutreeb), and co-organizes the parsing task in the evaluation campaigns for Italian NLP: http://www.evalita.it/. She is member of IEEE, ACL and AI*IA. bosco@di.unito.it



Viviana Patti PhD in Computer Science, is an assistant professor at the Department of Computer Science of the University of Torino (Italy). She is author of more than 50 scientific publications: www.di.unito.it/~patti.
Her research interests include: KR in MAS, Social Semantic Web, Ontology-driven Sentiment Analysis, Service-oriented Computing. Member of REWERSE (EU FP6 NoE). Vice-president at AC Arsmeteo developing the Arsmeteo art portal: http://www.arsmeteo.it. She is member of IEEE, AI*IA, and GULP. patti@di.unito.it

Andrea Bolioli is a computational linguist. Co-founder of the company CELI srl (Torino, Italy), which develops software solutions using NLP technologies: http://www.celi.it/en/. He works on BlogMeter (CELI and Me-Source), an Italian social media monitoring service based on a proprietary listening platform, which delivers accurate classification and sentiment analysis. Member of the management team of Cross Library Services. abolioli@celi.it