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

Cristina Bosco, *University of Torino*

Viviana Patti, *University of Torino*

Andrea Bolioli, *CELI srl*

Senti-TUT—an ongoing Italian project that investigates sentiment and irony in online political discussions—illustrates how to develop corpora for mining and analyzing opinion and sentiment in social media.

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. The growth of the social Web and the availability of a dynamic corpus of user-generated content—such as product reviews and statistical polling data—makes it 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—such as Social Mention (http://socialmention.com), TwitterSentiment (www.sentiment140.com), Twendz (http://twendz.waggeneredstrom.com), and Twitrratr (http://twitrratr.com)—which let users enter a term to locate the negative and positive posts that contain it. However, recent approaches are designed 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] multifaceted, and semantic notion of text.[2] Among them, several are based on statistical and machine-learning natural-language processing (NLP) and assume 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) benefits from two decades of advances in corpus-based NLP, where linguistic databases are crucial. However, OM&SA faces several new challenges, because it involves particular linguistic and nonlinguistic knowledge, new languages, text styles, and domains. Additionally, we must explore new concept-level approaches, which foresee the use of semantic and affective resources for annotation.

In this article, we discuss the problems underlying the development of written-text

1541-1672/13/$31.00 © 2013 IEEE
Published by the IEEE Computer Society

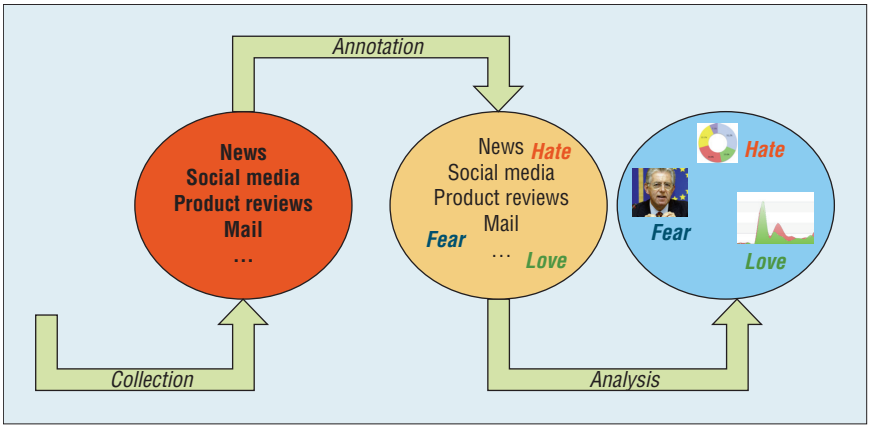

Figure 1. Steps in developing a corpus: collection, annotation, and analysis. The steps are interrelated, with each influencing the others.

corpora for 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 texts. As a case study, we present the Senti-Turin University Treebank (Senti-TUT) Twitter corpus that was designed to study irony for Italian, a language currently under-resourced for OM&SA.

Developing Corpora for Opinion and Sentiment Analysis

Developing a corpus consists of three main steps: collection, annotation, and analysis (see Figure 1). Each step is strongly influenced by the others. For instance, analyzing and exploiting a corpus can reveal the limits of the annotation or data sampling, which can then be addressed by improving annotation and collecting more adequate data.

Data Collection: What, From Where, and How?

Collection-related issues refer mainly to selecting the data and composing the corpus (what), choosing the data source (where), and collecting the methodologies applied (how). The task for which the resource is developed usually drives the decisions about the *what* and *where*. Most of the corpora designed for OM&SA are collected from Web services that provide comments on commercial products, such as reviews posted on Amazon.[3,4] Others are extracted from blogs and micro-blogs such as Facebook and Twitter to provide insights about people's sentiments on celebrities or politics, such as in the US,[5] German,[6] and UK elections.[7] Other kinds of texts are collected less frequently, such as corpora and tasks about OM&SA in emails[8] and suicide notes.[9]

Often, OM&SA corpora are the result of sampling and filtering a particular target or source, in contrast to resources for other tasks such as parsing, where the focus is mainly on building larger and balanced collections of texts as they spontaneously occur (unrestricted). Data selection and filtering are usually based on keywords such as named entities or metadata released by micro-blog authors, who sometimes exploit hashtags for irony and sarcasm.[3,10] Moreover, metadata on time and geolocations; the user's age, gender, background, and social environment; and communicative goals enable the detection of sentiment variations or trends.

We must also account for text genre during collection, because each genre is typically characterized by a particular manner of expressing opinions and sentiments[11] and exploits particular linguistic structures and devices. For instance, blog texts are highly subjective, while those from newspapers have a more objective tone. Limits imposed by social media on the message length usually influence 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.

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.[4,12]

Annotation: What to Annotate and How?

The annotation step includes a scheme's definition and its application to the collected data, but it also assesses the material by evaluating the inter-annotator agreement.

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. 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.[13] Accordingly, the kinds of knowledge usually annotated are the sentiment's category (hate versus love), polarity (positive versus negative), the source and target toward which the sentiment is directed, and the intensity. Annotations can be based on simple broad polarity labels, possibly equipped with intensity ratings, which also helps us classify texts where mixed sentiments are expressed.[14] They can also be based on labels representing different emotions.[11] When complex knowledge is involved, as in the case of emotional categories, it can be quite helpful to rely on structured knowledge of affective information, such as affective categorization models expressed by ontologies. An even better (more helpful) approach is if the affective categorization models are psychologically motivated—as they are, for example, in *Hourglass of Emotions*,[2,15] which organizes and blends 24 emotional categories into four affective dimensions that it can use to form compound emotions. An ontology that encodes knowledge about emotions can work as a guideline to be shared by the annotators to develop a common understanding about emotions and their relationships.[11] Also, it can support comparison and aggregation among results of emotional analysis, as in the hourglass model.[15]

Because opinions and sentiments are often expressed implicitly through context- and domain-dependent concepts, we must rely on approaches that go beyond the syntactic level, which is what the sentic computing approach strives for in OM&SA.[2] Most 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.[4,6] Different annotations for context-dependent and context-independent opinions also are useful.[6] Meanwhile, analysis at the subdocument level is concerned with distinguishing the portions of text (words, phrases, or more complex structures) containing sentiment expressions. It presupposes that texts have been tokenized with the parts of speech (PoS) tagged and syntactically analyzed. However, the results of such analyses are often limited by the text's ungrammaticality.

Online social data remain largely inaccessible to classic NLP techniques. Such data are specifically meant for human consumption, and their automatic analysis requires a deep understanding of natural language text by machines—an understanding from which we're 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 commonsense knowledge resource that infers both conceptual and emotional information associated with natural language opinions. It thus more easily extracts concept-level sentiments conveyed in word-level natural language texts.[16] Features that vary from one language to another—such as word order and morphological richness in the German language—also dramatically decrease tool portability and annotation suitability.[14,17] Also, most available resources are in English; the few exceptions include the multilingual Multi-Perspective Question Answering (MPQA) dataset, which is automatically annotated for subjectivity.[18]

These 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, however, can improve classifications based on document-level annotation.

Applying the annotation scheme to the data (that is, *how* it's annotated) is usually supported by semiautomatic tools, and necessarily involves more than one annotator to release reliable and unbiased data within the limits of a task inherently affected by subjectivity. The proper number of annotators depends on the task's difficulty.[6] The resulting inter-annotator disagreement is measured[14,17] and sometimes solved. The most commonly applied measures are those inspired by the Cohen's κ coefficient.[19] Best practices to limit and solve the disagreement consist of setting up guidelines shared among the annotators, or annotating and discussing portions of data collectively.[11]

Analyzing and Exploiting a Corpus

Annotated corpora for OM&SA are useful to train and test machine-learning statistical tools for classifying emotions and sentiments. Both the quantity and quality of data strongly influence the results. Error detection and quality-control techniques have been developed, and often the exploitation of the data discloses possible errors. A strategy that can give useful hints about the annotated data's reliability is to compare the results of automated classification and human annotation.[11]

Labeling schemes are always the outcome of a tension between simplicity

and complexity. However, instead of investing efforts in a minimal labeling, we recommend constructing richer labels that support different uses of the annotated material, as shown by Roddie Cowie and his colleagues.[13] Reusability and portability are indeed important measures for datasets that strive for suitability in developing integrated emotion-oriented computing systems. This motivates the efforts devoted to defining and disseminating standards for annotating data for several NLP tasks; those efforts include the Text Encoding Initiative (TEI; www.tei-c.org/index.html), Expert Advisory Group on Language Engineering Standards (EAGLES), and Corpus Encoding Standard (CES; www.cs.vassar.edu/CES/), as well as the more recent proposal for an Emotion Markup Language by Marc Schröder and his colleagues.[13]

The Senti-TUT Project

The Senti-TUT project (www.di.unito.it/~tutreeb/sentiTUT.html) acts as a case study for the issues raised in the previous section.[20] The project's major aims are to develop a resource currently missing for Italian and study a particular linguistic device: irony. The best sources of irony in social media are tweets expressing political opinions. Irony is recognized in literature as a specific phenomenon that can harm OM&SA systems.[3] To deal with this issue, we extended a traditional polarity-based framework with a new dimension that explicitly accounts for irony.

Irony and Sarcasm

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.[3,4,10,21] Before we delve into this further, let's briefly consider the theoretical issues and key aspects that we must account for when developing a corpus for irony detection.

Relevant contributions on irony can be found in a wide range of disciplines, from linguistics to psychology.[22] The rhetorical tradition treated irony as a figure of speech in which the intended meaning is the opposite of the literal meaning. Modern Gricean pragmatic theory hasn't radically departed from this view. Another interesting account within the relevance theory by Deirdre Wilson and Dan Sperber (see chapter 3 of *Irony in Language and Thought*)[22] suggests that irony is a type 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. In this perspective, it's clear that irony can play the role of a *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'll see in a bit.

Other factors to be considered are text context and common ground,[22] which, according to psychological models of language use, are often preconditions for understanding whether a text utterance is ironic. Consider, for instance, Facebook comment threads. Here, the dialogical context can be essential for detecting irony, because many threads often implicitly refer to a common ground restricted to a group of friends, thus making the irony recognition harder for others outside that group. In the case of Twitter, in contrast, posts don't follow a conversation thread and are therefore of a *contextless nature*.[3] Furthermore, even if identifying irony in tweets often requires world knowledge, the authors of posted comments usually refer to a *broader* common ground (for example, knowledge about news or important people), by expressing irony differently than in conversations among friends.

Another issue concerns boundaries between irony and other figurative devices, such as sarcasm, satire, or humor (see Carlo Strapparava and his colleagues' work in *Emotion-Oriented Systems*[13]). According to the literature, these boundaries in meaning between different types of irony are fuzzy.[22] This could be an argument in favor of annotation approaches where different types of irony aren't distinguished, such as the one adopted in Senti-TUT. However, as Antonio Reyes and his colleagues suggest,[10] in the case of figurative languages, the choice among coarse- or finer-grained annotation could lead to different outcomes in the analysis.

Psychological studies also underline the subjectivity of irony perception, regardless of the different world knowledge or limits 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 assigning intensity ratings to ironic annotations,[3] and also by implementing careful disagreement evaluation. Even if there is no agreement on a formal definition of irony—as is the case of most figurative devices—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.

Data Collection

Senti-TUT includes two Twitter corpora—TWNews and TWSpino—that focus

on politics, a domain where irony is frequently exploited by humans. Tweets are composed of less than 140 characters, distributed in one or more short sentences.

The TWNews corpus has been extracted by applying filters based on time and metadata, aimed at selecting posts that represent a variety of opinions about politics. For collection and filtering, we relied on the Blogmeter social media monitoring platform (www.blogmeter.eu), which exploits Twitter's API to extract the tweets. We collected Italian Twitter messages posted during election season in Italy, after Mario Monti was nominated to replace Silvio Berlusconi as the prime minister (from 6 October 2011 to 3 February 2012). We used the list of keywords and/or hashtags "mario monti/#monti," "governo monti/#monti," and "professor monti/#monti" (lowercase and capitalized) to select approximately 19,000 tweets on Monti's government. We then removed *retweets* (RT) because they weren't relevant to our task of irony and sentiment analysis; this pared our collection down to 11,000 tweets. We further discarded 70 percent of those tweets using annotators that deemed them ungrammatical, poorly written, duplicated (but not marked as RT), or incomprehensible without a context. (Even though tweets don't follow a conversation thread, a notion of context spreads in the data through repetitions and reprises of previous posts.) The final results are the 3,288 posts of TWNews.

The TWSpino corpus is composed of 1,159 messages from the Twitter section of *Spinoza* (www.spinoza.it), a very popular Italian blog of posts containing sharp satire on politics. We extracted posts published from July 2009 to February 2012 and removed advertising (1.5 percent). Because there's 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 without filtering.

```
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 Concorrerà (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]
```

Figure 2. An online post represented in the Turin University Treebank (TUT) format. It includes a detailed morphological tag set and a large inventory of grammatical relations labeling the dependency trees' edges.

Annotation

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

Annotating 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. Annotating at the subdocument level benefits from the experience gained in corpus-based NLP tasks such as PoS tagging and parsing. It's also in line with more recent work,[1] in which the task is both to find a piece of opinionated text and to extract a structured representation of the opinion (determining the holder and the target), inspired by experience in information extraction, semantic role labeling, and structured machine learning.

Morphological and syntactic annotation. We morphologically and syntactically annotated Senti-TUT according to the format developed and applied in the Turin University Treebank (TUT; www.di.unito.it/~tutreeb). This is a freely available resource—developed by the University of Turin's NLP group—that applies the Turin University Linguistic Environment (TULE; www.tule.di.unito.it), whose pipeline includes tokenization, morphological, and syntactic analysis. It has been successfully exploited as a testbed in the evaluation campaigns for Italian parsing (www.evalita.it).

Consider this post from TWSpino: "*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). The post is represented according to TUT's format (see Figure 2), which includes a detailed morphological tag set (an essential feature for describing a

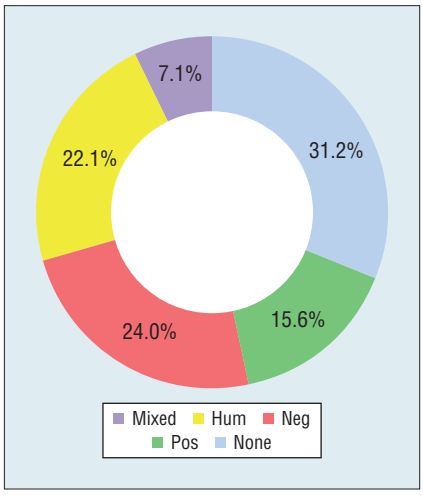

Figure 3. Distribution of Senti-TUT tags in the TWNews corpus.

language with a rich inflection), and a large inventory of grammatical relations labeling the dependency trees' edges (to describe the sentence's argument structure).

Tweet-level sentiment and irony annotation. We considered single tweets as individual documents and annotated one of the following sentiment tags for each tweet, by evaluating 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)

Here are 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 Berlusconi's resignation, but I have no illusion about Monti's government)

TWNews-123 (tagged as None)
Mario Monti premier? Tutte le indiscrezioni.
(Mario Monti premier? All the gossip.)

At first, we had five human annotators (two males and three females of varying ages) collectively annotate a small dataset (200 tweets) manually, attaining a general agreement on the labels' exploitation. Then we manually annotated the whole dataset, and we produced for each tweet no less than two independent annotations. The agreement calculated at this stage, according to the Cohen's κ score, was satisfactory: $\kappa = 0.65$. To extend our dataset, we applied a third independent annotation on any instances where disagreement was detected (about 25 percent of the data). After that, we sorted through the cases where disagreement persisted—for example, where every annotator had selected different tags. We discarded those as too ambiguous to be classified (this was around 2 percent, and it's an interesting sample to analyze for future work). Our final result for TWNews is 3,288 tweets.

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. Figure 3 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[21]—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 tweets. On those data, we detected a variation between BC and H classification, taken as an indicator of polarity reversing. We observed this variation in most of the selected tweets (68.5 percent). 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 → Htag denotes the direction of the polarity variation from the Blogmeter to the human classification. Although the dataset's limited size and its particular domain and text genre make our results preliminary, the theoretical accounts seem to be confirmed.

Table 1. Polarity variations in ironic tweets showing reversing phenomena.

	Polarity variation's direction	
PV's typology	**Positive → Negative (%)**	**Negative → Positive (%)**
Full reversal	33.6	3.7
	Positive → None (%)	**Negative → None (%)**
Attenuation	22.2	40.5

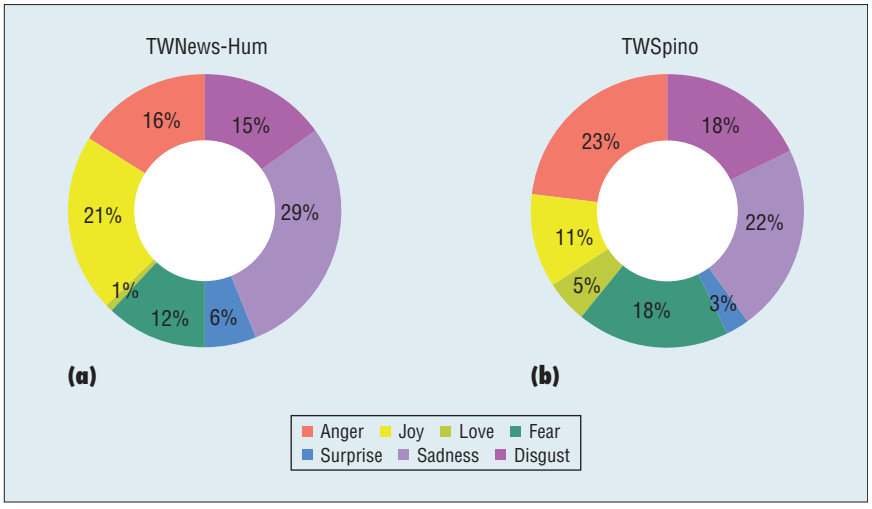

Figure 4. Emotion distribution in the ironic emotional tweets of (a) TWNews-Hum and (b) TWSpino. In TWNews-Hum, the most common emotions expressed were sadness and joy, while the most common emotions in TWSpino were anger and sadness.

Emotions in ironic tweets. Another interesting challenge is to apply our dataset to emotion-detection techniques (beyond positive or negative valence)—similar to the efforts of Reyes and his colleagues[10]—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).[11] These emotions are expressed in 20 percent of our dataset and distributed differently in the corpora, as Figure 4 shows.

In TWNews-Hum, the most common emotions were sadness (29.1 percent) and joy (20.9 percent), followed by anger, disgust, and fear. Surprise was rare, and love was almost nonexistent. TWSpino contains instead more negative emotions: anger (22.7 percent) and sadness (22.2 percent), 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" (rather than invoking a negative attitude, these tend to be playful and produce a comic or parodic effect, often to strengthen ties with others chatting online).

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. Moreover, Spinoza's editors 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.

Beyond developing a missing resource for Italian, the primary purpose of the Senti-TUT Twitter corpus is to study irony, rather than Twitter as a whole. 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 (see chapter 6 of *Irony in Language and Thought*[22]). 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 (which confirm the importance of shared knowledge in irony detection).

A formal account and a measure of these phenomena is a matter of future work. This work will require a finer granularity in text analysis, in line with Richard Johannson and Alessandro Moschitti's work.[1] It will also require the use of commonsense knowledge bases to extract the latent semantics from text—as hoped for in concept-level approaches to OM&SA[2]—and especially to measure incongruity and contextual imbalance in terms of the semantic relatedness of concepts expressed in ironic texts.[10] For this purpose, we're devoting our efforts to applying a semantic annotation based on the major semantic resources currently available for Italian (BabelNet and WordNet).

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;[22] and on the representative textual features that distinguish these phenomena.

For studying and identifying emotions, we propose a measure that relies on Blogmeter's emotion-annotation techniques applied to the ironic tweets of the Senti-TUT dataset. Blogmeter adopts a rule-based approach to sentiment analysis, which was tested recently in an experiment of automatic emotion annotation on a corpus of 31 million Italian tweets, with the set of emotions used in Kirk Roberts and his colleagues' work.[11] An interesting step forward would be to refer to a richer semantic model, as in the *Hourglass of Emotions*,[15] to enable reasoning about semantic relations among emotions (similarities, opposites, and intensities). Although we report here on a limited exploitation of our data in automatic classification tasks (these experiments are detailed elsewhere),[20] the lessons learned from this data analysis give useful hints about future directions. ■

Acknowledgments

This work was partially funded by the Portale per l'Accesso alle Risorse Linguistiche per l'Italiano (PARLI) project (MIUR PRIN 2008). We're grateful to our annotators and to the Language and Information Technology Center (CELI Torino) for providing the facilities offered by the Blogmeter platform.

References

1. R. Johansson and A. Moschitti, "Relational Features in Fine-Grained Opinion Analysis," *Computational Linguistics*, 2012; www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00141.
2. E. Cambria and A. Hussain, *Sentic Computing: Techniques, Tools, and Applications*, Springer, 2012.
3. D. Davidov, O. Tsur, and A. Rappoport, "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon," *Proc. Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Assoc. for Computational Linguistics (ACL), 2011, pp. 107–116.
4. E. Filatova, "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing," *Proc. Language Resources and Evaluation Conf.*, European Language Resources Assoc. (ELRA), 2012, pp. 392–398.
5. A. Tumasjan et al., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *Proc. Int'l Conf. Weblogs and Social Media*, Assoc. for the Advancement of Artificial Intelligence (AAAI), 2011, pp. 178–185.
6. H. Li et al., "Annotating Opinions in German Political News," *Proc. Language Resources and Evaluation Conf.*, ELRA, 2012, pp. 1183–1188.
7. Y. He et al., "Quantising Opinions for Political Tweets Analysis," *Proc. Language Resources and Evaluation Conf.*, ELRA, 2012, pp. 3901–3906.
8. S.M. Mohammad and T. Yang, "Tracking Sentiment in Mail: How Genders Differ on Emotional Axes," *Proc. 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, ACL, 2011, pp. 70–79.
9. J.P. Pestian et al., "Sentiment Analysis of Suicide Notes: A Shared Task," *Biomedical Informatics Insights*, 2012, vol. 5, no. 1, pp. 3–16.
10. A. Reyes, P. Rosso, and D. Buscaldi, "From Humor Recognition to Irony Detection: The Figurative Language of Social Media," *J. Data and Knowledge Eng.*, vol. 74, 2012, pp. 1–12.
11. K. Roberts et al., "Empatweet: Annotating and Detecting Emotions on Twitter," *Proc. Language Resources and Evaluation Conf.*, ELRA, 2012, pp. 3806–3813.
12. A. Wang, C. Hoang, and M.Y. Kan, "Perspectives on Crowdsourcing Annotations for Natural Language Processing,"

Language Resources and Evaluation, vol. 47, no. 1, 2013, pp. 9–31.
13. R. Cowie, C. Pelachaud, and P. Petta, Eds., *Emotion-Oriented Systems: The Humaine Handbook*, Springer-Berlin, 2011.
14. S. Momtazi, "Fine-Grained German Sentiment Analysis on Social Media," *Proc. Language Resources and Evaluation Conf.*, ELRA, 2012, pp. 1215–1220.
15. E. Cambria, A. Livingstone, and A. Hussain, *The Hourglass of Emotions*, LNCS 7403, Springer, 2012, pp. 144–157.
16. E. Cambria, C. Havasi, and A. Hussain, "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis," *Proc. 25th Int'l Florida Artificial Intelligence Research Society Conf.*, AAAI, 2012, pp. 202–207.
17. J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, 2005, vol. 39, nos. 2–3, pp. 165–210.
18. C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual Subjectivity: Are More Languages Better?" *Proc. 23rd Int'l Conf. Computational Linguistics*, ACL, 2010, pp. 28–36.
19. R. Artstein and M. Poesio, "Inter-Coder Agreement for Computational Linguistics," *Computational Linguistics*, vol. 34, no. 4, 2008, pp. 555–596.
20. A. Gianti et al., "Annotating Irony in a Novel Italian Corpus for Sentiment Analysis," *Proc. 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, ELRA, 2012, pp. 1–7.
21. D. Maynard, K. Bontcheva, and D. Rout, "Challenges in Developing Opinion Mining Tools for Social Media," *Proc. Language Resources and Evaluation Conf.*, ELRA, 2012, pp. 15–22.
22. R.W. Gibbs and H.L. Colston, eds., *Irony in Language and Thought*, Taylor and Francis, 2007, pp. 35–56.

THE AUTHORS

Cristina Bosco is an assistant professor in the Department of Computer Science at the University of Torino, and she's responsible for the TUT project (http://www.di.unito.it/~tutreeb). Her research interests include dependency and constituency parsing, linguistic resources with morphological and syntactic annotation, evaluation, and sentiment analysis. Bosco has PhD in computer science from the University of Torino. She's a member of IEEE, the Association for Computational Linguistics (ACL), and the Italian Association for Artificial Intelligence (AI*IA). Contact her at bosco@di.unito.it.

Viviana Patti is an assistant professor in the Department of Computer Science at the University of Torino. Her research interests include knowledge retrieval in multiagent systems, social Semantic Web, ontology-driven sentiment analysis, and service-oriented computing. Patti has a PhD in computer science from the University of Torino. She's a member of IEEE, AI*IA, and the Italian Association for Logic Programming (GULP). Contact her at patti@di.unito.it.

Andrea Bolioli is a computational linguist and is the co-founder of CELI srl, a company that develops software solutions using natural language-processing technologies. He also works on Blogmeter (CELI and Me-Source), an Italian social media monitoring service based on a proprietary listening platform that delivers accurate classification and sentiment analysis. His research interests include text mining, sentiment analysis, and computational narratology. Bolioli received his philosophy of language degree from the University of Torino. Contact him at abolioli@celi.it.