Detecting Happiness in Italian Tweets: Towards an Evaluation Dataset for Sentiment Analysis in Felicittà

C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, E. Sulis
Dipartimento di Informatica - Università degli Studi di Torino
Corso Svizzera 185 - 10149, Torino (Italy)
{bosco,patti,ruffo,msanguin,sulis}@di.unito.it, {leonardo.allisio,valeria.mussa}@studenti.unito.it

Abstract

This paper focuses on the development of a gold standard corpus for the validation of Felicittà, an online platform which uses Twitter as data source in order to estimate and interactively display the degree of *happiness* in the Italian cities. The ultimate goal is the creation of an Italian reference Twitter dataset for sentiment analysis that can be used in several frameworks aimed at detecting sentiment from big data sources. We will provide an overview of the reference corpus created for evaluating Felicittà, with a special focus on the issues raised from its development, on the inter-annotator agreement discussion and on implications for the further development of the corpus, considering that the assumption that a single right answer exists for each annotated instance cannot be done in several cases in the particular kind of data at issue.

Keywords: Sentiment analysis in Twitter, Corpus annotation, Italian

1. Introduction

In the last few years, the linguistic analysis of social media has become a relevant topic of research, and several frameworks for detecting sentiments and opinions in social media have been developed for different application purposes.

One of the possible applications of Sentiment Analysis (SA) is in the social and behavioral sciences field, where SA techniques could contribute to interpret the degree of well-being of a country. The studies concerning life satisfaction have grown substantially since the late 20th Century. New areas of research have arisen, such as the *Subjective Well-Being* (SWB) in Psychology (Diener, 2000) and the *Happiness economics* in Economy, within the debate on alternative measure to Gross Domestic Product (Helliwell et al., 2014). The rise of Big Data and the exponential growth of social media (e.g. Facebook, Twitter) has created vast opportunities and new challenges to the social sciences on this respect. In some pioneering work in this direction, extracting expressed sentiments – typically categorized as positive, negative or neutral – in short messages has been used for several purposes: to detect moods and happiness in a given geographical area from geotagged Tweets (Mitchell et al., 2013), to create a *hate map* based on expressions of homophobia and racism on Twitter[1], to show the correlation with traditional data (Bollen and Mao, 2011) and to measure the well-being of a population (Quercia et al., 2012).

It should also be observed that linguistic analysis of social media has gained in the last few years an increasing relevance in the detection of well-being or happiness (Mihalcea and Liu, 2006). However, various issues should be taken into account in the detection of sentiments and opinions in natural language texts. On the one hand, data on which SA is applied are from texts especially challenging for most Natural Language Processing (NLP) systems. Although, as observed in (Baldwin, 2012), social media texts can also be considered a valuable resource, rather than a foe, by virtue of the richness of non-textual data that can be exploited to enhance the robustness and accuracy of NLP techniques. As a matter of fact, hashtags, emoticons, emojis or links occurring in a post can be used to disambiguate the textual content. On the other hand, training and testing automatic systems requires the availability of several resources that may consist in large datasets of annotated posts or even in lexical databases where affective words are associated with polarity values. But their availability is currently very limited in particular for languages other than English.

In this paper, we would like to contribute to the debate in this area by describing our experience in the development of Felicittà, an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines a lexicon-based approach for SA and visualization techniques in order to provide users with an interactive interface for data exploration (Allisio et al., 2013). (Pianta et al., 2002; Strapparava and Valitutti, 2004).

In particular, we will report the most recent achievements in the development of the platform, especially focusing on the creation of a Twitter dataset for testing the sentiment algorithm in Felicittà. For what concerns the annotation schema and procedure, we rely on the research carried out within the Senti–TUT project (Bosco et al., 2013). The ultimate goal is the creation of an Italian reference corpus that can be used in several frameworks for detecting sentiments from big data sources, such as Twitter.

We will provide an overview of the reference corpus currently developed for Felicittà, by focusing in particular on the issues raised from annotator agreement analysis and their implications for the further development of the corpus.

2. Related Works

For what concerns the resources for SA, for English language, sentiment lexicons (listed in (Nakov et al., 2013)), Twitter datasets and gold standards for the sentiment analysis task on Twitter messages are now available [2], while

[1] http://users.humboldt.edu/mstephens/hate\/hate_map.html

[2] See the recent survey and comparison in (Saif et al., 2013).

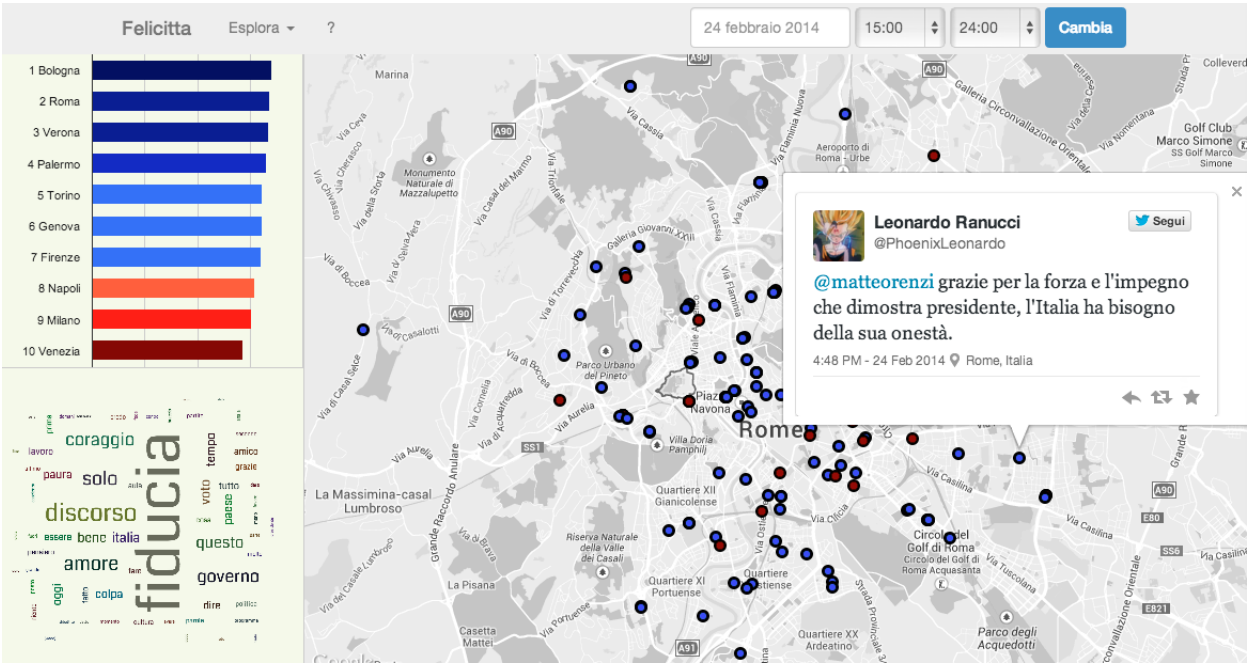

Figure 1: Felicittà: an interactive map displaying Tweets that convey negative or positive polarity and positioned within the area where they have been posted.

for several other languages, like Italian, the availability of such resources is currently very limited. Indeed, several resources are being developed by individual companies for their commercial use in sentiment monitoring services[3], but normally they are not shared nor publicly available.

For what concerns Italian, to the best of our knowledge, Senti–TUT is the first Italian gold corpus developed for Twitter SA (Bosco et al., 2013), which also includes ironic tweets. Irony detection is a hot topic in the SA research community indeed, and in particular the fact that Twitter messages include a high percentage of ironic messages cannot be neglected (González-Ibáñez et al., 2011; Reyes et al., 2013; Davidov et al., 2011; Hao and Veale, 2010). Platforms monitoring the sentiment in Twitter messages experience the phenomenon of wrong polarity classification in ironic messages. Indeed, the presence of ironic devices in a text can work as an unexpected ”polarity reverser” (one says something ”good” to mean something ”bad”, or vice versa), thus undermining systems’ accuracy[4].

Recent works (Caselli et al., 2012; Baldoni et al., 2012; Bertola and Patti, 2013) exploited WordNet-Affect (Strapparava and Valitutti, 2004), an affective lexicon which links synsets in the original Princeton WordNet (Fellbaum, 1998) to affects, but, being the affective extension of WordNet domains developed at irst-FBK and aligned with MultiWordNet, WordNet-Affect embeds information on the correlation between English and Italian terms. WordNet-Affect is freely available for research purposes. It is semi-automatically created, based on a manually realized core, and includes 4,787 affective words. Moreover, only very recently a new publicly available lexical resource for Italian has been developed, which is called Sentix (Sentiment Italian Lexicon) (Basile and Nissim, 2013) and is the result of the alignment of several existing lexical and affective resources: WordNet, MultiWordNet (Pianta et al., 2002), BabelNet (Navigli and Ponzetto, 2012) and SentiWordNet (Esuli et al., 2010).

It should also be observed that the development of corpora that can be usefully exploited in this kind of task is in itself very challenging. For other tasks the development of a corpus consists in creating an annotated human ground truth, assuming that for each annotated instance there is a single right answer and that the quality of the annotation can be measured in terms of inter-annotator agreement.

In the development of a corpus for SA this assumption cannot be done, and the disagreement reflects semantic ambiguity in the target instances, thus providing useful information. Under this respect, the annotation of a corpus for SA can be usefully compared to the development of corpora for clinical studies, see e.g. (Xia and Yetisgen-Yldiz, 2012), or those for co-reference where underspecified labels are adopted to cope with the vagueness of data (Versley, 2006). In corpora for SA the reasons for annotator disagreement are also related to the fact that *a*) there are many different ways to linguistically express the same polarity, and *b*) the same linguistic expression may be used for different polarities. This in turn makes context extremely important, for instance in case of humorous and ironic expressions. These factors create, in human understanding, a fairly wide range of possible, plausible interpretations of a post, and as a consequence a disagreement in the annotation.

[3]Think for instance to the affective Italian lexicon used in the social media monitoring platform Blogmeter (`http://www.blogmeter.eu/`), which includes about 10,000 entries (Bolioli et al., 2013).

[4]A pilot subtask concerning irony detection on Italian Tweets will be organised at Evalita: `http://www.di.unito.it/~tutreeb/sentipolc-evalita14/index.html`

3. Felicittà

Felicittà[5] is an online platform for estimating happiness in the Italian cities, which daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets, in order to enable the summarization of SA outcomes and the exploration of Twitter data (Allisio et al., 2013). Interactive maps offered by Felicittà provide users not only with the opportunity to have a comprehensive overview of the SA results about the main Italian cities, but also to zoom-in to a specific region to visualize a fine-grained map of the city or district and the location of the individual sentiment-labeled Tweets (Fig. 1). Interaction possibilities enabled by the platform allow users to tune their view on such huge amount of information and to interactively reduce the inherent complexity, possibly providing a help in the detection of meaningful patterns. Tag clouds highlighting the important words in the Tweets posted in a geographic area are daily generated and visualized together with the sentiment outcomes, with the aim of evoking possible correlations between mood and events.

The heart of the framework is a sentiment analyzer. By exploiting Twitter's APIs, the system collects every day all the Tweets freely downloadable (450,000), geo-located in the main Italian towns, and performs the analysis for each Tweet in order to classify it as positive or negative. This analysis includes, in particular, the application of Freeling[6], a multilingual open source tool for morpho-syntactic analysis, developed at the University of Catalunya (Spain). The grammatical category and lemma of each word is recognized, thus allowing a more efficient association with the lexical item to be searched in the affective lexicon. Finally, the polarity of all the Tweets is aggregated according to their geo-location and the happiness degree of each town and region is evaluated and made available in different visualization modes.

According to a lexicon-based approach, the polarity of each Tweet in Felicittà depends on the affective words detected within it and then found in the affective lexicon, i.e. in WordNet–Affect, that is the resource which most of the recent works for Italian currently exploit, see e.g. (Caselli et al., 2012; Baldoni et al., 2012; Bertola and Patti, 2013).

4. Data annotation for sentiment analysis

In order to validate our approach and to analyze the limits of the sentiment analyzer implemented in Felicittà, we have created a reference corpus including a set of Italian Tweets, called TW-FELICITTA.

4.1. Collection

1,500 Italian Tweets were randomly extracted from those collected by Twitter API, paying attention to avoiding geographic and temporal bias at different level of granularity. As a matter of fact, possible correlations have been observed between sentiment and time of the day or day of the week (weekdays or holidays), or between sentiment and geographical areas in a given time frame due to the occurrence of some special event. Furthermore, we gathered the Tweets for the collection in order to avoid a logical link between a Tweet and the next one, which is a typical situation where two users communicate with each other: this way, it is not possible to infer the discussion topic, unless this is explicitly mentioned; the principle that lies behind this choice is that of preventing both the system and the manual annotator from labeling the Tweets in a different way namely because of such inferred information. We therefore implemented an automatic algorithm for the collection which takes into account such issues.

4.2. Annotation schema

Sentiment annotation was manually performed at the Tweet level. This means that we considered single Tweets as individual documents and annotated them using one of the tags reported in Table 1 and previously applied to the annotation of the Senti–TUT Italian corpus for SA (Bosco et al., 2013)[7].

POS	positive
NEG	negative
NONE	objective (no sentiment expressed)
MIXED	mixed (POS and NEG both)
HUM	ironic
UN	unintelligible

Table 1: Tags annotated in TW-FELICITTA corpus.

The application on TW-FELICITTA has shown the suitability of this schema designed for the annotation also of mixed polarity and ironic expressions, exploiting the MIXED and HUM tags. Indeed, also because the sentiment annotation is performed at the Tweet level, it is often difficult to determine unambiguously the overall polarity of the sentiment expressed in it, especially in presence of irony and mixed sentiment. Ironic Tweets and Tweets containing parts expressing both positive and negative sentiment have recognized to be phenomena that strongly contribute to make the Tweet classification task harder (Nakov et al., 2013). In this context, the classical labels distinguishing only among positive, negative or neutral sentiment may not be sufficient; we thus extended the tag set by including:

- MIXED to mark the presence of more than one sentiment within a Tweet, which can be related to the expression of opinions on different targets or also to a contrast between polarity of the opinion conveyed and expressed mood, see also the gold standard presented in (Saif et al., 2013).
- HUM to mark the intention of the author of the post to express irony, which could be hardly classified as entirely positive or negative;
- UN to mark the difficulty experienced by the annotator due, e.g., to the incompleteness of the message or the absence of a context.

[5] http://felicitta.di.unito.it/

[6] http://nlp.lsi.upc.edu/freeling/

[7] http://www.di.unito.it/~tutreeb/sentiTUT.html

The following examples are applications of the above described labels.

TW-FELICITTA#504 (tagged as POS)
American Horror Story ti amo
#AmericanHorrorStory sei il telefilm che fa la differenza.
(I love you, American Horror Story
#AmericanHorrorStory you're the tv series that makes
the difference.)

TW-FELICITTA#518 (tagged as NEG)
Perche' non riesco a dimenticarla
(Why can't I just forget about her)

TW-FELICITTA#636 (tagged as NONE)
Accadde oggi: 1993: entra in vigore il Trattato di Maastricht, che stabilisce formalmente l'Unione Europea....
(Today in history: 1993: The Maastricht Treaty,
which formally establishes the European Union,
enters into force ...)

TW-FELICITTA#305 (tagged as MIXED)
E' stata una settimana perfetta
Ma questa domenica ha rovinato tutto Ma proprio tutto.
(It was a perfect week. But this Sunday has ruined everything
Absolutely everything)

TW-FELICITTA#683 (tagged as HUM)
RT@lddio:Letta:"I giovani senza lavoro sono l'incubo dell'Italia". Per non essere da meno, anche l'Italia e' l'incubo dei giovani.
(Letta: "Young people out of work are the nightmare of Italy."
Not to be outdone, Italy is the nightmare of young people.)

TW-FELICITTA#771 (tagged as UN)
@Caustica_mente ho detto che sono inconsistenti?
volevo capire i motivi dell'eventuale autogoal.
A leggerti, non e' alfine tale. Bene.
(@Caustica_mente Did I say that they are inconsistent?
I wanted to understand the reasons for an own goal.
By reading you, this is not finally such. All right.

For what concerns the last sample, the English translation was kept ungrammatical on purpose, in order to convey to the non-Italian reader as well the difficulty experienced by the annotator in inferring the meaning of the message.
For what concerns the label HUM, let us notice that, as also pointed out in the literature, there is no agreement on a formal definition of irony, as is the case of most figurative devices. Nonetheless, psychological experiments have given 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. Moreover, the boundaries between irony and other figurative devices, such as sarcasm, satire, or humor, are quite fuzzy (Strapparava et al., 2011). This made us lean on adopting the same approach proposed in Senti-TUT, where no distinction has been drawn among different types of irony.
Notice that, having a distinguished tag for irony do not prevent us to reconsider these Tweets at a later stage, and "force" their classification according to traditional annotation schemes for the SA task, as suggested for instance in (Bosco et al., 2013), where a similar approach has been applied to tackle with the polarity reversing phenomenon due to the presence of irony, and to measure how an automatic traditional sentiment classifier can be wrong. Similarly, identifying Tweets containing mixed sentiment can be useful in order to measure how the phenomenon impacts on the performances of sentiment classifiers[8].
Moreover, having distinguished tags for irony and mixed sentiment can be helpful for a better development of the corpus itself, in order to increase the inter-annotator agreement, since such cases, being typically source of disagreement on the polarity valence, are recognized and labeled apart.

4.3. Annotation process

The annotation process (together with the annotation guidelines) was developed through multiple stages. After a phase where four human annotators (A_1, A_2, A_3, A_4) (native-speakers, different genders, varying ages and background) collectively annotated a small set of data (i.e. 100 Tweets), results on the disagreement were discussed in order to both reach a better agreement on the exploitation of the labels on the entire corpus, and produce a document including annotation guidelines[9] shared by the annotators.
Then, A_1, A_2 and A_3 annotated all the data (i.e. 1,500 Tweets) producing for each Tweet not less than three independent annotations. The inter-annotator agreement has been calculated at this stage according to the Fleiss's Kappa (Fleiss, 1971) andthe measure obtained reached $\kappa = 0.51$. It can be observed that this rate positively compares to that described for the similar task in (Basile and Nissim, 2013), and it is a slightly lower rate with respect to the development of TW-NEWS (Bosco et al., 2013), where only two annotators were involved.
The agreement among the three annotators has been achieved in this step in 46% of cases, corresponding to 695 Tweets. On the remaining 805 Tweets, we can distinguish between Tweets in *hard disagreement*, when three different tags have been annotated, and those in *soft disagreement*, on whose polarity at least two annotators agreed. The former consist of 13% (191 Tweets), while the latter of 41% (614 Tweets) of the entire corpus.
In order to further extend our data set, we discarded the Tweets featured by hard disagreement, but we recovered the agreement on a large portion of those resulting in a soft disagreement after the first annotation step in two ways. First, we applied to this set a 4th independent annotation (by A_4), and we achieved in this way the agreement among three of the four annotations on further 433 Tweets of the

[8] Also in this case it could be interesting to reconsider Tweets tagged as MIXED at a second stage, by classifying them as either (mainly) positive or negative

[9] See: http://www.di.unito.it/˜tutreeb/AnnotationGuidelines.pdf

614 cited above. Second, at a last stage, the four annotators discussed the polarities of the remaining 181 Tweets (i.e. 29%), hypothesizing that the soft disagreement was persisting on them because of annotators' biases or errors. The discussion leaded to an updated version of the guidelines and to the ultimate version of the corpus where further 107 Tweets have been recovered in agreement, thus obtaining two sets: one set of Tweets in agreement composed of around 82% (1,235 Tweets), henceforth indicated as *A-set*, and one of those featured by an unsolvable disagreement composed of around 18% (265 Tweets) of the entire corpus, henceforth indicated as *D-set*.

Therefore we included in the final version of the TW-FELICITTA gold corpus only the 1,235 Tweets on which we achieved the agreement among the annotators, ready to be exploited for training and evaluation purposes. The final tags in the gold corpus are distributed as follows among the Tweets of the *A-set*: around 57% of them were classified as positive (338) or negative (364), 21% is classified as NONE (260), around 14% as HUM, and the remaining as MIXED (3%) or UN (5%), as shown in Figure 2.

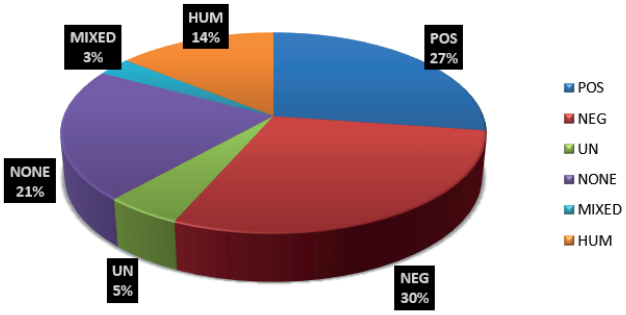

Figure 2: Distribution of the sentiment labels used by the annotators in the gold corpus of Felicittà.

For what concerns instead the remaining cases (around 18%), considered as too ambiguous to be classified according to the detected disagreement, we aim to define a framework to harness the analysis of the disagreement between the human annotators, in order to capture interesting features related to the sentiment and irony detection task. Our preliminary results in the definition of this frame can be seen in section 5.

5. Annotation Analysis

Annotation is an important task for NLP, and the traditional annotation pipeline, including writing detailed guidelines, trained annotators and disagreement calculation, has proved to work well in several projects. Other annotation strategies has been proposed for specific tasks, see e.g. (Xia and Yetisgen-Yldiz, 2012). On the one hand, annotation of polarity for SA is a task featured by specific peculiarities that can be made clear e.g. by observing the distribution of tags and disagreement calculation. On the other hand, the feature of each single corpus should be carefully taken into account and compared with those of other data sets.

For what concerns TW-FELICITTA, we first made a comparison with TW-NEWS (Bosco et al., 2013), a similar Italian corpus that includes Tweets collected in the time frame between October 2012 and February 2013 and that focuses on a specific topic (the past Montis government in Italy). Such comparison shows that in the former there is a meaningfully smaller amount of Tweets with neutral polarity with respect to the other data set we have previously annotated. This can be motivated by the larger frequency of emoticons and *emoji*[10], which are currently often used in social media and supported by smartphones interfaces, as observed also in (Suttles and Ide, 2013), but were very rarely used in 2012, when TW-NEWS has been collected. They are considered by the annotators as useful hints about the polarity of posts, and can also be used by automatic systems for a reliable detection of polarity. This is confirmed by the preliminary analysis performed by the sentiment analyzer implemented in Felicittà.

Second, considering the selection criteria (mentioned above) for the creation of the TW-FELICITTA corpus, there is a high variety in the topics addressed in the Tweets, and their independence with respect to the time frame and geographic area do not allow the annotator to trace back to the original communicative situation. This aspect, as also pointed out in (Basile and Nissim, 2013), together with the wider tag set used in our corpus (w.r.t. the classic annotation schemas for sentiment) and varying annotators' skills (depending, in their turn, on different genders and varying ages and background), is deemed to be a possible source of disagreement.

It should be observed that the final goal of the annotation of a corpus for SA is a consistent annotation rather than a full agreement. If we compare annotation for SA to that performed for other tasks, we can see relevant differences that should be dealt with in different ways with respect to e.g. co-reference annotation (Poesio and Artstein, 2008), where the use of underspecified representations is exploited as a means to cope with the inherent ambiguity of the data to be annotated. By contrast, according to the results of a fine-grained analysis of disagreements (see section 5.1.), for SA the occurrence of genuine ambiguities gives useful hints about what kind of annotation can be more suitable for the task. In particular, observing the features of the task, we investigated some directions of analysis, among which the detection of subjectivity of the sentiment tags according to different measures, and the detection of systematic differences among annotators, devoted to identify the peculiarities of this task.

5.1. Measuring disagreement

For what concerns the detection of the *subjectivity of the sentiment labels* in our annotation scheme, we hypothesized that when a sentiment label is more involved in the occurrence of disagreement, this is because it is more difficult to be annotated, as its meaning is less shared among the annotators and there is a larger range of subjectivity in its interpretation. This phenomenon can be modeled and described according to different perspective and with reference to different portions of the dataset.

[10]Emoji are an alternative for explicit, manual labels, see `http://en.wikipedia.org/wiki/Emoji`.

In order to calculate the subjectivity of each label *L* we propose the following measure: considering all the tags exploited by all the annotators during the annotation process (i.e. 4,936 for the 1,235 Tweets of the *A-set*, and 867 for the 265 Tweets of the *D-set*), we calculated for each *L* the percentage of cases where *L* has been annotated for a Tweet in the *A-set* or for one in the *D-set*. Table 2 shows therefore how much a label has been used in percentage to contribute to the definition of an agreed or disagreed annotation of the Tweets.

label	agreement	disagreement
POS	26.3	14.4
NEG	29.2	17.8
NONE	21.8	23.5
MIXED	3.3	8.8
HUM	11.9	13.0
UN	7.6	22.5

Table 2: A measure of subjectivity of tags annotated in TW-FELICITTA corpus: percentage of Tweets in agreement/disagreement where each label is involved.

It should be observed, in particular, that while POS and NEG labels seem to have a higher reference to the agreement, for UN and MIXED the opposite situation happens, confirming that the annotators are more troubled by the exploitation of the latter tags.

Assuming a perspective oriented to the single annotators and referring to all the annotated tags, as above, we also measured the *subjectiveness* of each *annotator involved in the task* according to the variation in the exploitation of the labels. For each label *L*, starting from the total amount of times when *L* has been annotated, we calculated the average usage of the label. Then we calculated the deviation with respect to the average and we observed how this varies among the annotators. In table 3 the labels are presented from the most to the least used, together with the percentage of positive and negative deviation with respect to the average number of times where they have been annotated.

label	total	average	deviation +	deviation -
NEG	1,592	398	15.32%	14.82%
POS	1,421	355.25	6.68%	5.13%
NONE	1,281	320.25	24.90%	16.31%
HUM	700	175	28.57%	31.42%
UN	569	142	73.94%	35.21%
MIXED	237	59.25	46.83%	80.18%

Table 3: A measure of variation among the exploitation of the labels in TW-FELICITTA corpus.

The deviation is maximum for the tags MIXED and UN, while is meaningfully lower for all the other tags, in particular for POS and NEG, showing that the annotators are more confident in exploiting these latter tags.

Focusing instead the analysis on the *A-set* only, and again assuming a perspective oriented to the single annotators, we can calculate a sort of precision of the annotation done by each of them. We calculated this measure by considering each annotator A as a system whose results should be evaluated against the gold standard represented by our *A-set*. Dividing the amount of Tweets annotated by A with the same tag exploited in the *A-set* over the amount of Tweets included in the *A-set*, we obtained the precision shown by A in the annotation task. The scores for our annotators vary from 0.801 to 0.911, confirming that they can be considered as skilled enough and featured by a limited bias.

On the same set of data, i.e. *A-set*, but focusing on the tags, for each polarity label *L* we calculated the amount of Tweets that contain in their annotation at least one occurrence of *L*, divided by the amount of Tweets whose final annotation has been done with that label. The value of this measure is 1, when *L* is highly precise, that is each time that *L* has been used by some annotator, the final annotation of the Tweet in the released corpus is exactly *L*; it is higher than 1 when *L* is less precise. As reported in table 4, the lower scores are referred for POS and NEG, while the higher for UN and MIXED, which are in effect the labels annotated when the polarity of the Tweet is more ambiguous.

label	precision
POS	1.2
NEG	1.2
NONE	1.5
MIXED	2.0
HUM	1.2
UN	3.5

Table 4: A measure of precision of tags annotated in TW-FELICITTA corpus.

We conclude with some observation on the tag HUM, which we would like to investigate in the future work. If we focus on the *A-set*, we can see that all the Tweets included in it are featured by three or four annotations done with the same tag. If we further limit our observation to the Tweets associated with only three annotations done with the same tag and the fourth different, we see that for more than a quarter of them the fourth annotation is done by the tag HUM.

Another aspect we investigated is related to the issue of which tags co-occur more frequently with the tag HUM in the Tweets. Comparing the distribution of the tags on tweets that were labeled as HUM at least by one of the annotators to the overall distribution of the tags (excluding the tweets containing in their annotation a tag HUM), it appears that HUM significantly co-occurs with the UN and MIXED tags. With regard to the co-occurrence of HUM and UN, this result can be explained with the importance of the con-

text and of common ground, which, according to functional psychological models of language use, are often preconditions for understanding if a text is ironic utterance. While with regard to the co-occurrence of HUM and MIXED, in many cases the misinterpretation takes place because a sarcastic expression has been used; as also noted in (Riloff et al., 2013), a common form of sarcasm on Twitter consists of a positive sentiment contrasted with a negative situation, therefore, even though a positive sentiment is expressed in the utterance, the overall perception of the ironic tweet is that it bears a negative polarity. This may lead in annotators that do not recognize the ironic intent (maybe, again, for the absence of a context) to the perception that the Tweet has a mixed polarity.

6. Conclusion and future work

We described a new corpus for SA developed within the context of a platform for the detection of happiness. The development resulted in both a data set for system training and testing (i.e. Tweets on which we achieved the agreement of the annotators), but it also provides the basis for a framework to capture and analyze the nature of the disagreement (i.e. Tweets on which the disagreement reflects semantic ambiguity in the target instances and provides useful information). We propose a new type of ground truth, which is richer in diversity of perspectives and interpretations, and reflects more realistic human knowledge. Moreover, we propose a framework to exploit such diverse human responses to annotation tasks for analyzing and understanding disagreement.

7. References

Allisio, L., Mussa, V., Bosco, C., Patti, V., and Ruffo, G. (2013). Felicittà: Visualizing and estimating happiness in Italian cities from geotagged Tweets. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096, pages 95–106. CEUR-WS.org.

Baldoni, M., Baroglio, C., Patti, V., and Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54.

Baldwin, T. (2012). Social media: friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 58–59.

Basile, V. and Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Bertola, F. and Patti, V. (2013). Emotional responses to artworks in online collections. In *UMAP Workshops*, volume 997 of *CEUR Workshop Proceedings*.

Bolioli, A., Salamino, F., and Porzionato, V. (2013). Social media monitoring in real life with blogmeter platform. In *ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 156–163. CEUR-WS.org.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 44(10):91–94.

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

Caselli, T., Russo, I., and Rubino, R. (2012). Assigning connotation values to events. In *Proc. of the 8th Language Resources and evaluation Conference, LREC'12*, pages 3082–3089.

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

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, 55(1):34–43.

Esuli, A., Baccianella, S., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh Language Resources and Evaluation Conference, LREC'10*. ELRA.

Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hao, Y. and Veale, T. (2010). An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.*, 20(4):635–650, November.

Helliwell, J., Layard, R., and Sachs, J. (2014). *World Happiness Report 2013*. UN Sustainable Development Solutions Network.

Mihalcea, R. and Liu, H. (2006). A corpus-based aproach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. ACL.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proc. of Int. Conference on Global WordNet*.

Poesio, M. and Artstein, R. (2008). Inter-coder agreement

for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Quercia, D., Crowcroft, J., Ellis, J., and Capra, L. (2012). Tracking "gross community happiness" from tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work)*, pages 965–968.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proc. of EMNLP*, pages 704–714. ACL.

Saif, H., Fernández, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 9–21. CEUR-WS.org.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA.

Strapparava, C., Stock, O., and Mihalcea, R. (2011). Computational humour. In Cowie, R., Pelachaud, C., and Petta, P., editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 609–634. Springer-Berlin.

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing, CICLing 2013*, volume 7817 of *LNCS*, pages 121–136. Springer.

Versley, Y. (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-)reference. In *Proceedings of ESSLI'06*.

Xia, F. and Yetisgen-Yldiz, M. (2012). Clinical corpus annotation: challenges and strategies. In *LREC 2012 Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*.