Towards an Italian Lexicon for Polarity Classification (polarITA): a Comparative Analysis of Lexical Resources for Sentiment Analysis

Delia Irazú Hernández Farías
PRHLT Research Center
Universitat Politècnica de València
dhernandez1@dsic.upv.es

Irene Laganà
Dipartimento di
Studi Umanistici
Università di Pavia
irene.lagana01@universitadipavia.it

Viviana Patti, Cristina Bosco
Dipartimento di Informatica
Università di Torino
{patti,bosco}@di.unito.it

Abstract

English. The paper describes a preliminary study for the development of a novel lexicon for Italian sentiment analysis, i.e. where words are associated with polarity values. Given the influence of sentiment lexica on the performance of sentiment analysis systems, a methodology based on the detection and classification of errors in existing lexical resources is proposed and an extrinsic evaluation of the impact of such errors is applied. The final aim is to build a novel resource from the filtering applied to the existing lexical resources, which can integrate them with missing lexical entries and more reliable associations of polarity with entries.

Italiano. *L'articolo descrive uno studio preliminare per lo sviluppo di una nuova risorsa lessicale per la sentiment analysis in italiano, i.e. dove alle parole sono associati valori di polarità. Data l'influenza dei lessici di sentiment sulle performance dei sistemi di sentiment analysis, viene proposta una metodologia basata sulla rilevazione e classificazione degli errori presenti nei lessici attualmente disponibili ed una valutazione estrinseca dell'impatto di tali errori sui sistemi. L'obiettivo finale è ottenere un nuovo lessico grazie ad un filtraggio applicato alle risorse lessicali disponibili, e a un'integrazione con le voci lessicali mancanti, ottenendo una maggiore affidabilità nell'associazione delle polarità alle voci.*

1 Introduction

Sentiment Analysis (SA), described as the task of automatically determine the polarity in a given piece of text (Mohammad, 2016), is currently among the most widely investigated topics within NLP. Overall, the approaches for addressing such task are mainly based on techniques ranging from traditional machine learning to novel deep learning ones, as it can be seen also in the context of shared tasks on sentiment polarity classification in Twitter recently proposed, respectively for English (Nakov et al., 2016) and Italian (Barbieri et al., 2016), within the SemEval and Evalita periodical evaluation campaigns. Moreover, the detection of specific words associated with polarity values or emotions has been considered as a powerful information source for identifying the sentiment behind a text. Among the resources which are more commonly exploited by SA systems for performing their task there are therefore sentiment lexica, i.e., lists of words with associated polarity values or emotions.

Several techniques have been applied for the development of lexical resources for SA: they can be built from scratch, manually or automatically, or extracted from corpora (Nissim and Patti, 2017). Nevertheless, the vast majority of these resources are written in English, and a lack of resources currently features several other languages. One of the most commonly applied alternatives for having resources in language other than English is to automatically translate some available English lexicon via tools such as Google translate[1]. But there are many constraints involved in this kind of process, such as handling synonyms and polysemous words, multi-word expressions, but also to deal with cultural differences between source and target language. Apart from this, possible variations of polarity across different contexts and languages should be carefully taken into account, while such approaches rely somehow on the assumption that affective norms related to sentiment are stable across languages.

---

[1] https://translate.google.com/

In this paper we are interested into evaluate the reliability of the lexical resources currently available for Italian SA and, providing that the most of them are obtained by translation, we will mainly focus on the reliability of automatically translating English resources to Italian language. For doing so, we carried out a methodology involving different facets. Our final aim is to develop a new SA resource for Italian, which comprises pre-existing translated lexical entries enriched with the manual correction of the polarity assigned, as resulting from our analysis, but also includes entries which are featured by a polarity but are missing in the available lexica.

The paper is organized as follows. In the next section, we describe our methodology which mainly consists in three steps: the selection of a sample of tweets from an Italian sentiment corpus and exploited as part of the gold standard in the Sentipolc@Evalita2016 shared task (Stranisci et al., 2016; Barbieri et al., 2016); automatic extraction of the lexical entries polarized according to a set of benchmark sentiment lexica for Italian; the analysis of these entries and the comparison with those expected by a human judge. Section three shows instead an extrinsic evaluation of the impact of the detected errors on the results of the SA system. Some hints about future development of this research are given in the conclusion.

2 Our Methodology

Given the relevance of affective lexica in SA and related tasks, our major aims in the current research are to detect the limits of the currently available lexical resources for Italian and to explore the possibility to develop a novel resource by correcting and extending them. In this paper we focus in particular on the detection of the deficiencies of existing resources and on their motivations. Our methodology consists therefore in: (i) selecting of a sample of tweets from an Italian sentiment corpus featured by political contents (Stranisci et al., 2016) and exploited as part of the gold standard in the Sentipolc@Evalita2016 shared task (Barbieri et al., 2016), with sentiment polarity annotation at the tweet level; (ii) automatically extracting the lexical entries polarized according to a set of benchmark sentiment lexica for Italian and (iii) manually checking the results for each expected lexical entry in the context of the whole tweet (i.e. if the polarity of the entry is that

expected by a human annotator or also if there are other entries in the tweet that should appear as polarized but are not in the lexicons).

We take as starting point the SA lexica exploited by (Hernández Farías et al., 2014) in the IRADABE system at Evalita2014's SENTIPOLC (Basile et al., 2014). The same resources where used also in the upgraded system that participated at the same task in Evalita2016 (Buscaldi and Hernández Farías, 2016).

In those works the lexicon AFINN, (Nielsen, 2011), the one developed by Hu and Liu (henceforth HaL) (Hu and Liu, 2004), and SentiWordNet (SWN) (Baccianella et al., 2010) were indeed automatically translated to Italian, to exploit obtained information as features in their supervised system, but no specific evaluation or refining of them was performed. In the present paper we extend our selection by considering, beyond these three, a further resource, i.e. Sentix (Basile and Nissim, 2013) (see Sec. 2.1) which has been developed following a semantics oriented strategy (see Sec. 2.1). Henceforth, we will use the expression *benchmark lexica*) for referring to the four resources. As reference corpus, we considered, instead, TwBuonaScuola (Stranisci et al., 2016), an Italian dataset manually annotated for sentiment polarity and irony, focused on the on-line debate regarding a controversial Italian political reform, which is part of the gold standard provided for the Sentipolc shared task (Barbieri et al., 2016) at Evalita 2016 (Basile et al., 2017).

Our methodology, whose results are shown in Sec. 2.2, includes the steps described below.

Given a random selection of 500 tweets from TwBuonaScuola (henceforth *ItalianTweets*) including 2,706 different words, we manually evaluated the coverage of the *benchmark lexica* for the words included in these tweets. In particular, for each tweet we extracted automatically all the words which are included in each of the *benchmark lexica* and its associated polarity.

Then, for each tweets belonging to *ItalianTweets*, we manually checked the obtained lists of words, considered in the context of the tweet, with a two-fold objective:

(i) To deduce which words in the *benchmark lexica* have a wrong polarity associated;

(ii) To identify those words that express certain polarity in the corpus but are not included in the *benchmark lexica*.

2.1 Sentiment Analysis Resources

In this section we describe the *benchmark lexica*.

AFINN (Nielsen, 2011) is an English lexicon composed of 2,477 words and 15 multi-word expressions. Each entry is associated with a score which varies from -5 to +5 in order to respectively introduce negative and positive polarity. The starting point for the development of this resource is a list of obscene words and some positive words; then the lexicon has been extended with words from a corpus of tweets and other lists of words from Urban Dictionary[2] for representing entries typical of Internet language (e.g. "WTF" and "LOL"). After the manual annotation of the entries the lexicon has been evaluated based on a corpus of tweets manually annotated for SA.

HaL, (Hu and Liu, 2004), has been built within a project for developing methods to deal with opinions expressed in reviews about various kinds of goods. A group of 30 adjectives featured by a single and stable polarity and manually annotated has been expanded by including the words which in WordNet's synsets are synonyms or antonyms of these seeds, providing that synonyms are featured by the same polarity and antonyms by the opposite one. The lexicon currently includes 6,800 entries classified as positive or negative.

SentiWordNet 3.0 (Baccianella et al., 2010) is among the larger and more used resources exploited for SA. The main goal of the SentiWordNet project is the fully automated annotation of the polarity of the WordNet's synsets using scores that vary from 0.0 to 1.0 to each of the three basic polarity values (positive, negative, neutral) in order to obtain 1 as the sum of them. By contrast with the other resources, SentiWordNet takes into account different possible senses for each word.

As far as Italian is concerned, only a few resources exist, such as Sentix (Basile and Nissim, 2013) and SABRINA (Borzì et al., 2015). Sentix is the result of the alignment of four semantic database, namely WordNet (Fellbaum, 1998), SentiWordNet, MultiWordNet (Pianta et al., 2002) and Babelnet (Navigli and Ponzetto, 2012). The methodology consists in transferring to the Italian section of WordNet the information about polarity encoded in the English SentiWordNet's synsets, thus aligning Italian and English synsets. The development of SABRINA instead is based on the application of a prior polarity method on two sets of Italian words, the first composed of 277,000 entries with associated inflexion. However the lexicon is not publicly available.

Finally let us mention ItEM (Passaro et al., 2015), an Italian emotive lexicon which aims at offering information about affect expressed in text according to finer levels of granularity, i.e. referring not simply to positive or negative sentiment polarity but to emotional categories. In ItEM each word is tagged with an emotional label from the height basic emotions of the Plutchik's psychological model (Plutchik, 1980).

Several scholars are devoting their efforts to the development of resources for other languages, by applying translation or other methodologies. Let us cite e.g. FEEL (Abdaoui et al., 2017), a French lexicon where words are associated with polarity and emotions obtained thanks to the application of translation tools to NRC-EmoLEx[3] and a manual validation of results.

2.2 Qualitative Analysis of *Benchmark Lexica*

In order to detect the coverage and correctness of each *benchmark lexicon*, we selected from our reference sample corpus the list of words that according to a human judge are featured by some affective value in the context of the tweet where they appear. Then, for each entry of this list and for each *benchmark lexicon*, we observed if the word is represented in the resource and featured by the same polarity.

Given the preliminary nature of this investigation only a couple of researchers have been involved in the task. Moreover, a further limit of our current research approach depends on the reference to a given context (that determined by our sample corpus); issues related to the context will be accounted for in future investigations.

We observed different coverages of the *benchmark lexica* on our Twitter corpus, first of all in terms of numbers of affective words occurring in the tweets for each lexicon. The full vocabulary of the tweets is composed of 2,706 different words. Only some of these words are featured by some affective value, and focusing on them only we observed the following occurrences: 160 words in AFINN, 190 words in HaL, 302 words in SWN and 551 in Sentix. These word sets are partially overlapped, since 69 words are included in all the

---

[2]http://www.urbandictionary.com

[3]http://www.saifmohammad.com/WebPages/lexicons.html

lexica.

Resource	Error			
	(i)	(ii)	(iii)	(iv)
AFINN	1.2	2.5	16.8	8.7
HaL	1.5	1.0	12.6	12.6
SWN	5.9	1.6	15.5	13.2
Sentix	5.9	2.1	15.2	16.6

Table 1: Distribution of different errors in the *benchmark lexica* (percentage wrt the coverage of the lexicon).

The total amount of words missing or with an attributed erroneous polarity in the *benchmark lexica* is 388. As far as the erroneous polarization concerns, as summarized in Table 1, these words are featured by four different kinds of errors: (i) a positive word is annotated as negative; (ii) a negative word is annotated as positive; (iii) a neutral[4] word is annotated as positive; and (iv) a neutral word is annotated as negative. The values are expressed in percentage with respect to the coverage of the lexica. As far as the distribution of errors in the four classes, they are for all lexica prevailingly distributed in the last two classes, i.e. iii and iv, laying foundation for the hypothesis that in the automatic transition between English and Italian several non (clearly) polarized Italian words were instead polarized.

Nevertheless, observing Table 1, we can see also that all the lexica are featured by very similar amounts of errors, regardless of the methodology applied for their development (i.e. translation or extraction from semantic databases). Several errors, in particular for what concerns the polarity associated to specific words, can be generated during translation, and a portion of them is therefore motivated by the application of translation tools mainly because they do not consider context where each word occurs. But observing the results extracted from Sentix, which is not obtained simply by translation, and weighting the larger coverage that features this resource, we can see that errors occurs in a percentage that positively compares with that of the other resources. In this case the problem probably depends on misalignment of synsets for different languages. For example, the Italian word "istituto", whose meaning can be

---
[4]We considered neutral a word which is featured by a polarity which may vary across contexts, indicated by None in Table 2.

"school" or "institution", is aligned with "prison" and "house/prison", with a negative polarity which is not appropriate for the Italian word.

Several errors could be probably avoided in the transition among languages by applying a pre-processing including Part of Speech tagging and considering the grammatical category of the source and target terms. See for instance, the word *tagliando* (cutting) that occurs in the corpus as a Verb and in the *benchmark lexica* is instead aligned with the corresponding noun with the meaning of voucher/coupon. This motivates our decision about the attribution of PoS tags to the words in the first nucleus of a novel resource obtained by extending and correcting the existing ones. The overall impression is that, a manual check, even is a very time-consuming task, is always necessary and unavoidable, both when the new lexicon is obtained by translation, and when it is obtained relying on synset alignment.

3 Lost in Translation: Impact of the Errors

The methodology even if applied on a small set of tweets and based on a manual check of the *benchmark lexica*, confirms the hypothesis that many directions can be followed to improve the quality of existing lexical resources. The first result of this preliminary analysis is the collection of a list of words with associated polarity which will be the nucleus of the novel resource, i.e. polarITA. Each of the words in polarITA has been annotated with an overall polarity value (i.e., positive, negative, or none), and its corresponding Part-Of-Speech (POS) label. Table 2 summarizes the distribution of the words in polarITA in terms of polarity and POS labels.

Experiments on a larger corpus and a quantitative analysis based on a more formal classification of errors is needed for the development of a fully developed reliable lexical resource, together with an in-depth investigation of the relevance of context in the attribution of polarity, which is a very important issue. A comparison of the results that a given SA engine exploiting features extracted from sentiment lexica, for instance IRADABE (Hernández Farías et al., 2014; Buscaldi and Hernández Farías, 2016), obtains using each of the *benchmark lexica* and using polarITA is planned as future work for the evaluation of the novel lexicon, which is not currently suitable because the

limited size of our reference corpus and the consequent partial coverage of errors.

Considering the current preliminary stage of development of polarITA, we tried an extrinsic evaluation for detecting the impact on the performance of SA systems of the errors currently featuring the *benchmark lexica* and corrected in the novel lexicon. We compared the words which are missing or assigned to erroneous polarity in the *benchmark lexica* with the Italian words more commonly used and understood by native speakers, whose collection is available in the *Vocabolario di base della lingua italiana (vocItalian)*[5] recently newly released. Like the first version of this resource, published in 1980, (De Mauro, 1980), it includes three word classes: 2,999 High Usage words (HU), 2,231 High Availability words (HA) and 1,979 Foundational words (FO). In polarITA we collected until now 284 words of the vocItalian, whose distribution across the three classes is shown in Table 2. Among the words in the FO category we found "bene" (good), "mentire" (lie), and "giustizia" (justice). While words like "assassino" (killer), "preoccupato" (worried), and "entusiasta" (enthusiastic) are part of the HU category. Finally, in the HA category it is possible to find words such as "dannoso" (harmful) and "emozionante" (exciting).

This analysis suggests some hints for further investigation, showing that the failures of lexica currently available for Italian SA affect words very commonly used in communication and therefore the improvement of these resources may hopefully result in an advancement for SA and related tasks.

4 Conclusions and Future Work

In this paper we propose the preliminary investigation about a methodology for the development of a novel lexical resource for Italian SA, namely polarITA, which takes advantage of the analysis and filtering of errors occurring in the available lexical resources. We carried out a manual analysis of a set of tweets for determining the reliability of sentiment-related lexica, showing that, even if the transfer of lexical information between two different languages is a common practice to address the lack of resources, information related to sentiment is lost during it. The identified errors are then ex-

Total words	388	
Polarity		
Positive	*Negative*	*None*
225	140	23
Part-of-speech labels		
Adjective		84
Adjective/Noun		1
Adjective/Pronoun		2
Adverb		16
Interjection		3
Noun		187
Noun/Adverb		1
Preposition		1
Pronoun		1
Verb		92
vocItalian		
FO		187
HU		86
HA		11

Table 2: Distribution of the words in polarITA in terms of polarity, POS labels, and vocItalian.

ploited as a starting point for developing the novel resource.

As future work, we are planning to extend the resource in several directions: by investigating multi-word expressions, extending the coverage to a larger corpus, exploring the impact of figurative language devices such as irony and sarcasm in the use of certain polarized words (Hernández Farías et al., 2016). Moreover, our future effort will be oriented to the automatization of a larger part of the methodology and its application to other languages currently under resourced.

Acknowledgements

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. FEEL: a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51:833–855, September.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lex-

---

[5]`https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana`

ical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Basile, Cutugno, Nissim, Patti, and Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings.

Valerio Basile and Malvina Nissim. 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, USA. Association for Computational Linguistics.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa, Italy.

Pierpaolo Basile, Francesco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2017. Evalita goes social: Tasks, data, and community at the 2016 edition. *IJCoL - Italian Journal of Computational Linguistics*, 3(1):93–127.

Valeria Borzì, Simone Faro, Arianna Pavone, and Sabrina Sansone. 2015. Prior Polarity Lexical Resources for the Italian Language. *CoRR*, abs/1507.00133.

Davide Buscaldi and Delia Irazú Hernández Farías. 2016. IRADABE2: Lexicon Merging and Positional Features for Sentiment Analysis in Italian. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. aAcademia University Press.

Tullio De Mauro. 1980. *Guida all'uso delle parole Num. 3 dei Libri di base*. Editori Riuniti, Roma.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Delia Irazú Hernández Farías, Davide Buscaldi, and Belém Priego-Sánchez. 2014. IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task. In *First Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA 2014*, pages 75–81.

Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Saif M. Mohammad. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, Heraklion, Crete, Greece. CEUR-WS.org.

Malvina Nissim and Viviana Patti. 2017. Semantic aspects in sentiment analysis. In Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 31–48. Morgan Kaufmann, Boston.

Lucia Passaro, Laura Pollacci, and Alessandro Lenci. 2015. ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon. volume II.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of International Conference on Global WordNet*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pages 3–33. Academic press, New York.

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).