IRADABE2: Lexicon Merging and Positional Features for Sentiment Analysis in Italian

Davide Buscaldi
LIPN, Université Paris 13
Villetaneuse, France
buscaldi@lipn.univ-paris13.fr

Delia Irazú Hernandez-Farias
Dipartimento di Informatica
Università degli studi di Torino
Turin, Italy

PRHLT group
Universitat Politècnica de València
Valencia, Spain
dhernandez1@dsic.upv.es

Abstract

English. This paper presents the participation of the IRADABE team to the SENTIPOLC 2016 task. This year we investigated the use of positional features together with the fusion of sentiment analysis resources with the aim to classify Italian tweets according to subjectivity, polarity and irony. Our approach uses as starting point our participation in the SENTIPOLC 2014 edition. For classification we adopted a supervised approach that takes advantage of support vector machines and neural networks.

Italiano. *Quest'articolo presenta il lavoro svolto dal team IRADABE per la partecipazione al task SENTIPOLC 2016. Il lavoro svolto include l'utilizzo di caratteristiche posizionali e la fusione di lessici specialistici, finalizzato alla classificazione di tweet in italiano, secondo la loro soggettività, polarità ed ironia. Il nostro approccio si basa sull'esperienza acquisita nel corso della partecipazione all'edizione 2014 di SENTIPOLC. Per la classificazione sono stati adottati dei metodi supervisionati come le macchine a supporto vettoriale e le reti neurali.*

1 Introduction

Sentiment analysis (SA) related tasks have attracted the attention of many researchers during the last decade. Several approaches have been proposed in order to address SA. Most of them have in common the use of machine learning together with natural language processing techniques. Despite all those efforts there still many challenges left such as: multililngual sentiment analysis, i.e, to perform SA in languages different from English (Mohammad, 2016). This year for the second time a sentiment analysis task on Italian tweets has been organized at EvalIta, the Sentiment Polarity Classification (SENTIPOLC) task (Barbieri et al., 2016).

In this paper we study the effect of positional features over the sentiment, irony and polarity classification tasks in the context of SENTIPOLC 2016 task. We propose a revised version of our IRADABE system (Hernandez-Farias et al., 2014), which participated with fairly good results in 2014. The novelties for this participation are not only in the positional features, but also in a new sentiment lexicon that was built combining and expanding the lexicons we used in 2014.

The rest of the paper is structured as follows: in Section 2 we describe the steps we took to build an enhanced sentiment dictionary in Italian from existing English resources; in Section 3 we describe the new positional features of the IRADABE system.

2 Building a unified dictionary

In sentiment analysis related tasks, there are several factors that can be considered in order to determine the polarity of a given piece of text. Overall, the presence of positive or negative words is used as a strong indicator of sentiment. Nowadays there are many sentiment analysis related resources that can be exploited to infer polarity from texts. Recently, this kind of lexicons has been proven to be effective for detecting irony in Twitter (Hernańdez Farías et al., 2016). Unfortunately, the majority of available resources are in English. A common practice to deal with the lack of resources in different languages is to automatically translate it from English.

However, the language barrier is not the only drawback for these resources. Another issue is

the limited coverage of certain resources. For instance, AFINN (Nielsen, 2011) includes only 2477 words in its English version, and the Hu-Liu lexicon (Hu and Liu, 2004) contains about 6800 words. We verified on the SENTIPOLC14 training set that the Hu-Liu lexicon provided a score for 63.1% of training sentences, while the coverage for AFINN was of 70.7%, indicating that the number of items in the lexicons is not proportional to the expected coverage; in other words, although AFINN is smaller, the words included are more frequently used than those listed in the Hu-Liu lexicon. The coverage provided by a hypothetical lexicon obtaining by the combination of the two resources would be 79.5%.

We observed also that in some cases these lexicons provide a score for a word but not for one of their synonyms: in the Hu-Liu lexicon, for instance, the word 'repel' is listed as a negative one, but 'resist', which is listed as one of its synonym in the Roget's thesaurus[1], is not. SentiWordNet (Baccianella et al., 2010) compensates some of the issues; its coverage is considerably higher than the previously named lexicons: 90.6% on the SENTIPOLC14 training set. Its scores are also assigned to synsets, and not words. However, it is not complete: we measured that a combination of SentiWordNet with AFINN and Hu-Liu would attain a coverage of 94.4% on the SENTIPOLC14 training set. Moreover, the problem of working with synsets is that it is necessary to carry out word sense disambiguation, which is a difficult task, particularly in the case of short sentences like tweets. For this reason, our translation of SentiWordNet into Italian (Hernandez-Farias et al., 2014) resulted in a word-based lexicon and not a synset-based one.

Therefore, we built a sentiment lexicon which was aimed to provide the highest possible coverage by merging existing resources and extending the scores to synonyms or quasi-synonyms. The sentiment lexicon was built following a three-step process:

1. Create a unique set of opinion words from the AFINN, Hu-Liu and SentiWordNet lexicons, and merge the scores if multiple scores are available for the same word; the original English resources were previously translated into the Italian language for our participation in SENTIPOLC 2014;

2. Extend the lexicon with the WordNet synonyms of words obtained in step 1;

3. Extend the lexicon with pseudo-synonyms of words obtained in step 1 and 2, using word2vec for similarity. We denote them as "pseudo-synonyms" because the similarity according to word2vec doesn't necessarily means that the words are synonyms, only that they usually share the same contexts.

The scores at each step were calculated as follows: in step 1, the weight of a word is the average of the non-zero scores from the three lexicons. In step 2, the weight for a synonym is the same of the originating word. If the synonym is already in the lexicon, then we keep the most polarizing weight (if the scores have the same sign), or the sum of the weights (if the scores have opposed signs). For step 3 we previously built semantic vectors using word2vec (Mikolov et al., 2013) on the ItWaC[2] corpus (Baroni et al., 2009). Then, we select for each word in the lexicon obtained at step 2 the 10 most similar pseudo-synonyms having a similarity score ≥ 0.6. If the related pseudo-synonym already exists in the lexicon, its score is kept, otherwise it is added to the lexicon with a polarity resulting from the score of the original word multiplied by the similarity score of the pseudo-synonym. We named the obtained resource the 'Unified Italian Semantic Lexicon', shortened as UnISeLex. It contains $31,601$ words. At step 1, the dictionary size was $12,102$; at step 2, after adding the synonyms, it contained $15,412$ words.

In addition to this new resource, we exploited *labMT-English words*. It is a list (Dodds et al., 2011) composed of 10,000 words manually annotated with a happiness measure in a range between 0 up to 9. These words were collected from different resources such as Twitter, Google Books, music lyrics, and the New York Times (1987 to 2007).

3 Positional Features

It is well known that in the context of opinion mining and summarization the position of opinion words is an important feature (Pang and Lee, 2008), (Taboada and Grieve, 2004). In reviews,

[1] http://www.thesaurus.com/Roget-Alpha-Index.html

[2] http://wacky.sslmit.unibo.it

users tend to summarize the judgment in the final sentence, after a comprehensive analysis of the various features of the item being reviewed (for instance, in a movie review, they would review the photography, the screenplay, the actor performance, and finally provide an overall judgment of the movie). Since SENTIPOLC is focused on tweets, whose length is limited to 140 characters, there is less room for a complex analysis and therefore it is not clear whether the position of sentiment words is important or not.

In fact, we analyzed the training set and noticed that some words tend to appear in certain positions when the sentence is labelled with a class rather than the other one. For example, in the subjective sub-task, 'non' (not), 'io' (I), auxiliary verbs like 'potere' (can), 'dovere' (must) tend to occur mostly at the beginning of the sentence if the sentence is subjective. In the positive polarity subtask, words like 'bello' (beautiful), 'piacere' (like) and 'amare' (love) are more often observed at the beginning of the sentence if the tweet is positive.

We therefore introduced a positional Bag-of-Words (BOW) weighting, where the weight of a word t is calculated as:

$$w(t) = 1 + pos(t)/len(s)$$

where $pos(t)$ is the *last* observed position of the word in the sentence, and $len(s)$ is the length of the sentence. For instance, in the sentence "I *love* apples in fall.", $w(love) = 1 + 1/5 = 1.2$, since the word *love* is at position 1 in a sentence of 5 words.

The Bag of Words was obtained by taking all the lemmatized forms w that appeared in the training corpus with a frequency greater than 5 and $I(w) > 0.001$, where $I(w)$ is the informativeness of word w calculated as:

$$I(w) = p(w|c^+)\left(\log(p(w|c^+)) - \log(p(w|c^-))\right)$$

where $p(w|c^+)$ and $p(w|c^-)$ are the probabilities of a word appearing in the tweets tagged with the positive or negative class, respectively. The result of this selection consisted in 943 words for the $subj$ subtask, 831 for pos, 991 for neg and 1197 for iro.

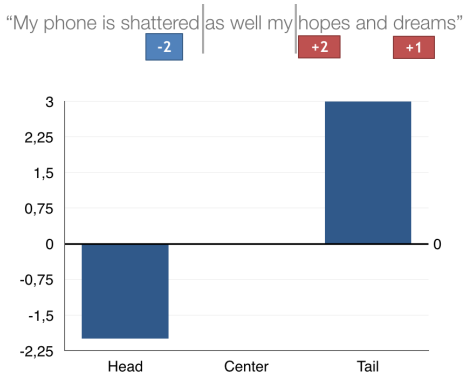
The results in Table 3 show a marginal improvement for the polarity and irony classes, while in subjectivity the system lost 2% in F-measure. This is probably due to the fact that the important words that tend to appear in the first part of the sentence may repeat later, providing a wrong score for the feature.

	Subj	***Pol(+)***	***Pol(-)***	***Iro***
pos. BOW	0.528	0.852	0.848	0.900
std. BOW	0.542	0.849	0.842	0.894

Table 1: F-measures for positional and standard BOW models trained on the train part of the dev set; results are calculated on the test part of the dev set.

With respect to the 2014 version of IRADABE, we introduced 3 more position-dependent features. Each tweet was divided into 3 sections, *head*, *centre* and *tail*. For each section, we consider the sum of the sentiment scores of the included words as a separate feature. Therefore, we have three features, named in Table 3.1 as $headS$, $centreS$ and $tailS$.

Figure 1: Example of lexicon positional scores for the sentence "My phone is shattered as well my hopes and dreams".



3.1 Other features

We renewed most of the features used for SENTIPOLC 2014, with the main difference that we are now using a single sentiment lexicon instead than 3. In IRADABE 2014 we grouped the features into two categories: *Surface Features* and *Lexicon-based Features*. We recall the ones appearing in Table 2, directing the reader to (Hernandez-Farias et al., 2014) for a more detailed description. The first group comprises features such as the presence of an URL address ($http$), the length of the tweet ($length$), a list of swearing words ($taboo$), and the ratio of uppercase characters ($shout$). Among the features extracted from dictionaries, we used the sum of polarity scores ($polSum$), the sum of only negative or pos-

itive scores ($sum(-)$ and $sum(+)$), the number of negative scores ($count(-)$) on UnISeLex, and the average and the standard deviation of scores on labMT (avg_{labMT} and std_{labMT}, respectively). Furthermore, to determine both polarity and irony, a subjectivity indicator (*subj*) feature was used; it is obtained by identifying first if a tweet is subjective or not. Finally, the $mixed$ feature indicates is the tweet has mixed polarity or not.

Subj	***Pol(+)***	***Pol(-)***	***Iro***
$http$	$subj$	$subj$	$subj$
$shout$	avg_{labMT}	$sum(-)$	$http$
$sum(-)$	'grazie'	$count(-)$	'governo'
$count(-)$	$smileys$	avg_{labMT}	$mixed$
$headS$	$polSum$	$length$	$shout$
$pers$	$http$	$polSum$	'Mario'
'!'	'?'	$http$	'che'
avg_{labMT}	$sum(+)$	$centreS$	'#Grillo'
'mi'	'bello'	$taboo$	$length$
$taboo$	'amare'	std_{labMT}	$sum(-)$

Table 2: The 10 best features for each subtask in the training set.

4 Results and Discussion

We evaluated our approach on the dataset provided by the organizers of SENTIPOLC 2016. This dataset is composed by up to 10,000 tweets distributed in training set and test set. Both datasets contain tweets related to political and socio-political domains, as well as some generic tweets[3].

We experimented with different configurations for assessing subjectivity, polarity and irony. We sent two runs for evaluation purposes in SENTIPOLC-2016:

- *run 1*. For assessign the subjectivity label a Tensorflow[4] implementation of Deep Neural Ngetwork (DNN) was applied, with 2 hidden layers with 1024 and 512 states, respectively. Then, the polarity and irony labels were determined by exploiting a SVM[5].

- *run 2*. In this run, the bag-of-words were revised to remove words that may have a different polarity depending on the context (. Classification was carried out using a SVM (radial basis function kernel) for all subtasks, including $subj$.

From the results, we can observe that the DNN obtained an excellent precision (more than 93%) in $subj$, but the recall was very low. This may indicate a problem due to the class not being balanced, or an overfitting problem with the DNN, which is plausible given the number of features. This may also be the reason for which the SVM performs better, because SVMs are less afflicted by the "curse of dimensionality".

run 1				
	Subj	***Pol(+)***	***Pol(-)***	***Iro***
Precision	0.9328	0.6755	0.5161	0.1296
Recall	0.4575	0.3325	0.2273	0.0298
F-Measure	0.6139	0.4456	0.3156	0.0484
run 2				
	Subj	***Pol(+)***	***Pol(-)***	***Iro***
Precision	0.8714	0.6493	0.4602	0.2078
Recall	0.6644	0.4377	0.3466	0.0681
F-Measure	0.7539	0.5229	0.3955	0.1026

Table 3: Official results of our model on the test set.

5 Conclusions

As future work, it could be interesting to exploit the labels for exact polarity as provided by the organizers. This kind of information could help in some way to identify the use of figurative language. Furthermore, we are planning to enrich IRADABE with other kinds of features that allow us to cover more subtle aspects of sentiment, such as emotions. The introduction of the "happiness score" provided by labMT was particularly useful, with the related features being critical in the subjectivity and polarity subtasks. This motivates us to look for dictionaries that may express different feelings than just the overall polarity of a word. We will also need to verify the effectiveness of the resource we produced automatically with respect to other hand-crafted dictionaries for the Italian language, such as Sentix (Basile and Nissim, 2013)

We plan to use a more refined weighting scheme for the positional features, such as the locally-weighted bag-of-words or LOWBOW (Lebanon et

[3]Further details on the datasets can be found in the task overview (Barbieri et al., 2016)

[4]http://www.tensorflow.org

[5]As in IRADABE-2014 version, the subjectivity label influences the determination of both the polarity values and the presence of irony.

al., 2007), although it would mean an increase of the feature space of at least 3 times (if we keep the head, centre, tail cuts), probably furtherly compromising the use of DNN for classification.

About the utility of positional features, the current results are inconclusive, so we need to investigate further about how the positional scoring affects the results. On the other hand, the results show that the merged dictionary was a useful resource, with dictionary-based features representing 25% of the most discriminating features.

Acknowledgments

This research work has been supported by the "Investissements d'Avenir" program ANR-10-LABX-0083 (Labex EFL). The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Faríias (Grant No. 218109/313683 CVU-369616).

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical 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, may. 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 Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele 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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta, Springer, and Science+business Media B. V. 2009. The wacky wide web: A collection of very large linguistically processed webcrawled corpora. language resources and evaluation.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *WASSA 2013*, Atlanta, United States, June.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6(12).

Irazú Hernandez-Farias, 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 EVALITA2014*, Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA2014, pages 75–81, Pisa, Italy, December.

Delia Irazú Hernańdez 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, July.

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, Seattle, WA, USA. ACM.

Guy Lebanon, Yi Mao, and Joshua Dillon. 2007. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, 8(Oct):2405–2441.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif M. Mohammad. 2016. Challenges in sentiment analysis. In *A Practical Guide to Sentiment Analysis*. Springer.

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.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161, Stanford, US.