10.12871/clicit2014213

IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task

Irazú Hernández Farias
Pattern Recognition and
Human Language Technology
Universitat Politècnica de València
Spain
dhernandez1@dsic.upv.es

Davide Buscaldi
Laboratoire d'Informatique de Paris Nord
CNRS (UMR 7030)
Université Paris 13, Sorbonne Paris Cité
France
buscaldi@lipn.univ-paris13.fr

Belém Priego Sánchez
Laboratoire de Lexiques, Dictionnaires, Informatique de Paris Nord
CNRS(UMR 7187)
Université Paris 13, Sorbonne Paris Cité
France
belemps@gmail.com

Abstract

English. Interest in the Sentiment Analysis task has been growing in recent years due to the importance of applications that may benefit from such kind of information. In this paper we addressed the polarity classification task of Italian tweets by using a supervised machine learning approach. We developed a set of features and used them in a machine learning system in order to decide if a tweet is subjective or objective. The polarity result itself was then used as an additional feature to determine whether a tweet contains ironical content or not. We faced the lack of resources in Italian by translating (mostly automatically) existing resources for the English language. Our model obtained good results in the SentiPolC 2014 task, being one of the best ranked systems.

Italiano. *L'interesse nell'analisi automatica dei sentimenti è continuamente cresciuto negli ultimi anni per via dell'importanza delle applicazioni in cui questo tipo di analisi può essere utilizzato. In quest'articolo descriviamo gli esperimenti portati a termine nel campo della classificazione di polarità di tweets scritti in italiano, usando un approccio basato sull'apprendimento automatico. Un certo numero di criteri è stato utilizzato come features per assegnare una polarità e quindi determinare se i tweets contengono dell'ironia o meno. Per questi esperimenti, la mancanza di risorse (in particolare di dizionari specializzati) è stata compensata adattando, in gran parte utilizzando delle tecniche di traduzione automatica, delle risorse esistenti per la lingua inglese. Il modello così ottenuto è stato uno dei migliori nel task SentiPolC a Evalita 2014.*

1 Introduction

Sentiment Analysis has been defined by (Liu, 2010) as "the computational study of opinions, sentiments and emotions expressed in text"; social media is a rich source of data that can be processed in order to detect subjectivity and classify the sentiments expressed by users. The effective extraction of such information is the main challenge in this research field. Sentiment analysis is an opportunity for researchers in Natural Language Processing (NLP) to make tangible progress on all fronts of NLP, and potentially have a huge practical impact. (Cambria et al., 2013)

In this paper we describe our participation to the SentiPolC task in polarity and irony classification of tweets in Italian. The paper is organized as follows: in Section 2 we briefly describe the related works in order to understand how they influenced our choices. In Section 3 we describe the features and the classification system used. Results obtained from our proposed model are shown in Section 4. Finally in Section 5 we draw some conclusions based on the early analysis of the results.

2 Related Work

Sentiment Analysis approaches are mainly based on machine learning and lexicons. There is a considerable amount of works related to sentiment analysis and opinion mining ((Liu, 2010), (Pang and Lee, 2008) in particular), all of them can be classified in one of the general approaches presented by Cambria et. al in (Cambria et al., 2013): keyword spotting, lexical affinity, statistical methods, and concept-based techniques. *Keyword spotting* consists in classifying text by affect categories based on the presence of unambiguous affect words such as *happy* , *sad*, *afraid*, and *bored*. *Lexical affinity* does not only detects obvious affect words, but also assigns to arbitrary words a probable "affinity" to particular emotions. *Statistical methods* are semantically weak, which means that individually — with the exception of obvious affect keywords — a statistical model's other lexical or co-occurrence elements have little predictive value. *Concept-based approaches*: relying on large semantic knowledge bases, such approaches step away from blindly using keywords and word co-occurrence counts, and instead rely on the implicit meaning/features associated with natural language concepts, superior to purely syntactical techniques; concept-based approaches can detect subtly expressed sentiments.

Respect to irony detection, Carvalho (Carvalho et al., 2009) developed a system able to detect irony using punctuation marks and emoticons in Portuguese. Veale and Hao (Veale and Hao, 2010) present a linguistic approach that takes into account the presence of heuristic clues in sentences (e.g. "about as" as indicator of ironic simile). Reyes et al. (Reyes et al., 2013) propose a model based on four dimensions (signatures, unexpectedness, style, and emotional scenarios) that support the idea that textual features can capture patterns used in this kind of utterances.

3 Features and Classification Framework

In order to address the tasks of subjectivity/polarity/ironic classification, we decide taking into account a statistical method that includes several features: structural, syntactical and lexicon based. We think that tweets belonging to the same class can share this kind of features, below we describe briefly each one. In parentheses, we provide the related id used in Table 4 and Table 5.

3.1 Surface Features

- *nGrams features.* We extracted the most frequent unigrams, bigrams and trigrams from the training corpus in order to have three different Bag of Words representations. This is a simple feature widely used in text classification. Only unigrams were finally used for our participation in SentiPolC.

- *Emoticons frequency.* (*emo*) By using emoticons, with few characters is possible to display one's true feeling. Emoticons are virtually required under certain circumstances in text-based communication, where the absence of verbal and visual cues can otherwise hide what was originally intended to be humorous, sarcastic, ironic, and some times negative (Wolf, 2000). We manually defined three different sets of emoticons for the detection of subjectivity, positiveness and negativeness, then we extracted the frequency of each one in tweets.

- *Negative Words frequency.* (*neg*) Handling negation can be an important concern in sentiment analysis, one of the main difficulties is that negation can often be expressed in a rather subtle way. We analyzed the training set and selected some words that triggers negation (*mai* (never), *non/no* (not/no)), avversative conjunction or adverbs (*invece* (instead), *ma* (but)). We extracted their frequency in each tweet. There are other ways to deal with negations, for example to reverse the polarity of the text if a negation word is found, but we did not employ this technique.

- *URL information frequency.* (*http*) We analyzed the training set and we found that most not-subjective, not-ironic tweets contained a hyperlink, so we decided to take into account this characteristic as a feature. In some cases this kind of information is also present in ironic tweets because users made an evaluation of some content (text, video, image, etc.) that they consider ironic and try to share with others in order to express themselves.

- *POS-based features.* (*pps*) We decided to use Part-of-speech (POS) tagging (the TreeTagger[1] implementation) to extract additional in-

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

formation to determine the subjectivity of tweets; in particular, we took into account the presence of verbs conjugated at the first and second persons (those endings in "-o", "-i", "-amo", "-ate/ete") and personal pronouns ("io", "tu", "noi", "voi", and their direct and indirect object versions).

- *Tweet Length and Uppercase ratio.* (*len*, *shout*) Although text in tweets only can contain maximum 140 characters, we decided to use the length in words of each tweet like a feature, trying to reflect the fact that ironic comments are often short. We took into account also the ratio between the uppercase words and length of the tweet, given that many subjective and/or ironic comments use uppercase words in order to express radical opinions about something, highlighting it with the use of uppercase.

3.2 Lexicon-based Features

Many state-of-the-art works are based on lexicons that assign to each words an empirical measure of their polarity. Most lexicons however are available only in English. We decided to use different lexicons and automatically translate them to Italian; a thoroughful description of each one is out of the scope of the present work and we refer the reader to the relative existing literature. We found that in some cases an Italian word can be translated in different ways in English. We tested on the dev set two possibilities: to keep for the Italian word the max of the scores of the English translations or their average. The test showed that the max allowed to obtain a slightly better accuracy than the average.

- *SentiWordNet* (*SWN*). Assigns to each synset of WordNet three sentiment scores: positivity, negativity and objectivity. We used only the positive and negative scores to derive six features: positive/negative words count (*SWN+/-c*), the sum of the positive scores in the tweet (*SWN+s*), the sum of negative scores in the tweet (*SWN-s*), the balance (positive-negative) score of the tweet (*SWNb*), and the standard deviation of SentiWN scores in the tweet (*SWNdev*).

- *Hu-Liu Lexicon*[2]. (*HL*) We derived three features from this lexicon: positive (*HL+c*) and negative (*HL-c*) words count, balance (sum of positive-negative words - *HLb*).

- *AFINN Lexicon*[3]. (*AF*) This lexicon contains two word lists labeled with polarity valences from -5 (negative) to +5 (positive). We derived 5 features from this lexicon: positive/negative word count (*AF+/-c*), sum of positive and negative scores (*AF+/-s*); overall balance of scores in the tweet (*AFb*).

- *Whissel Dictionary* (Whissell, 2009). (*WH*) Our translation of this lexicon contains 8700 Italian words with values of Activation, Imagery and Pleasantness related to each one. Range of scores go from 1 (most passive) to 3 (most active). We derived six features: average activation, imagery and pleasantness (*WH[aip]avg*), and the standard deviation of the respective scores (*WH[aip]dev*). We thought that an elevate score in one of these features may indicate an out-of-context word, thus indicating a possibly ironic comment.

- *Italian "Taboo Words"*. (*TAB*) Knowing the function of taboo words to trigger humor, catharsis, or to boost opinions (Zhou, 2010), we decided to use a list of taboo italian words that we extracted from Wiktionary[4].

- *Counter-Factuality* (Reyes et al., 2013). (*CF*) We use the frequency of discursive terms that hint at opposition or contradiction in a text such as "about" and "nevertheless".

- *Temporal Compression* (Reyes et al., 2013). (*TC*) We use the frequency of terms that identify elements related to opposition in time, i.e. terms that indicate an abrupt change in a narrative.

Moreover, in the irony subtask we used as features our results of the subjectivity (*subj*) and polarity (*pol*) classification subtasks.

3.3 Classification Framework

We used the nu-SVM (Schölkopf et al., 2000) implementation by LibSVM (Chang and Lin, 2011),

[2] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[3] https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

[4] http://it.wiktionary.org/wiki/Categoria:Parole_volgari-IT

with the nu parameter set to the standard value (0.5), with a RBF kernel. The classification was carried out in three steps: in the first one, the system classifies the tweet into subjective or not. The result of the subjectivity is passed as a feature to the second classification step that classifies the tweets as positive or negative. Finally, the results of subjectivity and polarity classification are passed to the final classifier that is used to detect irony. In the constrained run, we used the full SentiPolC training set (Basile et al., 2014). In the unconstrained run, we integrated into the training set 493 additional tweets that include the hashtag *#ironia* or were published on an ironical/satirical account (for instance, the *@spinozait* account[5]). We randomly subsampled the training set in order to obtain a balanced training set (with $50\%/50\%$ ratio for the ironic/not ironic tweets).

The additional tweets retrieved from *@spinozait* and those including the hashtag *#ironia* were automatically assigned the labels "1" for subjectivity and irony. The labels for polarity were automatically assigned using the model trained on the devset. This means that in some cases the combination of labels does not correspond to the labels allowed by the task guidelines (there are ironic tweets with mixed or neutral polarity). Therefore, we did not use the polarity information as feature for the unconstrained run.

4 Results

We evaluated our approach on the SentiPolC datasets, composed by approximately 4,000 italian tweets for training and 1,700 for test; each tweet on the training subset was labeled as objective/subjective, positive/neutral/negative/mixed, ironic/non-ironic and finally if the topic of the tweet was concern to politics. In Table 4 we show the results obtained on the training set using 10-fold cross validation. The official results are shown in Table 4 (Basile et al., 2014). The differences between the results obtained for the training and the test set can be explained by the fact that our system was not able to retrieve 186 tweets. Our evaluation on Weka on the partial set shows 80% F-measure in irony detection. However, we suppose that the other participants had similar problems. The results in Table 4 have been calculated only on the retrieved tweets of the training set.

[5] https://twitter.com/spinozait

	Subj	***Pol(+)***	***Pol(-)***	***Iro***
Precision	0.765	0.767	0.668	0.820
Recall	0.777	0.774	0.670	0.828
F-Measure	0.764	0.743	0.668	0.824

Table 1: Results of our model on training set

		Constrained			
		Subj	***Pol(+)***	***Pol(-)***	***Iro***
"1"	P	0.8284	0.7265	0.6822	0.2400
	R	0.7862	0.2998	0.5213	0.2521
	F-m	0.8067	0.4245	0.5910	0.2459
Comb F-m		0.6706	0.6347		0.5415

Table 2: Results of our model on test set Constrained Run (official results).

We carried out an analysis of the features using the information gain feature selection algorithm provided by Weka. We show in Table 4 and Table 5 the ten best dictionary-based features, in the test and training set respectively.

From these results we can see that SentiWordNet-based features worked very well in subjectivity prediction, more than features like the emoticons which we expected to have an important role. In the positive polarity task, emoticons were an important feature however, together with the positive word counts (or sum of positive scores) for AFINN, Hu-Liu and SentiWordNet lexicons. The respective negative word based features worked well also in the negative polarity prediction task. In the irony task we observed some discrepancies between the results obtained in the training set and those obtained in the test set. In fact, our intuition of finding "anomalies" using standard deviation of Whissell-based features worked particularly well in the training set, but we did not found the same kind of "anomalies" in the test set. In the test set we found instead a prevalence of features that

		Unconstrained			
		Subj	***Pol(+)***	***Pol(-)***	***Iro***
"1"	P	0.8955	0.4565	0.6266	0.2387
	R	0.5989	0.5556	0.5040	0.4202
	F-m	0.7178	0.5012	0.5587	0.3044
Comb F-m		0.6464	0.6108		0.5513

Table 3: Results of our model on test set Unconstrained Run(official results).

	Subj	*Pol(+)*	*Pol(-)*	*Iro*
1	*http*	*SWNb*	*SWN-s*	*subj*
2	*SWN+c*	*AFb*	*SWN-c*	*http*
3	*SWN-s*	*emo*	*HL-c*	*HL-c*
4	*SWN+s*	*AF+s*	*AF-s*	*pol*
5	*SWN-c*	*HLb*	*SWNb*	*AF-c*
6	*SWNdev*	*SWN+s*	*HLb*	*HLb*
7	*AFb*	*AF+c*	*AF-c*	*SWN-s*
8	*neg*	*WHidev*	*neg*	*AFb*
9	*AF+s*	*HL+c*	*CF*	*AF-s*
10	*pps*	*WHpdev*	*AFb*	*SWNb*

Table 4: Best ranked dictionary-based features for each subtask, according to their information gain values (test set).

	Subj	*Pol(+)*	*Pol(-)*	*Iro*
1	*http*	*AFb*	*SWN-s*	*subj*
2	*SWN+c*	*AF+s*	*AF-s*	*http*
3	*SWN+s*	*SWNb*	*HL-c*	*pol*
4	*SWNdev*	*emo*	*SWN-c*	*WHpdev*
5	*SWN-c*	*SWN+s*	*AF-c*	*WHadev*
6	*SWN-s*	*HLb*	*SWNb*	*WHidev*
7	*AFb*	*AF+c*	*AFb*	*len*
8	*SWNb*	*HL+c*	*SWNdev*	*SWN+c*
9	*AF+s*	*http*	*SWN+c*	*SWN-c*
10	*shout*	*len*	*HLb*	*TAB*

Table 5: Best ranked dictionary-based features for each subtask, according to their information gain values (training set).

indicates negative words (*HL-c*, *AF-c*, *SWN-s*, *AF-s*). In both train and test set we observed that the most important features that characterize irony were subjectivity and mixed polarity, while the presence of web addresses was a strong clue to the tweet being not ironic, or objective. The importance of web related features was indicated also by the high information gain of fragments of web addresses (not included in the tables), such as "http", "ly", "it", "fb", etc. Further analysis of the results showed that Italian politics have a great weight in the training set, with keywords like "governo" or "Monti" conveying a high predictive power.

5 Conclusions and Future Work

An analysis of the features using information gain showed that SentiWordNet was an important resource for the detection of subjectivity, and in general the translated lexicons were very useful. Many of the features related to the detection of web addresses were also very important, indicating that the training and test sets were flawed by the presence of such addresses. Finally, we noticed that the lexicon-based features using standard deviation performed particularly well on the irony detection task, at least in the training set, indicating that our intuition of finding "anomalies" was right. We plan to work furtherly in this direction as to detect anomalies in content or changes in polarity from one fragment of text to another and integrate them as further features.

Acknowledgments.

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the progam "Investissements d'Avenir" (reference: ANR-10-LABX-0083). The National Council for Science and Technology (CONACyT-Mexico) has funded the research work of the first author (218109/313683 grant).

References

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'14)*, Pisa, Italy.

Erick Cambria, B. Schuller, Yunqing Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, March.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, New York, NY, USA. ACM.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. 2000. New support vector algorithms. *Neural computation*, 12(5):1207–1245.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. *Frontiers in Artificial Intelligence and Applications: ECAI*, 215:765–770.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language 1, 2. *Psychological reports*, 105(2):509–521.

Alecia Wolf. 2000. Emotional expression online: Gender differences in emoticon use. In *CyberPsychology & Behavior*, volume 3.

Ningjue Zhou. 2010. Taboo language on the internet : An analysis of gender differences in using taboo language.