10.12871/clicit2014215

The CoLing Lab system for Sentiment Polarity Classification of tweets

Lucia C. Passaro♣, Gianluca E. Lebani♣, Laura Pollacci♣, Emmanuele Chersoni♣♠, Alessandro Lenci♣
♣CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, University of Pisa (Italy)
♠Laboratoire Parole et Langage, Aix-Marseille University
{lucia.passaro|gianluca.lebani}@for.unipi.it, laurapollacci.pl@gmail.com,
emmanuelechersoni@gmail.com, alessandro.lenci@ling.unipi.it

Abstract

English. This paper describes the CoLing Lab system for the EVALITA 2014 *SENTIment POLarity Classification* (SENTIPOLC) task. Our system is based on a SVM classifier trained on the rich set of lexical, global and twitter-specific features described in these pages. Overall, our system reached a 0.63 weighted F-score on the test set provided by the task organizers.

Italiano. *Questo contributo descrive il sistema CoLing Lab sviluppato per il task di SENTIment POLarity Classification (SENTIPOLC) organizzato nel contesto della campagna EVALITA 2014. Il nostro sistema è basato su un classificatore SVM addestrato sulle feature lessicali, globali e specifiche del canale twitter descritte in queste pagine. Il nostro sistema raggiunge uno score di circa 0.63 nel test set fornito dagli organizzatori del task.*

1 Introduction

Nowadays social media and microblogging services are extensively used for rather different purposes, from news reading to news spreading, from entertainment to marketing. As a consequence, the study of how sentiments and emotions are shown in such platforms, and the development of methods to automatically identify them, has emerged as a great area of interest in the Natural Language Processing community.

In this context, the research on sentiment analysis and detection of speaker-intended emotions from Twitter messages (tweets) appears to be a task on its own, rather distant from the previous sentiment classification research that focused on classifying longer pieces of texts, such as movie reviews (Pang and Lee, 2002).

As a medium, Twitter presents many linguistic and communicative peculiarities. A tweet, in fact, is a really short informal text (140 characters), in which the frequency of creative punctuation, emoticons, slang, specific terminology, abbreviations, links and hashtags is higher than in other domains. Twitter users post messages from many different media, including their cell phones, and they "tweet" about a great variety of topics, unlike what can be observed in other sites, which appear to be tailored to a specific group of topics (Go et al., 2009).

In this paper we describe the system we developed for the participation in the constrained run of the EVALITA 2014 *SENTIment POLarity Classification* Task (SENTIPOLC: Basile et al., 2014). The report is organized as follows: Section 2 describe the CoLing Lab system, starting from data preprocessing and annotation, to the adopted classification model. Section 3 shows the results obtained by our system.

2 System description

The CoLing Lab system for polarity classification of tweets includes the following three basic steps, that will be described in this section:

1. a **preprocessing** phase, aimed at the separate annotation of the linguistic and nonlinguistic elements in the target tweets;
2. a **feature extraction** phase, in which the relevant characteristics of the tweets are identified;
3. a **classification** phase, based on a Support Vector Machine (SVM) classifier with a linear kernel.

2.1 Data preprocessing and annotation

The aim of the preprocessing phase is the identification of the linguistic and nonlinguistic elements in the tweets and their annotation.

While the preprocessing of nonlinguistic elements such as links and emoticons is limited to their identification and classification (see section 2.2 for the complete list), the treatment of the linguistic material required the development of a dedicated rule-based procedure, whose output is a normalized text that is subsequently feed to a pipeline of general-purpose linguistic annotation tools. In details, the following rules applies in the linguistic preprocessing phase:

- Emphasis: tokens presenting repeated characters like *bastaaaa* are replaced by their most probable standardized form (i.e. *basta*).
- Links and emoticons: they are identified and removed.
- Punctuation: linguistically irrelevant punctuation marks are removed.
- Usernames: they are identified and normalized by removing the @ symbol and capitalizing the entity name.
- Hashtags: they are identified and normalized by simply removing the # symbol.

The output of this phase are “linguistically-standardized” tweets, that are subsequently POS tagged with the Part-Of-Speech tagger described in Dell’Orletta (2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009).

2.2 Feature extraction

By exploiting the linguistic and non-linguistic annotations obtained in the preprocessing, a total of 1239 features have been extracted to be feed to the classifier. The inventory of features can be organized into the five classes described in this subsection.

2.2.1 Lexical features

Lexical features represent the occurrence of bad words or of words that are either highly emotional or highly polarized. Relevant lemmas were identified from two in-house built lexica (cf. below), and from Sentix (Basile and Nissim, 2013), a lexicon of sentiment-annotated Italian words.

ItEM. Lexicon of 347 highly emotional Italian words built by exploiting an online feature elicitation paradigm. Native speakers were requested to list nouns, adjectives or verbs that are strongly associated with the eight basic positive and negative emotions defined in Plutchik (2001): joy, trust, surprise, sadness, anger, disgust, fear and anticipation.

In our model, we used ItEM to compute, for each of the above mentioned emotions, the total count of strongly emotional tokens in each tweet.

Bad words lexicon. By exploiting an in house built lexicon of common Italian bad words, we reported, for each tweet, the frequency of bad words belonging to a selected list, as well as the total amount of these lemmas.

Sentix. Sentix (Sentiment Italian Lexicon: Basile and Nissim, 2013) is a lexicon for Sentiment Analysis in which 59,742 lemmas are annotated for their polarity and intensity, among other information. Polarity scores range from -1 (totally negative) to 1 (totally positive), while Intensity scores range from 0 (totally neutral) to 1 (totally polarized). Both these scores appear informative for our purposes, so that we derived, for each lemma, a Combined score C_{score}:

$$C_{score} = Intensity * Polarity$$

on the basis of which we organized the selected lemmas into the following five groups:

- strongly positives: $1 \leq C_{score} < 0.25$
- weakly positives: $0.25 \leq C_{score} < 0.125$
- neutrals: $0.125 \leq C_{score} \leq -0.125$
- weakly negatives: $-0.125 < C_{score} \leq -0.25$
- highly negatives: $-0.25 < C_{score} \leq -1$

Since Sentix relies on WordNet sense distinctions, it is not uncommon for a lemma to be associated with more than one $< Intensity, Polarity >$ pair, and consequently to more than one C_{score}. We decided to handle this phenomenon by identifying three different ambiguity classes and treating them differently. Lemmas with only one entry or whose entries are all associated with the sameC_{score} value, are marked as “Unambiguous” and associated with that C_{score}. Ambiguous cases were treated by inspecting, for each lemma, the distribution of the associated C_{scores}.

Lemmas which had a Majority Vote[1] (MV) were marked as “Inferable” and associated with the C_{score} of the MV. If there was no MV, but the

[1] For each lemma a Majority Vote occurs when a class (strongly positive, weakly positive, etc) scores the greatest number of entries in Sentix. When two or more classes have the highest number of entries, the lemma has no MV.

highest number of senses in Sentix occurred simultaneously in both the positive or negative groups, lemmas were marked as “Inferable” and associated with the mean of the C_{scores}. All other cases were marked as “Ambiguous” and associated with the mean of the C_{scores}. To isolate a reliable set of polarized words, we focused only on the “Unambiguous” or “Inferable” lemmas and selected only the 250 topmost frequent according to the PAISÀ corpus (Lyding et al., 2014), a large collection of Italian web texts.

Other Sentix-based features in our model are: the number of tokens for each C_{score} group, the C_{score} of the first token in the tweet, the C_{score} of the last token in the tweet and the count of lemmas that are represented in Sentix.

2.2.2 Negation

Negation features have been developed to encode the presence of a negation and the morphosyntactic characteristics of its scope.

To count the negative tokens, we extracted from Renzi et al. (2001) an inventory of negative lemmas (e.g. “*non*”) and patterns (e.g. “*non…mai*”), and counted the occurrence of these lemmas and structures in every tweet.

We then relied on the dependency parses produced by DeSR to characterize the scope of each negation, by assuming that the scope of a negative element is its syntactic head or the predicative complement of its head, in the case the latter is a copula.

Clearly, this has been a simplifying assumption, but in our preliminary experiments it shows to be a rather cost-effective strategy in the analysis of linguistically simple texts like tweets.

We included this information in our model by counting the number of negation pattern encountered in each tweet, where a negation pattern is composed by the PoS of the negated element plus the number of negative token depending from it and, in case it is covered by Sentix, either its Polarity, its Intensity and its C_{score} value. For instance, the negation pattern instantiated in the phrase *non tornerò mai* (“I will never come back”) has been encoded, as “neg-negV$_{\text{POSPOL}}$”, “neg-negV$_{\text{HIGHINT}}$” and “neg-negV$_{\text{POSCOMB}}$”, meaning that a verb with high positive polarity, high intensity and a high C_{score} token is modified by two negative tokens.

2.2.3 Morphological features

The linguistic annotation produced in the preprocessing has been exploited also in the population of the following morphological statistics:

- number of sentences in the tweet;
- number of linguistic tokens;
- proportion of content words (nouns, adjectives, verbs and adverbs);
- number of tokens for Part-of-Speech.

2.2.4 Shallow features

This group of features has been developed to describe some distinctive characteristic of the web communication.

Emoticons. We built EmoLex, an inventory of common emoticons, such as :-(and :-), marked with their polarity score: 1 (positive), -1 (negative), 0 (neutral). In our system, EmoLex is used both to identify emoticons and to annotate their polarity.

In our model, emoticon-related features are the total amount of emoticons in the tweet, the polarity of each emoticon in sequential order and the polarity of each emoticon in reversed order. For instance, in the tweet *:-(quando ci vediamo? mi manchi anche tu! :*:** (“:-(when are we going to meet up? I miss you, too :*:*”) there are three emoticons, the first of which is negative while the others are positive. Accordingly, we feed our classifier with the information that the polarity of the first emoticon is -1, that of the second emoticon is 1 and the same goes for the third emoticon.

We additionally specified that the polarity of the last emoticon is 1, as it goes for that of the last but one emoticon, while the last but two has a polarity score of -1.

Links. We have performed a shallow classification of links using simple regular expressions applied to URLs. In particular, links are classified as following: video, images, social and other. For example, URLs containing substrings such as “youtube.com” or “twitcam” are classified as “video”. Similarly URLs containing substrings such as “imageshack”, or “jpeg” are classified as “images”., and URLs containing “plus.google” or “facebook.com” are classified as “social”. Unknown links are inserted in the residual class “other”.

We also use as feature the absolute number of links for each tweet.

Emphasis. The features report the number of emphasized tokens presenting repeated characters like *bastaaaa*, the average number of repeated characters in the tweet, and the cumulative number of repeated characters in the tweet.

For instance, in the message *Bastaaa! Sono stufaaaaaaaaa* ("Stop! I had enough"), there are 2 empathized tokens, the average number of repeated characters is 5, and the cumulative number of repetitions is 10.

Creative Punctuation. Sequences of contiguous punctuation characters, like "!!!", "!?!?!?!!?" or "……", are identified and classified as a sequence of dots, exclamations marks, question marks or mixed.

For each tweet, we mark the number of sequences belonging to each group and their average length in characters.

Quotes. The number of quotations in the tweet.

2.2.5 Twitter features

This group of features describes some Twitter-specific characteristics of the target tweets.

Topic. This information marks if a tweet has been retrieved via a specific political hashtag or keywords.

Usernames. The number of @username in the tweet.

Hashtags. We tried to infer the polarity of an hashtag by generalizing over the polarity of the tweets in the same thread. In other words, we used every hashtags we encountered as a search key [2] to download the most recent tweets in which they occur and inferred the polarity of the retrieved tweets by simply counting the number of positive and negative words in them.

In doing so, we made the assumption that the polarity of an hashtag is likely to be the same of the words it typically co-occurs with.

This, of course, does not take into account any kind of contextual variability of words meaning. We are aware that this is an oversimplifying assumption; nevertheless, we are confident that, in most cases, the polarity of the hashtag will reflect the polarity of its typical word contexts.

Moreover, tweets were assumed to be positive if they contained a majority of positive words, negative if they contained a majority of negative words, neutral otherwise.

In order to determine the polarity of a word, we used the scores of the Sentix lexicon. Words with a positive score ≤ 0.7got a score of 1, while words with a negative score ≤ -0.7received the score of -1. All the other words got a score of 0 (neutrality).

Unfortunately, for many hashtags in the corpus we have been able to retrieve just a small number of tweets, so that we chose to filter out those below a frequency threshold of 20 tweets, leaving us with 279 polarity-marked hashtags.

By relying on this hashtag-to-polarity mapping, the hashtag-related features in our model consisted in the total amount of hashtag for tweet, the polarity of each hashtag in sequential order and the polarity of each hashtag in reversed order.

2.3 Classification

Due to the better performance of SVM-based systems in analogue tasks (e.g. Nakov et al., 2013), we chose to base the CoLing Lab system for polarity classification on the SVM classifier with a linear kernel implementation available in Weka (Witten et al., 2011), trained with the Sequential Minimal Optimization (SMO) algorithm introduced by Platt (1998).

The classification task proposed by the organizers could be approached either by building two separate binary classifiers relying of two different models (one judging the positiveness of the tweet, the other judging its negativeness), or by developing a single multiclass classifier where the possible outcomes are Positive Polarity (Task POS:1, Task NEG:0), Negative Polarity (Task POS:0, Task NEG:1), Mixed Polarity (Task POS:1, Task NEG:1) and No Polarity (Task POS:0, Task NEG:0).

We tried both approaches in our development phase, and found no significant difference, so that we opted for the more economical setting, i.e. the multiclass one.

3 Experiments and Results

The evaluation metric used in the competition is the macro-averaged F_1-score calculated over the positive and negative categories. Our model obtained a macro-averaged F_1-score of 0.6312 on the test set and was ranked 3rd among 11 submissions. Table 2 reports the results of our model.

In addition, we present here two additional configurations (L and S) of our system, both of them using a smaller number of features.

The Lexical Model (L) is trained only on lexical features (see section 2.2.1), negation (see section2.2.2) and hashtags. This last group of features is used to train this model because the polarity of a thread is inferred from Sentix (see section 2.2.5).

The Shallow Model (S) is trained using only the non lexical features described in sections 0, 2.2.4, 2.2.5 (topic and usernames).

[2] We use the Python-Twitter library to query the Twitter API (https://code.google.com/p/python-twitter.)

Table 1 summarizes the features used to train the different models (F(ull), L(exical), S(hallow)), showing for each model the number of features:

Group	**Features**	**#**	**F**	**L**	**S**
Lexical	Badwords	28	∨	∨	
Lexical	ItEM	9	∨	∨	
Lexical	Sentix	1023	∨	∨	
Negation	Negation	53	∨	∨	
Morphol. features	Morphol. features	18	∨		∨
Shallow	Emoticons	17	∨		∨
Shallow	Emphasis	3	∨		∨
Shallow	Links	5	∨		∨
Shallow	Punctuation	6	∨		∨
Shallow	Quotes	1	∨		∨
Shallow	Slang	10	∨		∨
Twitter	Hashtags	63	∨	∨	
Twitter	Topic	1	∨		∨
Twitter	Usernames	2	∨		∨
Total number of features		**1239**	**1239**	**1176**	**63**

Table 1: Features used to train the models.

The Full model is trained on all the features described in the previous sections (1239 features).

Table 2 shows the detailed scores for each class both in the Positive and Negative tasks. It also points out the aggregate scores for each task and the overall scores.

Task	Class	Precision	Recall	F-score
POS	0	**0.7976**	0.7806	0.789
POS	1	0.581	**0.4109**	**0.4814**
POS task		0.6893	**0.5957**	**0.6352**
NEG	0	0.6923	0.6701	**0.681**
NEG	1	**0.6384**	0.5201	0.5732
NEG task		**0.6654**	0.5951	0.6271
GLOBAL		0.6774	**0.5954**	**0.6312**

Table 2: CoLing Lab system results

Table 3 shows the results obtained by the Lexical model, with 1176 features.

Task	Class	Precision	Recall	F-score
POS	0	0.7599	0.7755	0.7676
POS	1	0.4913	0.2981	0.371
POS task		0.6256	0.5368	0.5693
NEG	0	0.66	**0.6861**	0.6728
NEG	1	0.6218	0.4522	0.5237
NEG task		0.6409	0.5692	0.5983
GLOBAL		0.6333	0.553	0.5838

Table 3: CoLing Lab Lexical (L) system results

Table 4 reports the results obtained by the Shallow model, trained using non lexical information only, for a total of 63 features.

Task	Class	Precision	Recall	F-score
POS	0	0.7578	**0.8679**	**0.8092**
POS	1	**0.7184**	0.2205	0.3374
POS task		**0.7381**	0.5442	0.5733
NEG	0	**0.7369**	0.5174	0.608
NEG	1	0.5778	**0.6582**	**0.6154**
NEG task		0.6574	0.5878	0.6117
GLOBAL		**0.6978**	0.566	0.5925

Table 4: CoLing Lab Shallow (S) system results

4 Discussion

The best model to predict the polarity of a tweet is the one that combines lexical and shallow information (Full model).

Even though it achieves a better F_1-score, the global precision of the Shallow model is higher than the precision of the Full Model, despite the much smaller numbers of features. In particular, the Shallow model recognizes positive tweet more accurately. It is worth noticing that the class of positive tweets is the one in which our systems score worst. Besides the fact that the tweet class distribution is unbalanced in the training corpus, positive lexical features are likely to be not as able to predict tweets positivity, as negative features are with respect to negative tweets.

To sum up, on the one hand the three experiments demonstrate that significant improvements can be obtained by using lexical information. On the other hand the results highlight that the lexical coverage of the available resources such as Sentix and ItEM must be increased in order to obtain a more accurate classification.

5 Conclusion and future work

The CoLing Lab system participated in SENTIment POLarity Classification (SENTIPOLC) in EVALITA 2014 using a Support Vector Machine approach. The system combines lexical and shallow features achieving an overall F_1-score of 0.6312. Future developments of the system include refining the preprocessing phase, increasing the coverage of the lexical resources, improving the treatment of negation, and designing a more sophisticated way to exploit the information coming from the tweet thread. In particular, we are confident that a better preprocessed text and larger lexical resources will significantly enhance our system's performance.

Acknowledgments

Lucia C. Passaro received support from the Project SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE), funded by Regione Toscana (POR CReO 2007-2013), Gianluca E. Lebani works in the context of the PRIN grant 20105B3HE8, funded by the Italian Ministry of Education, University and Research; Emmanuele Chersoni is supported by the University Foundation A*MIDEX.

Reference

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proceedings of EVALITA 2009*.

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*: 100-107.

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)*.

Felice Dell'Orletta(2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

Alec Go, Richa Bhayani and Lei Huang (2009). *Twitter Sentiment Classification using Distant Supervision*. CS224N Project Report, Stanford.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*: 36-43.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov and Theresa Wilson (2013). Semeval-2013 Task 2: Sentiment Analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: 79-86.

John C. Platt (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola (eds.) *Advances in Kernel Methods*: 185-208.

Robert Plutchik (2001). The Nature of Emotions. In *American Scientist*, 89: 344-350.

Lorenzo Renzi Gianpaolo Salvi and Anna Cardinaletti (2001). *Grande grammatica italiana di consultazione*. Il Mulino: Bologna.

Ian H. Witten, Elibe Frank, E and Mark A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.