Convolutional Neural Networks for Sentiment Analysis on Italian Tweets

Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, Federica Semplici
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy
{attardi, sartiano}@di.unipi.it,
{c.alzetta, f.semplici}@studenti.unipi.it

Abstract

English. The paper describes our submission to the task 2 of SENTIment POLarity Classification in Italian Tweets at Evalita 2016. Our approach is based on a convolutional neural network that exploits both word embeddings and Sentiment Specific word embeddings. We also experimented a model trained with a distant supervised corpus. Our submission with Sentiment Specific word embeddings achieved the first official score.

Italiano. *L'articolo descrive la nostra partecipazione al task 2 di SENTIment POLarity Classification in Italian Tweets a Evalita 2016. Il nostro approccio si basa su una rete neurale convoluzionale che sfrutta sia word embeddings tradizionali che sentiment specific word embeddings. Abbiamo inoltre sperimentato un modello allenato su un corpus costruito mediante tecnica distant supervised. Il nostro sistema, che utilizza Specific Sentiment word embeddings, ha ottenuto il primo punteggio officiale.*

1 Introduction

The paper describes our submissions to the Task 2 of SENTiment POLarity Classification at Evalita 2016 (Barbieri et al. 2016).

In Sentipolc the focus is the sentiment analysis of the in Italian tweets, it is divided in three subtasks:

- Task 1: Subjectivity Classification: identify the subjectivity of a tweet.
- Task 2: Polarity Classification: classify a tweet as positive, negative, neutral or mixed (i.e. a tweet with positive and negative sentiment).
- Task 3: Irony Detection: identify if is present the irony in a tweet.

The state of the art on the polarity classification of tweets is the application of Deep Learning methods (Nakov et al., 2016), like convolutional neural network or recurrent neural networks, in particular long short-term memory networks (Hochreiter, and Schmidhuber, 1997).

We explored Deep Learning techniques for the sentiment analysis of English tweets at Semeval 2016 with good results, where we noticed that use of convolutional neural network and Sentiment Specific word embeddings was promising.

We applied a similar approach for the Italian language, building word embeddings from a big corpus of Italian tweets, sentiment specific word embeddings from positive and negative tweets, using a convolutional neural network as classifier. We also introduced a distant supervised corpus as silver training set.

We report the results of our experiments with this approach on the task Evalita 2016 Sentipolc Task 2 Polarity classification.

2 Description of the System

The architecture of the system consists of the following steps:

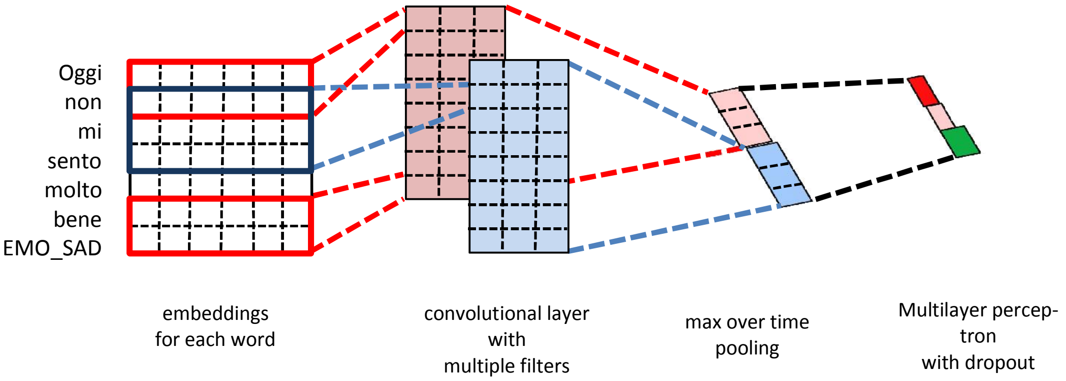
Oggi
non
mi
sento
molto
bene
EMO_SAD
embeddings for each word
convolutional layer with multiple filters
max over time pooling
Multilayer perceptron with dropout

Figure 1. The Deep Learning classifier.

- build word embeddings from a collection of 167 million tweets collected with the Twitter API over a period of May to September 2016, preprocessed as described later.
- build Sentiment Specific word embeddings using a portion of these tweets split into positive/negative by distant supervision.
- train a convolutional neural network classifier using one of the above word embeddings

The convolutional neural network classifier exploits pre-trained word embeddings as only features in various configurations as described below. The architecture of the classifier consists of the following layers described in Figure 1: a lookup layer for word embeddings, a convolutional layer with a ReLU activation function, a maxpooling layer, a dropout layer, a linear layer with tanh activation and a softmax layer. This is the same classifier described in (Attardi and Sartiano, 2016), that achieved good results at the SemEval 2016 task 4 on Sentiment Analysis in Twitter (Nakov et al., 2016). Here we test it on a similar task for Italian tweets.

2.1 Data Preprocessing

In order to build the word embeddings we preprocessed the tweets using tools from the Tanl pipeline (Attardi et al., 2010): the sentence splitter and the specialized tweet tokenizer for the tokenization and the normalization of tweets. Normalization involved replacing the mentions with the string “@mention”, emoticons with their name (e.g. “EMO_SMILE”) and URLs with “URL_NORM”.

2.2 Word Embeddings and Sentiment Specific Word Embeddings

We experimented with standard word embeddings, in particular building them with the tool word2vec[1] (Mikolov, 2013), using the skip-gram model. These word embeddings though do not take into account semantic differences between words expressing opposite polarity, since they basically encode co-occurrence information as shown by (Levy and Goldberg, 2014). For encodes sentiment information in the continuous representation of words, we use the technique of Tang et al. (2014) as implemented in the DeepNL[2] library (Attardi, 2015). A neural network with a suitable loss function provides the supervision for transferring the sentiment polarity of texts into the embeddings from generic tweets.

2.3 Distant supervision

The frequency distribution of classes in the dataset, as shown in Table 1, seems skewed and not fully representative of the distribution in a statistical sample of tweets: negative tweets are normally much less frequent than positive or neutral ones (Bravo-Marquez, 2015). To reduce this bias and to increase the size of the training set, we selected additional tweets from our corpus of Italian tweets by means distant supervision. In the first step we selected the tweets belonging to a class (positive, negative, neutral, mixed) via regular expressions. In the second step, the selected tweets are classified by the classifier trained using the task trainset. The silver corpus is built taking the tweets with the matched class between the regular expression system and the classifier.

[1] https://code.google.com/archive/p/word2vec/

[2] https://github.com/attardi/deepnl

3 Experiments

The plain word embeddings were built applying vord2vec to a collection of 167 million Italian unlabeled tweets, using the the skip gram model, and the following parameters: embeddings size 300, window dimension 5, discarding word that appear less than 5 times. We obtained about 450k word embeddings.

The sentiment specific word embeddings (SWE) were built with DeepNL, starting from the word embeddings built at the previous step and tuning them with a supervised set of positive or negative tweets, obtained as follows from 2.3 million tweets selected randomly from our corpus of collected tweets:

- Positive tweet: one that contains only emoticons from a set of positive emoticons (e.g. smiles, hearts, laughs, expressions of surprise, angels and high fives).
- Negative tweet: one that contains only emoticons from a set of negative emotions (e.g. tears, angry and sad).

Integris srl cooperated to the task providing a set of 1.3 million tweets, selected by relying on a lexicon of handcrafted polarized words. This resource is also added to the corpus.

We split the training set provided for the Evalita 2016 SentiPolc Task into a train set (5335 tweets), a validation set (592 tweets) and a test set (1482 tweets). This dataset was tokenized and normalized as described in Section 2.1.

For the take of participating to subtask 2, polarity classification, the 13-value annotations present in the datasets were converted into four values: "neutral", "positive", "negative" and "mixed" depending on the values of the fields "opos" and "oneg", which express the tweet polarity, according to the task guidelines[3]. We did not take into account the values for "lpos" and "lneg".

The frequency distribution of these classes turns out to be quite unbalanced, as shown in Table 1.

Class	Train set	Validation set
Neutral	2262	554
Negative	2029	513
Positive	1299	312
Mixed	337	103

Table 1. Task dataset distribution

The training set is still fairly small, compared for example to the size of the corpus used in SemEval 2106. The "mixed" class in particular is small in absolute numbers, even though not in percentage value, which makes hard to properly train a ML classifier.

Therefore we tried to increase the training set by means of the distant supervision as described above: we selected a maximum of 10,000 tweets for class via regular expressions, then we classified them with the classifier trained with the gold training set. We chose for addition into a silver training set, the tweets which were assigned by the classifier the same class of the regular expression. As reported in Table 2, the silver dataset remains unbalanced; in particular, no "mixed" example was added to the original train set.

Class	Train set	Dev set
Neutral	8505	554
Negative	5987	513
Positive	6813	312
Mixed	337	103

Table 2. Distant supervised dataset distribution.

Table 3 shows the common settings used for training the classifier. We used the same parameters as SemEval-2016.

Word Embeddings Size	300
Hidden Units	100
Dropout Rate	0.5
Batch size	50
Adadelta Decay	0.95
Epochs	50

Table 3. Network Common Settings

We performed extensive experiments with the classifier in various configurations, varying the number of filters; the use of skip-gram word embeddings or sentiment specific word embeddings; different training sets, either the gold one or the silver one. Results of the evaluation on the validation set allowed us to choose the best settings, as listed in the Table 4. Best Settings.

	Run1		**Run2**	
Embeddings	WE skipgram		SWE	
Training set	Gold	Silver	Gold	Silver
Filters	2,3,5	4,5,6,7	7,7,7,7,8,8,8,8	7,8,9,10

Table 4. Best Settings

4 Results

We submitted four runs for the subtask 2 "polarity classification":

[3] http://www.di.unito.it/~tutreeb/sentipolc-evalita16/sentipolc-guidelines2016UPDATED130916.pdf

- UniPI_1.c: gold training set, word embeddings with skip-gram model, filters: “2,3,5”.
- UniPI_1.u: silver corpus as training set, word embeddings with skip-gram model, filters: “4,5,6,7”.
- UniPI_2.c: gold training set, sentiment specific word embeddings, filters: “7,7,7,7,8,8,8,8”.
- UniPI_2.u: silver corpus as training set, sentiment specific word embeddings, filters: “7,8,9,10”.

The following table reports the top official results for the subtask 2:

System	Positive F-score	Negative F-score	Combined F-score
UniPI_2.c	**0.685**	0.6426	**0.6638**
team1_1.u	0.6354	**0.6885**	0.662
team1_2.u	0.6312	0.6838	0.6575
team4_.c	0.644	0.6605	0.6522
team3_.1.c	0.6265	0.6743	0.6504
team5_2.c	0.6426	0.648	0.6453
team3_.2.c	0.6395	0.6469	0.6432
UniPI_1.u	0.6699	0.6146	0.6422
UniPI_1.c	0.6766	0.6002	0.6384
UniPI_2.u	0.6586	0.5654	0.612

Table 5. Top official results for SentiPolc subtask 2.

The run UniPI_2.c achieved the top overall score among a total of 26 submissions to task 2. This confirms the effectiveness of sentiment specific word embeddings in sentiment polarity classification also for Italian tweets.

The use of an extended silver corpus did not provide significant benefits, possibly because the resulting corpus was still unbalanced.

In addition to the subtask 2, we submitted one run for the Task 1 “Subjectivity Classification”: given a message, decide whether the message is subjective or objective. We used the same classifier for the subtask 2, using only two classes (subjective, objective), with the same skip-gram word embeddings used for the other task and the configuration listed in Table 3, using the following filters: “7,8,9,10”, without performing extensive experiments. The following table reports the top official results for the subtask 1:

system	Objective F-score	Subjective F-score	Combined F-score
team1_1.u	**0.6784**	**0.8105**	**0.7444**
team1_2.u	0.6723	0.7979	0.7351
team2_.1.c	0.6555	0.7814	0.7184
team3_.2.c	0.6733	0.7535	0.7134
team4_.c	0.6465	0.775	0.7107
team5_2.c	0.6671	0.7539	0.7105
team6_.c	0.6623	0.755	0.7086
team1_.c	0.6499	0.759	0.7044
UniPI_1	0.6741	0.7133	0.6937
team3_.1.c	0.6178	0.735	0.6764
team8_.c	0.5646	0.7343	0.6495
team5_1.c	0.6345	0.6139	0.6242

Table 6 Top official results for SentiPolc subtask 1.

5 Discussion

We confirmed the validity of the convolutional neural networks in the twitter sentiment classification, also for the Italian language.

The system achieved top score in the task 2 of SENTiment POLarity Classification Task of Evalita 2016.

Acknowledgments

We gratefully acknowledge the support of the University of Pisa through project PRA and NVIDIA Corporation for a donation of the Tesla K40 GPU used in the experiments.

Integris srl cooperated by providing a corpus of sentiment annotated tweets.

References

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. *The Tanl Pipeline*. In Proc. of LREC Workshop on WSPP, Malta.

Giuseppe Attardi. 2015. DeepNL: a Deep Learning NLP pipeline. *Workshop on Vector Space Modeling for NLP.* Proc. of NAACL HLT 2015, Denver, Colorado (June 5, 2015).

Giuseppe Attardi and Daniele Sartiano. 2016. UniPI at SemEval-2016 Task 4: Convolutional Neural Networks for Sentiment Classification. *Proceedings of SemEval*, 220-224.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In *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 (EVALI-*

TA 2016). Associazione Italiana di Linguistica Computazionale (AILC).

Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. 2015. Positive, negative, or neutral: Learning an expanded opinion lexicon from emoticon-annotated tweets. *IJCAI 2015*. AAAI Press.

Sepp Hochreiter, and Jürgen Schmidhuber. (1997). Long short-term memory. *Neural computation 9.8* 1735-1780.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in neural information processing systems*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. arXiv:1310.4546

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), San Diego, US (forthcoming).

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014, June. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL (1)* (pp. 1555-1565).