Multi-Task Learning in Deep Neural Network for Sentiment Polarity and Irony classification

Lorenzo De Mattei[1,2], Andrea Cimino[2], and Felice Dell'Orletta[2]

[1] Dipartimento di Informatica, Università di Pisa
lorenzo.demattei@di.unipi.it
[2] Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa
ItaliaNLP Lab - *www.italianlp.it*
{andrea.cimino,felice.dellorletta}@ilc.cnr.it

Abstract. We study the impact of a new multi-task learning approach in deep neural network for polarity and irony detection in Italian Twitter posts. We compare this approach with traditional single-task learning models. The different behavior of the two approaches shows the effectiveness of the proposed method that is able to combine the information from the two tasks improving the accuracy in both tasks. This is particularly true on edge cases in which knowledge about the two tasks is needed to classify a tweet, this is the case, for example, when the literal polarity of a tweet is inverted by irony.

Keywords: Deep neural network · Multi-Task learning · Sentiment analysis.

1 Introduction

During the last years Sentiment Analysis and related tasks have attracted a lot of attention in the research community. Several works have been published on these topics, and with the rising of deep learning the performances of the systems have considerably increased. Despite these performances improvements, machine learning based systems still struggle to perform well in edge cases such as when literal polarity is inverted by irony, especially when these cases are underrepresented in the training data. Such cases were annotated for the SENTIPOLC 2016 shared task [2]: consider the tweet from the dataset *"Ho molta fiducia nel nuovo Governo Monti. Più o meno la stessa che ripongo in mia madre che tenta di inviare un'email"* ("I have a lot of faith in the new Monti government. More or less the same thing that I have in my mother who tries to send an email"): this tweet has literal positive polarity, but irony changes the final polarity annotation.

Previous works on neural networks already shown issues on learning such difficult cases: [10] pointed out a set of 10 criticisms of deep neural networks like the inability to deal with hierarchical structure, the limited capacity for transfer learning, the impossibility to integrate prior knowledge or lack of systematic compositional skills. Despite these issues, previous works [14] have shown that multi-task learning (MTL) is an appealing idea compared to single-task learning

(STL) since it allows to incorporate previous knowledge about tasks hierarchy into neural networks architectures. [12] have shown that MTL is useful to combine even loosely related tasks, letting the networks automatically learn the tasks hierarchy.

To study the effectiveness of MTL on Sentiment Analysis tasks, in this paper we present a mixed MTL/STL approach (named MIX) based on deep bi-directional recurrent neural networks [13] applied to polarity and irony detection on Italian tweets. We modeled our networks to solve three binary tasks: positive, negative and ironic tweet identification. We tested the performances of our system on the most recent datasets available for Italian. We show that our system outperforms the state of the art for Italian for what concerns polarity and irony classification. Furthermore, we show that the proposed mixed approach outperforms both our STL and MTL approaches.

To our knowledge, this is the first work that shows the effectiveness of MTL combining irony and polarity detection. A previous work on this topic [6] has been presented at EVALITA 2016, but the authors proposed an approach that is more similar to a multi-label classification method based on a single classifier for all the labels, rather than a MTL in which different loss functions are used for the different tasks.

We present an in-depth analysis on the results obtained by our method showing how the proposed multi-task learning approach is able to compose the information coming from the different tasks.

Our contributions: (i) to our knowledge this is the first work that presents a MTL system for polarity and irony detection; (ii) we introduce a novel mixed MTL and STL approach; (iii) we present an error analysis that suggests that the proposed multi-task learning approach is able to combine the information extracted from sentiment polarity and irony classification training sets and improves the performance on both the tasks. This is particularly true on edge cases in which knowledge about the two tasks are needed to classify a tweet.

2 Dataset

For the Italian polarity and irony classification tasks we relied on the dataset provided for the SENTIPOLC event which made part of EVALITA 2016, the periodic evaluation campaign NLP and speech tools for the Italian language. The SENTIPOLC dataset contains a training set made of 7,410 tweets and a test set of 2,000 tweets. Each tweet was labeled with a set of 6 binary labels that define if a tweet is subjective (*subj*), positive (*pos*), negative (*neg*), ironic (*iro*), literally positive (*lpos*) and literally negative (*lneg*). We performed our experiments only on positive, negative and ironic classes, but we still used the other labels to perform a comparative analysis between the performances of the system trained in the single-task and in the multi-task models.

Table 1 reports the distributions of labels in the data set.

Label combination						train	test
subj	pos	neg	iro	lpos	lneg		
0	0	0	0	0	0	2,312	695
1	0	0	0	0	0	504	219
1	1	0	0	1	0	1,488	295
1	0	1	0	0	1	1,798	520
1	1	1	0	1	1	440	36
1	0	1	1	0	0	210	73
1	1	0	1	1	0	62	8
1	0	1	1	1	0	239	66
1	1	0	1	0	0	29	3
1	0	1	1	0	1	225	53
1	1	0	1	0	1	22	4
1	0	1	1	1	1	71	22
1	1	0	1	1	1	10	6
total						7,410	2,000

Table 1. Distribution of label combinations in the SENTIPOLC 2016 data set

3 Architecture and Training

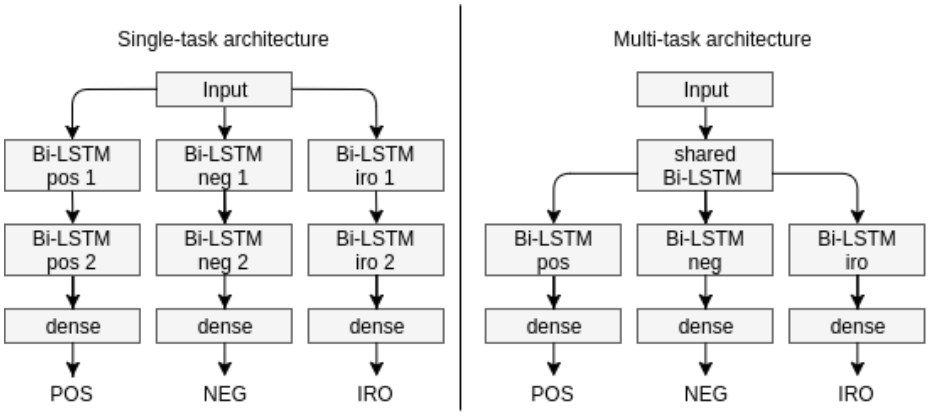

Fig. 1. STL and MTL neural networks architectures.

Figure 1 reports the architectures of the MTL and STL neural networks that we designed. Both the architectures are based on bidirectional long short-term memory networks (Bi-LSTM) [8, 7]. The STL architecture is composed of two stacked Bi-LSTM layers and a dense layer for each task. The MTL architecture is composed by a *shared Bi-LSTM*, three task specific Bi-LSTMs, and three task specific dense layers specialized in recognizing respectively positive, negative and

ironic inputs. We introduce in this work a new method (named MIX) to combine these two architectures using a two stage training approach in which a layer is shared in just one stage of the training phase.

Features: We built two sets of word embeddings with 128 dimensions using word2vec [11]. The first set of word embeddings was generated starting from the itWac Corpus [3], while the second was built exploiting approximatively 25 millions of Italian tweets. Both the corpora were postagged using the postagger by [5] and the word embeddings were computed using the combination of the word and its part of speech. The generated itWac and Twitter embeddings provided a coverage of 91.5% and 96.6% on the SENTIPOLC dataset. In addition, for each word its sentiment polarity is used as feature exploiting the sentiment polarity lexicon by [9].

Each token of a tweet is represented by a vector resulting from the concatenation of the described features.

Training: To train the STL networks, we performed three different training steps, one for each task. To train the MTL architecture, we run a shared training by iteratively optimizing at each step a loss function for each task. For the MTL the global loss function is given by the sum of the three individual loss functions. In STL and MTL architectures, we stopped the training after 50 epochs without improvements of the loss function on the validation set, choosing the parameters with the best performances.

To mix the MTL and STL approaches we used a two stage training. In the first stage we trained the MTL network as described above. In the second stage we initialized the weights of the three first Bi-LSTM layers of the STL architecture using the weights of the MTL network's *shared Bi-LSTM* and the second level Bi-LSTM using the weights learned in the first stage. We then run a specific training for each task. We used the same stopping criteria as for STL and MTL training.

Since in the dataset all the tweets are labeled with their polarity and irony labels and the number of ironic tweets is extremely unbalanced w.r.t. the non-ironic ones, we oversampled the ironic examples by replicating them in the dataset. The oversampling technique has been showed to improve classification performance on unbalanced datasets [4].

4 Results

Table 2 reports the performances on the test set achieved by our baselines and multi-task learning models. The scores are calculated accordingly to the official metrics adopted by the task organizers. Since random initialization lead to different performances in different runs, we repeated the experiments 10 times and the tables report the average scores. In addition, the tables report the performances obtained by the best systems that participated to SENTIPOLC 2016. To study the impact of multi-task learning across irony and polarity tasks, we also tested a MIX model trained only on positive and negative labels (PMIX) without using irony information.

System	POS	NEG	Polarity	IRO
STL	.641	.665	.653	.608
PMIX	.670	.699	.684	-
MTL	.674	.700	.674	.586
MIX	.660	**.736**	**.698**	**.622**
SwissCheese.c	.653	.713	.683	.536
UniPI.2.c	**.685**	.643	.664	-
tweet2check16.c	-	-	-	.541

Table 2. F1-Scores obtained on the SENTIPOLC 2016 dataset by the different systems. *Polarity* is the official metric for the polarity detection task and it is a combination of POS and NEG accuracies.

System	POS		NEG		Polarity	
	Iro	l_Pol	Iro	l_Pol	Iro	l_Pol
STL	.115	.105	0.11	.090	.080	.085
PMIX	.143	.044	.093	.075	.093	.049
MTL	.104	.069	.086	.075	.086	.061
MIX	**.539**	**.567**	**.553**	**.492**	**.553**	**.500**

Table 3. Polarity F1-Scores for ironic tweets (Iro) and for tweets in which irony modifies literal polarity (l_Pol) in the Italian test set.

As we can see in Table 2, in the polarity detection tasks the MTL, PMIX, and MIX models all outperform the best SENTIPOLC system that used a single task approach [1] (UniPI.2.c row), while only the MIX model performed better than the [6] system (SwissChese.c row), that used a multi-label classifier for the subjectivity, polarity and irony identification tasks.

For what concerns Irony detection, we observe that all our networks outperform the best SENTIPOLC system, probably thanks to the usage of oversampling (the F-score of our STL model without oversampling is only 0.473). More importantly, we observe that MIX model significantly outperforms the STL baseline, while the standard MTL does not.

These results show that MIX model brings improvement in both polarity and irony detection tasks.

To study the impact of multi-task learning in Polarity and Irony detection, we conducted an in-depth error analysis to investigate the performance of our models on edge cases. We studied the behavior of the models for a selected subsets of the test set. Table 3 reports the polarity detection accuracies of our models on Italian ironic tweets (columns *Iro* in the table) and on tweets for which irony changes the literal polarity (*l_Pol*). We can clearly observe how the MIX model brings great improvements for polarity detection in *l_Pol* tweets while the standard MTL does not. The improvements are clear for both positive and negative tweets. This result suggests that the MIX model is able to compose information coming from different examples of different tasks and to obtain

Label combination						Freq	IRO Accuracy		
subj	pos	neg	lpos	lneg	iro		MIX	MTL	STL
1	1	0	0	0	1	8	**37.50**	0.00	0.00
1	1	0	0	0	1	3	**33.33**	0.00	0.00
1	1	0	0	1	1	4	**50.00**	0.00	0.00
1	1	0	1	1	1	6	16.67	0.00	**33.33**
1	0	1	1	1	1	22	**27.27**	4.55	13.64
1	0	1	1	0	1	66	**21.21**	6.06	4.55
1	0	1	0	0	1	73	**32.88**	4.11	8.22
1	0	1	0	1	1	53	**26.42**	5.66	5.66
1	1	0	1	0	0	295	**94.58**	93.22	91.19
1	1	1	1	1	0	36	80.56	**97.22**	**97.22**
1	0	1	0	1	0	520	89.81	**97,31**	94.04
0	0	0	0	0	0	695	**98.27**	95.83	93.38
1	0	0	0	0	0	219	92.24	**95.43**	93.61

Table 4. Irony accuracy of our models for the different combinations of labels in the SENTIPOLC 2016 test set.

better results on edge cases. This is also shown by the results obtained in the polarity detection task on ironic tweets (*Iro*).

Table 4 reports the accuracy of our systems in the irony detection task for all the different label combinations in the test set. We can observe that the STL and the MTL models show the same behavior while the MIX model significantly outperforms the other two in mostly all kinds of ironic instances (rows 1-8) and not ironic positive instances (row 9). Vice versa, MTL and STL outperform MIX in the negative not ironic comments (rows 10-11). Given that the MIX approach brings impressive improvements for edge-cases (especially rare ones), it is likely that it overestimates the correlation between irony and negativity.

5 Conclusion

We conducted a study on the effectiveness of multi-task learning approaches in sentiment polarity and irony classification. We presented a mixed single- and multi-task learning approach, that is able to improve the performance both in polarity and irony detection with respect to single-task and standard multi-task learning approaches. In particular, our approach led to substantial improvements on edge cases in which knowledge about the two tasks are needed to classify a tweet. This is particularly true, when these cases are under-represented in the training data. An example is the case when a the literal polarity of a tweet is inverted by irony.

References

1. Attardi, G., Sartiano, D., Alzetta, C., Semplici, F.: Convolutional neural networks for sentiment analysis on italian tweets. 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 (EVALITA 2016) (2016)
2. Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: 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 (EVALITA 2016) (2016)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Journal of Language Resources and Evaluation **43**(3), 209–226 (2009)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
5. Cimino, A., Dell'Orletta, F.: Building the state-of-the-art in pos tagging of italian tweets. 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 (EVALITA 2016) (2016)
6. Deriu, J.M., Cieliebak, M.: Sentiment analysis using convolutional neural networks with multi-task training and distant supervision on italian tweets. In: Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Napoli, Italy, December 5-7, 2016. Italian Journal of Computational Linguistics (2016)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
9. Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Azpeitia, A., Vossen, P.: Generating polarity lexicons with wordnet propagation in five languages. Proceedings of LREC2014, Reykjavik (2014)
10. Marcus, G.: Deep learning: A critical appraisal. Computing Research Repository **abs/1801.00631** (2018), http://arxiv.org/abs/1801.00631, version 2
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Ruder, S., Bingel, J., Augenstein, I., Søgaard, A.: Sluice networks: Learning what to share between loosely related tasks. arXiv preprint arXiv:1705.08142 (2017)
13. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
14. Søgaard, A., Goldberg, Y.: Deep multi-task learning with low level tasks supervised at lower layers. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 231–235 (2016)