On the generalization of figurative language detection: the case of irony and sarcasm

Lorenzo Famiglini[1], Elisabetta Fersini[1⋆], and Paolo Rosso[2]

[1] University of Milano-Bicocca, Milano, Italy
[2] PRHLT Research Center, Universitat Politcnica de Valncia, Valencia, Spain

Abstract. The automatic detection of figurative language, such as irony and sarcasm, is one of the most challenging tasks of Natural Language Processing (NLP). In this paper, we investigate the generalization capabilities of figurative language detection models, focusing on the case of irony and sarcasm. Firstly, we compare the most promising approaches of the state of the art. Then, we propose three different methods for reducing the generalization errors on both in- and out-domain scenarios.

Keywords: Irony and Sarcasm Detection · Generalization Capabilities

1 Introduction

During the last decade, several models have been introduced in the research panorama to recognize few rhetorical figures, and in particular to identify those elements that discriminate, in a significant way, what is sarcastic or ironic from what is not. In particular, sarcasm and irony detection has been defined as a classification problem, where the ground truth is a dichotomy variable 0 and 1, where 0 means that text is not a rhetorical figure, otherwise is an ironic or sarcastic statement. The irony and sarcasm detection problem has been widely addressed in the literature, where a plethora of computational approaches have been proposed ranging from the earlier techniques based on linguistic patterns [3, 12], to the more recent ones based on neural architecture [6, 14] or combination of both [4]. Although all of these approaches in the state of the art represent a fundamental step towards the modeling of irony and sarcasm, less effort has been dedicated to measure and improve the generalization capabilities of the models when considering both in- and out-domain vocabularies. In order to address this problem, we investigate three main research questions:

(R1) What are the most representative linguistic features for identifying sarcasm and irony patterns?
(R2) How can we exploit transformer-based and emotional-based embeddings to train accurate irony and sarcasm detection models? In particular, are pure neural models better than approaches based also on linguistic features?
(R3) What are the generalisation capabilities of the developed models?

⋆ Corresponding author: elisabetta.fersini@unimib.it

Contribution. We addressed the above-mentioned research questions, by comparing the most promising approaches of the state of the art, and by proposing several approaches, based on embeddings and ensembles methods, for reducing the generalization errors on both in- and out-domain scenarios. In particular, the main contributions of the paper are:

1. A comparative analysis of the state of the art models for irony/sarcasm detection to determine their generalising capabilities, highlighting the most representative features for discriminating irony and sarcasm from others;
2. A novel methodology, based on the combination of multiple output encoder layers of the BERTweet model [7], for creating a more contextualized sentence embeddings, called BERTweet-Features based;
3. A novel model based on the emotional features of DeepMoji [2], built on the concept of self-attention layer, called DeepMoji Features-based;
4. A novel model, called Ensemble of Ensembles, where machine learning classifiers trained on several aspects of the text identify various patterns of irony and sarcasm.

All the developed models are available at `https://github.com/MIND-Lab/GIS`.

2 State-of-the-art Models for Irony and Sarcasm Detection

The first objective of this paper is to carry out a comparative analysis of different models, which are the most promising approaches in the state of the art for irony and sarcasm detection. To this purpose, we considered the following models:

- **Machine Learning classifiers**, i.e. XGBoost, AdaBoost, HistGradientBoosting, Logistic Regression and Random Forest trained on embeddings (extracted from BertTweet and reduced on the basis of Principal Component Analysis maintaining 95% of the variance) together with a set of hand-crafted features. In particular, we considered Part-OF-Speech (POS) tags, pragmatic particles (PP), including emoji, punctuation, initialisms and onomatopoeic figures, and finally the polarity of the text (POL). All the possible combinations of these features have been evaluated.
- **Bayesian Model Averaging (BMA)**, initially presented in [9], which combines the models introduced above to finally derive an ensemble of traditional classifiers.
- **DeepMoji**, presented in [2], focused emotional information encoded by a recent transformer-based architecture named RCNN-Roberta.
- **RCNN-Roberta**, presented in [8], consists of a RoBERTa pre-trained transformer followed by a bidirectional LSTM layer (BiLSTM).

3 Proposed Models

3.1 BERTweet Features-based Model (BERTweet-FB)

We firstly introduce in Fig. 1 the proposed BERTweet Features-based model, which is based on the outputs encoding layers of the original BERTweet model. The BERTweet Features-based model[3] stems from the following question: how can we exploit, in a

[3] Sarcasm task: batch size 64, learning rate 0.0001, optimizer AdamW and 80 epochs. Irony task: batch size 32, learning rate 0.00002, optimizer AdamW, and 100 epochs

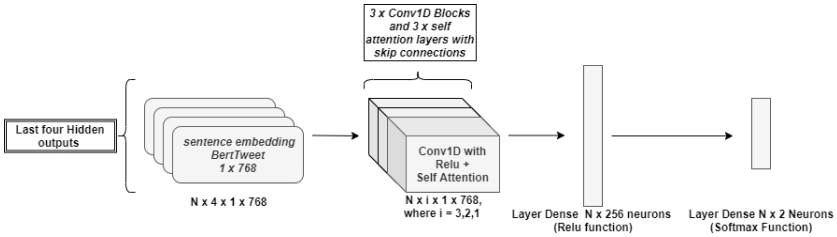
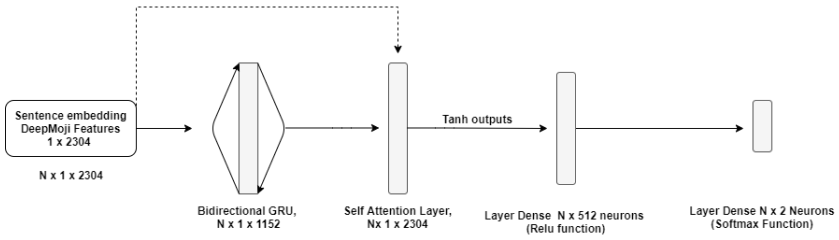


Fig. 1: BERTweet Features-based model.

flexible way, the sentence embeddings of the outputs of each encoding layer of a Transformer? To address this question, an architecture has been developed that focuses on the last four output layers of BERTweet. Instead of concatenating the various layers, they are joined by inserting the concept of flexibility, i.e. contextualised weights for the task to be analysed. In this case, the input tensor has a size of N x 4 x 1 x 768, where N denotes the number of training examples and the second dimension is associated with each input layer of the model. The next layers are based on the reduction of the number of channels to obtain a single one in order to merge the different information obtained from the different features' levels. They are developed on the basis of 1D Convolutions, self-attention layers and residual connections.

3.2 DeepMoji Features-based Model (DeepMoji-FB)

The DeepMoji Features-based model[4], presented in Fig. 2, takes as input a tensor of a dimension N x 1 x 2304. Each instance is the emotional embedding generated by the original DeepMoji model.



(2).png

Fig. 2: DeepMoji Features-based

In this case, different information is used for developing a new model for irony and sarcasm detection. The architecture of the DeepMoji Features-based model is slightly

[4] Sarcasm task: batch size 32, learning rate 0.00001, optimizer Adam and 25 epochs. Irony task: batch size 32, learning rate 0.0002, optimizer Adam, and 35 epochs

different from the BERTweet Features-based model. Indeed, the sentence embedding is fed into a Bidirectional GRU [1] layer. The most important aspect is that a new sentence embedding is generated using a skip connection between the input embedding and the output embedding generated by the BiGRU.

3.3 Ensemble of Ensembles (EoE)

The last model that we proposed is based on the combination of Bayesian Model Averaging, DeepMoji-FB and BERTweet-FB, by means of Soft/Hard classification. We will call this model 'Ensemble of ensembles' (EoE). The proposed EoE relies on a simple concept: exploiting several models, trained on different aspects of the text, to create a composition of models that better identifies the meaningful pattern of irony and sarcasm. Therefore, we created an ensemble that includes BMA, BERTweet-FB and DeepMoji-FB. Two different *classification strategies* have been evaluated: hard classification and soft classification. In particular, hard classification determines the final label of each testing instance by using the most frequent predicted label (i.e. majority voting), while soft classification selects the final label according to the sum of the marginal probability distributions given by each model.

4 Experimental Settings

In order to understand if the compared models are characterized by good generalization capabilities, we created the training and the test set (for both irony and sarcasm detection tasks), to make possible two different experimental scenarios: (1) train and test models using posts drawn from the same dataset to investigate the in-domain performance and (2) train and test models using posts coming from two different datasets, to estimate out-domain capabilities.

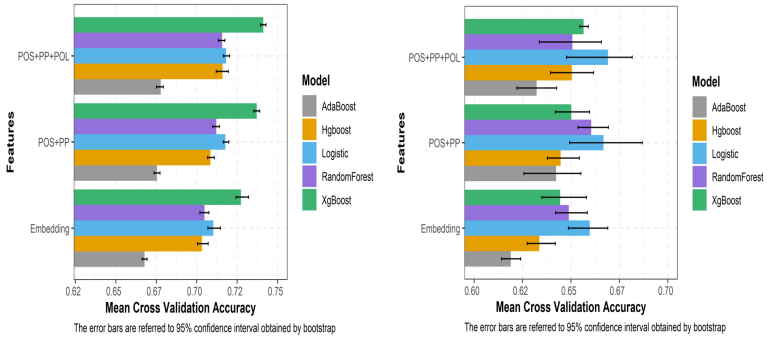
Training Set. In order to address **sarcasm** detection, three different datasets have been used for creating the training set to be supplied to the compared models: (1) Ptacek [10], composed of 14.070 sarcastic and 16.718 not sarcastic tweets; (2) Fersini [3], composed of 8.000 tweets, perfectly balanced between sarcastic and not sarcastic and (3) Gosh [5], that consists of 21.292 not sarcastic and 18.488 sarcastic tweets.

Regarding **irony** detection, two main datasets have been used for creating the training set to be used by the considered models: (1) SemEval-2018 Task 3A [13], specifically task 3A, composed of 1898 ironic and 1904 not ironic tweets; (2) Reyes [11], which consists of 10,000 ironic tweets, and 30,000 non-ironic posts about Politics, Humour, and Education. For irony detection, in order to compare the results of the proposed models with the state of the art, we considered the constrained and unconstrained settings defined at SemEval-2018 Task 3A. For the unconstrained scenario, we created a training set composed of the training released for SemEval-2018 Task 3A and the training of the Reyes dataset. The unconstrained settings will allow us to understand if, by introducing more variance in the training data (SemEval 2018 + Reyes), the models will maintain/improve their prediction capabilities on the test set (SemEval 2018). For the constrained settings, only the training set of the SemEval-2018 challenge has been used to train the models, and to be then validated on the SemEval 2018 test set.

Test Set. As far as **sarcasm** is concerned, two different test sets were selected: (1) Ghosh [5], which consists of 1975 samples, i.e. 975 labelled as non-sarcastic and 1000 labelled as sarcastic. This test set is used for in-domain validations; (2) Riloff [12], composed of 1956 tweets, i.e. 1648 non-sarcastic and 308 sarcastic posts. This test set is used for out-domain validations. Concerning **irony** detection, due to the limited number of available datasets, only the test set of [13] Task 3 A was chosen, with a total of 784 tweets, of which 473 as non-ironic and 311 as ironic. This test set is used for both constrained and unconstrained experimental settings. In the experiments, *Accuracy*, *Precision*, *Sensitivity*" and $F_1 - Measure$ are reported as the main measures of comparison among the models.

5 Results and Discussion

The first experiment regards the identification of the most representative features for identifying sarcasm and irony patterns (R1). To this purpose, Machine Learning classifiers introduced in Section 2, have been trained considering both embeddings (extracted from BERTweets and reduced by means of PCA) and hand-crafted features. The hyperparameters of each model have been optimized using a k-folds cross-validation based on random search and considering accuracy as the target metric to optimize.



(a) Sarcasm (b) Irony

Fig. 3: Comparison of feature contribution. The accuracy achieved by traditional machine learning models are reported, together with their confidence interval at 95%.

In Fig. 3, we report the most significant combinations of features considered by the traditional Machine Learning models. It is interesting to note that, for sarcasm detection, adding hand-creafted features related to pragmatic particles, part of speech and polarity, to the embeddings leads the models to achieve a significant improvement of F1 score with a 95% confidence level. However, this improvement emerges only in the case of sarcasm, while for the irony detection task, adding these features to the baseline of the embeddings, does not seem to discriminate better the information related to irony.

Regarding the remaining two research questions (R2 and R3), we compared the results of all the considered models, focusing on both in- and out-domain distributions. Fig. 4, reports the results achieved in terms of F1-Measure.

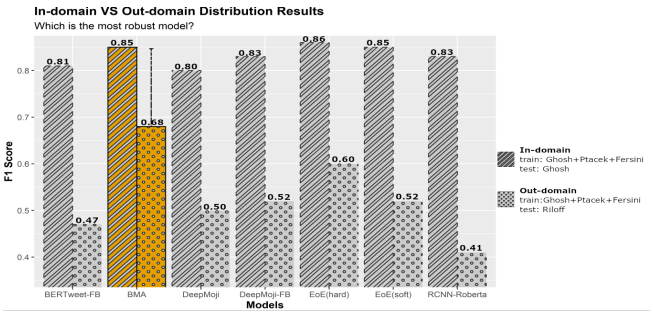


Fig. 4: Generalization abilities for sarcasm detection.

It is important to underline that the state of the art models, i.e. BMA, DeepMoji and RCNN-Roberta, achieve very good performance in the case of an in-domain distribution of the test set. However when processing a test set sampled from an out-domain distribution, the F1-Measure decreases by 40%. This suggests that the state of the art models focus on particular characteristics of the training set, and are not able to identify generic patterns for sarcasm that still hold for unseen data. The only exception is represented by BMA, which is much more robust when the unseen test data come from an out-domain distribution.

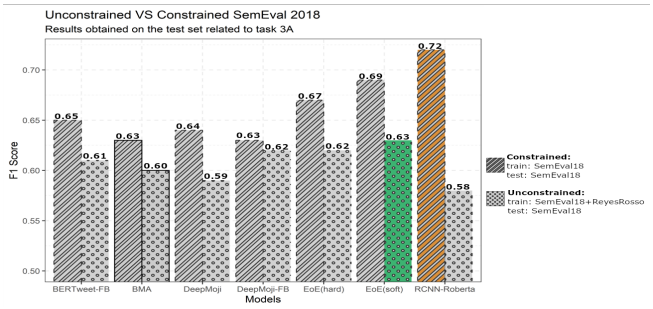


Fig. 5: Generalization abilities for irony detection.

Concerning the irony detection task, since only in-domain data are available for testing (SemEval 2018 test set), we compared the models in terms of constrained and unconstrained settings. When addressing an unconstrained task, where the training set

is composed of tweets from different datasets (i.e. SemEval and Reyes-Rosso), the performance of the various models deteriorates significantly with respect to the constrained task, where the training and the testing data come from the same (SemEval 2018) distribution. By comparing the results reported in Fig. 5, it emerges that all the models are not able to capture the features that can discriminate what is irony from what is not, denoting therefore reduced generalization capabilities. We can also highlight that even if RCNN-Roberta is the best performing model for the constrained task, when introducing more variance in the training set, the model is no longer able to generalize well. On the contrary, the proposed EoE model emerges as more robust than others.

Regarding irony detection both in a constrained and unconstrained settings, we report in Tables 1 and 2 the comparison of our best performing model (EoE with soft classification) with the systems ranked in the official SemEval 2018 competition. We can highlight that the proposed model, in the constrained case (Table 1), is ranked third (the rank of the constrained task was based on F1-Measure).

	Accuracy	Precision	Sensitivity	F_1-Measure
UCDCC	**0.797**	**0.788**	0.669	**0.724**
THUNGN	0.735	0.630	0.801	0.705
Ensemble of Ensembles (soft)	0.693	0.681	0.692	0.690
NTUA-SLP	0.732	0.654	0.691	0.672
WLV	0.643	0.532	**0.836**	0.650

Table 1: Ranking SemEval Task 3A, constrained

For the unconstrained task (Table 2), the results obtained by our EoE model are much better, highlighting that the proposed model outperforms the other teams that participated in the challenge (also in this case the rank of the unconstrained task was based on F1-Measure).

	Accuracy	Precision	Sensitivity	F_1-Measure
Ensemble of Ensembles (soft)	0.612	**0.661**	0.653	**0.631**
NonDicevo-SulSerio	**0.679**	0.583	0.666	0.622
INAOE-UPV	0.651	0.546	0.714	0.618
RM@IT	0.649	0.544	0.714	0.618
ValenTO	0.598	0.496	**0.781**	0.607

Table 2: Ranking SemEval Task 3A, unconstrained

The results reported above highlight that the proposed EoE model is quite robust even when considering more variance in the training data. In fact, EoE is not only ranked third in the constrained settings (with 0.69 of F_1-Measure), but it is placed first in the unconstrained scenario with a reduced drop of performance with respect to the constrained one.

6 Conclusions

The proposed models and the comparative analysis presented in this paper about irony and sarcasm has provided several insights. For the case of sarcasm, the models that achieved the best generalization are based on linguistic features, showing their robustness in the case of out-domain scenarios. Regarding irony, as the sample size increases, the performance of the models are reduced significantly. This shows that the structures of these models are not able to identify general information related to irony, but only focus on specific in-domain aspects.

References

1. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)
2. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proc. of Empirical Methods in Natural Language Processing. pp. 1615–1625 (2017)
3. Fersini, E., Pozzi, F.A., Messina, E.: Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–8. IEEE (2015)
4. Ghanem, B., Karoui, J., Benamara, F., Rosso, P., Moriceau, V.: Irony detection in a multilingual context. In: European Conference on Information Retrieval. pp. 141–149. Springer (2020)
5. Ghosh, A., Veale, T.: Fracking sarcasm using neural network. In: Proc. of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis. pp. 161–169 (2016)
6. Kayalvizhi, S., Thenmozhi, D., Kumar, B.S., Aravindan, C.: Ssn_nlp@ idat-fire-2019: Irony detection in arabic tweets using deep learning and features-based approaches. In: FIRE (Working Notes). pp. 439–444 (2019)
7. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14 (2020)
8. Potamias, R.A., Siolas, G., Stafylopatis, A.G.: A transformer-based approach to irony and sarcasm detection. Neural Computing and Applications **32**(23), 17309–17320 (2020)
9. Pozzi, F.A., Fersini, E., Messina, E.: Bayesian model averaging and model selection for polarity classification. In: International Conference on Application of Natural Language to Information Systems. pp. 189–200. Springer (2013)
10. Ptáček, T., Habernal, I., Hong, J.: Sarcasm detection on czech and english twitter. In: Proc. of the 25th international conference on computational linguistics. pp. 213–223 (2014)
11. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language resources and evaluation **47**(1), 239–268 (2013)
12. Riloff, E., Qadir, A., Surve, P., Silva, L.D., Gilbert, N., Huang, R.: Sarcasm as contrast between a positive sentiment and negative situation. In: Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 704–714 (2013)
13. Van Hee, C., Lefever, E., Hoste, V.: Semeval-2018 task 3: Irony detection in english tweets. In: Proc. of The 12th International Workshop on Semantic Evaluation. pp. 39–50 (2018)
14. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony detection via sentiment-based transfer learning. Information Processing & Management **56**(5), 1633–1644 (2019)