Irony Detection: from the Twittersphere to the News Space

Alessandra Cervone, Evgeny A. Stepanov, Fabio Celli, Giuseppe Riccardi
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, Italy
{alessandra.cervone,evgeny.stepanov}@unitn.it
{fabio.celli,giuseppe.riccardi}@unitn.it

Abstract

English. Automatic detection of irony is one of the hot topics for sentiment analysis, as it changes the polarity of text. Most of the work has been focused on the detection of figurative language in Twitter data due to relative ease of obtaining annotated data, thanks to the use of hashtags to signal irony. However, irony is present generally in natural language conversations and in particular in online public fora. In this paper, we present a comparative evaluation of irony detection from Italian news fora and Twitter posts. Since irony is not a very frequent phenomenon, its automatic detection suffers from data imbalance and feature sparseness problems. We experiment with different representations of text – bag-of-words, writing style, and word embeddings to address the feature sparseness; and balancing techniques to address the data imbalance.

Italiano. *Il rilevamento automatico di ironia è uno degli argomenti più interessanti in sentiment analysis, poiché modifica la polarità del testo. La maggior parte degli studi si sono concentrati sulla rilevazione del linguaggio figurativo nei dati di Twitter per la relativa facilità nell'ottenere dati annotati con gli hashtags per segnalare l'ironia. Tuttavia, l'ironia è un fenomeno che si trova nelle conversazioni umane in generale e in particolare nei forum online. In questo lavoro presentiamo una valutazione comparativa sul rilevamento dell'ironia in blogs giornalistici e conversazioni su Twitter. Poiché l'ironia non è un fenomeno molto frequente, il suo rilevamento automatico risente di problemi di mancanza di bilanciamento nei dati e feature sparseness. Per ovviare alla feature sparseness proponiamo esperimenti con diverse rappresentazioni del testo – bag-of-words, stile di scrittura e word embeddings; per ovviare alla mancanza di bilanciamento nei dati utilizziamo invece tecniche di bilanciamento.*

1 Introduction

The detection of irony in user generated content is one of the major issues in sentiment analysis and opinion mining (Ravi and Ravi, 2015). The problem is that irony can flip the polarity of apparently positive sentences, negatively affecting the performance of sentiment polarity classification (Poria et al., 2016). Detecting irony from text is extremely difficult because it is deeply related to many out-of-text factors such as context, intonation, speakers' intentions, background knowledge and so on. This also affects interpretation and annotation of irony by humans, often leading to low inter-annotator agreements.

Twitter posts are frequently used for the irony detection research, since users often signal irony in their posts utilizing hashtags such as *#irony*, *#justjoking*, etc. Despite the relative ease of collecting the data, Twitter is a very particular kind of text. In this paper we experiment with different representations of text to evaluate the utility of Twitter data for the detection of irony in text coming from other sources such as news fora. The representations of text – bag-of-words, writing style, and word embeddings – are chosen such that they are not dependent on the resources available for the language. Due to the fact that irony is less frequent than literal meaning, the data is usually imbalanced. We experiment with balancing techniques such as random undersampling, random oversampling and cost-sensitive training to observe its effects on a supervised irony detection.

The paper is structured as follows. In Section 2 we introduce related work on irony. In Section 3 we describe the corpora used throughout experiments. In Sections 4 and 5 we describe the methodology and the result of the experiments. In Section 6 we provide concluding remarks.

2 Related Works

The detection of irony in text has been widely addressed. Carvalho et al. (2009) showed that in Portuguese news blogs, pragmatic and gestural text features such as emoticons, onomatopoeic expressions and heavy punctuation marks work better than deeper linguistic information such as n-grams, words or syntax. Reyes et al. (2013) addressed irony detection in Twitter, using complex features like temporal expressions, counterfactuality markers, pleasantness or imageability of words, and pair-wise semantic relatedness of terms in adjacent sentences. This rich feature set enabled the same authors to detect 30% of the irony in movie and book reviews in (Reyes and Rosso, 2014).

Ravi and Ravi (2016), on the other hand, exploited resources such as LIWC (Tausczik and Pennebaker, 2010) to analyze irony in two different domains: satirical news and Amazon reviews; and found out that LIWC's words related to sex or death are good indicators of irony.

Charalampakis et al. (2016) addressed irony detection in Greek political tweets comparing semi-supervised and supervised approaches, with the aim to analyze whether irony predicts election results or not. In order to detect irony, they use as features: spoken style words, word frequency, number of WordNet SynSets as a measure of ambiguity, punctuation, repeated patterns and emoticons. They found that supervised methods work better than semi-supervised in the prediction of irony (Charalampakis et al., 2016).

Poria et al. (2016) developed models based on pre-trained convolutional neural networks (CNNs) to exploit sentiment, emotion and personality features for a sarcasm detection task. They trained and tested their models on balanced and unbalanced sets of tweets retrieved searching the hashtag #sarcasm. They found that CNNs with pre-trained models perform very well and that, although sentiment features are good also when used alone, emotion and personality features help in the task (Poria et al., 2016).

Sulis et al. (2016) investigated a new set of features for irony detection in Twitter with particular regard to affective features; and studied the difference between irony and sarcasm. Barbieri et al. (2014) were the first ones to propose an approach for irony detection in Italian.

Irony detection is a popular topic for shared tasks and evaluation campaigns. Among others, SemEval-2015 (Ghosh et al., 2015) task on sentiment analysis of figurative language in Twitter, and SENTIPOLC 2014 (Basile et al., 2014) and 2016 (Barbieri et al., 2016) tasks on irony and sentiment classification in Twitter. SemEval considered three broad classes of figurative language: irony, sarcasm and metaphor. The task was cast as a regression as participants had to predict a numeric score (crowd-annotated). The best performing systems made use of manual and automatic lexica, term-frequencies, part-of-speech tags, and emoticons.

The SENTIPOLC campaigns on Italian tweets, on the other hand, included three tasks: subjectivity detection, sentiment polarity classification and irony detection (binary classification). The best performing systems utilized broad sets of features ranging from the established Twitter-based features, such as URL links, mentions, and hashtags, to emoticons, punctuation, and vector space models to spot out-of-context words (Castellucci et al., 2014). Specifically, in SENTIPOLC 2016, the best performing system exploited lexica, handcrafted rules, topic models and Named Entities (Di Rosa and Durante, 2016). In this paper, on the other hand, we address irony detection from features not dependent on language resources such as manually crafted lexica and source-dependent features such as hashtags and emoticons.

3 Data Set

The experiments reported in this paper make use of two data sets: SENTIPOLC 2016 (Barbieri et al., 2016) and CorEA (Celli et al., 2014). While SENTIPOLC is a corpus of tweets, CorEA is a data set of news articles and related reader comments collected from the Italian news website *corriere.it*. The two corpora consist of inherently different types of text. While tweets have a limit on the length of the post, news articles comments are not constrained. The length limitation does not only impact the number of tokens per post, but also the style of writing, since in Tweets authors

SENTIPOLC 2016	CorEA
@gadlernertweet Se #Grillo fosse al governo, dopo due mesi lo Stato smetterebbe di pagare stipendi e pensioni. E lui capeggerebbe la rivolta	bravo, escludi l'universitá .... restare ignoranti non fa male a nessuno, solo a sé stessi. questi sono i nostri.... geni. non mi meraviglierei se votasse grillo
#Grillo,fa i comizi sulle cassette della frutta,mentre alcune del #Pdl li fanno senza,cassetta...solo sulle banane. #ballaró @Italialand	beh dipende da come la guardi..A campagna elettorale all'inverso: rispettano ció che avevano promesso
@MissAllyBlue Non mi fido della compagnia.. meglio far finta di stare sveglio.. sveglissimo O_o	Saranno solo 4 milioni (comunque dimentichi i 42 mil di rimborsi) peró pochi o tanti li hanno restituiti. Gli altri invece , probabilmente politici a te "simpatici" continuano a gozzovigliare con i soldi tuoi . Sveglia volpone

Table 1: Examples of ironic posts from SENTIPOLC 2016 and CorEA.

naturally try to squeeze as much content as possible within the limits.

This difference can be seen also in the type of irony used across the two corpora, as shown in the examples reported in Table 1. While in Tweets we observe much more the presence of external 'sources' (such as URL links, mentions, hashtags and emoticons) to signal the irony and make it interpretable (for example by disambiguating entities using hashtags); news fora users tend to use style much more similar to natural language, where entities are not specifically signaled and there are no emojis to mark the non-literal meaning of a sentence. Thus, CorEA presents a more difficult, but also a more interesting, dataset for automatic irony detection, given the closer similarity to the language used in other genres.

Both corpora have been annotated following a version of the scheme of SENTIPOLC 2014 (Basile et al., 2014). According to the scheme, the annotator is asked to decide whether the given text is subjective or not, and in case it is considered subjective, to annotate the polarity of the text and irony as binary values. The CorEA corpus (Celli et al., 2014) was annotated for irony by three annotators specifically for this paper, and has an inter-annotator agreement of $\kappa = 0.57$.

Since SENTIPOLC 2016 is composed of different data sets, which used various agreement metrics (Barbieri et al., 2016), it is not possible to directly compare the inter-annotator agreements between the corpora. The two component data sets of SENTIPOLC 2016 for which a comparable metric is reported have an inter-annotator agreement of $\kappa = 0.538$ (TW-SENTIPOLC14) and $\kappa = 0.492$ (TW-BS) (Stranisci et al., 2016).

Despite the differences in the number of posts (9,410 for SENTIPOLC and 2,875 for CorEA; see Table 2); due to the length constraint of the former, the corpora have comparable numbers of tokens:

	Non-Ironic		Ironic		Total
SENTIPOLC 2016					
Training	6,542	(88%)	868	(12%)	7,410
Test	1,765	(88%)	235	(12%)	2,000
CorEA	2,299	(80%)	576	(20%)	2,875

Table 2: Counts and percentages of ironic and non-ironic posts in SENTIPOLC 2016 training and test set and CorEA corpus.

159K for SENTIPOLC and 164K for CorEA. Consequently, there are drastic differences in the average number of tokens per post: 21 for SENTIPOLC and 57 for CorEA. As shown in Table 2, we also observe a major difference in the percentages of ironic posts between the corpora: 12% for SENTIPOLC and 20% for CorEA.

4 Methodology

In this paper we address irony detection in Italian making use of source independent and 'easily' obtainable representations of text such as lexical (bag-of-words), stylometric, and word embedding vectors. The models are trained and tested using Support Vector Machines (SVM) (Vapnik, 1995) with linear kernel and defaults parameters, implemented in the scikit-learn (Pedregosa et al., 2011) python library.

To obtain the desired representations of text, the data is pre- For the bag-of-word representation, the data is lowercased, and all source-specific entities, such as emoji, URL, Twitter hashtags, and mentions are mapped to a single entity (e.g. $\langle$H$\rangle$ for hashtags); as the objective is to use Twitter models to detect irony in news fora and other kinds of textual data, where presence of such entities is less likely. We also apply a cut-off frequency and remove all the tokens that appear in a single document only.

For the style representation, we use the lexical richness metrics based on type and token frequen-

cies such as type-token ratio, entropy, Guiraud's R, Honores H, etc. (Tweedie and Baayen, 1998) (22 features); and character-type ratios, (including specific punctuation marks) (46 features) that previously were successfully applied to tasks such as agreement-disagreement classification (Celli et al., 2016) and mood detection (Alam et al., 2016).

To extract the word embedding representation (Mikolov et al., 2013), we use skip-gram vectors (size: 300, window: 10) pre-trained on Italian Wikipedia, and a document is represented as a term-frequency weighted average of per-word vectors.

Since our goal is to analyze utility of Twitter data for irony detection in Italian news fora, we first experiment with the text representations and chose models that behave above chance-level baseline on per-class F_1 scores and Micro-F_1 score using a 10-fold stratified cross-validation setting. Even though on imbalanced data the frequently used evaluation metric is Macro-F_1 score, e.g. (Barbieri et al., 2016), which we report for comparison purposes; it is misleading as it does not reflect the amount of correctly classified instances. The majority baseline, on the other hand, is very strong for highly imbalanced data sets, and is provided for reference purposes only.

As data imbalance has been observed to adversely affect irony detection performance (Poria et al., 2016; Ptacek et al., 2014), we experiment with simple balancing techniques such as random under- and oversampling and cost sensitive training. While undersampling balances the data set by removing majority class instances, oversampling achieves that by replicating (copying) minority class instances. Undersampling is often reported as a better option, as oversampling may lead to overfitting problems (Chawla et al., 2002). In cost-sensitive training, on the other hand, the performance on minority class is improved by higher misclassification costs for it. In the paper, the selected representations are analyzed in terms of balancing effects and cross-source performance (Twitter - news fora).

5 Results and Discussion

The results of experiments comparing different document representations – bag-of-words, writing style, and word embeddings – are presented in Table 3 for stratified 10-fold cross-validation on both corpora (SENTIPOLC and CorEA). The

Model	NI	I	Mic-F_1	Mac-F_1
SENTIPOLC: Training				
BL: Chance	0.8783	0.1183	0.7862	0.4983
BL: Majority	0.9378	0.0000	0.8829	0.4689
BoW	0.8979	**0.2112**	0.8207	**0.5546**
Style	0.8817	0.0892	0.7612	0.4605
WE	0.9361	0.0044	0.8799	0.4702
CorEA				
BL: Chance	0.7952	0.1895	0.6733	0.4923
BL: Majority	0.8886	0.0000	0.7996	0.4443
BoW	0.8414	**0.2951**	0.7411	**0.5682**
Style	0.7116	0.1688	0.6186	0.4402
WE	0.8811	0.1447	0.7912	0.5129

Table 3: Average per-class, micro and macro-F_1 scores for stratified 10-fold cross-validation on SENTIPOLC 2016 training set and CorEA for different **document representations**: bag-of-words (*BoW*), stylometric features (*Style*) and word embeddings (*WE*). *BL: Chance* and *BL: Majority* are chance-level and majority baselines. **NI** and **I** are non-ironic and ironic classes, respectively.

document representations behave similarly across corpora, and the only representation that achieves above chance-level per-class and micro-F_1 scores is the bag-of-words. At the same time, it achieves the highest macro-F_1 score. However, none of the representations is able to surpass the majority baseline in terms of micro-F_1.

The performance of the bag-of-words representation on data balancing techniques is presented in Table 4. The training with natural distribution (*BoW: ND*) yields the best performance across the corpora. For SENTIPOLC data, it is the only model that produces above chance-level (Table 3: *BL: Chance*) performances for per-class and micro-F_1 scores.

Cost-sensitive training (*BoW: CS*) and random oversampling (*BoW: RO*) perform very close. For CorEA corpus, all balancing techniques except random undersampling (*BoW: RU*) yield above chance-level performances. Random undersampling, however, yields the highest F_1 score for the irony class, which unfortunately comes at the expense of the overall performance. This verifies previous observations in the literature that undersampling leads to negative effect on novel imbalanced data (Stepanov and Riccardi, 2011). Since cost-sensitive training achieves the best performance in terms of macro-F_1 score, which was used as official evaluation metrics in SENTIPOLC 2016 (Barbieri et al., 2016), it is retained for SENTIPOLC training-test and cross-corpora (SEN-

Model	NI	I	Mic-F_1	Mac-F_1
SENTIPOLC: Training				
BoW: ND	**0.8979**	0.2112	**0.8207**	0.5546
BoW: CS	0.8732	0.2493	0.7861	**0.5612**
BoW: RO	0.8737	0.2375	0.7857	0.5555
BoW: RU	0.7270	**0.2679**	0.6115	0.4974
CorEA				
BoW: ND	**0.8414**	0.2951	**0.7411**	0.5682
BoW: CS	0.8331	0.3202	0.7321	**0.5766**
BoW: RO	0.8302	0.3138	0.7279	0.5720
BoW: RU	0.6882	**0.3599**	0.5810	0.5241

Table 4: Average per-class, micro and macro-F_1 scores for stratified 10-fold cross-validation on SENTIPOLC 2016 training set and CorEA for **balancing techniques**: cost-sensitive training (*CS*), random oversampling (*RO*) and random undersampling (*RU*). *ND* is training with natural distribution of classes (*BoW* in Table 3). **NI** and **I** are non-ironic and ironic classes, respectively.

Model	NI	I	Mic-F_1	Mac-F_1
SENTIPOLC: Training - Test Split				
BL: Chance	0.8826	0.1155	0.7927	0.4990
BL: Majority	0.9376	0.0000	0.8825	0.4688
SoA	0.9115	0.1710	–	0.5412
BoW: ND	**0.9330**	0.1678	**0.8760**	0.5504
BoW: CS	0.9245	**0.2023**	0.8620	**0.5634**
SENTIPOLC - CorEA: 10-fold testing				
BL: Chance	0.8393	0.1213	0.7286	0.4803
BL: Majority	0.8886	0.0000	0.7996	0.4443
BoW: ND	**0.8164**	0.1755	**0.7001**	0.4959
BoW: CS	0.8109	**0.2020**	0.6945	**0.5065**

Table 5: Average per-class, micro and macro-F_1 scores for SENTIPOLC Training-Test split and 10-fold testing of SENTIPOLC models on CorEA for bag-of-words representation with imbalanced (*ND*) and cost-sensitive (*CS*) training. *SoA* are the state-of-the-art results for SENTIPOLC 2016: the system of (Di Rosa and Durante, 2016). *BL: Chance* and *BL: Majority* are chance-level and majority baselines. **NI** and **I** are non-ironic and ironic classes, respectively.

TIPOLC - CorEA) evaluation along with the models trained on natural imbalanced distribution with equal costs.

The final models make use of bag-of-words representation and are trained on SENTIPOLC training set in cost-sensitive and insensitive settings. The evaluation of models is performed on SENTIPOLC 2016 test set and CorEA's 10-folds. This setting allows us to compare our results to the state of the art on SENTIPOLC data and CorEA's cross-validation setting. From the results in Table 5, we observe that on the SENTIPOLC test set both models outperform the state of the art in terms of macro-F_1 score. The model with cost-sensitive training additionally outperforms it in terms of irony class F_1 score. However, both models fall slightly short of outperforming the majority baseline in terms of micro-F_1.

In the cross-corpora setting the behavior of models is similar – cost-sensitive training favors minority class F_1 and macro-F_1 scores. While both models perform worse than the chance-level baseline generated using the label distribution of SENTIPOLC data in terms of micro-F_1, they both outperform it in terms of irony class F_1 score. However, only the model with cost-sensitive training yields statistically significant difference using paired two-tail t-test with $p = 0.05$.

6 Conclusion

We have presented experiments on irony detection in Italian Twitter and news fora data comparing different document representations – bag-of-words, writing style as stylometric features, and word embeddings. The objective is to evaluate the suitability of Twitter data for detecting irony in news fora. The models were compared for balanced and imbalanced training, as well as cross-corpora performance. We have observed that the bag-of-words representation with imbalanced cost-insensitive training produces the best results (micro-F_1) across settings, closely followed by cost-sensitive training.

The models outperform the results on irony detection in Italian tweets (Di Rosa and Durante, 2016) in terms of macro-F_1 scores reported for SENTIPOLC 2016 (Barbieri et al., 2016). However, micro-F_1 is the most informative metric for the downstream application of irony detection, as it considers the total amount of true positives. Given that the highest micro-F_1 is attained by the majority baselines for both corpora (0.8829 for SENTIPOLC and 0.7996 for CorEA), the task of irony detection is far from being solved.

Acknowledgments

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 610916 – SENSEI.

We would like to thank Paolo Rosso and Mirko Lai for their help in annotating CorEA.

References

F. Alam, F. Celli, E.A. Stepanov, A. Ghosh, and G. Riccardi. 2016. The social mood of news: Self-reported annotations to design automatic mood detection systems. In *PEOPLES @COLING*.

F. Barbieri, F. Ronzano, and H. Saggion. 2014. Italian irony detection in twitter: a first approach. In *CLiC-it 2014 & EVALITA*.

F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *CLiC-it - EVALITA*.

V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. In *EVALITA*.

P. Carvalho, L. Sarmento, M.J. Silva, and E. De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-. In *Topic-sentiment analysis for mass opinion*.

G. Castellucci, D. Croce, and R. Basili. 2014. Context-aware convolutional neural networks for twitter sentiment analysis in italian. In *EVALITA*.

F. Celli, G. Riccardi, and A. Ghosh. 2014. CorEA: Italian news corpus with emotions and agreement. In *CLIC-it*.

F. Celli, E.A. Stepanov, and G. Riccardi. 2016. Tell me who you are, I'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blogs. In *NLPJ @IJCAI*.

B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence*, 51:50–57.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

E. Di Rosa and A. Durante. 2016. Tweet2check evaluation at evalita sentipolc 2016. In *CLiC-it - EVALITA*.

A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *SemEval*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

S. Poria, E. Cambria, D. Hazarika, and P. Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv:1610.08815*.

T. Ptacek, I. Habernal, and J. Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING*.

K. Ravi and V. Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*.

K. Ravi and V. Ravi. 2016. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*.

A. Reyes and P. Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*.

A. Reyes, P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

E.A. Stepanov and G. Riccardi. 2011. Detecting general opinions from customer surveys. In *SENTIRE @ICDM*.

M. Stranisci, C. Bosco, D.I. Hernández Farías, and V. Patti. 2016. Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *LREC*.

E. Sulis, D.I. Hernández Farías, P. Rosso, V. Patti, and G. Ruffo. 2016. Figurative messages and affect in Twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143.

Y.R. Tausczik and J.W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*.

F.J. Tweedie and R.H. Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*.

V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.