#NonDicevoSulSerio at SemEval-2018 Task 3: Exploiting Emojis and Affective Content for Irony Detection in English Tweets

Endang Wahyu Pamungkas, Viviana Patti
Dipartimento di Informatica, University of Turin
{pamungka,patti}@di.unito.it

Abstract

This paper describes the participation of the #NonDicevoSulSerio team at SemEval2018-Task3, which focused on *Irony Detection in English Tweets* and was articulated in two tasks addressing the identification of irony at different levels of granularity. We participated in both tasks proposed: Task A is a classical binary classification task to determine whether a tweet is ironic or not, while Task B is a multi-class classification task devoted to distinguish different types of irony, where systems have to predict one out of four labels describing verbal irony by clash, other verbal irony, situational irony, and non-irony. We addressed both tasks by proposing a model built upon a well-engineered features set involving both syntactic and lexical features, and a wide range of affective-based features, covering different facets of sentiment and emotions. The use of new features for taking advantage of the affective information conveyed by emojis has been analyzed. On this line, we also tried to exploit the possible incongruity between sentiment expressed in the text and in the emojis included in a tweet. We used a Support Vector Machine classifier, and obtained promising results. We also carried on experiments in an unconstrained setting.

1 Introduction

The use of creative language and figurative language devices such as irony has been proven to be pervasive in social media (Ghosh et al., 2015). The presence of these devices makes the process of mining social media texts challenging, especially because they can influence and twist the sentiment polarity of an utterance in different ways. Glossing over differences across different theoretical accounts proposed in the context of various disciplines (Gibbs and Colston, 2007; Grice, 1975; Wilson and Sperber, 1992; Attardo, 2007; Giora, 2003), irony can be defined as an incongruity between the literal meaning of an utterance and its intended meaning (Karoui et al., 2017). The term irony covers mainly two phenomena: verbal and situational irony (Attardo, 2006). Situational irony refers to events or situations which fail to meet expectations, such as for instance "warnings the dangerous effect of smoking on the cigarette advertisement", while verbal irony occurs when the speaker intend to communicate a different meaning w.r.t what he/she is literally saying. Most of the time it involves the intention of communicating an *opposite meaning*, and this kind of opposition can be expressed by polarity contrast. However this is not the only possibility, and social media messages well reflect such variety, including different expressions of verbal irony and descriptions of situational irony (Van Hee et al., 2016a; Sulis et al., 2016). Automatic irony detection is an important task to improve sentiment analysis (Reyes et al., 2013; Maynard and Greenwood, 2014). However, detecting irony automatically from textual messages is still a challenging task for scholars (Joshi et al., 2017). The linguistic and social factors which impact on the perception of irony contribute to make the task complex.

In this paper, we will describe the irony detection systems we developed for participating in SemEval2018-Task3: *Irony Detection in English Tweets* (Van Hee et al., 2018). Our systems used a support vector classifier model by exploiting some novel and well-handcrafted features including lexical, syntactical and affective based features. We participated in 3 different scenarios (Task A constrained, Task A unconstrained, and Task B unconstrained). The official results show that our system outperformed all systems in the unconstrained setting on both tasks and was able to achieve a reasonable score in Task A constrained, ranking in the top ten out of 44 submissions.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 649–654
New Orleans, Louisiana, June 5–6, 2018.

2 The #NonDicevoSulSerio System

We performed our experiments using a support vector machine classifier, with *radial basis function* kernel. We exploited different kind of features (lexical, syntactical and affective-based), which has been proven effective in literature to identify ironic phenomena. In addition, we also investigated the use of novel features aimed at exploiting information conveyed by emojis, studying in particular sentiment incongruity between sentiment expressed in the text and in the emojis of a tweet.

2.1 Structural Features

Structural features consist of lexical and syntactical features which characterize Twitter data. Such kind of features has been proven beneficial in several tasks dealing with Twitter data, and we selected the most relevant ones for irony detection.
Hashtag Presence: binary value 0 (if no hashtag in tweet) and 1 (if hashtag contained in tweet). **Hashtag Count**: number of hashtags contained in tweet. **Mention Count**: number of mentions contained in tweet. **Exclamation Mark Count**: number of exclamation marks contained in tweet. **Upper Case Count**: number of upper case characters in tweet. **Link Count**: number of links (http) contained in tweet. **Link Presence**: binary value 0 (if no link in tweet) and 1 (if at least one link found in tweet). **Has Quote**: binary value 0 (if quote (“ ” or ’ ’) not found in tweet) and 1 (if at least one pair of quote (“ ” or ’ ’) found in tweet). **Intensifiers & Overstatement Words Count**: number of intensifiers and words typically used in ironic overstatements [1] found in tweet. **Emoji Presence**: binary value 0 (if no emoji found in tweet) and 1 (if at least one emoji found in tweet). **Repeated Character**: binary value 0 (if there is no repeated character found in tweet) and 1 (if at least three characters repeated consequently in one word found in tweet). **Text Length**: the length of characters in each tweet. **Conjunction Count**: the number of conjunctions found in tweet. **Verb Count**: the number of verbs found in tweet. **Noun Count**: the number of nouns found in tweet. **Adjective Count**: the number of adjectives found in tweet. We use Standford PoS-Tagger [2] to get the count of conjunctions, verbs, nouns, and adjectives.

[1] love, really, lovely, like, great, brilliant, perfect, thank, glad.

[2] https://nlp.stanford.edu/software/tagger.shtml

2.2 Affective-Based Features

Affective features were proven effective in prior work to detect irony in tweets (Farías et al., 2016). We exploited available affective resources to extract affective information trying to capture multiple facets of affects -sentiment polarity and emotions- by selecting a few resources developed for English, which refers to both categorical and dimensional models of emotions.
AFINN.: AFINN is a sentiment lexicon consisting of English words labeled with valence score between -5 and 5. We used the normalized version of AFINN in (Farías et al., 2016), where the valence score was already normalized to the range between 0 and 1.
Emolex. Emolex (Mohammad and Turney, 2013) was developed by using crowdsourcing. Emolex contains 14,182 words associated with eight primary emotion based on (Plutchik, 2001).
EmoSenticNet. EmoSenticNet(EmoSN) (Poria et al., 2013) is an enriched version of SenticNet, where emotion labels were added by mapping WordNet-Affect labels to the SenticNet concepts. WordNet-Affect labels refers to six Ekman’s basic emotions.
Linguistic Inquiry and Word Count (LIWC). LIWC dictionary (Pennebaker et al., 2001) has 4,500 words distributed into 64 different emotional categories including positive and negative. Here we only use the positive (PosEMO) and negative emotion (NegEMO) categories.
Dictionary of Affect in Language (DAL). DAL was developed by (Whissell, 2009) and composed of 8,742 English words. These words were labeled by three scores representing the emotion dimensions *Pleasantness*, *Activation*, and *Imagery*.
Emoji Sentiment Ranking. Since we observed the presence of a lot of emojis in Twitter data, we used the emoji sentiment ranking lexicon by (Novak et al., 2015) to get the sentiment score of each emoji in the tweet. We also tried to detect *sentiment incongruity* between text and emoji in the same tweet. We used VADER (Hutto and Gilbert, 2014) to extract the polarity score of the text.

3 Experiment and Results

3.1 Task Description and Dataset

SemEval2018-Task3’s organizers proposed two subtasks related to the topic of detecting irony in Twitter automatically (Van Hee et al., 2018). SubTask A is a binary classification task, where every

SubTask A				
	Irony		Not Irony	
Training	1911		1923	
Testing	311		473	
SubTask B				
	0	**1**	**2**	**3**
Training	1923	1390	316	205
Testing	473	164	85	62

Table 1: Dataset Distribution on Both Tasks.

0 : Not irony
1 : Verbal irony by polarity contrast
2 : Others irony
3 : Situational irony

SubTask A		
Amt.	**HashTag**	**Source**
500	#irony	(Barbieri et al., 2014)
400	#sarcasm	(Riloff et al., 2013)
100	#sarcasm	(Barbieri et al., 2014)
500	#not	(Sulis et al., 2016)
500	non-irony	(Mohammad et al., 2015)
500	non-irony	(Ptáček et al., 2014)
500	non-irony	(Riloff et al., 2013)
SubTask B		
Amt.	**HashTag**	**Source**
867	#irony	(Barbieri et al., 2014)

Table 2: Additional data on Unconstrained Scenario.

system should determine whether a tweet is ironic or not ironic. Meanwhile, SubTask B is defined as a multi-class classification problem, where the aim is to classify each tweet into four different categories including: verbal irony by polarity contrast, other verbal irony, situational irony, and not irony. In both tasks, organizers allowed submissions in two scenarios: constrained and unconstrained. In unconstrained settings, participants were allowed to exploit external data from other corpora annotated with irony labels in the training phase. Standard evaluation metrics were proposed for the task, including, precision, recall, accuracy, and F_1-score.

Dataset The organizers provided 3,834 training data and 784 test data for both tasks. Table 1 shows the dataset distribution. Data were collected by using three irony-related hashtags: #irony, #sarcasm, and #not. Datasets for both tasks were manually labeled by using the fine-grained annotation scheme in (Van Hee et al., 2016b). A two-layer annotation has been applied on the same tweets, one concerning the presence and absence of irony, the second one identifying different types of irony, when irony is present. As a consequence, as Table 1 shows, there is a class imbalance on SubTask B dataset in favor of non-ironic class (50%), verbal irony by polarity contrast (25%), other verbal irony (13%) and situational irony (12%). The irony-related hashtags were removed from the final dataset release.

3.2 Experimental Setup

We built our supervised systems based on available training data. In this phase performances were evaluated based on the mean of F_1-score, by using 10-fold cross validation. We chose an SVM classifier with radial basis function kernel[3]. Our system implementation is free available for research purpose in GitHub page [4]. Therefore, we lean on feature selection process to improve the system performance. We carried on an ablation test on our feature sets to get the highest F_1-score. We decided to participate in three different scenarios: SubTask A constrained, SubTask A unconstrained, and SubTask B unconstrained.

For unconstrained scenario in SubTask A, we used the available corpora from previous work. We tried to add new data with balance proportion (1500 ironic and 1500 non-ironic). We also added a balance proportion of ironic data based on different hashtag (500 #irony, 500 #sarcasm, and 500 #not) from three different corpora, with the aim of enriching the training data with ironic samples of various provenance and trying to avoid biases. The distribution and source of our additional data can be seen in Table 2.

In SubTask B, we proposed to use a pipeline approach in three-steps classification scenario. First, we classify the ironic and non-ironic (similar configuration with SubTask A). Second, we classify the ironic data from step one into two categories, *verbal_irony_by_polarity_contrast* and the rest (other_verbal_irony+situational_irony). In the second step, we add more training data on the *other_verbal_irony+situational_irony class* to

[3]SVM good performances for similar tasks were recognized (Joshi et al., 2017). We built our system by using scikit-learn Python Library http://scikit-learn.org/.

[4]https://github.com/dadangewp/SemEval-2018-Task-3

Structural Features	**System 1**	**System 2**	**System 3**	**System 4**
	Task A (C)	**Task A (U)**		
	Task B (U)-1		**Task B (U)-2**	**Task B (U)-3**
Hashtag Count	✓	✓	✓	✓
Hashtag Presence	✓	✓	✓	-
Mention Count	✓	✓	-	-
Exclamation Mark	-	-	✓	-
UpperCase Count	-	✓	-	-
Link Count	✓	✓	-	✓
Link Presence	✓	✓	-	-
Has Quote	-	-	✓	-
Intensifiers/Overstatement	✓	✓	✓	-
Emoji Presence	-	-	✓	✓
Repeated Chars	-	-	✓	✓
Text Length	✓	-	✓	✓
Conjunction Count	✓	-	✓	✓
Noun Count	✓	✓	-	✓
Adjective Count	-	-	✓	✓
Verb Count	-	-	✓	✓
Affective Features				
AFINN Score	-	-	✓	✓
DAL Pleasantness	✓	-	-	-
DAL Activation	✓	-	-	-
DAL Imagery	-	-	✓	-
Emolex Surprise	-	-	✓	-
Emolex Trust	✓	-	-	-
Emolex Positive	-	-	✓	-
Emolex Negative	✓	-	✓	-
Emolex Anticipation	-	-	✓	-
Emolex Fear	✓	✓	✓	✓
LIWC Positive	-	-	✓	-
LIWC Negative	-	-	-	✓
EmoSenticNet Disgust	✓	✓	✓	✓
EmoSenticNet Fear	-	-	-	✓
EmoSenticNet Joy	-	✓	✓	✓
EmoSenticNet Sad	-	✓	-	✓
EmoSenticNet Surprise	-	-	✓	-
Vader Sentiment Score	-	-	✓	-
Emoji Incongruity	✓	✓	-	-

Table 3: Feature Selection on each System.

overcome the imbalance issue. We decided to use only additional tweets marked with #irony hashtags, relying on the analysis in (Sulis et al., 2016) suggesting that the polarity reversal phenomenon seems to be relevant in messages marked with #sarcasm and #not, but less relevant for messages tagged with #irony. In the last step, we classify between *other_verbal irony* and *situational_irony*. Table 3 shows selected features on each submitted system based on our ablation test.

3.3 Result and Analysis

Table 4 shows our experimental results based on four different metrics including accuracy, precision, recall, and F_1-score. For experiments on the training set we used 10-fold cross validation, and we report the score for each metric. However, F_1-score has been used as the criterion to tune the configuration. Official Codalab results show that our system ranked 10^{th} out of 44 submissions on SubTask A and 9^{th} out of 32 on SubTask B. We obtained F_1-score 0.6216 (Best system: 0.7054) on SubTask A and 0.4131 (Best system: 0.5074) on SubTask B. However, our system outperformed all systems in the unconstrained setting on both tasks.

Based on our analysis, several stylistic features were very effective in Task A (both in constrained and unconstrained settings). Especially, Twitter specific symbols such as hashtags, mentions, and URLs were very useful to discriminate non ironic tweets. In addition, we found that affective resource were very helpful in the Step 2 and Step 3 of Task B, especially Emolex (Step 2) and EmoSenticNet (Step 3). Another important finding is that additional data on Task A did not improve the classifier performance. Instead, additional tweets marked with #irony on Task B were very useful to handle the imbalance dataset in Step-2 (*verbal_irony_by_polarity_contrast* vs *other_verbal_irony+situational_irony*). Our clas-

SubTask A Constrained				
	Acc	Prec	Rec	F_1
Training	0.630	0.616	0.664	0.638
Testing	0.666	0.562	0.717	0.630
SubTask A Unconstrained				
	Acc	Prec	Rec	F_1
Training	0.617	0.615	0.646	0.630
Testing	0.679	0.583	0.666	0.622
SubTask B Unconstrained				
	Acc	Prec	Rec	F_1
Training-1	0.630	0.616	0.664	0.638
Training-2	0.702	0.671	0.779	0.720
Training-3	0.689	0.651	0.488	0.544
Testing	0.555	0.409	0.441	0.413

Table 4: Results on Training and Test sets.

sifier was able to achieve a high F_1 score on the training phase in this case. Furthermore, we also found that our new features for capturing affective information in emojis (e.g. emoji incongruity) were very helpful in classifying between ironic and not ironic data.

Table 5 shows the confusion matrix of our classification result on SubTask B. Our system performed quite well in Step 1 (*irony* vs *non-irony*) and Step 2 (*verbal_irony_by_polarity_contrast* vs *other_verbal_irony+situational_irony*). However, our system was struggling in distinguishing between *other_verbal_irony* and situational_irony (Step 3). Our system got very low precision in detecting situational irony, and this has a huge impact on macro average F-score. The difficulties to find an important feature to discriminate other verbal and situational irony was, indeed, for us the main challenge in Task B. A qualitative error analysis was conducted. We found a lot of tweets which where difficult to understand without the context), like:

(tw1)*"Produce Mobile Apps http://t.co/3OV57ZhqcH http://t.co/wX1DbI8W9M"*

(tw2) *"#Consensus of Absolute Hilarious - #MichaelMann to lecture on #Professional #Ethics for #Climate #Scientists? http://t.co/pD0TEMq1Z0"*

The first tweet is featured by situational irony and was originally including a #not hashtag before the link. Also for humans it is very difficult to get the ironic intention behind the tweet when the #not hashtag is removed and without having access to the information in the URL, which was anyway inactive. The second example was labelled as *other_verbal_irony*. Although it is very difficult to resolve the context of this tweet, accessing to the URL contained was helpful in understanding the ironic intent.

	0	**1**	**2**	**3**
0	**285**	80	75	33
1	19	**96**	31	18
2	29	7	**38**	11
3	29	13	12	**8**

Table 5: Confusion Matrix SubTask B.
0 : Not irony
1 : Verbal irony by polarity contrast
2 : Others irony
3 : Situational irony

4 Conclusion

This paper described the participation of the #NonDicevoSulSerio[5] team at SemEval2018-Task3: *Irony Detection of English Tweets*. We proposed to use several stylistic features and exploited several affective resources to deal with this task. Based on our evaluation and analysis, classifying irony into its several types (verbal irony by polarity contrast, other verbal irony, and situational irony) is a very challenging task. Especially, getting relevant features to discriminate between other verbal irony and situational irony will become our main focus on the future research direction. In this case, capturing semantic incongruity by exploiting word embedding semantic similarity is an issue worth to be explored (Joshi et al., 2015).

References

Salvatore Attardo. 2006. Irony. *Encyclopedia of Language & Linguistics*, 6:26–28.

Salvatore Attardo. 2007. Irony as relevant inappropriateness. In H. Colston and R. Gibbs, editors, *Irony in language and thought: A cognitive science reader*, pages 135–172. Lawrence Erlbaum.

[5]"Non dicevo sul serio" (Italian) means: 'I didn't mean that".

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. In *Proc. of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58.

Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):19.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478.

Raymond W. Gibbs and Herbert L. Colston, editors. 2007. *Irony in language and thought*. Routledge (Taylor and Francis), New York.

R. Giora. 2003. *On Our Mind: Salience, Context, and Figurative Language*. Oxford University Press.

H. P. Grice. 1975. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5):73:1–73:22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proc. of the 53rd Annual Meeting of the ACL and the 7th Int. Joint Conference on NLP*, pages 757–762, Beijing, China. ACL.

Jihen Karoui, Benamara Farah, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 262–272.

Diana Maynard and Mark A. Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proc. of the 9th Int. Conference on Language Resources and Evaluation*, pages 4238–4243. ELRA.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014*, pages 213–223.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language resources and evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proc. of EMNLP 2013*, pages 704–714.

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2016a. Exploring the realization of irony in Twitter data. In *Proc. of the 10th Int. Conference on Language Resources and Evaluation (LREC 2016)*. ELRA.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2016b. Guidelines for annotating irony in social media text, version 2.0. *LT3 Technical Report Series*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, SemEval-2018, New Orleans, LA, USA. ACL.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521.

Deirdre Wilson and Dan Sperber. 1992. On Verbal Irony. *Lingua*, 87(1-2):53–76.