SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter

Aniruddha Ghosh
University College Dublin, Ireland.
arghyaonline@gmail.com

Tony Veale
University College Dublin, Ireland.
Tony.Veale@UCD.ie

Ekaterina Shutova
University of Cambridge.
Ekaterina.Shutova@cl.cam.ac.uk

John Barnden
University of Birmingham, UK
J.A.Barnden@cs.bham.ac.uk

Guofu Li
University College Dublin, Ireland.
li.guofu.l@gmail.com

Paolo Rosso
Universitat Politècnica de València, Spain.
prosso@dsic.upv.es

Antonio Reyes
Instituto Superior de Intérpretes y Traductores
Mexico
antonioreyes@isit.edu.mx

Abstract

This report summarizes the objectives and evaluation of the SemEval 2015 task on the sentiment analysis of figurative language on Twitter (Task 11). This is the first sentiment analysis task wholly dedicated to analyzing figurative language on Twitter. Specifically, three broad classes of figurative language are considered: *irony*, *sarcasm* and *metaphor*. Gold standard sets of 8000 training tweets and 4000 test tweets were annotated using workers on the crowdsourcing platform *CrowdFlower*. Participating systems were required to provide a fine-grained sentiment score on an 11-point scale (-5 to +5, including 0 for neutral intent) for each tweet, and systems were evaluated against the gold standard using both a Cosine-similarity and a Mean-Squared-Error measure.

1 Introduction

The limitations on text length imposed by micro-blogging services such as Twitter do nothing to dampen our willingness to use language creatively. Indeed, such limitations further incentivize the use of creative devices such as metaphor and irony, as such devices allow strongly-felt sentiments to be expressed effectively, memorably and concisely. Nonetheless, creative language can pose certain challenges for NLP tools that do not take account of how words can be used playfully and in original ways. In the case of language using figurative devices such as irony, sarcasm or metaphor – when literal meanings are discounted and secondary or extended meanings are intentionally profiled – the affective polarity of the literal meaning may differ significantly from that of the intended figurative meaning. Nowhere is this effect more pronounced than in ironical language, which delights in using affirmative language to convey critical meanings. Metaphor, irony and sarcasm can each sculpt the affect of an utterance in complex ways, and each tests the limits of conventional techniques for the sentiment analysis of supposedly literal texts.

Figurative language thus poses an especially significant challenge to sentiment analysis systems, as standard approaches anchored in the dictionary-defined affect of individual words and phrases are often shown to be inadequate in the face of indirect figurative meanings. It would be convenient if such language were rare and confined to specific genres of text, such as poetry and literature. Yet the reality is that figurative language is pervasive in almost any genre of text, and is especially commonplace on the texts of the Web and on social media platforms such as Twitter. Figurative language often draws attention to itself as a creative artifact, but is just as likely to be viewed as part of the general fabric of human communication. In any case, Web users widely employ figures of speech (both old and new) to project their personality through a text, especially when their texts are limited to the 140 characters of a tweet.

Natural language researchers have attacked the problems associated with figurative interpretations

at multiple levels of linguistic representation. Some have focused on the conceptual level, of which the text is a surface instantiation, to identify the schemas and mappings that are implied by a figure of speech (see e.g. Veale and Keane (1992); Barnden (2008); Veale (2012)). These approaches yield a depth of insight but not a robustness of analysis in the face of textual diversity. More robust approaches focus on the surface level of a text, to consider word choice, syntactic order, lexical properties and affective profiles of the elements that make up a text (e.g. Reyes and Rosso (2012, 2014)). Surface analysis yields a range of discriminatory features that can be efficiently extracted and fed into machine-learning algorithms.

When it comes to analyzing the texts of the Web, the Web can also be used as a convenient source of ancillary knowledge and features. Veale and Hao (2007) describe a means of harvesting a common-sense knowledge-base of stereotypes from the Web, by directly targeting simile constructions of the form "*as X as Y*" (e.g. "*as hot as an oven*", "*as humid as a jungle*", "*as big as a mountain*", etc.). Though largely successful in their efforts, Veale and Hao were surprised to discover that up to 20% of Web-harvested similes are ironic (examples include "*as subtle as a freight train*", "*as tanned as an Irishman*", "*as sober as a Kennedy*", "*as private as a park bench*"). Initially filtering ironic similes manually – as irony is the worst kind of noise when acquiring knowledge from the Web – Hao & Veale (2010) report good results for an automatic, Web-based approach to distinguishing ironic from non-ironic similes. Their approach exploits specific properties of similes and is thus not directly transferrable to the detection of irony in general. Reyes, Rosso and Veale (2013) and Reyes, Rosso and Buscaldi (2012) thus employ a more general approach that applies machine learning algorithms to a range of structural and lexical features to learn a robust basis for detecting humor and irony in text.

The current task is one that calls for such a general approach. Note that the goal of Task 11 is not to detect irony, sarcasm or metaphor in a text, but to perform robust sentiment analysis on a fine-grained 11-point scale over texts in which these kinds of linguistic usages are pervasive. A system may find detection to be a useful precursor to analysis, or it may not. We present a description of Task 11 in section 2, before presenting our dataset in section 3 and the scoring functions in section 4. Descriptions of each participating system are then presented in section 5, before an overall evaluation in reported in section 6. The report then concludes with some general observations in section 7.

2 Task Description

The task concerns itself with the classification of overall sentiment in micro-texts drawn from the micro-blogging service Twitter. These texts, called tweets, are chosen so that the set as a whole contains a great deal of irony, sarcasm or metaphor, so no particular tweet is guaranteed to manifest a specific figurative phenomenon. Since irony and sarcasm are typically used to criticize or to mock, and thus skew the perception of sentiment toward the negative, it is not enough for a system to simply determine whether the sentiment of a given tweet is positive or negative. We thus use an 11-point scale, ranging from –5 (very negative, for tweets with highly critical meanings) to +5 (very positive, for tweets with flattering or very upbeat meanings). The point 0 on this scale is used for neutral tweets, or those whose positivity and negativity cancel each other out. While the majority of tweets will have sentiments in the negative part of the scale, the challenge for participating systems is to decide just how negative or positive a tweet seems to be.

So, given a set of tweets that are rich in metaphor, sarcasm and irony, the goal is to determine whether a user has expressed a positive, negative or neutral sentiment in each, and the degree to which this sentiment has been communicated.

3 Dataset Design and Collection

Even humans have difficulty in deciding whether a given text is ironic or metaphorical. Irony can be remarkably subtle, while metaphor takes many forms, ranging the dead to the conventional to the novel. Sarcasm is easier for humans to detect, and is perhaps the least sophisticated form of non-literal language. We sidestep problems of detection by harvesting tweets from Twitter that are *likely* to contain figurative language, either because they have been explicitly tagged as such (using e.g. the hashtags *#irony*, *#sarcasm*, *#not*, *#yeahright*) or because they use words commonly associated with the use of metaphor (ironically, the words

"*literally*" and "*virtually*" are reliable markers of metaphorical intent, as in "*I literally want to die* ").

Datasets were collected using the Twitter4j API (*http://twitter4j.org/en/index.html*), which supports the harvesting of tweets in real-time using search queries. Queries for hashtags such as #*sarcasm*, #*sarcastic* and #*irony*, and for words such as "*figuratively*", yielded our initial corpora of candidate tweets to annotate. We then developed a Latent Semantic Analysis (LSA) model to extend this seed set of hashtags so as to harvest a wider range of figurative tweets (see Li. *et. al*., 2014). This tweet dataset was collected over a period of 4 weeks, from June 1st to June 30th, 2014. Though URLs have been removed from tweets, all other content, including hashtags – even those used to retrieve each tweet – has been left in place. Tweets must contain at least 30 characters when hashtags are *not* counted, or 40 characters when hashtags *are* counted. All others are eliminated as too short.

3.1 Dataset Annotation on an 11-point scale

A trial dataset, consisting of 1025 tweets, was first prepared by harvesting tweets from Twitter users that are known for their use of figurative language (e.g. comedians). Each trial tweet was annotated by seven annotators from an internal team, three of whom are native English speakers, the other four of whom are competent non-native speakers. Each annotator was asked to assign a score ranging from -5 (for any tweets conveying disgust or extreme discontent) to +5 (for tweets conveying obvious joy and approval or extreme pleasure), where 0 is reserved for tweets in which positive and negative sentiment is balanced. Annotators were asked to use ±5, ±3 and ±1 as scores for tweets calling for strong, moderate or weak sentiment, and to use ±4 and ±2 for tweets with nuanced sentiments that fall between these gross scores. An overall sentiment score for each tweet was calculated as a weighted average of all 7 annotators, where a double weighting was given to native English speakers.

Sentiment was assigned on the basis of the perceived meaning of each tweet – the meaning an author presumably intends a reader to unpack from the text – and not the superficial language of the tweet. Thus, a sarcastic tweet that expresses a negative message in language that feigns approval or delight should be marked with a negative score (as in "*I just love it when my friends throw me under the bus.*"). Annotators were explicitly asked to consider *all* of a tweet's content when assigning a score, including any hashtags (such as #*sarcasm*, #*irony*, etc.), as participating systems are expected to use all of the tweet's content, including hashtags.

Tweets of the training and test datasets – comprising 8000 and 4000 tweets respectively – were each annotated on a crowd-sourcing platform, *CrowdFlower.com,* following the same annotation scheme as for the trial dataset. Some examples of tweets and their ideal scores, given as guidelines to *CrowdFlower* annotators, are shown in Table 1.

Tweet Content	Score
@ThisIsDeep_ you are about as deep as a turd in a toilet bowl. Internet culture is #garbage and you are bladder cancer.	-4
A paperless office has about as much chance as a paperless bathroom	-3
Today will be about as close as you'll ever get to a "PERFECT 10" in the weather world! Happy Mother's Day! Sunny and pleasant! High 80.	3
I missed voting due to work. But I was behind the Austrian entry all the way, so to speak. I might enter next year. Who knows?	1

Table 1: Annotation examples, given to Annotators

Scammers tend to give identical or random scores for all units in a task. To prevent scammers from abusing the task, trial tweets were thus interwoven as test questions for annotators on training and test tweets. Each annotator was expected to provide judgments for test questions that fall within the range of scores given by the original members of the internal team. Annotators are dismissed if their overall accuracy on these questions is below 70%. The standard deviation $std_u(u_i)$ of all judgments provided by annotator u_i also indicates that u_i is likely to be a scammer when $std_u(u_i)=0$. Likewise, the standard deviation $std_t(t_j)$ of all judgments given for a tweet t_j allows us to judge that annotation $A_{i,j}$ as given by u_i for t_j is an outlier if:

$$\left| A_{i,j} - \underset{i'}{\text{avg}}(A_{i',j}) \right| > std_t(t_j)$$

If 60% or more of an annotator's judgements are judged to be outliers in this way then the annotator is deemed a scammer and dismissed from the task.

Each tweet-set was cleaned of all annotations provided by those deemed to be scammers. After cleaning, each tweet has 5 to 7 annotations. The

ratio of in-range judgments on trial tweets, which was used to detect scammers on the annotation of training and test data, can also be used to assign a reliability score to each annotator. The reliability of an annotator u_i is given by $R(u_i)=m_i/n_i$, where n_i is the number of judgments contributed by u_i on trial tweets, and m_i is the number of these judgments that fall within the range of scores provided by the original annotators of the trial data. The final sentiment score for tweet $S(t_j)$ is the weighted average of scores given for it, where the reliability of each annotator is used as a weight.

$$S(t_j) = \frac{\sum_i R(u_i) \times A_{i,j}}{\sum_i R(u_i)}$$

The weighted sentiment score is a real number in the range [-5 … +5], where the most reliable annotators contribute most to each score. These scores were provided to task participants in two CSV formats: tweet-ids mapped to real number scores, and tweet-ids to rounded integer scores.

3.2 Tweet Delivery

The actual text of each tweet was not included in the released datasets due to copyright and privacy concerns that are standard for use of Twitter data. Instead, a script was provided for retrieving the text of each tweet given its released tweet-id.

Tweets are a perishable commodity and may be deleted, archived or otherwise made inaccessible over time by their original creators. To ensure that tweets did not perish in the interval between their first release and final submission, all training and test tweets were re-tweeted via a dedicated account to give them new, non-perishable tweet-ids. The distributed tweet-ids refer to this dedicated account.

Type	# Tweets	Mean Sentiment
Sarcasm	746	-1.94
Irony	81	-1.35
Metaphor	198	-0.34
Overall	1025	-1.58

Table 2: Overview of the Trial Dataset

3.3 Dataset Statistics

The trial dataset contains a mix of figurative tweets chosen manually from Twitter. It consists of 1025 tweets annotated by an internal team of seven members. Table 2 shows the number of tweets in each category. The trial dataset is small enough to allow these category labels to be applied manually.

The training and test datasets were annotated by CrowdFlower users from countries where English is spoken as a native language. The 8,000 tweets of the training set were allocated as in Table 3. As the datasets are simply too large for the category labels *Sarcasm*, *Irony* and *Metaphor* to be assigned manually, the labels here refer to our expectations of the kind of tweets in each segment of the dataset, which were each collated using harvesting criteria specific to different kinds of figurative language.

Type	# Tweets	Mean Sentiment
Sarcasm	5000	-2.25
Irony	1000	-1.70
Metaphor	2000	-0.54
Overall	8000	-1.75

Table 3: Overview of the Training Dataset

To provide balance, an additional category *Other* was also added to the Test dataset. Tweets in this category were drawn from general Twitter content, and so were not chosen to capture any specific figurative quality. Rather, the category was added to ensure the ecological validity of the task, as sentiment analysis is never performed on texts that are wholly figurative. The 4000 tweets of the Test set were drawn from four categories as in Figure 4.

Type	# Tweets	Mean Sentiment
Sarcasm	1200	-2.02
Irony	800	-1.87
Metaphor	800	-0.77
Other	1200	-0.26
Overall	4000	-1.21

Table 4: Overview of the Test Dataset

4 Scoring Functions

The Cosine-similarity scoring function represents the gold-standard annotations for the Test dataset as a vector of the corresponding sentiment scores. The scores provided by each participating system are represented in a comparable vector format, so that the cosine of the angle between these vectors captures the overall similarity of both score sets. A score of 1 is achieved only when a system provides

all the same scores as the human gold-standard. A script implementing this scoring function was released to all registered participants, who were required in turn to submit the outputs of their systems as a tab-separated file of tweet-ids and integer sentiment scores (as systems may be based either on a regression or a classification model).

A multiplier p_{cos} is applied to all submissions, to penalize any that do not give scores for *all* tweets.

Thus, $$p_{\cos} = \frac{\#\text{submitted-entries}}{\#\text{all-entries}}$$

E.g., a cherry-picking system that scores just 75% of the test tweets is hit with a 25% penalty.

Mean-Squared-Error (MSE) offers a standard basis for measuring the performance of predictive systems, and is favored by some developers as a basis for optimization. When calculating MSE, in which lower measures indicate better performance, the penalty-coefficient p_{MSE} is instead given by:

$$p_{MSE} = \frac{\#\text{all-entries}}{\#\text{submitted-entries}}$$

5 Overview of Participating Systems

A total of 15 teams participated in Task 11, submitting results from 29 distinct runs. A clear preference for supervised learning methods can be observed, with two types of approach – SVMs and regression models over carefully engineered features – making up the bulk of approaches.

Team ***UPF*** used regression with a Random-Sub-Space using M5P as a base algorithm. They exploited additional external resources such as SentiWordnet, Depeche Mood, and the American National Corpus. Team ***ValenTo*** used a regression model combined with affective resources such as *SenticNet* (see Poria *et al.*, 2014) to assign polarity scores. Team ***Elirf*** used an SVM-based approach, with features drawn from character *N*-grams ($2 < N < 10$) and a bag-of-words model of the tf-idf coefficient of each N-gram feature. Team ***BUAP*** also used an SVM approach, taking features from dictionaries, POS tags and character n-grams. Team ***CLaC*** used four lexica, one that was automatically generated and three than were manually crafted. Term frequencies, POS tags and emoticons were also used as features. Team ***LLT_PolyU*** used a semi-supervised approach with a Decision Tree Regression Learner, using word-level sentiment scores and dependency labels as features. Team ***CPH*** used ensemble methods and ridge regression (without stopwords), and is notable for its specific avoidance of sentiment lexicons. Team ***DsUniPi*** combined POS tags and regular expressions to identify useful syntactic structures, and brought sentiment lexicons and WordNet-based similarity measures to bear on their supervised approach. Team ***RGU***'s system learnt a sentiment model from the training data, and used a linear Support Vector Classifier to generate integer sentiment labels. Team ***ShellFBK*** also used a supervised approach, extracting grammatical relations for use as features from dependency tree parses.

Team ***HLT*** also used an SVM-based approach, using lexical features such as negation, intensifiers and other markers of amusement and irony. Team ***KElab*** constructed a supervised model based on term co-occurrence scores and the distribution of emotion-bearing terms in training tweets. Team ***LT3*** employed a combined, semi-supervised SVM- and regression-based approach, exploiting a range of lexical features, a terminology extraction system and and both WordNet and DBpedia. Team ***PRHLT*** used a deep auto-encoder to extract features, employing both words and character 3-grams as tokens for the autoencoder. Their best results were obtained with ensembles of Extremely Random Trees with character n-grams as features.

6 Results and Discussions

For comparison purposes, we constructed three baseline systems, each implemented as a naïve classifier with shallow bag-of-word features. The results of these baseline systems for both the MSE and Cosine metrics are shown in Table 5.

Baseline	Cosine	MSE
Naïve Bayes	0.390	5.672
MaxEnt	0.426	5.450
Decision Tree	0.547	4.065

Table 5: Performance of Three Baseline approaches

Table 6 shows the results for each participating system using these metrics. Team ***CLaC*** achieves the best overall performance on both, achieving **0.758** on the Cosine metric and 2.117 on the MSE

metric. Most of the other systems also show a clear advantage over the baselines reported in Table 5.

Team	Cosine	MSE
CLaC	**0.758**	**2.117**
UPF	0.711	2.458
LLT_PolyU	0.687	2.6
elirf	0.658	3.096
LT3	0.658	2.913
ValenTo	0.634	2.999
HLT	0.63	4.088
CPH	0.625	3.078
PRHLT	0.623	3.023
DsUniPi	0.602	3.925
PKU	0.574	3.746
KELabTeam	0.552	4.177
RGU	0.523	5.143
SHELLFBK	0.431	7.701
BUAP	0.059	6.785

Table 6: Overall results, sorted by cosine metric. Scores are for last run submitted for each system.

The best performance on *sarcasm* and *irony* tweets was achieved by teams ***LLT_PolyU*** and ***elirf***, who ranked 3rd and 4th respectively. Team ***ClaC*** came first on tweets in the *Metaphor* category. One run of team ***CPH*** excelled on the *Other* (non-figurative) category, but scored poorly on figurative tweets. Most teams performed well on *sarcasm* and *irony* tweets, but the *Metaphor* and *Other* categories prove more of a challenge. Table 7 presents the Spearman's rank correlation between the ranking of a system overall, on all tweet categories, and its ranking of different categories of tweets. The right column limits this analysis to the top 10 systems.

	Spearman Correl – All	Spearman Correl – Top10
Sarcasm	0.854	0.539
Irony	0.721	0.382
Metaphor	0.864	0.939
Other	0.857	0.624

Table 7. How well does overall performance correlate with performance on different kinds of tweets?

When we consider all systems, their performance on each category of tweet is strongly correlated to their overall performance. However, looking only at the top 10 performing systems, we see a strikingly strong correlation between performance overall and performance on the category *Metaphor*. Performance on *Metaphor* tweets is a bellwether for performance on figurative language overall. Then category *Other* also plays an important role here. Both the trail data and the training datasets are heavily biased to negative sentiment, given their concentration of ironic and sarcastic tweets. In contrast, the distribution of sentiment scores in the test data is more balanced due to the larger proportion of Metaphor tweets and the addition of non-figurative *Other* tweets. To excel at this task, systems must not treat all tweets as figurative, but learn to spot the features that cause figurative devices to influence the sentiment of a tweet.

7 Summary and Conclusions

This paper has described the design and evaluation of Task 11, which concerns the determination of sentiment in tweets which are likely to employ figurative devices such as irony, sarcasm and metaphor. The task was constructed so as to avoid questions of what specific device is used in which tweet: a glance at Twitter, and the use of the #*irony* hashtag in particular, indicates that there are as many folk theories of irony as there are users of the hashtag #*irony*. Instead, we have operationalized the task to put it on a sound and more ecologically valid footing. The effect of figurativity in tweets is instead measured via an extrinsic task: measuring the polarity of tweets that use figurative language.

The task is noteworthy in its use of an 11-point sentiment scoring scheme, ranging from -5 to +5. The use of 11 fine-grained categories precludes the measurement of inter-annotator agreement as a reliable guide to annotator/annotation quality, but it allows us to measure system performance on a task and a language type in which negativity dominates. We expect the trial, training and test datasets will prove useful to future researchers who wish to explore the complex relation between figurativity and sentiment. To this end, we have taken steps to preserve the tweets used in this task, to ensure that they do not perish through the actions of their original creators. Detailed results of the evaluation of all systems and runs are shown in Tables 9 and 10, or can be found online here:

http://alt.qcri.org/semeval2015/task11/

Team Name	Name of Run	Rank	Overall	Sarcasm	Irony	Metaphor	Other
ClaC		1	**0.758**	0.892	0.904	**0.655**	0.584
UPF		2	0.711	0.903	0.873	0.520	0.486
LLT_PolyU		3	0.687	0.896	**0.918**	0.535	0.290
LT3	*run 1*	4	0.6581	0.891	0.897	0.443	0.346
	run 2		0.648	0.872	0.861	0.355	0.357
elirf		5	0.6579	**0.904**	0.905	0.411	0.247
ValenTo		6	0.634	0.895	0.901	0.393	0.202
HLT		7	0.630	0.887	0.907	0.379	0.365
CPH	*ridge*	8	0.625	0.897	0.886	0.325	0.218
	ensemble		0.623	0.900	0.903	0.308	0.226
	special-ensemble		0.298	-0.148	0.281	0.535	**0.612**
PRHLT	*ETR-ngram*	9	0.623	0.891	0.901	0.167	0.218
	ETR-word		0.611	0.890	0.901	0.294	0.129
	RFR-word		0.613	0.888	0.898	0.282	0.170
	RFR-ngram		0.597	0.888	0.898	0.135	0.192
	BRR-word		0.592	0.883	0.880	0.280	0.110
	BRR-ngram		0.593	0.886	0.879	0.119	0.186
DsUniPi		10	0.601	0.87	0.839	0.359	0.271
PKU		11	0.574	0.883	0.877	0.350	0.137
KELabTeam			0.531	0.883	0.895	0.341	0.117
	content based	12	0.552	0.896	0.915	0.341	0.115
	emotional pattern based		0.533	0.874	0.900	0.289	0.135
RGU	*test-sent-final*	13	0.523	0.829	0.832	0.291	0.165
	test-sent-warppred		0.509	0.842	0.861	0.280	0.090
	test-sent-predictions		0.509	0.842	0.861	0.280	0.090
SHELLFBK	*run3*	14	0.431	0.669	0.625	0.35	0.167
	run2		0.427	0.681	0.652	0.346	0.146
	run1		0.145	0.013	0.104	0.167	0.308
BUAP		15	0.058	0.412	-0.209	-0.023	-0.025

Table 9. Detailed evaluation of each submitted run of each system (using the **Cosine similarity** metric).

Key: ***CLaC****= Concordia University;* ***UPF****= Universitat Pompeu Fabra;* ***LLT_PolyU****=Hong Kong Polytechnic University;* ***LT3****= Ghent University;* ***elirf****= Universitat Politècnica de València;* ***ValenTo****= Universitat Politècnica de València;* ***HLT****= FBK-Irst, University of Trento;* ***CPH****= Københavns Universitet;* ***PRHLT****= PRHLT Research Center;* ***DsUniPi****= University of Piraeus;* ***PKU****= Peking University;* ***KELabTeam****= Yeungnam University;* ***RGU****= Robert Gordon University;* ***SHELLFBK****= Fondazione Bruno Kessler;* ***BUAP****= Benemérita Universidad Autónoma de Puebla*

Team Name	Name of Run	Rank	Overall	Sarcasm	Irony	Metaphor	Other
ClaC		1	**2.117**	1.023	0.779	**3.155**	**3.411**
UPF		2	2.458	**0.934**	1.041	4.186	3.772
LLT_PolyU		3	2.600	1.018	**0.673**	3.917	4.587
LT3	run1		3.398	1.287	1.224	5.670	5.444
	run2	4	2.912	1.286	1.083	4.793	4.503
elirf		8	3.096	1.349	1.034	4.565	5.235
ValenTo		5	2.999	1.004	0.777	4.730	5.315
HLT		11	4.088	1.327	1.184	6.589	7.119
CPH	ridge		3.079	1.041	0.904	4.916	5.343
	ensemble	7	3.078	0.971	0.774	5.014	5.429
	special-ensemble		11.274	19.267	9.124	7.806	7.027
PRHLT	ETR-ngram	6	3.023	1.028	0.784	5.446	4.888
	ETR-word		3.112	1.041	0.791	5.031	5.448
	RFR-word		3.107	1.060	0.809	5.115	5.345
	RFR-ngram		3.229	1.059	0.811	5.878	5.243
	BRR-word		3.299	1.146	0.934	5.178	5.773
	BRR-ngram		3.266	1.100	0.941	5.925	5.205
DsUniPi		10	3.925	1.499	1.656	7.106	5.744
PKU		9	3.746	1.148	1.015	5.876	6.743
KELabTeam			5.552	1.198	1.255	7.264	9.905
	content based		6.090	1.756	1.811	8.707	11.526
	emotional pattern	12	4.177	1.189	0.809	6.829	7.628
RGU	test-sentfinal	13	5.143	1.954	1.867	8.015	8.602
	test-sent-warppred		5.323	1.855	1.541	8.033	9.505
	test-sent-predictions		5.323	1.855	1.541	8.033	9.505
SHELLFBK	run3	15	7.701	4.375	4.516	9.219	12.16
	run2		9.265	5.183	5.047	11.058	15.055
	run1		10.486	12.326	9.853	10.649	8.957
BUAP		14	6.785	4.339	7.609	8.93	7.253

Table 10. Detailed evaluation of each submitted run of each system (using the **Mean-Squared-Error** metric).

Key: ***CLaC****= Concordia University;* ***UPF****= Universitat Pompeu Fabra;* ***LLT_PolyU****=Hong Kong Polytechnic University;* ***LT3****= Ghent University;* ***elirf****= Universitat Politècnica de València;* ***ValenTo****= Universitat Politècnica de València;* ***HLT****= FBK-Irst, University of Trento;* ***CPH****= Københavns Universitet;* ***PRHLT****= PRHLT Research Center;* ***DsUniPi****= University of Piraeus;* ***PKU****= Peking University;* ***KELabTeam****= Yeungnam University;* ***RGU****= Robert Gordon University;* ***SHELLFBK****= Fondazione Bruno Kessler;* ***BUAP****= Benemérita Universidad Autónoma de Puebla*

Acknowledgements

The authors gratefully acknowledge the support of the following projects funded by the European Commission: *PROSECCO* (Grant No. 600653), *WIQ-EI IRSES* (Grant No. 269180) and *MICINN DIANA-Applications* (TIN2012-38603-C02-01). We are also grateful for the support of the *CNGL Centre for Global Intelligent Content*, funded by Science Foundation Ireland (SFI).

References

Barnden, J.A. (2008). Metaphor and artificial intelligence: Why they matter to each other. In R.W. Gibbs, Jr. (Ed.), The Cambridge Handbook of Metaphor and Thought, 311-338. Cambridge, U.K.: Cambridge University Press.

Hao, Y., Veale, T. (2010). An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. Minds and Machines 20(4):635–650.

Li G., Ghosh A., Veale T. (2014). Constructing A corpus of Figurative Language for a Tweet Classification and Retrieval Task. In the preceedings of FIRE 2014, the 6th Forum for Information Retrieval Evaluatio. Bengaluru, India.

Poria, S., Cambria, E., Winterstein, G. and Huang, G.B. (2014). Sentic patterns Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69, pp. 45-63.

Reyes A., Rosso P. (2014). On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. Knowledge and Information Systems. 40(3): 595-614. DOI: 10.1007/s10115-013-0652-8.

Reyes A., Rosso P., Veale T. (2013). A Multidimensional Approach for Detecting Irony in Twitter. Languages Resources and Evaluation 47(1): 239-268.

Reyes A., Rosso P. (2012). Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. Journal on Decision Support Systems 53(4): 754–760.

Reyes A., Rosso P., Buscaldi D. (2012). From Humor Recognition to Irony Detection: The Figurative Language of Social Media. Data & Knowledge Engineering 74:1-12.

Shutova, E., L. Sun, A. Korhonen. (2010). Metaphor identification using verb and noun clustering. Proceedings of the 23rd International Conference on Computational Linguistics.

Veale, T., Keane, M. T. (1992). Conceptual Scaffolding: A spatially founded meaning representation for metaphor comprehension. Computational Intelligence 8(3): 494-519.

Veale, T. (2012). Detecting and Generating Ironic Comparisons: An Application of Creative Information Retrieval. AAAI Fall Symposium Series 2012, Artificial Intelligence of Humor. Arlington, Virginia.

Veale, T., Hao, Y. (2007). Comprehending and Generating Apt Metaphors: A Web-driven, Case-based Approach to Figurative Language. In proceedings of AAAI 2007, the 22nd AAAI Conference on Artificial Intelligence. Vancouver, Canada.