Overview of the Task on Irony Detection in Spanish Variants

Reynier Ortega-Bueno[1], Francisco Rangel[2,3], Delia Irazú Hernández Farías[4],
Paolo Rosso[3], Manuel Montes-y-Gómez[4], and José E. Medina-Pagola[5]

[1] Center for Pattern Recognition and Data Mining, University of Oriente, Cuba
reynier.ortega@cerpamid.co.cu
[2] Autoritas Consulting, S.A., Spain
francisco.rangel@autoritas.es
[3] PRHLT Research Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es
[4] Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
dirazuherfa@inaoep.mx, mmontesg@inaoep.mx
[5] University of Informatics Science, Havana, Cuba
jmedinap@uci.cu

Abstract. This paper introduces IroSvA, the first shared task fully dedicated to identify the presence of irony in short messages (tweets and news comments) written in three different variants of Spanish. The task consists in: given a message, automatic systems should recognize whether the message is ironic or not. Moreover, with respect to the previous tasks on irony detection, the messages are not considered as isolated texts but together with a given context (e.g. a headline or a topic). The task is comprised by three different subtasks: i) irony detection in tweets from Spain, ii) irony detection in Mexican tweets, and iii) irony detection in news comments from Cuba. These subtasks aim at studying the way irony changes across the distinct Spanish variants. We received 14 submissions from 12 teams. Participating systems were evaluated against the test dataset using F1 macro averaged. The highest classification scores obtained for the three subtasks are F1=0.7167, F1=0.6803, and F1=0.6596, respectively.

Keywords: Irony Detection · Spanish Variants · Spanish datasets · Cross-variant

1 Introduction

From its birth in the Ancient Greek to the present times *irony* has been a complex, controversial, and intriguing issue. It has been studied from many disciplines such as philosophy, psychology, rhetoric, pragmatics, semantics, etc. However, irony is not only enclosed to specialized theoretical discussions, this phenomenon appears in everyday conversations. As human beings, we appeal to irony for expressing in effective way something distinct to what we utter. Thus, understanding irony speech requires a more complex set of cognitive and linguistics abilities than direct and literal speech [1]. Despite that the term seems familiar for all of us, the mechanics underlying in ironic communication continues to be a challenging issue. The benefits of detecting and understanding irony computationally, have caused that irony oversteps its theoretical and philosophical perspective, and attracted the attention of both artificial intelligence researchers and practitioners [64]. Although, a well-established definition of irony still lacks in the literature, many authors appear to agree with two points: i) by using irony, the author does not intend to communicate with what she appears to be putting forward, the real meaning is evoked implicitly and differs from what she utters; and, ii) irony is closely connected with the expression of a feeling, emotion, attitude, or evaluation [20,27,57].

Due to its nature, irony has important implications in natural language processing tasks, and particularly in those that require semantic processing. A representative case is the well-known task of sentiment analysis which aims at automatically assess the underlying sentiments expressed in a text [42,50]. Interesting evidences about the impact of irony in sentiment analysis have been widely discussed in [7,26,30,44,58]. Systems dedicated to sentiment analysis struggle when facing ironic texts because the intentional meaning of the text is expressed implicitly. Taking into account words and statistical information derived from text is not enough to deal with the sentiment expressed when ironic devices are used for communication purposes. Therefore, the systems require to recall contextual, commonsense, and world-knowledge for disentangling the right meaning. Indeed, in sentiment analysis irony plays a role of "*implicit valence shifter*", and ignoring it, cause an abrupt drop in systems' accuracy [58].

Automatic irony detection has gained popularity and importance in the research community, paying special attention to social media content in English. Several shared tasks have been proposed to tackle this issue, such as: SemEval 2018 Task 3 [63], SemEval 2015 Task 11 [21], and PAKDD 2016 contest[6]. Also, parallel tasks have been proposed for addressing irony in Italian: SentiPOLC tasks at EVALITA in 2014 [5] and 2016 [3], IronITA task at EVALITA 2018 [15]. However, for Spanish, the availability of datasets is scarce, which limits the amount of research done for this language.

In this sense, we propose a new task, *IroSvA* (Irony Detection in Spanish Variants), which aims at investigating whether a short message, written in Spanish language, is ironic or not with respect to a given context. Particularly, we aim at studying the way irony changes in distinct Spanish variants.

[6] https://pakdd16.wordpress.fos.auckland.ac.nz/technical-program/contests/

1.1 Task Description

The task consists in automatically classifying short messages from Twitter and news comments for irony. It is structured in three independent subtasks.

- **Subtask A**: Irony detection in Spanish tweets from Spain.
- **Subtask B**: Irony detection in Spanish tweets from Mexico.
- **Subtask C**: Irony detection in Spanish news comments from Cuba

The three subtasks are centered on the same objective: systems should determine whether a message is ironic or not according to a specified context (by assigning a binary value 1 or 0). The following examples present an ironic and non-ironic tweet from Spanish users, respectively:

Given the context: *The politician of the Podemos party, Pablo Iglesias, appears in the Hormiguero TV program teaching to Spanish people to change baby diapers* (**PañalesIglesias**)

1) (Sp.) *@europapress Pues resulta que @Pablo_Iglesias_ es el primer papá que cambia pañales*
(En.) *@europapress It seems that @Pablo_Iglesias_ is the first daddy that changes baby diapers.*

2) (Sp.) *Como autónomo, sin haber disfrutado prácticamente de días de baja cuando nacieron mis hijos, y habiendo cambiado muchos más pañales que tú, te digo: eres tonto.*
(En.) *A self-employed person, without having practically enjoyed days off when my children were born, and having changed many more diapers than you, I tell you: you are stupid.*

The main difference with previous tasks on irony detection at SemEval 2018 Task 3 and IronITA 2018 is that messages are not considered as isolated texts but together with a given context (e.g. a headline or a topic). In fact, the context is mandatory for understanding the underlying meaning of ironic texts. This task provided a first dataset manually annotated for irony in Spanish social media and news comments.

Additionally, and in unofficial way, participants were asked to evaluate their systems in a cross-variant setting. That is, to test each trained model on the test datasets of the other two variants. For example, to train the model on the Mexican dataset and validate it on the Spanish and Cuban datasets (and so on for the rest). The participants were allowed to submit one run for each subtask (exceptionally, two runs). No distinction between constrained and unconstrained systems was made, but the participants were asked to report what additional resources and corpora they have used for each submitted run.

2 Automatic Irony Detection

With the increasing in the use of social media, user-generated content in those platforms has been considered as an interesting source of data for studying the use of irony. Data coming from different platforms such as Amazon reviews [17], comments from debate sites such as 4forums.com [43], Reddit [66], and Twitter (it has been without doubts the most exploited one [35]) have been considered in order to detect irony. Such an interesting and challenging task has been tackled as a binary classification problem.

Automatic irony detection has been addressed from different perspectives. Exploiting textual-based features from the text on its own (such as n-grams, punctuation marks, part-of-speech labels, among others) has been widely used for irony detection [12,16,25,41,53]. Irony is strongly related to subjective aspects, in such a way some approaches have been proposed in order to take advantage of affective information [4,27,29,57]. In a similar fashion, in [67] the authors proposed a transfer learning approach that takes advantage of sentiment analysis resources.

Information regarding to the context surrounding a given comment has been exploited in order to determine whether or not it has an ironic intention [2,40,65]. There are some deep learning-based approaches for dealing with irony detection. Word-embeddings and convolutional neural networks have been exploited for capturing the presence of irony in social media texts [22,23,31,36,49,52]. As in other natural language processing tasks, most of the research carried out on irony detection has been done in English. Notwithstanding, there have been some efforts to investigate such figurative language device in other languages such as: Chinese [62], Czech [53], Dutch [41], French [38], Italian [9], Portuguese [12], Spanish [33], and Arabic [39].

The strong relation between irony detection and sentiment analysis has derived in the emergence of some evaluation campaigns focused on sentiment analysis where the presence of ironic content was considered to assess the performance of the participating systems. The 2014 [5] and 2016 [3] editions of SENTIPOLC (SENTIment POLarity Classification) in the framework of EVALITA included a set of ironic tweets written in Italian. A drop in the performance of the systems in the task was observed when ironic instances are involved, confirming the important role of irony for carrying out sentiment analysis. In 2015, the first shared task dedicated to sentiment analysis on figurative language devices in Twitter [21] was organized. The first shared task considering the presence of ironic content with sentiment analysis in Twitter data written in French was organized in 2017 [6]. The participating systems proposed supervised methods to address the task by taking advantage of standard classifiers together with n-grams, word-embeddings, as well as lexical resources. In a similar fashion, in [28] the authors proposed a pipeline approach that incorporates two modules: one for irony detection and the other one for polarity assignment.

In addition to this, some shared tasks fully dedicated to irony detection have been organized. On 2018, in the framework of SemEval-2018 the first shared task aimed to detect irony in Twitter was organized (SemEval-2018 Task 3:

Irony Detection in English Tweets) [63]. The task is composed by two subtasks: i) to determine whether a tweet is ironic or not (Task A), and ii) to identify which type of irony is expressed (Task B). The participating systems used a wide range of features (such as n-grams, syntactic, sentiment-based, punctuation marks, word-embeddings, among others) together with different classification approaches: ensemble-based classifiers, Logistic Regression (LR), Support Vector Machines (SVMs), as well as Long Short Term Memory Neural Networks (LSTMs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). A shared task on irony detection in Italian tweets, denoted as IronITA, was organized in the framework of the EVALITA 2018 evaluation campaign [15]. Two subtasks were proposed: i) determining the presence of irony, and ii) identifying different types of irony (with a special attention to recognize instances expressing sarcasm[7]). Traditional classifiers (such as SVM and Naive Bayes (NB)) as well as deep learning techniques were used for addressing irony detection. Word-embeddings, n-grams, different lexical resources, as well as stylistic and structural features were exploited to characterize the presence of ironic intention. At the moment, a new shared task on irony detection in Arabic tweets (IDAT 2019)[8] has been organized. The aim of the competition is to determine whether or not an Arabic tweet is ironic. IDAT task provides a useful evaluation framework for comparing the performance of Arabic irony detection methods with respect to those results reported in recent shared tasks.

Analyzing the differences among diverse types of ironic devices has been also investigated. In the framework of SemEval-2018 Task 3 and IronITA-2018 subtasks aimed to identify ironic instances in a finer-grained way. In [4] the authors attempted to distinguish between ironic and sarcastic tweets. An analysis on the multi-faceted affective information expressed in tweets labeled with ironic hashtags (#irony, #sarcasm, and #not) was carried out in [61] where the authors identified some interesting differences among such figurative linguistic devices. However, it has been recognized that such a challenging task is still very difficult [15,63].

3 Datasets Description

In this section we describe the datasets proposed for evaluation, how they were collected, the labeling process and the inter-annotator agreement (IAA).

3.1 Annotation Guidelines

For creating our multi-variant dataset for irony detection in short messages written in Spanish language we decided not to use any kind of standard guideline in

[7] From a computational linguistics perspective, irony is often considered as an umbrella term covering sarcasm. However, there are theoretical foundations on the separation of both concepts. Sarcasm involves a negative evaluation towards a particular target with the intention to offend [1]

[8] https://www.irit.fr/IDAT2019/

the annotation process. However, two important aspects were considered: i) the annotators for each variant must be native speakers, and they do not annotate messages in other Spanish variants different from theirs. For instance, Mexican annotators do not label messages from Cuba and Spain. This constraint was defined because there are significant bias in irony and sarcasm labeling when cultural and social knowledge are required to understand the underlying meaning of messages [34]; ii) we asked annotators for labeling each message as ironic/non-ironic, given an specific "context", based only on their own concept of irony. They made use of their own world-knowledge and linguistics skills. Also, no differentiation among any type of irony (situational, dramatically or verbal) was made; in the case of sarcasm, the annotators assumed that it is a special case of irony.

3.2 Cuban Variant

In Cuba, the popularity of the social platforms (Twitter, Facebook, WhatsApp, etc.) is now increasing due to the technological advances in the communication sector, however the number of people that actually access to them continues to be limited with respect to other countries such as Mexico or Spain. For this reason, it is difficult to retrieve many tweets posted by Cuban users. As an alternative to this problem, we aim to explore other sources with similar characteristics. In particular, we envisaged the news comments as an interesting textual genre that shares characteristics with tweets.

To collect the news comments were identified three popular Cuban news sites (Cubadabate[9], OnCuba[10], CubaSí[11]). In concordance with the idea presented in [56], we had the intuition that some topics or headlines are more controversial than others and they generate major discussion threats. In this scenario, the readers spontaneously express their judgments, opinions, and emotions about the discussed news. This enables the possibility to obtain diverse points of view about the same topic, where irony device is often used.

In this way, we manually chose 113 polemic headlines about social, economic, and political issues concerning Cuban people. We noted that those news with a fast and huge increase in the number of comments is correlated with controversial topics. This observation helped us to increase the speed of the selection process. Afterwards, the 113 headlines were grouped manually in 10 coarse topics which can be considered as context:

- Digital Television, TV Decoders, Cuban Television and Audiovisuals (DigitalTV).
- Sports Scandals, Cuban National Baseball League and Football (Sports).
- ETECSA, Quality and Service (E-Quality).
- ETECSA, Internet, and Mobile Data (E-Mobile).
- Transport, Bus Drivers, Taxi Drivers, Buses and Itineraries (Transport).

[9] http://www.cubadebate.cu/

[10] https://oncubanews.com/

[11] http://cubasi.cu/

- Advanced Technologies and Computerization of Society (TechSociety).
- Intra-Cuban Trade, Prices, Shops and Markets (IC-Trade).
- Economy, Hotels and Tourism (Economy).
- Science, Education and Culture (Science).
- Others.

Once we defined both the topics and the headlines, we extracted and filtered all the comments. News comments do not have any restriction about the maximum number of characters as imposed by Twitter. With the purpose of providing a dataset with short messages like tweets, we filtered out text with more than 300 characters. A final dataset composed of 5507 comments was obtained.

The annotation process over the dataset was performed by three annotators simultaneously. All of them having a degree in Linguistics. In a first stage, only 100 instances were labeled by the three annotators. Based on them, an initial IAA was computed in terms of Cohen's Kappa κ, between pairs of annotators; the averaged value was $\kappa = 0.39$. All cases of disagreement were discussed in order to establish a consensus in the annotation process. Later, a second stage of annotation was carried out, all instances, including the previous ones, were labeled by the annotators. At this time, an averaged $\kappa = 0.67$ was reached. This value reflects a good agreement and it is close to the results achieved in [63] for the English language. Finally, we considered as "ironic"/"non-ironic" instances those in which at least two annotators agreed, respectively. Considering this criterion we obtained a corpus with 1291 and 4216 "ironic"/"non-ironic" comments respectively.

The official dataset to be provided for evaluation purposes consists of 3000 news comments distributed across the 9 distinct topics. We do not consider the topic "Others" because it is very broad and no "context" was provided for it. Then, the data were divided into two partitions considering the 80% for training and the rest for the test. Table 1 shows the distribution of comments for each topic in the training and test data.

Table 1: Training and Test partitions distribution on the Cuban variant data.

	Training		**Test**	
Topic	**Ironic**	**Non-ironic**	**Ironic**	**Non-ironic**
DigitalTV	137	275	32	65
Sports	108	219	28	55
E-Quality	100	201	25	51
E-Mobile	92	185	23	47
Transport	91	184	23	46
TechSociety	85	172	22	44
IC-Trade	74	150	19	38
Economy	57	103	14	26
Science	56	111	14	28
Total	**800**	**1600**	**200**	**400**

3.3 Mexican Variant

In a first attempt to build a dataset of tweets written in Mexico, we tried to collect ironic data from Twitter by applying a well-known strategy, i.e., by relying on the users' intent to self-annotate their tweets using specific hashtags: "#ironia" and "#sarcasmo" ("#irony" and "#sarcasm", respectively). However, we were able to retrieve only a few tweets with such a methodology, i.e., it seems that those labels are not commonly used by Twitter users in Mexico in order to self-annotate their intention of being ironic. Thus, an alternative approach was followed. We have the intuition that, in controversial tweets generated by accounts with solid reputation for information disclosure, Twitter users express their opinions about a certain topic. In this way, it is possible to capture different points of view (including of course ironic statements) around the same topic. In other words, we are establishing a "context" in which a set of tweets are generated.

First, we selected a set of Twitter accounts belonging to well-known journalists, newspapers, newsmedia, and alike. In the second step, we defined nine controversial topics in Mexico to be considered as "context":

- Divorce of the Former President of Mexico Enrique Peña Nieto (DivorceEPN).
- "Rome" movie during the Academy Awards 2019 (RomeMovie).
- Process of selection of the head of the Mexico's Energy Regulatory Commission (CRE).
- Fuel shortage occurred in Mexico in January 2019 (F-Shortage).
- Funding cuts for children day-care centers (Ch-Centers).
- Issues related to the new government in Mexico (GovMexico).
- Issues related to the new government in Mexico City (GovCDMX).
- Issues related to the National Council of Science and Technology (CONACYT).
- Issues related to the Venezuela government (Venezuela).

Once defined both the accounts and topics, we manually collected a set of tweets regarding the aforementioned topics posted by the selected accounts. A total of 54 tweets (denoted as *tweetsForContext*) were used as a starting point in order to retrieve the data. Then, for each tweet in *tweetsForContext* we retrieved those tweets posted as answers to the tweet in hand. The final step consisted in to filter out those instances composed by less than four words and also those containing only emojis, links, hashtags or mentions. Additionally, with the intention of having a topic in common with the context considered in the data collected in Spain, we also consider one more topic: "People supporting the Flat Earth Theory" (FlatEarth). The data belonging to this theme were retrieved according to two criteria: i) by exploiting specific terms to perform Twitter queries: "*tierraplanistas*" and "*tierra plana*"; and ii) by verifying that the geo-localization of the tweets corresponds to any place in Mexico.

The final set of collected tweets is composed by 5442 instances. We perform an annotation process over the retrieved data involving three people. We did not provide any kind of guideline for annotation purposes. Instead, we ask the annotators to rely on their own definition of what irony is. In the first stage, the data were annotated by two independent annotators. Then, only for those instances where a disagreement exists, we asked for a third annotation. The inter-annotator agreement in terms of Cohen's kappa between the first two annotators is $\kappa = 0.1837$ (this value reflects a *slight agreement*). The obtained IAA value validates the inherent complexity involved in the annotation of ironic data [54]. After the first annotation, we achieved a total of 771 ironic tweets. The second stage of the annotation process involved 2015 instances that were labeled by the third annotator. Finally, a set of 1180 tweets were annotated as ironic while 4262 as non-ironic. The final dataset to be provided for evaluation purposes consists of 3000 tweets distributed across the 10 different topics. Then the data were divided into two partitions considering the 80% for training and the rest for test. Table 2 shows the distribution of tweets for each topic in the corresponding data partition.

Table 2: Training and Test partitions distribution on the Mexican variant data.

	Training		**Test**	
Topic	**Ironic**	**Non-ironic**	**Ironic**	**Non-ironic**
DivorceEPN	46	90	11	23
RomeMovie	78	161	19	41
CRE	123	146	31	37
F-Shortage	111	128	28	33
Ch-Centers	80	159	21	40
GovtMexico	114	125	29	32
GovCDMX	54	156	14	39
CONACYT	139	210	33	49
Venezuela	20	220	5	55
FlatEarth	35	205	8	52
Total	**800**	**1600**	**199**	**401**

3.4 Spanish Variant

For building the Spanish dataset a similar process to the Cuban and Mexican variants was adopted. Guided by the idea that controversial and broad discussed topics are a potential source of spontaneous content where several points of view are exposed about a particular topic, resulting this scenario an attractive way for capturing figurative language usages such as irony. Firstly, a set of 10 controversial topics for Spanish users were identified. For each topic, several queries were defined with the purpose of retrieval messages from Twitter about the same topic. Table 3 shows the query terms and the topics defined.

After that, all tweets were manually labeled by two annotators. In this case, the annotators labeled tweets until the amount of 1000 and 2000 ironic and non-ironic was reached. For this dataset, the Cohen's Kappa κ was not computed, because only those tweets in which both annotators agreed the corresponding label was assigned.

Table 3: Topics and query terms defined in the Spanish dataset variant.

Topic	**Description/*Query Terms***
Tardà	Declaration of the Catalan politician in the procès trial. *(joan tardá)*
Relator	Relator (teller or rapporteur) figure to mediate in negotiations between the Spanish government and Catalonia. *(relator)*
LibroSánchez	Launching of the book "I will resist" (Resistiré) written by President Pedro Sánchez. *(Pedro Sánchez & libro), (@sanchezcastejon & libro)*
Franco	Exhumation process of the dictator Franco from Valle de los Caídos *(exhumación & franco)*
Grezzi	Valencian politician of Mobility. (*Grezzi*)
SemáforosA5	Start-up of traffic lights on the A5 motorway entering Madrid. (*#semaforosA5*)
TierraPlanistas	Referring to the current tendency of freethinkers in favor of Earth is flat. (*tierraplanistas & tierra plana*)
VenAcenar	Reality show where a group of people alternate in whose house to dine, episode 289. (*#VenACenar289*)
YoconAlbert	Hashtag of the political campaign of Albert Rivera, member of the Citizens party, applying for the presidency. (*#yoconalbert*)
PañalesIglesias	The politician Pablo Iglesias of Podemos party appears in the Hormiguero TV program teaching Spaniards to change diapers. (*@Pablo_Iglesias_ AND pañales*)

The official dataset to be provided for evaluation purposes consists of 3000 tweets distributed across the 10 distinct topics. Then, the data were divided into two partitions considering the 80% for training and the rest for test. Table 4 shows the distribution of tweets for each topic in the training and test data.

Table 4: Training and Test partitions distribution on the Spain variant data.

	Training		**Test**	
Topic	**Ironic**	**Non-ironic**	**Ironic**	**Non-ironic**
Tardà	32	240	8	64
Relator	112	75	19	15
Librosánchez	162	90	19	12
Franco	52	240	10	86
Grezzie	54	182	20	36
SemáforosA5	48	215	12	54
Tierraplanistas	86	191	31	40
Venacenar	91	113	19	29
Yoconalbert	55	150	12	38
PañalesIglesias	108	104	50	26
Total	**800**	**1600**	**200**	**400**

4 Evaluation Measures and Baselines

As we consider the task of irony detection as a binary classification problem, we used the standard metrics for evaluating a classifier performance. For the three subtasks, participating systems were evaluated using precision, recall and F1 measure, calculated as follows:

$$Precision_{class} = \frac{\#correct\_classified}{\#total\_classified} \tag{1}$$

$$Recall_{class} = \frac{\#correct\_classified}{\#total\_instances} \tag{2}$$

$$F1_{class} = 2 \times \frac{Precision_{class} \times Recall_{class}}{Precision_{class} + Recall_{class}} \tag{3}$$

The metrics will be calculated per class label and macro-averaged. The submissions were ranked according to *F1-Macro*. This overall metric implies that all class labels have equal weight in the final score, resulting interesting in imbalanced datasets. Participating teams were restricted to submit only one run for each subtask.

In order to assess the complexity of the task per language variant and the performance of the participants' approaches, we propose the following baselines:

- *BASELINE-majority.* A statistical baseline that always predicts the majority class in the training set. In case of balanced classes, it predicts one of them.
- *BASELINE-word n-grams,* with values for n from 1 to 10, and selecting the 100, 200, 500, 1000, 2000, 5000, and 10000 most frequent ones.
- *BASELINE-W2V [46,47].* Texts are represented with two word embedding models: *i)* Continuous Bag of Words (CBOW); and *ii)* Skip-Grams.
- *BASELINE-LDSE [55].* This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: human/bot, male/female. The distribution of weights for a given document should be closer to the weights of its corresponding category. LDSE takes advantage of the whole vocabulary.

For all the methods we have experimented with several machine learning algorithms (below) and will report in the following the best performing one in each case. For each method we used the default parameters setting provided by Weka tool[12].

- Bayesian methods: Naive Bayes, Naive Bayes Multinomial, Naive Bayes Multinomial Text, Naive Bayes Multinomial Updateable, and BayesNet.

[12] https://www.cs.waikato.ac.nz/ml/index.html

- Logistic methods: Logistic Regression and Simple Logistic.
- Neural Networks: Multilayer Perceptron and Voted Perceptron.
- Support Vector Machine.
- Rule-based method: Decision Table.
- Trees: Decision Stump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree, and REP Tree.
- Lazy method: KStar.
- Meta-classifiers: Bagging, Classification via Regression, Multiclass Classifier, Multiclass Classifier Updateable, and Iterative Classifier Optimize.

Finally, we have used the following configurations:

- BASELINE-word n-grams:
  - CU: 10000 words 1-grams + SVM
  - ES: 200 words 1-grams + BayesNet
  - MX: 2000 words 1-grams + SVM
- BASELINE-W2V:
  - CU: Fasttext-Wikipedia + Logistic Regression
  - ES: Fasttext-Wikipedia + Voted Perceptron
  - MX: Fasttext-Wikipedia + BayesNet
- BASELINE-LDSE:
  - CU: LDSE.v1 (MinFreq=10, MinSize=1) + Random Forest
  - ES: LDSE.v2 (MinFreq=5, MinSize=2) + SVM
  - MX: LDSE.v1 (MinFreq=2, MinSize=2) + BayesNet

5 Participating Systems

A total of 12 teams participated simultaneously in the three subtasks (A,B, and C) on binary irony classification. Table 5 shows each team's name, institutions and country. As can be observed in the table, teams from five countries where motivated by the challenge, specifically 4 teams from Spain, 3 teams from Mexico, 3 teams from Italy, one team from Cuba, and another from Brazil.

Generally, the participating systems employed machine learning-based approaches ranging from traditional classifiers (being the SVM the most popular one) to complex neural network architectures [24,59]; only one approach [13] addressed the challenge by using a pattern matching strategy, and one more exploited the impostor method [60]. Regarding the features used, we identified word embeddings (different models were employed such as Word2Vec, FastText, Doc2Vec, Elmo, and Bert) [19] as well as n-grams (in terms of words and characters) [32,48]. Only a few approaches took advantage of affective and stylistic features [10,18]. It is worthy to notice the use of features extracted from universal syntactic dependencies [14], which proved to be useful for detecting irony.

Although we suggested to consider the given context for identifying irony, only three approaches took it into account [10,18,32]. In general, no strong evidence was shed about the impact of context for understanding irony on short

Table 5: Paricipating teams

Team name	Institution	Country
ELiRF-UPV	Universitat Politècnica de València (UPV)	Spain
CIMAT	Mathematics Research Center (CIMAT),	Mexico
JZaragoza	Universitat Politècnica de València (UPV),	Spain
ATC	Università degli Studi di Torino (UniTO)	Italy
CICLiku	Computing Research Center, National Polytechnic Institute (CIC-IPN)	Mexico
LabGeoCi	Mathematics Research Center (CIMAT), Center for Research in Geospatial Information Sciences A.C. (CentroGeo)	Mexico
SCoMoDI	Università degli Studi di Torino (UniTO)	Italy
LaSTUS/TALN	Universitat Pompeu Fabra (UPF),	Spain
VRAIN	Universitat Politècnica de València (UPV)	Spain
Aspie96	Università degli Studi di Torino (UniTO)	Italy
UO	Center for Pattern Recognition and Data Mining (CERPAMID)	Cuba
UFPelRules	Universidade Federal de Pelotas (UFPel)	Brazil

Spanish messages. We are aware that modeling the context is still really difficult. Moreover, when we compare constrained systems to unconstrained ones, we noted that only two systems included additional data.

Table 6 shows the performance of each participant in terms of F1 in each subtask and F1-Average (AVG) according to all subtasks. Systems were ranked according to the last global score F1-Average. As can be observed in Table 6, all systems outperform the *Majority class* baseline, six overpass the *Word N-gram* baseline whereas three systems achieved better results than the *Word2Vec baseline* and only two outperform *LDSE* baseline. The last mentioned two baselines clearly perform well in the three subtasks and generally they can be considered as strong.

Below we discuss the top five best-performing teams, which all built a constrained (i.e., only the provided training data were used) and supervised system. The best system, developed by [24] [26], achieved an AVG= 0.6832. Their proposal computes vector representations combining the encoder part of a Transformer Model and word embeddings extracted from a skip-gram model trained with the 87 million tweets by using Word2Vec tool [45]. The messages were represented in a d-dimensional fixed embedding layer, which was initialized with the weights of the word embedding vectors. After that, transformer encoders are applied relaying on the multi-head scaled dot-product attention. A global average pooling mechanism was applied to the output of the last encoder, that it is used as input to a feed-forward neural network, with only one hidden layer, whose output layer computes a probability distribution over the the two classes of each subtask.

In the top five systems it is possible to find also the teams *CIMAT* [48] (AVG=0.6585), *JZaragoza* (AVG=0.6490), *ATC* (AVG=0.6302) [14], and *CICLiku*

Table 6: Macro F-measure per language and global ranking as average per variant.

Ranking	Team	CU	ES	MX	AVG
1	ELiRF-UPV	0.6527	**0.7167**	**0.6803**	**0.6832**
2	CIMAT	**0.6596**	0.6449	0.6709	0.6585
*	BASELINE-LDSE	0.6335	0.6795	0.6608	0.6579
3	JZaragoza	0.6163	0.6605	0.6703	0.6490
*	BASELINE-W2V	0.6033	0.6823	0.6271	0.6376
4	ATC	0.5941	0.6512	0.6454	0.6302
5	CICLiku	0.5621	0.6875	0.641	0.6302
6	LabGeoCi	0.6396	0.6251	0.6121	0.6256
*	BASELINE-word n-grams	0.5684	0.6696	0.6196	0.6192
7	SCoMoDI	0.6338	0.6652	0.5574	0.6188
8	LASTUS-UPF_method1	0.6017	0.6606	0.5933	0.6185
9	VRAIN	0.5204	0.6842	0.6476	0.6174
10	LASTUS-UPF_method2	0.5737	0.6493	0.6218	0.6149
11	Aspie96	0.5388	0.5935	0.5747	0.5690
12	UO_run2	0.5930	0.5445	0.5353	0.5576
13	UFPelRules	0.5620	0.5088	0.5464	0.5391
14	UO	0.4996	0.5110	0.4890	0.4999
	BASELINE-majority	0.4000	0.4000	0.4000	0.4000
	Min	0.4996	0.5088	0.4890	0.4999
	Q1	0.5620	0.6014	0.5617	0.5805
	Median	0.5936	0.6502	0.6169	0.6187
	Mean	0.5891	0.6288	0.6061	0.6080
	SDev	0.0492	0.0653	0.0584	0.0496
	Q3	0.6294	0.6641	0.6471	0.6302
	Max	0.6596	0.7167	0.6803	0.6832
	Skewness	-0.2438	-0.8078	-0.4794	-0.7328
	Kurtosis	2.0663	2.4494	2.1610	2.8608
	Normality (p-value)	0.9119	0.0343	0.5211	0.0984

[10] (AVG=0.6302). The *CIMAT* system considers vectors by concatenating features built from three distinct representations: i) based on words embeddings leaned by Word2Vec on huge corpus, ii) based on a deep representation leaned by LSTMs neural networks, and iii) based on n-grams at character and word level. The first representation uses traditional pre-trained Word2Vec and average the word vectors of the tokens contained in each document. The second considers only the last hidden state of an LSTMs with 256 units. The third is a set of 2-3-4 grams at character and word levels, which are selected (the top 5000) by using the Chi-square metric implemented in sklearn tool[13]. All representations were concatenated and fed into a SVM with a linear kernel.

The third best system presented by the team *JZaragoza* addressed the challenge by using a character and word-based n-grams representation and a SVM as classifier with a radial kernel. The team *ATC* ranked fourth and it faced the task of irony detection by a shallow machine learning approach. The most salience and novel contribution is based on representing the messages by morphological and dependency-based features. It also worth noting that the proposed model trained a SVM on the three datasets altogether (7,200 texts) and tested the same model on the three different test sets, regardless of the three variants of Spanish. The fifth best system was presented by the *CICLiku* team. The proposed

[13] https://scikit-learn.org/

model is based on embeddings based on the FastText [8] trained on Spanish Billion Words [11] and the emotion-levels as features, and AdaBoost M1 function on Random Forest as classifier. Considering the role of affective information in irony detection, in this work the messages were represented by the six main emotions (love, joy, surprise, sadness, anger, and fear), with the particularity of taking into account intensities of such emotions learned from the text. The emotion based representation (with only six features) achieved competitive results compared with the embedding based representation.

The remainder systems obtained results very close to the *Word N-gram* baseline. All of them except one (*UFPelRules*) tackled the irony detection task using supervised approaches, however the nature and complexity of their architectures and features vary significantly. The *LabGeoCi* proposal [19] uses a distributed representation of the texts; i.e., the deep contextualized word representations ELMo (Embeddings from Language Models)[51]. The *SCoMoDI* system [18] uses a SVM with radial kernel approach and stylistic, semantic, emotional, affective and lexical features. In [59] the *LaSTUS/TALN* team trained the models for the different languages simultaneously and considered data from other IberLEF 2019 shared tasks, as a technique for data augmentation. It uses word embeddings (FastText) built by using external data from the other IberLEF 2019 shared tasks; besides, it uses a neural network model based on a simple bidirectional LSTM (biLSTM) networks.

The *VRAIN* system [32] uses vectors of counts of word n-grams and an ensemble of the SVM and Gradient Tree Boosting model. The work presented by the *Aspie96* team addressed the task by using character-level neural network, representing each character as an array of binary flags. The network is composed of some convolutional layers, followed by a bidirectional GRU layer (BiGRUs). The *UO* team [13] uses an adaptation of the impostors method and bag-of-words, punctuation marks, and stylistic features for building vector representation. They submitted the results of two runs, the first one considering as features the token extracted by the Freeling NLP tokenizer, the second one considering the lemmas extracted by the FreeLing tool[14]. It is worth to notice that the *UO* team tackled the problem from one-class classification perspective (commonly used for verification tasks). Finally, the last ranked system *UFPelRules* [37], which was the single unsupervised system, uses several linguistic patterns in order to trained the models: syntactic rules, static expressions, lists of laughter expressions, specific scores, and symbolic language.

6 Evaluation and Discussion of the Results

In this section we present and discuss the results obtained by the participants. Firstly, we show the final ranking. Then, we analyse the error in the different variants. Finally, a cross-variant analysis is presented.

[14] http://nlp.lsi.upc.edu/freeling/node/1

6.1 Global Ranking

A total of 12 teams have participated in the shared task, submitting a total of 14 runs. In Table 6 the overall performance per language variant and users' ranking are shown. The highest results have been obtained for the Spanish variant (0.7167), followed by the Mexican (0.6803) and the Cuban (0.6527) one. The best results for the Cuban variant have been obtained by [48]. The best results for the other variants, as well as the best average results, have been achieved by [24].

In average, the systems obtained better results for the Spanish variant (0.6288) than for the Mexican (0.6061) and Cuban (0.5891) ones. In the case of the Spanish variant, the distribution is also narrower than for the other variants (see Figure 1). This is reflected in their inter-quartile ranges (ES: 0.0627; MX: 0.0854; CU: 0.0674), although the standard deviation in the case of Spanish (0.0653) is higher than for the other variants (CU: 0.0492; MX: 0.0584). This is due to some systems with high performance (far from the average, albeit not enough to be considered as outliers) that stretch the median up with respect to the mean (ES: 0.6502 vs. 0.6288; CU: 0.5936 vs. 0.5891; MX: 61.69 vs. 0.6061).

It can be observed in Figure 1 that the Spanish variant has two peaks, the highest one around 0.68 and the other one around 0.52. This is reflected in the ranking with two groups of systems with F-measures between 0.6251 and 0.7167, and between 0.5088 and 0.5445, respectively. Furthermore, the lowest p-value for this variant (0.0343) indicates a restraint from the normal distribution.

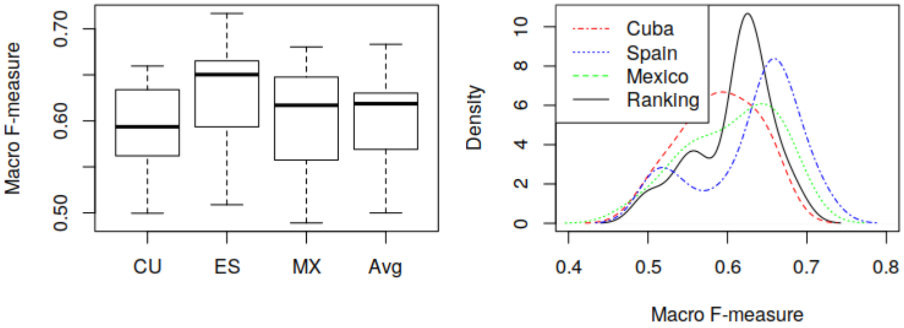

Fig. 1: Distribution and density of the results in the different variants.

6.2 Results per Topic in each Variant

In this section we analyse the achieved results per topic. We have aggregated all the systems predictions, except baselines, and calculated the F-measure per topic in each variant. Then, the distribution of F-measures have been plotted in Figures 2, 3, and 4 respectively for Cuba, Spain, and Mexico.

Regarding Cuba, it can be observed that the topic with the systems performing better refers to "Economy", although with similar median than "E-Quality" and “DigitalTV”. On the contrary, there are several topics where the systems performed worst, although with a different behaviour. For example, the median value is similar for “Sports” and “TechSociety”. Nevertheless, the sparsity is much higher in the last case, with even an outlier system which failed in most cases.

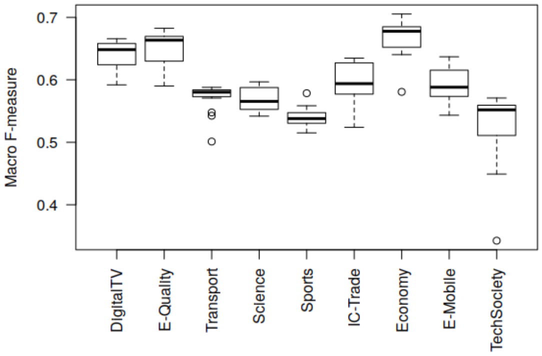

Fig. 2: Distribution of results per topic in the Cuban variant.

Regarding Spain, the topic where the systems performed better was “El Relator” (The Relator), with a high median and not very large inter-quartile range (sparsity). Furthermore, this is the topic with the highest F-measure, with a median about 0.75. The topic with the worst performance is “VenACenar” (the reality show), where there are also two outliers with F-measures close to 0.45. There are two topics with similar maximum, minimum and inter-quartile range, but with inverted medians: “Franco” and “YoconAlbert”. We can also highlight the obtained results in the “Tierraplanistas” (Flatearthers) topic due to its low sparsity: most systems behaved similarly, albeit the overall performance was not very high, contrary to what could be expected due to the topic.

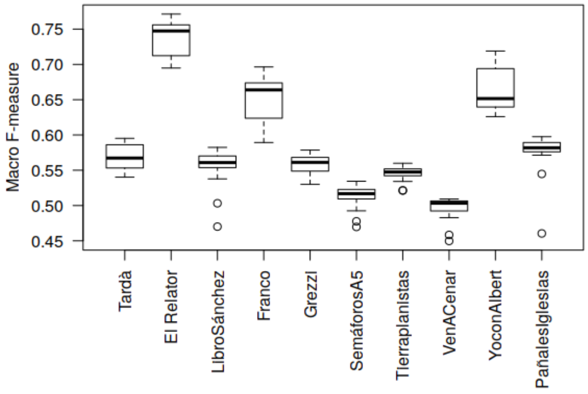

Fig. 3: Distribution of results per topic in the Spanish variant.

Regarding Mexico, the topics with the highest performance are "Funding cuts for children day-care centers" and "CRE", although the second one with lowest sparsity. The topic with the lowest performance is "Venezuela", with average values around 0.50. Similar to the Spanish variant, the topic with the lowest sparsity is 'FlatEarth", although the performance of the systems is higher in average (0.60 vs. 0.55), probably meaning that irony is easier to be identified in Mexico for this particular topic.

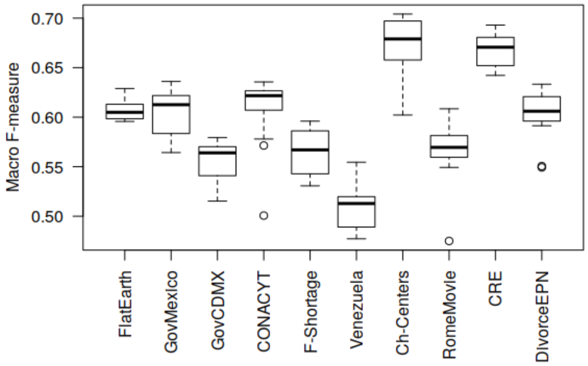

Fig. 4: Distribution of results per topic in the Mexican variant.

6.3 Error Analysis

We have aggregated all the participants' predictions for the different variants, except baselines, and plotted the respective confusion matrices in Figures 5, 6 and 7, respectively for Cuba, Spain, and Mexico. In all the variants, the highest confusion is from Ironic to Non-Ironic texts (0.5338, 0.4963, and 0.5263 respectively for Cuba, Spain, and Mexico). As can be seen, the error is similar in the three variants, ranging from 0.4963 to 0.5338, a difference of 0.0375. Regarding the confusion from Non-Ironic to Ironic texts, the difference among variants is also similar (0.2761, 0.2357, and 0.2579), although with a slightly larger range of 0.0404.

As a consequence, the highest results are obtained in the case of Ironic texts (0.7239, 0.7643, and 0.7421, respectively for Cuba, Spain, and Mexico), whereas they are significantly lower in case of Non-Ironic texts (0.4662, 0.5037, and 0.4737). As can be seen, in the case of Cuba and Mexico, the accuracy in Non-Ironic texts is below the 50%.

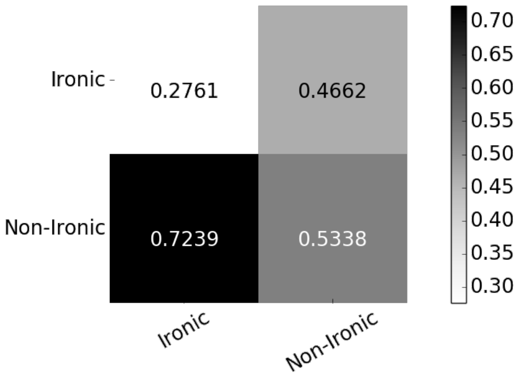

Fig. 5: Aggregated confusion matrix for the Cuban variant

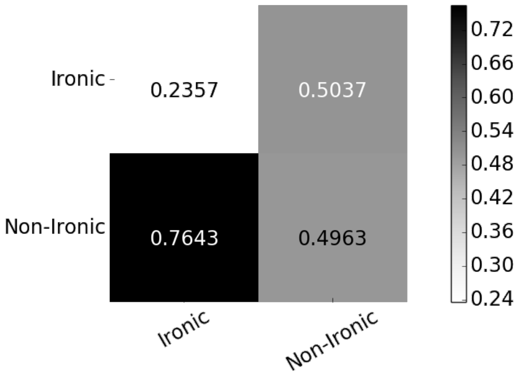

Fig. 6: Aggregated confusion matrix for the Spanish variant

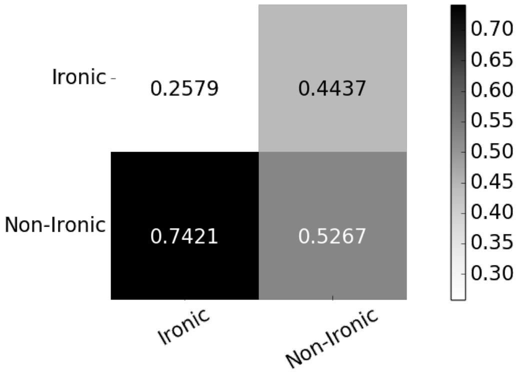

Fig. 7: Aggregated confusion matrix for the Mexican variant

6.4 Cross-Variant Evaluation

In this section we analyse the performance of the systems when they are trained on one variant and tested on a different one. Looking at Table 7, we can see that the highest performance was achieved by CIMAT when trained their system in the Cuban variant and tested it on the Spanish one (0.6106). Nevertheless, we can observe that the average performance is very similar in all cases (see Figure 8), ranging from 0.5078 in case of Spain to Cuba, to 0.5451 in case of Cuba to Mexico. Similarly, the median ranges from 0.5145 in case of Mexico to Cuba, to 0.5511 also in case of Cuba to Mexico.

Table 7: Cross-variant raking.

TEAM	MX->ES	CU->ES	ES->MX	CU->MX	ES->CU	MX->CU	AVG
JZaragoza	0.4904	0.5846	**0.5734**	**0.5741**	0.5216	0.5263	**0.5451**
ELiRF	**0.5359**	0.5442	0.5595	0.5733	0.4978	0.5585	0.5449
CIMAT	0.5070	**0.6106**	0.4944	0.5632	0.5187	0.5593	0.5422
LabGeoCi	0.5328	0.4825	0.5464	0.5663	0.5218	**0.5648**	0.5358
LASTUS-UPF_method1	0.5350	0.5183	0.5329	0.5404	**0.5225**	0.4842	0.5222
CICLiku	0.5238	0.5551	0.5100	0.5502	0.4841	0.5028	0.5210
SCoMoDI	0.4677	0.5333	0.5599	0.5519	0.5062	0.4866	0.5176
VRAIN	0.5198	0.5086	0.5422	0.5034	0.4683	0.5485	0.5151
LASTUS-UPF_method2	0.5176	0.4523	0.5516	0.5478	0.5207	0.4712	0.5102
UO	0.4626	0.3574	0.4891	0.4806	0.5166	0.4965	0.4671
Min	0.4626	0.3574	0.4891	0.4806	0.4683	0.4712	0.4671
Q1	0.4945	0.4890	0.5157	0.5423	0.4999	0.4891	0.5157
Median	0.5187	0.5258	0.5443	0.5511	0.5177	0.5145	0.5216
Mean	0.5093	0.5147	0.5359	0.5451	0.5078	0.5199	0.5221
SDev	0.0270	0.0720	0.0289	0.0306	0.0188	0.0358	0.0233
Q3	0.5305	0.5524	0.5575	0.5655	0.5214	0.5560	0.5406
Max	0.5359	0.6106	0.5575	0.5741	0.5225	0.5648	0.5451
Skewness	-0.7474	-0.8930	-0.5195	-1.1419	-1.1158	0.0464	-1.2446
Kurtosis	2.1136	3.4012	1.9277	3.1084	2.8864	1.3756	4.1749
Normality (p-value)	0.1146	0.5322	0.2960	0.0517	0.0160	0.1798	0.1008

Looking at Figure 8, we can highlight the similar inter-quartile range (sparsity) in case of Cuba to Spain (from 0.4890 to 0.5524), and in case of Mexico to Cuba (from 0.4891 to 0.5560), even with a small difference in their median (0.5258 vs. 0.5145).

In Figure 9, the distribution of the results for the cross-variant scenario is shown without outliers. This reshapes the figures and highlights some insights. For example, in the case of systems tested on Spanish from Spain, they have similar median (0.5187 in case of Mexican as training set; 0.5258 in case of Cuban). However, the inter-quartile range is much higher in the second case (0.0634 vs. 0.0360). In the case of Mexico as test variant, the systems performed better when trained on the Cuban variant than on the Spanish one (0.5451 vs. 0.5359 in average; 0.5511 vs. 0.5443 in median), and also the sparsity is lower (0.0232 vs. 0.0418 in terms of inter-quartile range). Finally, with respect to Cuban as testing variant, the results are better with the Mexican variant as training in terms of maximum accuracy (0.5648 vs. 0.5225), Q3 (0.5560 vs. 0.5214) and mean (0.5199 vs. 0.5078). However, with the Spain variant as training the sparsity is lower (0.0215 vs. 0.0669) as well as the median (0.5177 vs. 0.5145) is slightly higher.

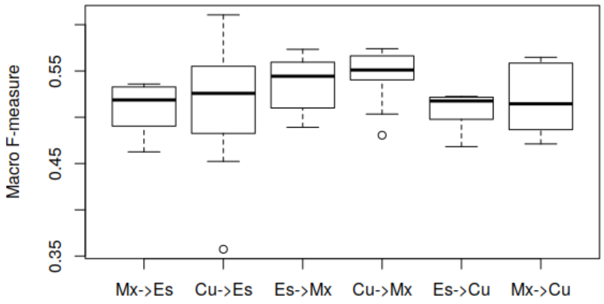

Fig. 8: Distribution of results cross-variant.

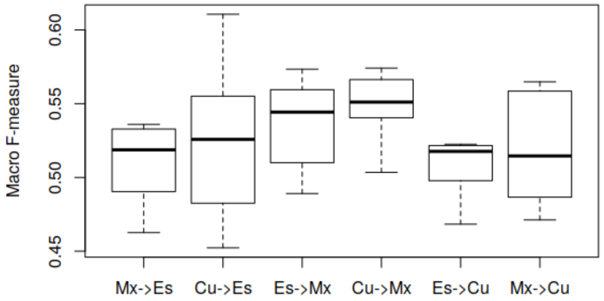

Fig. 9: Distribution of results cross-variant (without outliers).

6.5 Intra-Variant vs. Cross-Variant

In this section the obtained results are compared with the results in the cross-variant scenario. As can be seen, there is a considerable decrease in the performance for all the statistical variables, specially in the case of the best performing system where the F-measure decreases from 0.6832 to 0.5451 (a drop of 0.1381).

Table 8: Statistics Intra-Variant vs. Cross-Variant.

Statistics	Intra-Variant	Cross-Variant	Diff
Min	0.4999	0.4671	0.0328
Q1	0.5805	0.5157	0.0648
Median	0.6187	0.5216	0.0971
Mean	0.6080	0.5221	0.0859
SDev	0.0496	0.0233	0.0263
Q3	0.6302	0.5406	0.0896
Max	0.6832	0.5451	0.1381
Skewness	-0.7328	-1.2446	0.5118
Kurtosis	2.8608	4.1749	-1.3141
Normality (p-value)	0.0984	0.1008	-0.0024

As can be seen in Figure 10, the intra-variant results are closer to a normal distribution with the average performance around 0.6080, whereas the cross-variant results contain two clear peaks, one around the median value of 0.5216 and the other one around the minimum value of 0.4671. Nevertheless, the systems' behavior in the cross-variant scenario is more homogeneous: most of them obtained results around the mean and their inter-quartile range is half (0.0249 vs. 0.0497).

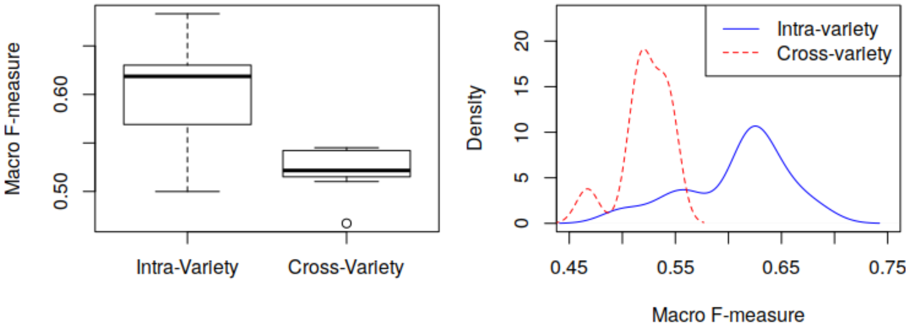

Fig. 10: Distribution and density of results intra-variant vs. cross-variant.

7 Conclusions

This paper describes IroSvA (Irony Detection in Spanish Variants), the first shared task fully dedicated to irony detection in short messages written in Spanish. The task was composed of three subtasks aiming to identify irony in user-generated content written by Spanish speaking users from Spain, Mexico, and Cuba. Unlike related competitions, participating systems to this task were asked to determine the presence of ironic content considering not only isolated texts but also the "context" to which each text belongs to. Datasets from each variant were developed considering diverse contexts according to controversial topics at each country. Aiming to investigate their performance in a cross-variant setting, the participating systems were asked to train their models in a given variant and evaluated it on the two remainings.

A total of 12 teams participated in the shared task. Several approaches were proposed by participants, ranging from traditional strategies exploiting n-grams (at both word and character levels), stylistic and syntactic features to deep learning models using different word embeddings representations (such as Word2Vec, FastText, and ELMo), convolutional layers, autoencoders, and LSTM. The performance of the systems was ranked considering as evaluation metric the F1-Average (it takes into account the F1 score obtained in each subtask). Overall, participating systems achieved a higher performance in F1 terms for the Spanish variant. The best-ranked team, *ELiRF-UPV*, achieved an F1-Average of 0.6832 by exploiting a deep learning-based approach. Regarding the cross-variant evaluation, the best result (0.6106 in F1 terms) was obtained by *CIMAT* when their system was trained on the Cuban variant and then applied over the one coming from Spain. It is important to highlight that, the results achieved by the participating systems are similar to the ones obtained in other shared tasks on irony detection focused on different languages.

Broadly speaking, IroSvA serves to establish a common framework for the evaluation of Spanish irony detection models. Furthermore, the datasets developed for this task could serve to foster the research on irony detection when the instances are related to a defined context.

Acknowledgments

The work of the fourth author was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The third and fifth authors were partially supported by CONACYT-Mexico project FC-2410.

References

1. Attardo, S.: Irony as Relevant Inappropriateness. Journal of Pragmatics **32**(6), 793–826 (2000). https://doi.org/10.1016/S0378-2166(99)00070-3
2. Bamman, D., Smith, N.A.: Contextualized Sarcasm Detection on Twitter. In: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015. pp. 574–577. AAAI, Oxford, UK (2015)
3. Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: Overview of the Evalita 2016 SENTIment POLarity Classification Task. In: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016. CEUR Workshop Proceedings, vol. 1749. CEUR-WS.org (2016)
4. Barbieri, F., Saggion, H., Ronzano, F.: Modelling Sarcasm in Twitter, a Novel Approach. In: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 50–58. Association for Computational Linguistics, Baltimore, Maryland, USA (June 2014)
5. Basile, V., Bolioli, A., Nissim, M., Patti, V., Rosso, P.: Overview of the Evalita 2014 SENTIment POLarity Classification Task. In: Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014. pp. 50–57 (2014)
6. Benamara, F., Grouin, C., Karoui, J., Moriceau, V., Robba, I.: Analyse d'Opinion et Langage Figuratif dans des Tweets : Présentation et Résultats du Défi Fouille de Textes DEFT2017. In: Actes de l'atelier DEFT2017 Associé à la Conférence TALN. Orléans, France (June 2017)

7. Bharti, S.K., Vachha, B., Pradhan, R.K., Babu, K.S., Jena, S.K.: Sarcastic Sentiment Detection in Tweets Streamed in Real Time: a Big Data Approach. Digital Communications and Networks (2016). https://doi.org/10.1016/j.dcan.2016.06.002
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the ACL. **5**, 135–146 (2017)
9. Bosco, C., Patti, V., Bolioli, A.: Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. IEEE Intelligent Systems **28**(2), 55–63 (2013)
10. Calvo, H., Juárez-Gambino, O.: Emotion-Based Cross-Variety Irony Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
11. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (2016). [Online]. Available: http://crscardellino.me/SBWCE/. Retrieved May 4, 2018, http://crscardellino.me/SBWCE/
12. Carvalho, P., Sarmento, L., Silva, M.J., de Oliveira, E.: Clues for Detecting Irony in User-generated Contents: Oh...!! it's "so easy" ;-). In: Proceedings of the 1st International Conference on Information Knowledge Management Workshop on Topic-Sentiment Analysis for Mass Opinion. pp. 53–56 (2009)
13. Castro, D., Benavides, L.: UO-CERPAMID at IroSvA: Impostor Method Adaptation for Irony Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
14. Cignarella, A.T., Bosco, C.: ATC at IroSvA 2019: Shallow Syntactic Dependency-based Features for Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
15. Cignarella, A.T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., et al.: Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA). In: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018). vol. 2263, pp. 1–6. CEUR-WS (2018)
16. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. pp. 107–116. CoNLL '10, Association for Computational Linguistics, Uppsala, Sweden (2010)
17. Filatova, E.: Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). pp. 392–398. European Language Resources Association (ELRA), Istanbul (May 2012)
18. Frenda, S., Patti, V.: SCoMoDI: Computational Models for Irony Detection in three Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
19. García, L., Moctezuma, D., Muñiz, V.: A Contextualized Word Representation Approach for Irony Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society

for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
20. Garmendia, J.: Irony. Cambridge University Press, New York, USA, first edn. (2018). https://doi.org/10.1017/9781316136218
21. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation. pp. 470–478. Association for Computational Linguistics, Denver, Colorado (2015)
22. Ghosh, A., Veale, T.: Fracking Sarcasm using Neural Network. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 161–169. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/W16-0425
23. Ghosh, D., Richard Fabbri, A., Muresan, S.: The Role of Conversation Context for Sarcasm Detection in Online Interactions. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 186–196. Association for Computational Linguistics, Saarbrücken, Germany (Aug 2017). https://doi.org/"10.18653/v1/W17-5523"
24. González, J.A., Hurtado, L.F., Pla, F.: ELiRF-UPV at IroSvA: Transformer Encoders for Spanish Irony Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
25. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying Sarcasm in Twitter: A Closer Look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 581–586. HLT '11, Association for Computational Linguistics, Portland, Oregon (2011)
26. Gupta, R.K., Yang, Y.: CrystalNest at SemEval-2017 Task 4: Using Sarcasm Detection for Enhancing Sentiment Classification and Quantification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 626–633. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/S17-2103
27. Hernández Farías, D.I., Benedí, J.M., Rosso, P.: Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In: Paredes, R., Cardoso, J.S., Pardo, X.M. (eds.) Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, vol. 9117, pp. 337–344. Springer International Publishing, Santiago de Compostela, Spain (2015). https://doi.org/10.1007/978-3-319-19390-8_38
28. Hernández Farías, D.I., Bosco, C., Patti, V., Rosso, P.: Sentiment Polarity Classification of Figurative Language: Exploring the Role of Irony-Aware and Multifaceted Affect Features. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 46–57. Springer International Publishing, Cham (2018)
29. Hernández Farías, D.I., Patti, V., Rosso, P.: Irony Detection in Twitter: The Role of Affective Content. ACM Trans. Internet Technol. **16**(3), 19:1–19:24 (2016). https://doi.org/10.1145/2930663
30. Hernández Farías, D.I., Rosso, P.: Irony, Sarcasm, and Sentiment Analysis. In: Pozzi, F.A., Fersini, E., Messina, E., Liu, B. (eds.) Sentiment Analysis in Social Networks, pp. 113–128. Elsevier Science and Technology (2016), http://dx.doi.org/10.1016/B978-0-12-804412-4.00007-3
31. Huang, Y.H., Huang, H.H., Chen, H.H.: Irony Detection with Attentive Recurrent Neural Networks. In: Jose, J.M., Hauff, C., Altıngovde, I.S., Song, D., Albakour, D., Watt, S., Tait, J. (eds.) Advances in Information Retrieval. pp. 534–540. Springer International Publishing, Cham (2017)

32. Iranzo-Sánchez, J., Ruiz-Dolz, R.: VRAIN at IroSvA 2019: Exploring Classical and Transfer Learning Approaches to Short Message Irony Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
33. Jasso López, G., Meza Ruiz, I.: Character and Word Baselines Systems for Irony Detection in Spanish Short Texts. Procesamiento del Lenguaje Natural **56**, 41–48 (2016), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285
34. Joshi, A., Bhattacharyya, P., Carman, M.J.: Investigations in Computational Sarcasm. Springer Nature, Singapore (2018)
35. Joshi, A., Bhattacharyya, P., Carman, M.J.: Automatic Sarcasm Detection: A Survey. ACM Comput. Surv. **50**(5), 73:1–73:22 (Sep 2017). https://doi.org/10.1145/3124420
36. Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., Carman, M.J.: Are Word Embedding-based Features Useful for Sarcasm Detection? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November, 2016. pp. 1006–1011 (2016)
37. Justin Deon, D., de Freitas, L.A.: UFPelRules to Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
38. Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., Hadrich-Belguith, L.: Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 644–650. Association for Computational Linguistics (July 2015)
39. Karouia, J., Zitoune, F.B., Veronique Moriceau: SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media. In: 3rd International Conference on Arabic Computational Linguistics, ACLing 2017. pp. 161–168. Association for Computacional Linguistic (ACL), Dubai, United Arab Emirates (2017)
40. Khattri, A., Joshi, A., Bhattacharyya, P., Carman, M.: Your Sentiment Precedes You: Using an Author's Historical Tweets to Predict Sarcasm. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 25–30. Association for Computational Linguistics, Lisboa, Portugal (September 2015)
41. Kunneman, F., Liebrecht, C., van Mulken, M., van den Bosch, A.: Signaling Sarcasm: From Hyperbole to Hashtag . Information Processing & Management **51**(4), 500 – 509 (2015)
42. Liu, B.: Sentiment Analysis and Opinion Mining, vol. 5. Morgan & Claypool Publishers (2012). https://doi.org/10.2200/S00416ED1V01Y201204HLT016
43. Lukin, S., Walker, M.: Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue. In: Proceedings of the Workshop on Language Analysis in Social Media. pp. 30–40. Association for Computational Linguistics, Atlanta, Georgia (June 2013)
44. Maynard, D., Greenwood, M.A.: Who Cares about Sarcastic Tweets ? Investigating the Impact of Sarcasm on Sentiment Analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (2014)

45. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Nips pp. 1–9 (2013). https://doi.org/10.1162/jmlr.2003.3.4-5.951
46. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at International Conference on Learning Representations (ICLR'13) (2013)
47. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems pp. 3111–3119 (2013)
48. Miranda-Belmonte, H.U., López-Monroy, A.P.: Early Fusion of Traditional and Deep Features for Irony Detection in Twitter. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
49. Nozza, D., Fersini, E., Messina, E.: Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. pp. 68–76 (2016). https://doi.org/10.5220/0006052000680076
50. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval **2**(1-2), 1–135. https://doi.org/10.1561/1500000011
51. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202
52. Poria, S., Cambria, E., Hazarika, D., Vij, P.: A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1601–1612. Association for Computational Linguistics, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/C16-1151
53. Ptáček, T., Habernal, I., Hong, J.: Sarcasm Detection on Czech and English Twitter. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. pp. 213–223. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (August 2014)
54. Rangel, F., Hernández Farías, D.I., Rosso, P.: Emotions and Irony per Gender in Facebook. In: Proc. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD), LREC-2014. pp. 68–73. Reykjavík, Iceland (2014)
55. Rangel, F., Rosso, P., Franco-Salvador, M.: A Low Dimensionality Representation for Language Variety Identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), pp. 156-169 (2018)
56. Reyes, A., Rosso, P.: Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. pp. 118–124. WASSA '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2107653.2107668
57. Reyes, A., Rosso, P., Veale, T.: A Multidimensional Approach for Detecting Irony in Twitter. Language Resources and Evaluation **47**(1), 239–268 (2013)

58. Rosenthal, S., Ritter, A., Nakov, P., Stoyanov, V.: SemEval-2014 Task 9: Sentiment Analysis in Twitter. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 73–80. No. SemEval (2014)
59. Seda Mut Altin, L., Bravo, A., Saggion, H.: LaSTUS/TALN at IroSvA: Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR Workshop Proceedings. CEUR-WS.org, Bilbao, Spain (2019)
60. Seidman, S.: Authorship verification using the impostors method notebook for PAN at CLEF 2013. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-Seidman2013.pdf
61. Sulis, E., Hernández Farías, D.I., Rosso, P., Patti, V., Ruffo, G.: Figurative Messages and Affect in Twitter: Differences between #irony, #sarcasm and #not. Knowledge-Based Systems **108**, 132 – 143 (2016). https://doi.org/10.1016/j.knosys.2016.05.035, new Avenues in Knowledge Bases for Natural Language Processing
62. Tang, Y.j., Chen, H.H.: Chinese Irony Corpus Construction and Ironic Structure Analysis. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. pp. 1269–1278. Association for Computational Linguistics, Dublin, Ireland (2014)
63. Van Hee, C., Lefever, E., Hoste, V.: SemEval-2018 Task 3: Irony Detection in English Tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation. SemEval-2018, Association for Computational Linguistics, New Orleans, LA, USA (June 2018)
64. Wallace, B.C.: Computational Irony: A Survey and New Perspectives. Artificial Intelligence Review **43**(4), 467–483 (2015)
65. Wallace, B.C., Choe, D.K., Charniak, E.: Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1035–1044. Association for Computational Linguistics, Beijing, China (July 2015)
66. Wallace, B.C., Choe, D.K., Kertz, L., Charniak, E.: Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 512–516. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
67. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony Detection via Sentiment-based Transfer Learning. Information Processing & Management **56**(5), 1633 – 1644 (2019). https://doi.org/10.1016/j.ipm.2019.04.006