Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 319-322 recibido 04-12-2025 revisado 20-01-2026 aceptado 13-02-2026

New Avenues in Computational Irony Detection in Social Media

Nuevos Enfoques para la Detección Computacional de la Ironía en Redes Sociales

Reynier Ortega Bueno
Universitat Politècnica de Valencia, Spain
rortbue@alumni.upv.es

Abstract: PhD thesis in Computer Science focussed on irony detection in social media, written by Reynier Ortega Bueno under the supervision of *Prof.* Paolo Rosso, at the Universitat Politècnica de València. This thesis investigates irony detection as a multifaceted linguistic, computational, and social challenge, addressing multilingual variation, multimodality, and corpus bias. The work introduces an attentive LSTM architecture integrating linguistic and deep features for Spanish irony and satire detection, and proposes an end-to-end model combining textual and visual transformers for multimodal irony detection in social media content. This work further analyses topic bias in irony corpora, demonstrating its detrimental impact on model generalisation and showing gains achieved through bias identification and mitigation. The defense took place in Valencia, Spain, on July 25^{th}, 2025. The doctoral committee was composed by Rafael Berlanga Llavori, (Universitat Jaume I), Els Lefever, (Ghent University), and Tony Veale, (University College Dublin). The thesis received an international mention, an excellent qualification, and the distinction of *Cum Laude*.
Keywords: Irony Detection, Multimodal, Multilingual, Bias and Generalisation, Social Implications.

Resumen: Tesis doctoral en Informática, centrada en la detección de ironía, presentada por Reynier Ortega Bueno bajo la supervisión del *Prof.* Paolo Rosso en la Universitat Politècnica de València. Esta tesis aborda la detección de ironía como un desafío lingüístico, computacional y social multifacético, considerando la variación multilingüe, la multimodalidad y el sesgo en los corpus de entrenamiento. La tesis propone una arquitectura LSTM con mecanismos de atención que integra rasgos lingüísticos y representaciones profundas para detectar la ironía y la sátira en Español. Introduce un nuevo modelo que combina *transformers* textuales y visuales para el análisis multimodal. Además, analiza el sesgo en los corpus de ironía, mostrando su efecto negativo en la generalización y mostrando mejoras tras su identificación y mitigación. La defensa tuvo lugar Valencia, España el 25 de julio de 2025. El tribunal de defensa estuvo conformado por el *Prof.* Rafael Berlanga Llavori (Universitat Jaume I), la *Prof.* Els Lefever (Ghent University) y el *Prof.* Tony Veale (University College Dublin). La tesis obtuvo la mención internacional, la calificación de excelente y la distinción *Cum Laude*.
Palabras clave: Detección de Ironía, Multimodalidad, Multilingualidad, Sesgo y Generalización, Implicaciones Sociales.

1 Introduction

Irony is a figurative and pragmatic phenomenon (Grice, 1975) whose dependence on context, intent, and cultural cues makes it difficult to model computationally. These challenges intensify on social media, where irony appears in multilingual and multimodal forms (Cai, Cai, and Wan, 2019) and often overlaps with harmful rhetoric (Frenda et al., 2022). Detection requires models that can understand subtle text and visual cues. In addition, model generalisation across domains and languages remains limited (Ghanem et al., 2020). Given the role of irony and satire in preventing miscommunication and misinformation (Rubin et al., 2016), addres-

ISSN 1135-5948 DOI 10.26342/2026-76-24

sing these gaps is socially relevant.

This thesis aims to improve computational irony detection by analysing textual irony in Spanish variants, exploring multimodal irony detection, and investigating bias in irony corpora and its effect on model generalisation, while also examining how irony contributes to miscommunication on social media. The Research Questions (RQ) that we aim to answer in this thesis are:

RQ I Could irony and satire detection be enhanced by effectively combining linguistic-based representations, universal sentence encoder-based representations, and contextualised pre-trained embeddings to inform an attentive recurrent model? Does the use of multihead attention outperform single self-attention in capturing the complex word relationships inherent in such texts?

RQ II Could an end-to-end model that fuses textual transformers and visual transformers effectively combine text and image representations to improve the detection of irony in multimodal social media content? What impact does this integration have on classification accuracy?

RQ III To what extent are benchmark corpora for computational irony detection biased? How can bias mitigation enhance model generalisation across domains and languages, particularly in transferring irony knowledge from one domain or language to others?

RQ IV How does the sensitivity of transformer-based models vary in response to different linguistic perturbations in the irony detection task?

2 Thesis Overview

The research presented in the thesis was organised into **7** chapters. **Chapter 1** presents the central challenges of irony detection and outlines the new avenues through which this phenomenon should be examined, including multilingual variation, multimodality, and data biases. It also discusses potential computational strategies to address these issues automatically. The following 6 chapters are organised into 3 main parts.

Part I: Irony Detection from Text to Multimodality presents the research concerning to irony detection. This part focusses on the problem of textual irony detection and extends our research to multimodal scenarios, incorporating text enriched with visual context (**RQ I** and **RQ II**).

Chapter 2 introduces an attentive LSTM model with three complementary views —linguistic, MUSE-based, and BERT-based —to detect irony and satire in Spanish tweets. Two fusion strategies are evaluated across three corpora. The robustness of the model is further tested on humour recognition, where it obtains competitive results (Ortega-Bueno, Rosso, and Medina Pagola, 2022).

Chapter 3 focusses on multimodal irony detection in social media posts that combine text and images. This chapter introduces a transformer-based architecture that integrates textual and visual information through textual and visual transformers. Results show that in many cases the text-only model can detect irony without visual cues (Tomás et al., 2023).

Part II: Bias, Robustness, and Generalisation addresses the problem of bias and generalisation in irony detection models. Moreover, it explores the impact of bias mitigation in cross-language irony detection (**RQ III** and **RQ IV**).

Chapter 4 examines how bias in irony corpora impact model generalisation through identifying and mitigating topic bias in training data to improve generalisation and evaluating performance when bias is removed from training data. The chapter analyses cross-domain and cross-lingual irony detection, using the Weirdness Index and Polarized Weirdness Index to identify topic-related bias terms. Then a mitigation strategy is applied by masking these terms in the training corpora and retraining the models (Ortega-Bueno, Rosso, and Fersini, 2023).

Chapter 5 assesses the robustness of transformer-based irony detection models to linguistic perturbations, exposing biases in training data across ironic and non-ironic classes. The metodology considers three types of perturbation —word masking, word substitution, and paraphrasing— and finds strong model sensitivity, with substitution and paraphrasing generating the largest prediction shifts, especially for ironic cases. SHAP and LIME analyses further link attribution score variations to prediction errors (Ortega-Bueno, Fersini, and Rosso, 2025).

Part III: Summary extends the previous findings by examining the social impacts of irony and humour. Moreover, it presents the conclusions and future work.

Chapter 6 focusses on how ironic ex-

pressions facilitate stereotype propagation (Ortega-Bueno et al., 2022) and how humour serves as a vehicle for hurtful content (Labadie et al., 2021; Merlo et al., 2023). The analysis further deepens the examination of generalisation capabilities of transformer models for humour recognition across domains and languages (Labadie Tamayo et al., 2023).

Chapter 7 presents the general conclusions of the research, lists the contributions, and comments on the open research lines for possible future work.

3 Conclusions and Contributions

The thesis **proposes and validates the MvAttLSTM model** and shows **that integrating linguistic, emotional, multilingual, and contextual information enhances figurative language understanding**. MvAttLSTM is applied to detect irony in three Spanish variants (Castilian, Mexican, and Cuban), revealing notable differences even within the same language. The approach is further extended to satire detection in two Spanish variants (Castilian and Mexican), improving state-of-the-art results and highlighting variant-specific patterns. In addition, the robustness of MvAttLSTM in humour recognition in Spanish texts is investigated, showing adaptability across figurative phenomena. Experiments distinguishing satire from irony indicate solid but bias-affected performance, evidencing the influence of topic bias in the available corpora. Overall, findings show that linguistic cues and deep sentence encodings are particularly effective for irony, whereas contextual embeddings (BERT) better capture satire, and affective features consistently support both irony and satire. Moreover, the thesis examines multimodal irony detection by **integrating textual and visual data through a novel deep learning architecture based on textual and visual transformers (DT4MID)**. While DT4MID surpasses all baselines across corpora, the findings show that *text-only models proved effective in detecting irony through textual patterns*, indicating that, in many cases, visual information is not essential. These results highlight the strong discriminative power of linguistic cues and the uneven contribution of images to multimodal irony detection in the available corpora.

The thesis proposes a **methodological approach to assess the presence of bias in irony corpora and the effectiveness of mitigation strategies**, showing that benchmark corpora contain substantial topic-related biases that hinder model generalisation. Using Weirdness-based measures, the biased words are identified and demonstrate that **simple mitigation techniques improve out-of-domain performance and cross-lingual transfer**. The study offers new insights into how the reduction of bias affects the generalisation of irony detection models in different languages, confirming that cross-lingual knowledge transfer is feasible when bias is controlled. In addition, it introduces **a methodology to identify potentially biased words in irony corpora that influence model predictions**, revealing an asymmetry in robustness: models are consistently more vulnerable to perturbations in ironic instances. Paraphrasing proved the most disruptive, while masking had minimal impact on attribution distributions, exposing corpus-specific vulnerabilities and the instability of learned features.

The thesis proposes a **study to investigate humour recognition in cross-domain and cross-language scenarios** and **analysing humour that potentially conveys offensive content**. A key finding in the cross-language study was that while neural machine translation preserves core humorous semantics, multilingual transformer models exhibit significant vulnerability when recognising translated humour, particularly with language-dependent humour forms such as puns. The research establishes a practical solution: translating test samples into the language used during the model's fine-tuning process substantially improves cross-language humour recognition performance. Regarding offensiveness, the proposed method found that **affective features significantly impact offensiveness scoring**. In addition, through a linguistic study **key discriminators of offensive humour, such as negative stereotypes and swear words were identified**. Furthermore, the analysis revealed distinctive pronoun usage (more first-person for non-offensive, more third-person plural for offensive) and confirmed that hurtful humour contains measurably higher levels of negative emotional content (anger, disgust, fear, sadness), providing concrete, explainable markers for de-

tection systems. Addressing the social impact of harmful language, this thesis introduces a **corpus to profile Twitter users who spread stereotypes through irony** (Ortega-Bueno et al., 2022), contrasting them with users who employ literal language, which served as the standardised evaluation framework for the PAN 2022 shared task.

Our research demonstrated that the **incorporation of multiple views, including affective information, significantly enhanced the detection of irony and satire and the rating of offensiveness in humorous messages**. Furthermore, it confirmed that **combining visual and textual transformers improved multimodal irony detection**, although textual content alone often proved sufficient for accurate irony discrimination in some cases. Potential **bias in irony corpora is identified and confirmed their impact on model generalisation** across various dimensions, highlighting the need for robust methods to mitigate this bias.

Acknowledgments

This work was mainly supported by *Fairness and Transparency for equitable NLP applications in social media, funded by MCIN/AEI/10.13039/501100011033* and *MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech(PGC2018-096212-B-C31)*.

References

Cai, Y., H. Cai, and X. Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th ACL*, pages 2506–2515. ACL.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2022. The Unbearable Hurtfulness of Sarcasm. *Expert Syst. Appl*, 193:116398.

Ghanem, B., J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. 2020. Irony Detection in a Multilingual Context. In *Advances in Information Retrieval: ECIR 2020*, pages 141–149.

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

Labadie, R., M. J. Rodriguez, R. Ortega, and P. Rosso. 2021. RoMa at SemEval-2021 task 7: A transformer-based approach for detecting and rating humor and offense. In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, editors, *Proceedings of the SemEval-2021*, pages 297–305, Online, August. ACL.

Labadie Tamayo, R., R. Ortega-Bueno, P. Rosso, and M. Rodriguez Cisneros. 2023. On the poor robustness of transformer models in cross-language humor recognition. *PLN*, 70(0):73–83.

Merlo, L. I., B. Chulvi, R. Ortega-Bueno, and P. Rosso. 2023. When humour hurts: linguistic features to foster explainability. *PLN*, 70(0):85–98.

Ortega-Bueno, R., B. Chulvi, F. Rangel, R. Paolo, and E. Fersini. 2022. Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO). In *Proceedings of the Working Notes of CLEF 2022*. CEUR-WS.org.

Ortega-Bueno, R., E. Fersini, and P. Rosso. 2025. On the robustness of transformer-based models to different linguistic perturbations: A case of study in irony detection. *Expert Syst.*, 42(6):e70062.

Ortega-Bueno, R., P. Rosso, and E. Fersini. 2023. Cross-Domain and Cross-Language Irony Detection: The Impact of Bias on Models' Generalization. In *NLDB'2023*, pages 140–155. Springer.

Ortega-Bueno, R., P. Rosso, and J. E. Medina Pagola. 2022. Multi-view Informed Attention-based Model for Irony and Satire Detection in Spanish variants. *Knowl.-based Syst.*, 235:107597.

Rubin, V. L., N. J. Conroy, Y. Chen, and S. Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Workshop (NAACL-CADD'2016)*, pages 7–17, California, USA. ACL.

Tomás, D., R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella. 2023. Transformer-based Models for Multimodal Irony Detection. *JAIHC*, 14(6):7399–7410.