CIC@PAN: Simplifying Irony Profiling using Twitter Data

Notebook for PAN at CLEF 2022

Sabur Butt, Fazlourrahman Balouchzahi, Grigori Sidorov and Alexander Gelbukh

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico

Abstract

The article explains the model submission by the team CIC for "Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO)" at PAN 2022. Irony profiling can help in identifying stereotype spreaders and can enhance the understanding of author behaviours. We proposed a methodology focusing on feature engineering to classify irony for long texts based on multiple linguistic and emotion-based features. We also extensively discussed the shortcomings of the data and the proposed task to provide the future research direction. The paper reveals the impact of robust feature engineering with a machine learning approach on the long social media texts in the author profiles. Our method achieved an accuracy of 87.22% on the test set.

Keywords

Irony profiling, Feature Engineering, Figurative Language Processing, Machine Learning, Stereotype spreaders

1. Introduction

Irony as defined by Wilson and Sperber [1], is an act of communication that expresses an opposite meaning of the literal explanation. It is also defined [2] as a phrase referring to the worldly affairs, that should not be. Irony presents challenges in understanding at various pragmalinguistic levels. Especially, in texts, where our challenge is to automatically detect irony, a phrase that can often befuddle human beings without context, let alone machines. In psychological discourse, human beings can identify irony [3] in the following situations:

- when someone says something ironic according to the situation without the intention of being ironic.
- when irony is easily identifiable without excessive thinking on the ironic statement.
- when no intonational cues are required to understand irony.
- when statements echo societal expectations or norms
- when understanding irony does not violate norms of cooperative communication.

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2021, Bologna, Italy

sbutt2021@cic.ipn.mx (S. Butt); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@cic.ipn.mx (A. Gelbukh)

https://nlp.cic.ipn.mx/sabur/ (S. Butt); https://sites.google.com/view/fazlfrs/home (F. Balouchzahi); https://www.cic.ipn.mx/~sidorov/ (G. Sidorov); http://www.gelbukh.com/ (A. Gelbukh)

0000-0003-1937-3475 (F. Balouchzahi)

Meanwhile, stereotypes [4] are generalized beliefs about controversial topics in the society i.e sexism, misogyny, rumours. Ironic statements posits stereotypes communicating an inherent bias [5]. Hence, while identifying irony in profiles, it is also very useful to flag profiles that contain both irony and stereotypes towards a certain targets i.e. women and LGBTQ. Research in irony detection in textual data has advanced research in Natural Language Processing (NLP) tasks such as author profiling [6, 7], fake news detection [8, 9] and emotion detection [10, 11], where, the text remains ambiguous for models to detect the correct answer.

In this paper, we describe our model submitted to PAN 2022 shared task: "Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) 2022" [12, 13, 14]. The task required us to study user profiles to identify language cues for irony and stereotypes. In natural language processing, this task is referred to as author profiling. While the task description required participants to flag stereotypes using irony, no labelled data was provided to enable such methods (discussed in Section 5). The main goal of this paper (in the context of a classification task) is to experiment and prove that even a single ML classifier can obtain a reasonable and high performance if it gets support from a powerful feature engineering. Therefore, our proposed methodology, consisting a pre-processing step along with the described feature engineering (see Section 4) on a Logistic Regression (LR) classifier achieved a accuracy of 87.22% on the test set. We also explained in detail the shortcomings of the task and the dataset.

2. Related Work

In figurative language processing [15], irony is an umbrella term that might contain sarcasm. Sarcasm [16] is differentiated by an element of scorn that irony is void of. Irony and sarcasm detection have been popular tasks in Natural language processing with multiple approaches published in English [17], Spanish [18], Chinese [19], Portuguese [20] and Arabic [21]. However, in the literature review, our focus would be on scientific approaches for irony detection in English.

Though the research in irony detection for text spans to decades of scientific contribution, the recent methods are dominated by deep learning [22, 23, 24] and transformer methods [25, 26, 17]. The SemEval 2018 task 3 [17] presented an irony detection dataset using tweets challenging the researcher to devise methodology for binary classification and multi-class classification of irony (verbal irony, situational irony, verbal irony with polarity contrast). The leading approach [22] in the competition used Densely connected Bidirectional LSTM (D-BiLSTM) with concatenated features of Part of Speech Tags and word embeddings. The model used a late fusion of the generated vectors of the tweets with the sentiment features. Among the prominent works using transfomers, [25] used transformer encoders for contextualizing pre-trained Twitter word embeddings with multi-head self-attention based system and achieved promising results. Another study [26] used the outputs encoding layers (last four) of the original BERTweet model to generate the sentence embeddings and were fed to 1D Convolutions, self-attention layers and residual connections. The study also reported two additional models. The authors used DeepMoji Features-based sentence embeddings fed into Bi-directional GRU and later derived skip-gram connection between output and input embeddings generated by the Bi-directional GRU. The derived embeddings were then passed into the self attention layer similar to the BERTweet model. The last model used an ensemble of the first two models and BERTweet model using hard/soft classification. The ensemble model achieved the best results on unconstrained settings defined at SemEval-2018 Task 3A. Another model presented in [23] used two Attentional LSTM (Att-LSTM) models to create an ensemble with character level embeddings in one architecture and word embeddings in the other. The authors experimented with majority voting and unweighted average for the ensemble to attain the best results.

The unsupervised approach [24] towards irony detection is also an interesting direction to tackle the problem. The researchers used generative models (probabilistic topic model) combined with word embeddings derived from neural language lexicon. The model was tailored for domain-independent irony detection and achieved 85.81% accuracy on an unbalanced dataset. Many of the studies [27, 28, 29] used emotions and text polarities to identify irony, one of the study [30] used the polarity contrast between words, emojis and fragmented hashtags to identify contrasts in the tweets. The paper used temporal relations and discovered that among most ironic texts, negative polarity was followed by positive or neutral polarity. The polarity contrast, surface-level features and the word embeddings were concatenated to feed the ensemble of Support Vector Machine (SVM) and Logistic Regression (LR) to deliver the best results.

In the text-based studies many of the identifiers of the irony such as speech gestures [31] and kinesthetics [32] cannot be recorded, hence, limiting us to linguistic cues and features. Some of the relevant features used for irony include repetitions or punctuation marks [33], affective features [34], common sense knowledge [35], contextual features [36] and polarity contrast [27] etc. It has also been revealed that the best generalization in the irony detection task is achieved using linguistic features [26].

The profiling task, in general, is different from the binary classification of texts, as it analyses the greater size of the text from the profile which is computationally not efficient for transformers. Keeping the literature in mind, we attempted to create a model using the linguistic cues and emotion text features, focusing on robust feature engineering, to create a simple yet effective solution for efficient irony profiling.

3. Dataset

The dataset used in this paper was collected from English tweets by the IROSTEREO 2022 shared task organizers. The dataset is modeled in such a way that each user can be profiled as Irony spreader (I) and Not Irony spreader (NI). It contains 200 tweets per user (420 users) and the labels are distributed equally. The organizers also provided the participant with 180 blinded XML files (each contain 200 tweets) as test set for profiling Irony spreaders.

4. Methodology

The proposed methodology relies on feature engineering to solve the problem. A simple Logistic Regression (LR) is empowered with a feature engineering module comprising feature reduction and selection and a hyperparameter tuning step with Grid Search. This enabled the parameters to be set specifically for the features obtained from the feature engineering module. We explain the steps in detail in the following subsections.

4.1. Feature Engineering

The feature engineering phase started with preprocessing, where we converted the emojis to text using Emot [1] and removed stopwords, links, non-alphabet and non-digit characters from the text. The text was lowercased and passed for feature extraction. Inspired by [7], the character and word n-grams were combined with syntactic n-grams. We also generated emotions, sentiments and hate speech features using the pysentimiento [2] toolkit. The different types of features used in this study are presented in Table 1.

The huge number of features does not guarantee better performance in ML, on the contrary, it increases the feature dimensions and complexity of classification problem [37]. Hence, we initiated dimensionality reduction and reduced the number of features to 100,000 most frequent features. Later, inspired by [38], we used Shapely values to select the features that have higher contributions to the classification task. Lundberg and Lee [39] presented a framework called SHAP [3] that explains the contribution of features in a classification task by assigning a value of importance. The Figure 2 is borrowed from [38] that explains the procedure of obtaining relevant features from traditional and syntactic n-grams. Extreme Gradient Boosting (XGB) classifier was used as the primary classifier because SHAP is integrated into the tree-boosting frameworks. Therefore, the XGB classifier was evaluated on 0.25% of the train set to estimate the most efficient features by computing their feature importance. Better performance of XGB classifier resulted in an efficient feature selection and eventually better classification results.

Table 1
Feature types in the proposed methodology

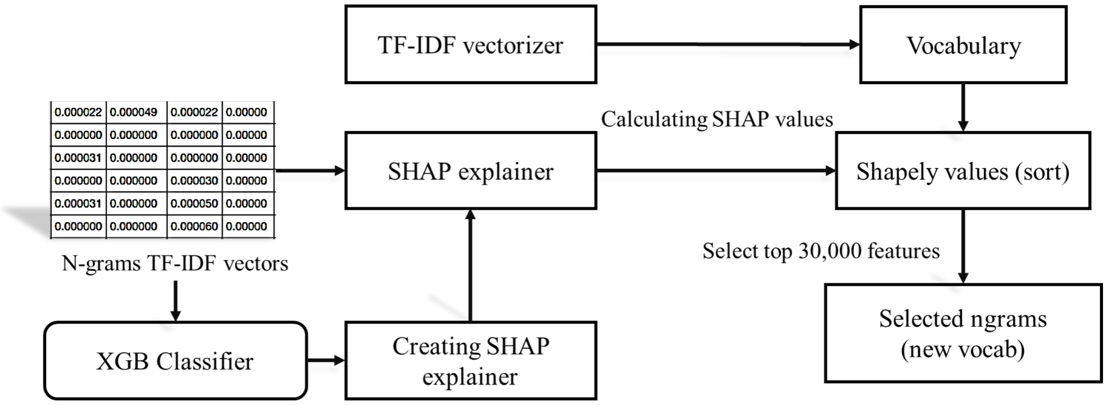
Feature types	Traditional n-grams	Syntactic n-grams	Sentiments	Hate Speech	Emotions
Features	Characters in range (2, 5) Words in range (1, 3)	Bi-grams Tri-grams	Positive Negative Neutral	Aggressive Hateful Targeted	Surprise Sadness Joy Fear Disgust Anger Other

The final selected features were combined with emotions, sentiments and hate speech features and used for classification. The workflow of feature engineering and statistics of features in the proposed methodology is presented in Figure 2.

4.2. Model Construction

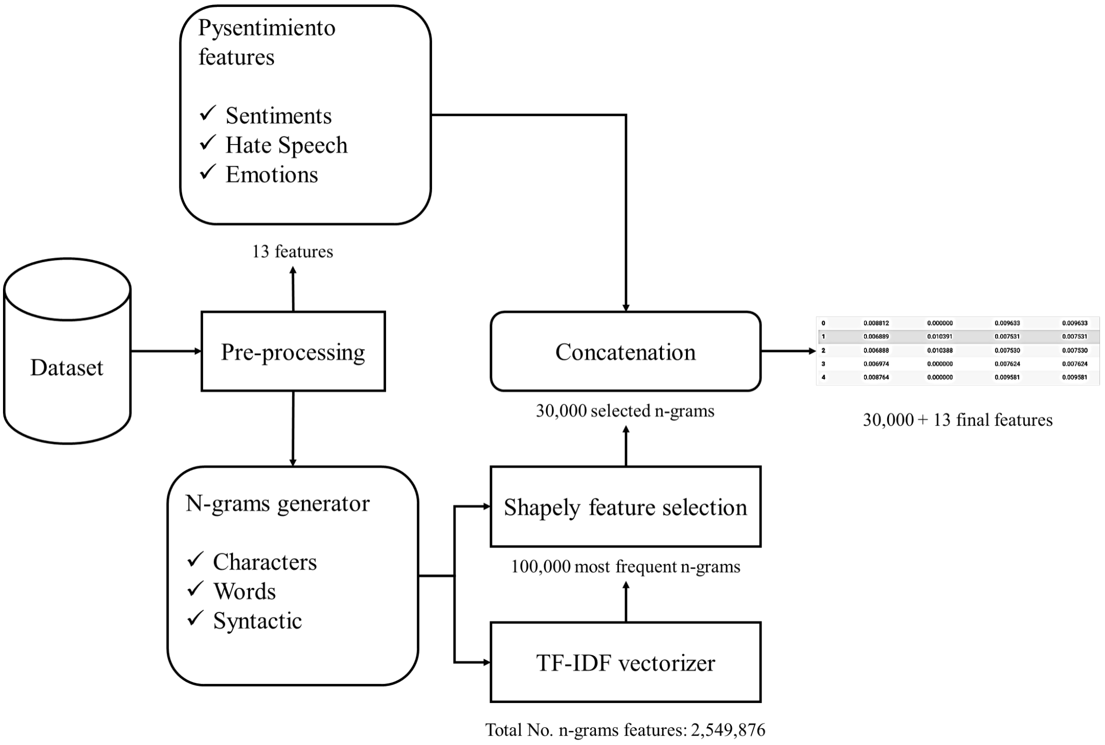
Since the main objective of the current methodology was based on powerful feature engineering and the simplicity of the model, we only experimented with an LR classifier with hyperparameter tuning to enhance efficiency. The hyperparameter tuning process was done using the Grid Search algorithm. Table 2 shows the default, candidate and tuned parameters for the LR classifier in this paper.

[1] https://github.com/NeelShah18/emot
[2] https://github.com/pysentimiento/pysentimiento
[3] https://shap.readthedocs.io/en/latest/index.html



0.000022	0.000049	0.000022	0.00000
0.000000	0.000000	0.000000	0.00000
0.000031	0.000000	0.000000	0.00000
0.000000	0.000000	0.000030	0.00000
0.000031	0.000000	0.000050	0.00000
0.000000	0.000000	0.000060	0.00000

Figure 1: Feature selection process



0	0.008812	0.000000	0.009633	0.009633
1	0.006889	0.010391	0.007531	0.007531
2	0.006888	0.010388	0.007530	0.007530
3	0.006974	0.000000	0.007624	0.007624
4	0.008764	0.000000	0.009581	0.009581

Figure 2: Feature engineering process

5. Criticism and Limitations

While working on the shared task, we felt a few shortcomings with the dataset and the general approach to the competition. The dataset provided several tweets per profile and required us to concatenate all the tweets per profile to identify the irony spreader. None of the tweets was marked individually as irony or non-irony, hence, amalgamating the tweets per profile was the

Table 2
Parameters for LR classifier

Type	Parameters

Default	solver= lbfgs
	C= 1.0
	tol= 0.0001
	n_jobs= None

Candidate	solver: [newton-cg, lbfgs, liblinear] C: [100, 10, 1.0, 0.1, 0.01] tol: [0.0001, 0.0003, 0.0005] n_jobs: [2, 4, 6]

Tuned	solver= newton-cg
	C= 100
	tol= 0.0001
	n_jobs= 2

only option to train the irony profiles. The problem with this approach is that there is no cut-off to decide how many tweets should be ironic tweets to mark the profile as an irony spreader. The profile could have eighty per cent of the tweets as neutral and twenty per cent of the collected tweets as ironic and still have the irony-based linguistic cues in the long concatenated tweet text. Concatenating the tweets per profile makes the author profiling task a text classification problem, instead of just author profiling. We also noted that the task description [4] mentions that "Special emphasis will be given to those authors that employ irony to spread stereotypes, for instance, towards women or the LGTB community", however, the data does not enable any stereotype identification in the profiles. Given the tweets were labelled as ironic / non-ironic, the profiles could have been labelled as stereotypes spreaders given that they mostly tweeted ironically about a certain topic i.e. women. This simply was not possible with the current information and resources provided by the competition. The PAN shared task was structured such that, neither the test set nor the validation set was available for the model construction. This is highly problematic as the constructed models cannot be seen and understood for their errors. Model explainability cannot be performed on the proposed methodology and that defeats the purpose of the shared tasks.

6. Conclusion and Future Work

The article presented a simple but robust solution for classifying irony spreaders. Our approach targeted extensive feature engineering that comprised of feature reduction and advanced feature selection. We used the XGB classifier for feature importance on a subset of the training set and later used SHAP to create shapely values for the selected linguistic features. The linguistic features are proven to be important for generalization of the task. The selected features were concatenated with features such as hate speech (aggressive, hateful, targeted), emotions (surprise, joy, sadness, fear, disgust, anger, other) and sentiments (positive, neutral, negative) that give another dimension to the model understanding. Our results achieved 87.22% accuracy on the

[4] https://pan.webis.de/clef22/pan22-web/author-profiling.html

task proving that robust feature engineering can produce potent results even without complex model construction. In future, we would like to experiment with the model building using the same feature engineering. This makes more sense once the test set is released to identify the strengths and weakness of each model and to present a complete ablation study with the error analyses.

7. Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] D. Wilson, D. Sperber, On Verbal Irony, Lingua 87 (1992) 53–76.

[2] S. Attardo, Irony as Relevant Inappropriateness, Journal of pragmatics 32 (2000) 793–826.

[3] R. W. Gibbs Jr, J. O'Brien, Psychological Aspects of Irony Understanding, Journal of pragmatics 16 (1991) 523–530.

[4] N. Dragan, M. L. Collard, J. I. Maletic, Automatic identification of class stereotypes, in: 2010 IEEE International Conference on Software Maintenance, IEEE, 2010, pp. 1–10.

[5] C. Burgers, C. J. Beukeboom, Stereotype Transmission and Maintenance through Interpersonal Communication: The Irony Bias, Communication Research 43 (2016) 414–441.

[6] H. Shashirekha, F. Balouchzahi, Ulmfit for twitter fake news spreader profiling., in: CLEF (Working Notes), 2020.

[7] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier, in: CLEF, 2021.

[8] S. Butt, S. Sharma, R. Sharma, G. Sidorov, A. Gelbukh, What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses, Computers in Human Behavior (2022) 107345.

[9] N. Ashraf, S. Butt, G. Sidorov, A. Gelbukh, Cic at checkthat! 2021: fake news detection using machine learning and data augmentation, in: CLEF, 2021–Conference and Labs of the Evaluation Forum, 2021.

[10] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label emotion classification of urdu tweets, PeerJ Computer Science 8 (2022) e896.

[11] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE Access 9 (2021) 97803–97812.

[12] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[13] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture,

in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.

[14] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: A. Barron-Cedeno, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.

[15] D. I. H. Farías, V. Patti, P. Rosso, Irony Detection in Twitter: The Role of Affective Content, ACM Transactions on Internet Technology (TOIT) 16 (2016) 1–24.

[16] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Systems with Applications 193 (2022) 116398.

[17] C. Van Hee, E. Lefever, V. Hoste, SemEval-2018 task 3: Irony Detection in English Tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 39–50.

[18] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y Gómez, J. E. Medina Pagola, Overview of the Task on 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-WS. org, volume 2421, 2019, pp. 229–256.

[19] R. Xiang, X. Gao, Y. Long, A. Li, E. Chersoni, Q. Lu, C.-R. Huang, Ciron: a New Benchmark Dataset for Chinese Irony Detection (2020).

[20] U. B. Corrêa, L. Coelho, L. Santos, L. A. d. Freitas, Overview of the IDPT Task on Irony Detection in Portuguese at IberLEF 2021 (2021).

[21] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, P. Rosso, Idat at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 10–13.

[22] T. Vu, D. Q. Nguyen, X.-S. Vu, D. Q. Nguyen, M. Catt, M. Trenell, Nihrio at SemEval-2018 Task 3: A Simple and Accurate Neural Network Model for Irony Detection in Twitter, arXiv preprint arXiv:1804.00520 (2018).

[23] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, A. Potamianos, Ntua-slp at SemEval-2018 task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning, arXiv preprint arXiv:1804.06658 (2018).

[24] D. Nozza, E. Fersini, E. Messina, Unsupervised Irony Detection: a Probabilistic Model with Word Embeddings, in: International Conference on Knowledge Discovery and Information Retrieval, volume 2, SCITEPRESS, 2016, pp. 68–76.

[25] J. Á. González, L.-F. Hurtado, F. Pla, Transformer based Contextualization of Pre-trained Word Embeddings for Irony Detection in Twitter, Information Processing & Management 57 (2020) 102262.

[26] L. Famiglini, E. Fersini, P. Rosso, On the Generalization of Figurative Language Detection:

The Case of Irony and Sarcasm, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2021, pp. 178–186.

[27] C. Van Hee, Can Machines Sense Irony?: Exploring Automatic Irony Detection on Social Media, Ph.D. thesis, Ghent University, 2017.

[28] H. Rangwani, D. Kulshreshtha, A. K. Singh, Nlprl-iitbhu at SemEval-2018 task 3: Combining Linguistic Features and Emoji Pre-trained CNN for Irony Detection in Tweets, in: Proceedings of the 12th international workshop on semantic evaluation, 2018, pp. 638–642.

[29] J.-Á. González, L.-F. Hurtado, F. Pla, ELiRF-UPV at SemEval-2018 Tasks 1 and 3: Affect and Irony Detection in Tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 565–569.

[30] O. Rohanian, S. Taslimipoor, R. Evans, R. Mitkov, Wlv at SemEval-2018 task 3: Dissecting Tweets in Search of Irony, Association for Computational Linguistics, 2018.

[31] F. Poyatos, La comunicación no Verbal, volume 13, Ediciones AKAL, 1994.

[32] D. C. Muecke, Irony Markers, Poetics 7 (1978) 363–375.

[33] P. Schoentjes, La Poetica de la Ironia., Cathedra„ 2003.

[34] S. Poria, E. Cambria, D. Hazarika, P. Vij, 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, 2016, pp. 1601–1612.

[35] C. Van Hee, E. Lefever, V. Hoste, We Usually Don’t Like Going to the Dentist: Using Common Sense to Detect Irony on Twitter, Computational Linguistics 44 (2018) 793–832.

[36] B. C. Wallace, E. Charniak, et al., 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), 2015, pp. 1035–1044.

[37] F. Balouchzahi, S. Butt, G. Sidorov, A. Gelbukh, CIC@ LT-EDI-ACL2022: Are Transformers the Only Hope? Hope Speech Detection for Spanish and English Comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 206–211.

[38] F. Balouchzahi, G. Sidorov, H. L. Shashirekha, Fake News Spreaders Profiling using N-grams of Various Types and SHAP-based Feature Selection, Journal of Intelligent & Fuzzy Systems (2022) 1–12.

[39] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, Advances in neural information processing systems 30 (2017).