

# VRAIN at IroSva 2019: Exploring Classical and Transfer Learning Approaches to Short Message Irony Detection

Javier Iranzo-Sánchez<sup>[0000-0002-4035-3295]</sup> and  
Ramon Ruiz-Dolz<sup>[0000-0002-3059-8520]</sup>

Valencian Research Institute for Artificial Intelligence (VRAIN)  
Camino de Vera s/n. 46022 Valencia (Spain)  
{jairsan,raruidol}@vrain.upv.es

**Abstract.** This paper describes VRAIN’s participation at *IroSvA 2019: Irony Detection in Spanish Variants* task of the Iberian Languages Evaluation Forum (IberLEF 2019). We describe the entire pre-processing, feature extraction, model selection and hyperparameter optimization carried out for our submissions to the shared task. A central part of our work is to provide an in-depth comparison of the performance of different classical machine learning techniques, as well as some recent transfer learning proposals for Natural Language Processing (NLP) classification problems.

**Keywords:** Natural Language Processing · Irony Detection · Transfer Learning

## 1 Introduction

From the linguistic point of view, irony is a very interesting property of language. As defined in [9], irony is the ability of expressing some specific meaning by the use of terms and words that, by its own, have the completely opposite meaning. On the other hand, from a computational viewpoint, irony can be seen as an important headache when performing natural language analysis tasks. For example, in sentiment analysis, some of the most important features to determine the text polarity are inferred from the words appearing in the text (e.g. negation, n-grams, POS tagging, etc.) [11]. With an ironic text, some of these features should be smoothed in order to perform correctly the sentiment analysis. Irony can be a problem when performing sentiment analysis from a text, this issue has been directly observed in past International Workshop on Semantic Evaluation (SemEval[1]) editions. In 2015, [17] two different datasets were considered, one without sarcastic tweets and other containing sarcastic tweets. Systems performance was considerably lower with the sarcastic dataset. In fact, irony detection

---

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IberLEF 2019, 24 September 2019, Bilbao, Spain.

was proposed as a task in 2015 [7] and tackled for English language tweets in 2018 [18].

This paper describes VRAIN’s participation at *IroSvA 2019: Irony Detection in Spanish Variants*[1] task of the Iberian Languages Evaluation Forum (IberLEF 2019) [13]. In this task we must identify ironic texts written in three Spanish variants (from Spain, Mexico and Cuba). For Spain and Mexico subtasks, we must detect ironic tweets and for Cuba subtask we are supposed to detect ironic comments from a news website.

We worked in each subtask in isolation (only the Spanish tweets from Mexico were used for the Mexico subtask and so on), but used the same approach and pipeline in all 3 subtasks. Model selection and hyper-parameter optimisation were individually carried out for each subtask.

The rest of the paper is structured as follows. In Section 2 we explain the feature extraction process carried out for this specific task. In Section 3 we briefly describe the system and present all the different approaches taken into account to perform our comparison. In Section 4 we present the evaluation made in order to compare the behaviour of the different models considered in this work. We also compare the results obtained by our approach with all the different baselines provided by the organisation. Finally, in Section 5 we summarise the conclusions and the most important features of our work.

## 2 Dataset and Feature Extraction

We will first begin by describing the structure of the competition’s dataset. Table 1 contains statistics of the training and test datasets. Both Spain (*es*) and Mexico (*mx*) variants have 10 different topics and the amount of tweets is different for each topic. On the other hand, Cuba (*cu*) variant has only 9 different topics and the amount of news per topic is also different. Regarding the training dataset, for all the Spanish variants we have 2400 samples, thus the corpus is balanced regarding each variant. On the other hand, the test dataset is made up of 600 samples for every Spanish variant.

**Table 1.** Statistics of the IroSvA 2019 dataset.

Variant	Topics	Train Samples (Ironic Samples)	Test Samples (Ironic Samples)
<i>es</i>	10	2400 (800)	600 (200)
<i>mx</i>	10	2400 (800)	600 (200)
<i>cu</i>	9	2400 (800)	600 (200)

Having described the most important features of the dataset, we will now focus on the feature extraction process. The text was tokenized using NLTK’s [3] TweetTokenizer. Additionally, we experimented with substituting all occurrences of hashtags, url, user mentions and numbers by a generic topic for each category, but we finally decided against it since it decreased the model’s performance.

Each tweet was represented by a vector of counts of word n-grams. Using counts directly instead of tf-idf performed better in our exploratory experiments. The dataset contains additional information apart from the tweets themselves. Specifically, we are given the corresponding topic for each of the tweets. We have tried two ways of leveraging this information. In the first approach, which we have called *global-model*, only one model is trained for each subtask, and a one-hot vector encoding the topic is appended to every sample. Therefore, in this approach, we have a single model per sub-task, trained with data from all the topics.

In the second approach, which we have called *topic-model*, we trained one model per topic. Thus, at training time, we trained each of the individual models using only data from one topic, and at inference time, for each of the tweets, we used the predictions of the model that has been trained using the data of the tweet’s topic. The results of both approaches are compared and evaluated in Section 4.

### 3 System Description

We will now describe the different models we tried for irony detection. In order to select appropriate values for the hyperparameters of each model, we carry out 5-fold Cross Validation, and select the configurations that obtained higher F1 (macro-averaged). Unless otherwise noted, methods are implemented using the sklearn toolkit [14].

#### 3.1 Classification approaches

- **Naive Bayes:** The Naive Bayes approach is a well-known technique for tackling many classification problems. A Multinomial distribution is used to model  $P(x_i|c)$ .
- **Support Vector Machines:** Support Vector Machines [5] are Maximum Margin Classifiers that have been shown to obtain good results in a variety of tasks. We use a linear kernel, that has been shown to outperform other non-linear kernels in text classification problems [20].
- **Gradient Tree Boosting:** Gradient Tree Boosting is a boosting technique that consists in an ensemble of tree models built in a sequential way from a set of weak learners. We have used the implementation available in the XGBoost toolkit [4].
- **Linear Models (fastText):** fastText [8] is a toolkit implementing a set of linear architectures for text classification. The model based on the CBOW architecture [12], has a word embedding matrix used to look up a representation of each word in the text. The embeddings are summed and averaged into a fixed-sized vector, which is then fed into a softmax classifier. Additionally, we have also trained a version using pre-trained word-embeddings, using a publicly available dataset of 200-dimensional word embeddings trained on Spanish tweets [2].

- **BERT**: BERT [6] is a pre-training methodology for Transformer models [19]. BERT models are pre-trained on massive amounts of unsupervised text data, and can then be used in a transfer-learning approach for other downstream tasks. For this task, we have used the pre-trained *BERT-Base Multilingual Cased* model, and fine-tuned it on the IroSvA data for 10 epochs.

## 4 Experimental Evaluation

The results obtained by the different models are shown in Table 2.

**Table 2.** Model performance measured with 5-fold CV performed over the training data (macro F1).

Model	Subtask		
	<i>es</i>	<i>mx</i>	<i>cu</i>
Naive Bayes (Topic model )	0.67	0.51	0.50
Naive Bayes	0.63	0.52	0.57
fastText	0.63	0.62	0.61
fastText(Tweeter pre-trained)	0.63	0.62	0.61
BERT	0.61	0.50	0.57
SVM (Topic model)	0.70	0.60	0.58
SVM	0.70	0.60	0.66
Gradient Tree Boosting (Topic)	0.52	0.50	0.45
Gradient Tree Boosting	0.69	0.60	0.66
Ensemble (SVM + Gradient Boosting)	0.71	0.65	0.66

We can see a number of interesting results from the table. First, except for a single case (Naive Bayes for the *es* variety), *topic models* obtain similar or worse results than their *global* counterparts. Most likely due to the reduced number of training data, the fine-grained approach of individually modelling each topic seems counterproductive.

In terms of the transfer learning approaches we tried, we have not been able to leverage the knowledge obtained from the pre-trained tasks. The fastText model using pre-trained embeddings does not improve the results of the base fastText model, and the BERT model obtains results similar to the Naive Bayes model.

Overall, the best results are obtained by the SVM and Gradient Boosting models. In order to further improve the results, we have constructed an Ensemble of the SVM and Gradient Boosting models, whose predictions are the average of the individual models’ predictions. This obtains additional improvements in the *es* and *mx* variants, and was the model submitted to the competition. Table 3 shows the performance of our model compared to the competition baselines.

The results obtained by our model present significant variations depending on the task. In the case of the *es* task, our model outperforms all baselines, and in the *mx*, our system comes in second place behind the LDSE [15] baseline.

**Table 3.** System comparison between our submission and the competition baselines (macro F1)

Model	Subtask			Average
	<i>es</i>	<i>mx</i>	<i>cu</i>	
LDSE [15]	0.6795	0.6608	0.6335	0.6579
W2V	0.6823	0.6271	0.6033	0.6376
Word nGrams	0.6696	0.6196	0.5684	0.6192
MAJORITY	0.4000	0.4000	0.4000	0.4000
VRAIN	0.6842	0.6476	0.5204	0.6174

However, in the case of the *cu* task, our model is only able to beat the majority baseline. We do not know the reasons for the significant performance drop in the *cu* task between our internal experiments and the competition results, although one possible culprit is the aforementioned domain mismatch between the  $\{es, mx\}$  and *cu* tasks.

## 5 Conclusions

This paper has described VRAIN’s submission to IroSvA 2019. The different experiments have shown that, under the current conditions, classical models have an edge over some of the recent transfer-learning techniques that we tested. We believe that the limiting factor is the lack of sufficient training data for the finetuning step.

Our submission, based on an Ensemble of SVM and Gradient Tree Boosting models, obtains good results across the board, although the performance could be improved in the *cu* case. This has been achieved using non-task-specific bag of n-gram features. It is expected that these results could be further improved with specific features for irony detection, such as those from [16,10].

**Acknowledgements** The research leading to these results has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 761758 (X5gon) and from the Valencian Government grant for excellence research groups PROMETEO/2018/002.

## References

1. IroSvA 2019: Irony detection in spanish variants. <http://www.autoritas.net/IroSvA2019/>, accessed: 2019-07-05
2. Word embeddings trained with word2vec on 200 million spanish tweets using 200 dimensions, <http://new.spinningbytes.com/resources/wordembeddings/>
3. Bird, S.: NLTK: the natural language toolkit. In: ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006 (2006), <http://aclweb.org/anthology/P06-4018>

4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 785–794 (2016), <https://doi.org/10.1145/2939672.2939785>
5. Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* **20**(3), 273–297 (1995), <https://doi.org/10.1007/BF00994018>
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J.A., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015. pp. 470–478 (2015), <http://aclweb.org/anthology/S/S15/S15-2080.pdf>
8. Grave, E., Mikolov, T., Joulin, A., Bojanowski, P.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers. pp. 427–431 (2017), <https://aclanthology.info/papers/E17-2068/e17-2068>
9. Grice, H.P., et al.: Logic and conversation
10. Hernández-Farías, I., Patti, V., Rosso, P.: Irony detection in twitter: The role of affective content. *ACM Trans. Internet Techn.* **16**(3), 19:1–19:24 (2016), <https://doi.org/10.1145/2930663>
11. Liu, B.: Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* **5**(1), 1–167 (2012)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013), <http://arxiv.org/abs/1301.3781>
13. Ortega-Bueno, R., Rangel, F., Hernández Farías, D.I., Rosso, P., Montes-y-Gómez, M., Medina Pagola, J.E.: 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 (2019)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011), <http://dl.acm.org/citation.cfm?id=2078195>
15. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2016). LNCS, vol. 9624, pp. 156–169. Springer-Verlag (2018)
16. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* **47**(1), 239–268 (2013), <https://doi.org/10.1007/s10579-012-9196-x>
17. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V.: Semeval-2015 task 10: Sentiment analysis in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). pp. 451–463 (2015)

18. 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. pp. 39–50 (2018)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 6000–6010 (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need>
20. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA. pp. 42–49 (1999), <https://doi.org/10.1145/312624.312647>