Emotion-Based Cross-Variety Irony Detection*

Hiram Calvo[1], Omar Juárez Gambino[1,2]

[1] Center for Computing Research (CIC)
[2] Escuela Superior de Cómputo (ESCOM)
Instituto Politécnico Nacional
J.D. Bátiz e/ M.O. de Mendizábal, 07738, Mexico City, Mexico
hcalvo@cic.ipn.mx, b150697@sagitario.cic.ipn.mx

Abstract. This work is centered on the data made available for the IroSvA challenge, consisting of three variants of Spanish language from three different countries. We propose a simple model for identifying irony, based on tweet embeddings, refraining from using of additional NLP techniques. We aim to find cues that are able to generalize the knowledge obtained from a language variant, and evaluate the ability to detect irony in different combinations of variants, from different countries and topics. For this purpose, we propose using six features based on the degree of emotion present in each tweet. These automatically tagged features include 5 levels of strength, ranging from none to very high, of six emotions: love, joy, surprise, sadness, anger, and fear. Experiments were carried out with different combinations of language variants. Obtained results show that exclusively using the information of the emotion levels (discarding the embeddings) could improve the irony detection in a language variant different from that used for training.

1 Introduction

Several resources have been used as features for detecting irony: from lexical, syntactic features, to polarity, or changes in polarity [1]. Other works pay special attention to the role of affective information involved in tweets [2] and have experimented with several emotion lexicons such as EMOLEX, EmoSN, SentiSense, LIWC, etc., obtaining state-of-the-art results. In this work, we experiment with the use of similar information, particularly automatically emotion-tagged tweets within the framework of the 6 main emotions described by Shaver (love, joy, surprise, sadness, anger and fear) [3], with the particularity of considering intensities of such emotions learned from text, ranging from N–none, to VH–very high), as described in [4].

* The authors wish to thank support by CONACyT-SNI and Instituto Politécnico Nacional (IPN), particularly through grants SIP20195402, SIP20195886, EDI, and COFAA-SIBE.

Our main goal is to determine to what extent the use of these tags allows irony identification in different corpora (with unrelated topics) of the same language (in this case, Spanish, with some regional variants). An F-measure around 70% has been reported for tests performed on the same kind of trained text [5] [6] and some works report up to 90% using affective content [7]. These tests have been carried out in the same language variant and same topics; however, are there more general cues of irony present that would allow to classify irony learning from one language variant, and testing with another? That is called cross-variety irony detection. For this purpose, a general feature representation that allows domain generalization is needed. A common solution to this, is to use embeddings for representing each tweet [8] [9]. In this work, we propose adding emotion labels as part of these features. Two main questions arise: (1) Can emotion intensity labels alone work as features for cross-variety irony detection? and (2) When used as a complement to an embeddings representation, the use of emotion-based features improves irony cross-variety classification?

To answer these questions, we focus on the corpora provided by the IroSvA challenge [10]. Within this context, our definition of what is ironic and what is not, is defined by the examples provided in the training datasets of this challenge.

The IroSvA challenge aims at investigating whether a short message, written in Spanish language, is ironic or not with respect to a given context. In particular, this challenge aims at studying the way irony changes in distinct Spanish variants. Concretely, it is focused on Spanish from Spain, Mexico and Cuba. Further details are given in [10].

In the next section we describe our classification scheme, along with description of features used. In Section 3 we provide details on our experiments and results, and finally in Section 4 we draw our conclusions.

2 Classification scheme

The same strategy was followed for the three subtasks (although some variants had improvements for some particular subtasks, we opted for using the same method). We performed a standard preprocessing consisting in fixing CR/LF lines, tweets running several lines, and topic names (removing numbers and multi-word names). Then we converted representation to one-hot (word space model, WSM) with no lemmatization, no stopwords handling, and without filtering the minimum number of occurrences of each word. Approximately 12,000 tokens were identified for each corpus. Finally, the WSM was converted to embeddings using FastText Embeddings [11] from SBWC (Spanish Billion-Words Corpus)[3]. The number of dimensions was 300 and a total of 855,380 vectors were used[4].

The IroSvA challenge has three corpora of distinct Spanish variants. Particularly Spanish from Spain, Mexico and Cuba. Each corpus was manually labeled,

[3] crscardellino.github.io/SBWCE/

[4] github.com/dccuchile/spanish-word-embeddings

and includes 1,600 examples of non-ironic texts, and 800 ironic texts. We randomly sampled 800 examples of non-ironic texts to have a balanced training data with 800 ironic texts and 800 non-ironic texts—800 non-ironic texts were discarded.

Models were trained using the AdaBoost M1 function [12] on Random Forest Classifiers [13] with parameters Bag Size=100%; Batch Size=100; and Unlimited depth trees.

3 Experiments and results

Table 1 shows the accuracy of 10-fold stratified cross-validation for each language variant separately. This test was run on the 1,600 tweets balanced corpora for each language variant (Spain '*es*', Mexico '*mx*' and Cuba '*cu*'). Tests were performed with features consisting of vectors of 300 dimensions for each language, as described in the previous section, along with the corresponding topic (as a nominal feature—see the first column of Table 1), and removing topic information (see second column).

Table 1. Accuracy for 10-fold stratified cross-validation on each language variant

language	using topics	without topics
es	82.83%	82.83%
mx	80.68%	79.23%
cu	78.45%	80.28%

The effect of considering topics is different for each language variant: for the *es* variant, there were no changes on performance, while for the *mx* variant, removing topics resulted on a small performance decrease. Finally, for the *cu* variant, not using topics represented a small performance increase. Therefore, we cannot conclude that adding or removing topic information could be of general benefit for this task. However, for the next series of experiments, topic information had to be removed, as topics among language variants are completely different. Results of Table 1 suggest that removing this information would not harm general performance for this task, so that for following experiments, only features of embeddings are used.

3.1 Cross-variety irony detection using embeddings

For this series of experiments, we considered the previously balanced corpora of 1,600 tweets each. Additionally, we built three new corpora by combining two language varieties in order to observe the capabilities of generalizing irony characterization from only one language variant vs. a different one, as well as two amalgamated language varieties against a different one. The new corpora were

named *esmx*, which combined the *es* and *mx* corpora; *escu* (*es* + *cu*), and *mxcu* (*mx* + *cu*). Table 2 shows accuracy of all possible combinations, including those which were tested against a subset of training. For example, for the third row (*esmx*) tested with the first colum (*es*), result was significantly higher (89.09%) because *es* was a subset of *esmx*, and, of course, its cases had been already seen in the training set.

Table 2. Accuracy of cross-variety irony detection – using word embeddings

train / test	es	mx	cu	esmx	escu	mxcu
es	-	*56.31%*	55.25%	73.64%	70.18%	*56.32%*
mx	*57.01%*	-	54.17%	72.79%	*55.78%*	68.70%
cu	*57.89%*	52.40%	-	*56.06%*	69.77%	69.14%
esmx	89.08%	88.70%	*55.88%*	-	73.56%	68.10%
escu	84.85%	*58.78%*	78.26%	76.20%	-	78.68%
mxcu	*58.84%*	82.58%	79.90%	69.03%	78.81%	-

From Table 2 more interesting values can be observed: for example, for the first quadrant (top-left), which compared simple (not combined) corpora, the best value was obtained when training with the *cu* variant, tested on the *es* variant. The inverse situation yielded the best results as well (training with *es* and testing with *cu*) compared with training with *mx* (and tested on *cu*). A similar situation happened for the *mx* variant: Training with *es* yielded better results than training with *cu*. Best results for each language variant (per row) are shown in italics for this quadrant.

For the second quadrant (top-right), when using the *es* variant for evaluating with the *mxcu* combined corpus, results were very similar to evaluating only with the *mx* corpus (56.32% vs. 56.31%). However, when training with *mx* on unseen varieties together (*escu*) results were lower than the previous best result (55.78% vs. 57.01% with *es*). The same happened for the *cu* variant evaluated on *esmx* and *es*, respectively (56.06% vs. 57.89%).

Finally, for the third quadrant (bottom-left), combined corpora were used to train, and they were evaluated with single corpora. Combinations not including the evaluation set in the training set are shown in italics. Compared with the first quadrant (training with simple corpora), all varieties were benefited. For example, *cu* increased from 55.25% with *es*, to 55.88% with *esmx*; *mx* increased from 56.31% with *es*, to 58.78% with *escu*; es increased from 57.89% with *cu*, to 58.84% with *mxcu*. This may suggest that, despite being different language varieties with different topics and ways of expression, amalgamating two corpora helped to predict irony on a different corpus.

The last quadrant (bottom-right) is also shown in Table 2; however, as all training sets are partially contained on all evaluation subsets, these results are not so interesting to discuss.

3.2 Cross-variety irony detection using emotion-levels

As mentioned in Section 1, a different set of features is proposed for this task: the use of 5 levels of emotions (None, Very Low, Low, High, and Very High) for a 6-tuple of emotions: (love, joy, surprise, anger, sadness, and fear) corresponding to the top level of emotions proposed by [3]. Another application of an automatic tagger for this kind of emotion-levels can be found in [4].

As an example of the obtained features, consider Table 3. The first tweet has a value of *None* for love, joy, and surprise, while *Very High* anger, and *Low* values of sadness and fear.

Table 3. Examples of emotion-level tagging. Tuple: (love, joy, surprise, anger, sadness, fear); values N=None, VL=Very Low, V=Low, H=High, VH=Very High.

Tweet	**Emotions tuple**
Como cuando cambias de personal porque hacen mal el trabajo encomendado y resulta que los reemplazos son piores	N,N,N,VH,L,L
El cine es subjetivo..... creo q es muy buena para los que vivimos en CDMX en esa época, es nostálgica... sí le faltó un poco más de historia... pero sí me gustó... pienso que dirigir una película sin actores profesionales es un gran mérito!! Felicidades @alfonsocuaron	L,VH,VL,N,N,N
Muy bien, ¡¡¡a comprar!!! Bueno si abre la pagina primero	N,VH,N,N,VL,N

Results for the set of experiments using only emotion features are shown in Table 4. As can be seen, this time experiments that involved the test set in the training set did not have a high accuracy; compare *esmx* with *mx*—embeddings: 88.70%, emotions: 57.68%. Yet interestingly, when evaluating the *cu* variant, both training with *es* or *mx*, results are higher than their embeddings counterpart (shown in bold). In overall, results using emotion-levels only are only 1.91% below their embeddings counterpart for single to single corpora (*es, mx, cu*, first quadrant—top-left), which is interesting, considering the reduction of 300 to only 6 features.

A general comparison of accuracies using embeddings or emotions as features is shown in Table 5. Quadrants are numbered as (1) top-left, (2) top-right, and (3) bottom-left. The first quadrant represents single vs. single varieties, i.e., no variant combinations were used. The second quadrant represents training with single varieties evaluated on their unseen combined variant, i.e. *es* vs *mxcu*, *mx* vs. *escu*, and *cu* vs. *esmx*. The third quadrant represents training with combined corpora, evaluated on their unseen single variant, i.e. *esmx* vs. *cu*; *escu* vs. *mx*; and *mxcu* vs. *es*. For calculating these averages, no overlapping combinations were considered (v.gr. *es* vs. *esmx*).

Table 4. Accuracy of Cross-variety irony detection – using emotions. Values improving embedding-based classification are shown in bold.

train / test	es	mx	cu	esmx	escu	mxcu
es	-	54.36%	**56.44%**	54.61%	55.33%	55.05%
mx	54.36%	-	**54.80%**	54.32%	54.73%	56.47%
cu	49.81%	51.83%	-	49.87%	52.90%	55.46%
esmx	55.62%	55.68%	55.18%	-	55.05%	55.46%
escu	55.43%	54.48%	56.50%	54.77%	-	56.69%
mxcu	51.70%	55.62%	56.12%	52.97%	55.15%	-

Table 5. Average accuracies per quadrant for experiments using embeddings and emotions. Train vs. test.

Quadrant	embeddings	emotions	difference
1-Single vs. Single	55.50%	53.59%	1.91%
2-Single vs. Combined	56.05%	53.22%	2.83%
3-Combined vs. Single	57.83%	53.79%	4.04%
average	56.46%	53.54%	2.93%

As can be seen from Table 5, cross-variety irony detection is performed better when using embeddings as features in average; however, difference found is relatively small, suggesting that emotion features could be used to improve or aid sentiment related tasks, such as irony detection.

Finally, to answer the second question posed in Section 1, we experimented with using emotions and embeddings altogether, obtaining only a slight increase for the *cu* dataset. Accuracies of using only embeddings to embeddings+emotions were: *es*:82.32 to 82.13%; *mx*:80.37 to 78.79%; *cu*:79.10 to 79.36%. From these results, we are not able to conclude that using both embeddings and emotions simultaneously would be of general benefit, at least for the language varieties and topics addressed in this task.

3.3 Results on test set

Finally, in this section we compare our results with other works. Particularly, we were provided with four different results, being majority voting, using word nGrams, Word2Vec features (no specific details provided), and using LDSE, as described in [14]. Accuracy results are shown in Table 6. For one language variant (*es*), our model was able to overcome the provided results, but in average both LDSE and Word2Vec systems presented better results.

4 Conclusions and Future Work

For this task, a relatively simple model was proposed to classify tweets as ironic or not ironic for three different language varieties. This model was mainly based

Table 6. Results with test set and comparison with other approaches

method	es	mx	cu	avg
CICLiku (us)	**68.75%**	64.10%	56.21%	63.02%
LDSE [14]	67.95%	**66.08%**	**63.35%**	**65.79%**
Word2Vec [8]	68.23%	62.71%	60.33%	63.76%
Word nGram	66.96%	61.96%	56.84%	61.92%
Majority voting	40.00%	40.00%	40.00%	40.00%

on embeddings as features. This representation allowed our model to learn features from a different language variant or varieties, and attempt to classify tweets from an unseen variant as ironic or not ironic.

A particular contribution of this work consisted on using emotion-levels as features to perform the same task. Interestingly, the classifiers were still able to classify tweets with a similar performance than when using tweet embeddings—less than 3% overall average difference in accuracy; and for some variant pairs (*es* vs. *cu* and *mx* vs. *cu*) performance was improved, compared to using embeddings only. This evidence suggests that using emotion levels as features could be used to aid sentiment-related classification tasks such as irony detection.

For this work, no additional information other than the embeddings and the emotion-level tagger was used. As a future work, we plan to include information on the context, as well as the possibility to perform opinion objects identification along with sentiment analysis to improve performance in this task.

References

1. Van Hee, C.: Can machines sense irony?: exploring automatic irony detection on social media. PhD thesis, Ghent University (2017)
2. Farías, D.I.H.: Irony and sarcasm detection in Twitter: the role of affective content. PhD thesis, Universitat Politècnica de València (2017)
3. Shaver, P., Schwartz, J., Kirson, D., O'connor, C.: Emotion Knowledge: Further Exploration of a Prototype Approach. Journal of personality and social psychology **52** (1987) 1061
4. Gambino, O.J., Calvo, H.: Predicting emotional reactions to news articles in social networks. Computer Speech & Language **58** (2019) 280–303
5. 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. (2018) 39–50
6. 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). Volume 2263., CEUR-WS (2018) 1–6
7. Farías, D.I.H., Patti, V., Rosso, P.: Irony detection in Twitter: The role of affective content. ACM Transactions on Internet Technology (TOIT) **16** (2016) 19
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

9. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.W.: Tweet2vec: Character-based distributed representations for social media. arXiv preprint arXiv:1605.03481 (2016)
10. 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)
11. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
12. Cortes, E.A., Martinez, M.G., Rubio, N.G.: Multiclass corporate failure prediction by adaboost. m1. International Advances in Economic Research **13** (2007) 301–312
13. Breiman, L.: Random forests. Machine learning **45** (2001) 5–32
14. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 9624, Springer (2018) 156–169