VRAIN at IroSva 2019: Exploring Classical and Transfer Learning approaches to Tweet Irony Detection

Javier Iranzo-Sanchez, Ramon Ruiz-Dolz

Institut Valencià d'Investigació en Intel·ligència Artificial

May 15, 2019

Methodology

- Text tokenized with NLTK's Tweet Tokenizer [2]
- Features: Bag of n-grams (1-3)
- Two types of models:
  - *Topic-model*: 1 model per topic
  - *Global-model*: 1 model per subtask
- Select Hyperparameters through 5-fold cross-validation
- Software: sklearn toolkit [8]

Models

- **Naive Bayes**: Multinomial distribution for $P(x_i|c)$.
- **Support Vector Machines**: Maximum Margin Classifiers [4]. Linear kernel outperforms others in text classification [10].
- **Gradient Tree Boosting**: Boosting, sequential ensemble of weak learners (classification trees). XGBoost toolkit [3].

Transfer Learning Models: fasttext

- fasttext toolkit [6]
- Based on CBOW [7]
- Softmax classifier. Features: Average of token embeddings.
- Pre-trained embeddings: 200 dimensional vectors from 200 million Spanish Tweets [1].

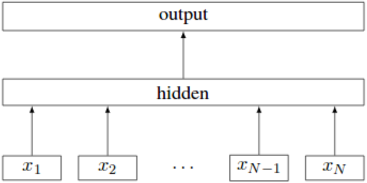

Figure 1: fasttext architecture

Transfer Learning Models: BERT

- BERT [5]: Pre-training for Transformer models [9].

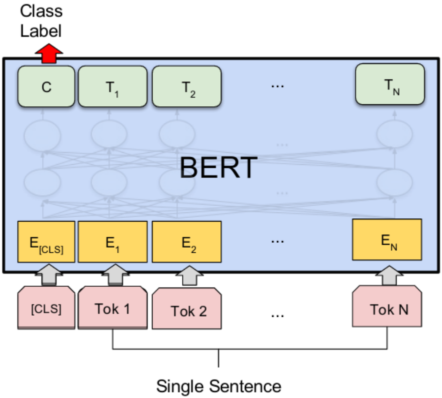

Figure 2: BERT architecture for single sentence classification

Table 1: Model performance measured with 5-fold CV (macro F1)

Model	Subtask *es*	Subtask *mx*	Subtask *cu*
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 (SMV + Gradient Boosting)	0.71	0.65	0.66

Key insights

- *Topic-model* obtains worse results than *Global-model* (Few tweets per topic)
- Transfer learning approaches (fasttext & BERT) underperform classical approaches. Not enough data for the finetuning step.
- Further improvements can be obtained by combining individual models. Final submission: Ensemble of SVM and Gradient Tree Boosting.
- Having used CV to select our hyperparameters, we train using all available data for the final submission.

Submission results

Table 2: System comparison between our submission and the competition baselines (macro F1)

	Subtask			
Model	*es*	*mx*	*cu*	Average
LDSE	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

Key insights(II)

- Results vary a lot depending on the subtask
- In *es*, our system beays all baselines. In *mx*, we come close second.
- However, in *cu*, results way below all relevant baselines except majority class.
- Competitive results using non-task-specific features. It is expected to improve if we use specific features for irony detection.

Bibliography I

Word embeddings trained with word2vec on 200 million spanish tweets using 200 dimensions.

http://new.spinningbytes.com/resources/wordembeddings/.

Steven Bird.

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.

Tianqi Chen and Carlos Guestrin.

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*, pages 785–794, 2016.

Bibliography II

Corinna Cortes and Vladimir Vapnik.
Support-vector networks.
Machine Learning, 20(3):273–297, 1995.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.

Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski.
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*, pages 427–431, 2017.

Bibliography III

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay.
Scikit-learn: Machine learning in python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

Bibliography IV

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.

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*, pages 6000–6010, 2017.

Yiming Yang and Xin Liu.

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*, pages 42–49, 1999.