

Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis^{*}

Franco Martín Luque

Universidad Nacional de Córdoba & CONICET
franco1q@famaf.unc.edu.ar

Abstract. In this article we describe our participation in TASS 2019, a shared task aimed at the detection of sentiment polarity of Spanish tweets. We combined different representations such as bag-of-words, bag-of-characters, and tweet embeddings.

In particular, we trained robust subword-aware word embeddings and computed tweet representations using a weighted-averaging strategy.

We also used two data augmentation techniques to deal with data scarcity: two-way translation augmentation, and instance crossover augmentation, a novel technique that generates new instances by combining halves of tweets.

In experiments, we trained linear classifiers and ensemble models, obtaining highly competitive results despite the simplicity of our approaches.

Keywords: Sentiment Analysis · Polarity Classification · Embeddings · Data Augmentation · Linear Models

1 Introduction

TASS is a shared task organized every year, since 2012, with challenges related to Sentiment Analysis in Spanish. In TASS 2019 [5], the proposed task is to label tweets according to the general sentiment polarity they express, classifying them into four classes: P (positive), N (negative), NEU (neutral, undecided) and NONE (no sentiment).

Five datasets are offered for the task, each one from a different Spanish speaking country: CR (Costa Rica), ES (Spain), MX (México), PE (Perú) and UY (Uruguay). Each corpus is divided into train, development and test sections. No other supervised datasets can be used, but external linguistic resources such as embeddings and lexicons are allowed.

The challenge is divided into two subtasks. In monolingual subtask 1, systems must be trained and tested on the same dataset. In cross-lingual subtask 2,

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.

^{*} This work was partially supported by a research grant from SeCyT, Universidad Nacional de Córdoba.

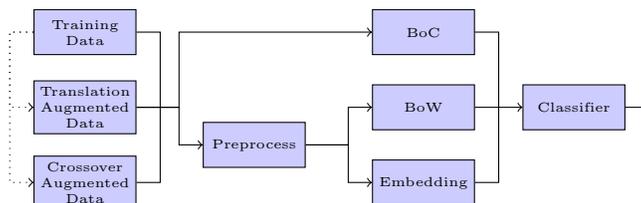


Fig. 1. General architecture of our systems, including the input used in training time. Dotted lines denote precomputed processes.

systems must be trained using datasets from countries others than the one used for testing.

In this article, we describe our participation in TASS 2019 as team Atalaya. We based our systems on our previous work [6] for TASS 2018 [7]. For this edition, we focused our work on data augmentation and robust representations.

To represent tweets, we used a combined approach of bag-of-words, bag-of-characters and tweet embeddings. Tweet embeddings were computed from word embeddings using a weighted averaging scheme. For word embeddings, we used fastText subword-aware vectors [3] specifically trained for sentiment analysis over Spanish tweets.

Our fastText embeddings are robust to noise since they can compute embeddings for unseen words by using subword embeddings. Moreover, we trained them using a database of 90M tweets from various Spanish-speaking countries, giving wide domain-specific vocabulary coverage. We achieved additional robustness by doing preprocessing using several text normalization and noise reduction techniques.

To cope with training data scarcity, we experimented with data augmentation techniques. As in our previous work, we did augmentation using machine translation to and from several other languages.

We also tried a novel augmentation technique we called instance crossover, loosely inspired by the crossover operation from genetic algorithms. This technique combines halves of tweets to generate new instances. Despite its simplicity, this idea showed to be useful in our experiments.

For the classifying models, we used logistic regressions and also bagging ensembles of logistic regressions.

The rest of the paper is as follows. The next section presents the components of our systems and the ideas we used to build them. Section 3 presents the experiments and results for both subtasks. Section 4 concludes the work with some observations about our experience.

2 Techniques and Resources

The components and general architecture of our systems is shown in Fig. 1. In this section, we describe the techniques and resources we used to build them.

2.1 Preprocessing

Preprocessing is important to reduce noise from tweets. We follow our previous work, applying two levels of preprocessing. Basic tweet preprocessing includes tokenization, replacement of handles, URLs, and e-mails, and shortening of repeated letters. Further preprocessing is done, aimed at semantic tasks. It includes removal of punctuation, stopword and numbers, lowercasing, lemmatization, and negation handling.

For negation handling, we followed a simple approach [4, 8]: We find negation words and add the prefix `NOT_` to the following tokens. Up to three tokens are negated, or less if a non-word token is found.

No treatment was performed to hashtags, emojis, interjections and onomatopoeias. Moreover, no spelling correction nor any other additional normalization was applied.

2.2 Bags of Words and Characters

A simple way to represent textual data as feature vectors is to use bag-of-words (BoWs). A bag-of-words represents a tweet as a vector with the counts of words occurring in it. Resulting vectors are high-dimensional and sparse. The BoW representation can be extended to count also word n-grams. In this work, we used BoWs, together with count binarization and TF-IDF re-weighting, both useful for semantic tasks such as sentiment analysis.

For more robustness, we also used a bag-of-characters (BoC) representation. BoCs have exactly the same properties and variants than BoWs but are applied to characters instead of word tokens. Character usage in tweets holds useful information for sentiment analysis. In our work, the BoC representation is computed over the original raw text of tweets, with no preprocessing at all.

2.3 Word Embeddings

A more interesting way to represent text is using embeddings. Word embeddings are low-dimensional dense vector representations of words. These representations are learned in an unsupervised fashion using large quantities of plain text, providing high vocabulary coverage.

For our systems, we used fastText embeddings [3], that introduces additional robustness by learning also subword-level embeddings and using them to compute vectors for unseen words. With subword-aware embeddings, the need for normalization of highly noisy text in preprocessing is greatly alleviated.

We did not use a pretrained fastText model but trained our own using a big preprocessed dataset of ~ 90 million tweets from various Spanish-speaking countries. This dataset is mostly composed of tweets we collected for previous work, and also includes the tweets from all sections of all TASS 2019 datasets.

2.4 Tweet Embeddings

To use word embeddings in sentiment analysis, the embeddings of the individual tokens must be aggregated in some way to obtain a complete tweet representation.

A simple approach is to do averaging to obtain a single vector. A bit more interesting is to add weights to the averaging scheme. This way, some words may be considered more relevant than others for the classification task.

In this work, we used Smooth Inverse Frequency (SIF), a simple weighted averaging scheme from [1] inspired by TF-IDF re-weighting. In SIF, words w are weighted with $\frac{a}{a+p(w)}$, where $p(w)$ is the word unigram probability, and $a > 0$ is a smoothing hyperparameter. Big values of a mean more smoothing towards plain averaging. We model the unigram probability using unigram counts from our preprocessed ~ 90 million tweets dataset.

In [1] a final transformation is applied to tweet embeddings by subtracting from them a common component shared by all the vectors. Preliminary experiments with this idea, however, showed it to be harmful to our systems, so we did not use it in our final experiments.

An important limitation of this tweet embedding scheme is that word order is completely ignored. Only preprocessing may allow the influence of ordering in the result. In particular, the negation handling trick from section 2.1 is a useful, although naive, way to let words be affected by previous negations.

2.5 Data Augmentation with Two-Way Translation

One of the main successful approaches from our previous work on TASS was the use of data augmentation techniques. Data augmentation helps to cope with training data scarcity. Augmentation aims at the introduction of data variability using label-preserving transformations on real data. When correctly used, it contributes to data robustness and acts as a regularizer for the models.

Our approach for TASS 2018 (also as team Atalaya) was to use two-way translation augmentation. In two-way translation, an external machine translation service is used to translate tweets to other “pivot” languages and then back to Spanish. This augmentation technique helps to introduce lexical and syntactical variations to tweets, most times preserving their meaning.

In [6] we used two-way translation to augment the training data using four pivot languages (English, French, Portuguese and Arabic). This augmentation was found to be useful for the ES and CR datasets, but not for PE.

In this work, we explored two-way translation further, applying it to all the datasets using 20 different pivot languages. To get translations, we used Google’s Cloud Translation API service.

Pivot languages were selected by hand from the list of available languages that the API can translate from/to Spanish. The selection was done trying to pick representative languages from different language families.

2.6 Data Augmentation with Instance Crossover

We also tried a new augmentation idea that aims at the generation of new data by combining pairs of instances with the same label. We call it the instance crossover augmentation technique, inspired by the chromosome crossover operation from genetic algorithms.

Our approach is simply to split tokenized tweets into two halves, and then randomly sample and combine first halves with second halves. Resulting instances will probably be ungrammatical and semantically unsound, but our hypothesis is that what is left of semantics, for instance at the lexical level, will preserve sentiment polarity most of the times.¹

Fig. 2 shows an example of instance crossover using two tweets with positive sentiment. In this example, crossover is successful in the sense that the resulting instances can be clearly judged as having a positive sentiment. In other cases, crossover may fail to preserve polarity, for instance, because of an unfortunate combination involving a negation. Resulting instances may even be completely nonsensical, introducing noise to the data.

For this work, we chose to directly validate in experiments this augmentation idea. In our experiments, we applied augmentation over the training tweets after basic preprocessing and before semantic preprocessing (as defined in section 2.1). We tried different levels of augmentation, multiplying the size of original training datasets by factors of 4, 8, 12, 16 and 20. We preserved the original distribution over labels and therefore the class imbalance.

Instance crossover is a very rough and naive augmentation technique. However, it may be useful to introduce more data variability than two-way translation. With translation, new data points may fall very close to the original ones, while crossover introduces new points in the “spaces” between the original ones. Moreover, this is done in a representation agnostic fashion. It can be used with bag-of-words, embeddings, or even neural based representations.

Another clear advantage of instance crossover is that it does not rely on any external resource or system. Unlike this, translation requires an external service, at a cost, and other techniques such as synonym replacement require thesauruses or word similarity databases.

3 Experiments

In this section, we describe our experiments. We implemented all our systems using scikit-learn [9]. In the preprocessing stage, we used an NLTK-based tokenizer [2] and TreeTagger for lemmatization [10].

3.1 System Development

For simplicity, most of our work was centered on subtask 1 and on the ES dataset, looking for model configurations and hyperparameter values that gave the best

¹ Grammaticality and semantic soundness are already rare in the original tweets, so it is not something we should worry about very much.

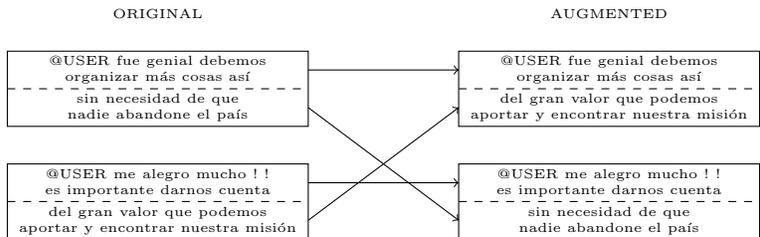


Fig. 2. Instance crossover augmentation example using two tweets with positive (P) sentiment polarity. The original tweets are on the left. The first half of one tweet is combined with the second half of the other, resulting in the new instances on the right. The dotted lines show the division in halves.

results over the development section of the ES dataset. The optimization process was done using a mixed approach of grid search and by-hand tuning.

We targeted the maximization of both macro-F1 and accuracy scores. The macro-F1 score is the main metric of TASS 2019, but it is very unstable and sensible to small changes in predictions for minority classes. On the other hand, accuracy is more stable and reliable for the development process.

As a starting point, we used our optimal model and configuration from TASS 2018 [6]. This model is a logistic regression over a combination of bag-of-words (BoW), bag-of-characters (BoC) and tweet embeddings as follows:

- Augmentation: Two-way translation with English, French, Portuguese and Arabic (EN+FR+PT+AR) as pivot languages.
- BoW: All word n-grams for $n \leq 5$.
- BoC: All character n-grams for $n \leq 6$.
- Tweet embeddings: 50 dimension fastText vectors. Weighted averaging with $\alpha = 0.1$.
- Logistic regression: liblinear solver with primal formulation, L2 regularization with inverse strength $C = 1.0$, and class-balanced reweighting.

Table 1 shows a detailed evaluation of this baseline model.

The first idea we explored was augmentation using two-way translation with the new 20 pivot languages. We tried adding all new data, but also adding some subsets of it, by grouping pivot languages in packs of four datasets. However, we could not find any improvement from the original EN+FR+PT+AR augmentation, sometimes finding important drops in model quality.

Next, we explored augmentation using instance crossover. We tried 4x, 8x, 12x, 16x and 20x factor augmentations from the original ES training corpus, with and without additional translation-based augmentation. In every case results were improved w.r.t. not using crossover augmentation. The best result was found for 8x augmentation.

Last, we tuned the hyperparameters of the logistic regression. The best configuration found was a liblinear solver with primal formulation, L2 regularization with inverse strength $C = 0.2$, with no class-balanced reweighting.

Table 1. Subtask 1: Detailed results for our baseline system on the ES development set. Left: Classification report. Right: Confusion matrix.

	Prec.	Rec.	F1		P	N	NEU	NONE
P	60.64	67.86	64.04	P	114	36	5	13
N	67.82	80.83	73.76	N	28	215	8	15
NEU	21.05	4.82	7.84	NEU	29	43	4	7
NONE	38.60	34.38	36.36	NONE	17	23	2	22
macro avg	47.03	46.97	47.00					
Accuracy:	61.10							

Table 2. Subtask 1: Detailed results for our best system on the ES development set. Left: Classification report. Right: Confusion matrix.

	Prec.	Rec.	F1		P	N	NEU	NONE
N	66.38	87.59	75.53	P	122	42	1	3
NEU	50.00	2.41	4.60	N	27	233	1	5
NONE	60.71	26.56	36.96	NEU	29	49	2	3
P	61.62	72.62	66.67	NONE	20	27	0	17
macro avg	59.68	47.30	52.77					
Accuracy:	64.37							

We also tried an ensemble of logistic regressions by using bagging. Bagging was found to be useful for the ES dataset. The best configuration found was using a bag of 40 logistic regressions.

Table 2 shows a detailed evaluation for our best model for ES found following the development process.

3.2 Subtask 1: Monolingual Experiments

To build a submission for subtask 1, we first ran the final test on the ES dataset using the best model described in the previous section.

To build submissions for the other datasets, we followed a similar development approach, but with most hyperparameters fixed with the optimal values for ES. We focused the optimization process in the usage of translation and crossover augmentations, in the logistic regression hyperparameters and in the usage of bagging. Tuning was done mostly by hand and sometimes using grid-search. Table 3 shows the optimal configurations found for each dataset.

The final results for the complete submission for subtask 1 are shown in Table 4.

3.3 Subtask 2: Crosslingual Experiments

To build submissions for subtask 2 we did a minimal set of experiments. For each language, we started from the optimal model configuration found for subtask 1 and then trained it using the union of the training datasets of every other language. We then proceeded to optimize the main hyperparameters of the logistic regression, mostly doing by-hand tuning.

We did some preliminary experiments with data augmentation and bagging for the ES dataset. However, results were not improved, so we didn't do further experimentation with the other datasets.

Table 3. Subtask 1: Best system configurations found for each dataset. Optimization was done using the development sections. Hyperparameters and their values are described in section 3.1.

	augmentation		logistic regression		
	translation	crossover	C	class-weight	bagging
ES	<small>EN+FR+PT+AR</small>	8x	0.2	no	40
PE	<small>EN+FR+PT+AR</small>	4x	0.22	balanced	no
CR	no	8x	1.15	balanced	no
UY	no	8x	0.6	no	no
MX	<small>EN+FR+PT+AR</small>	16x	0.125	balanced	no

Table 4. Subtask 1: Final results for each dataset, on development and test sections. Rank is the official final rank in the competition.

	dev		test		Rank
	Acc.	M-F1	Acc.	M-F1	
ES	64.37	52.77	60.67	48.42	2
PE	51.41	47.90	45.36	45.38	1
CR	61.28	53.36	57.20	46.91	3
UY	61.93	54.81	60.64	49.86	3
MX	67.65	53.88	68.87	48.46	4

The final results for the complete submission for subtask 2 are shown in Table 5. Results are surprisingly good, considering that we did limited experimentation because of lack of time.

3.4 Ablation Tests

As a complementary post-competition experiment, we performed ablation tests for each of the components of our systems, to assess the relevance of each of the techniques used in this work.

The ablation tests were done using the best system for subtask 1 on the ES dataset. The results are displayed in Table 6.

It can be seen that all the techniques have a positive impact. Among representations, tweet embedding is the most important representation, way above BoW and BoC representations. Also, it is interesting to observe that crossover augmentation has an impact on the F1 but not on the accuracy, indicating that it is helping mostly on the minority classes NEU and NONE.

4 Conclusions

Robust representations and data augmentation play a strong role in sentiment analysis with small-sized training datasets. As in our previous experience with TASS 2018, we are still able to obtain top ranking results without having to resort to complex models such as deep neural networks.

Table 5. Subtask 2: Final results for each dataset, on development and test sections. Rank is the official final rank in the competition.

	dev		test		Rank
	Acc.	M-F1	Acc.	M-F1	
ES	63.68	53.57	61.55	45.42	3
PE	41.97	43.71	54.64	47.42	1
CR	57.69	48.77	57.12	47.41	2
UY	60.91	53.15	61.20	51.35	1
MX	64.90	50.18	68.13	47.25	1

Table 6. Ablation tests for several techniques used in our final system. Results are for subtask 1 on the ES development set.

	Acc.	M-F1
full system	64.37	52.77
augmentation:	no translation	62.99 48.05
	no crossover	64.37 47.57
representation:	no BoW	62.99 49.89
	no BoC	62.65 50.80
	no BoW+BoC	62.13 48.26
	no embeddings	58.52 41.83
classifier:	no bagging	64.03 51.94

We observe that, for this edition of TASS, most of our work was on the application of general ML techniques, and not on particular task/domain specific engineering. In particular, we successfully tried instance crossover augmentation, a novel technique that, despite its simplicity, showed a positive impact on results. This idea can be useful to augment small datasets for other short text classification tasks without the need for external resources.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2017)
2. Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 31. Association for Computational Linguistics (2004)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Das, S.R., Chen, M.Y., Agarwal, T.V., Brooks, C., shee Chan, Y., Gibson, D., Leinweber, D., Martinez-jerez, A., Raghur, P., Rajagopalan, S., Ranade, A., Rubinstein, M., Tufano, P.: Yahoo! for amazon: Sentiment extraction from small talk on the web. In: 8th Asia Pacific Finance Association Annual Conference (2001)
5. Díaz-Galiano, M.C., et al.: Overview of TASS 2019. CEUR-WS, Bilbao, Spain (2019)
6. Luque, F.M., Pérez, J.M.: Atalaya at TASS 2018: Sentiment analysis with tweet embeddings and data augmentation. In: Martínez-Cámara, E., Almeida Cruz, Y.,

- Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J. (eds.) Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018). CEUR Workshop Proceedings, vol. 2172, pp. 29–35. CEUR-WS, Sevilla, Spain (2018)
7. Martínez-Cámara, E., Almeida Cruz, Y., Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J.: Overview of TASS 2018: Opinions, health and emotions. In: Martínez-Cámara, E., Almeida Cruz, Y., Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J. (eds.) Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018). CEUR Workshop Proceedings, vol. 2172. CEUR-WS, Sevilla, Spain (September 2018)
 8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. pp. 79–86. Association for Computational Linguistics (July 2002)
 9. 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)
 10. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: Proceedings of the ACL SIGDAT-Workshop. pp. 47–50 (1995)