TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 45-49

INGEOTEC solution for Task 1 in TASS'18 competition

Solución del grupo INGEOTEC para la tarea 1 de la competencia TASS'18

Daniela Moctezuma[1], **José Ortiz-Bejar**[3], **Eric S. Tellez**[2], **Sabino Miranda-Jiménez**[2], **Mario Graff**[2]
[1]CONACYT-CentroGEO
[2]CONACYT-INFOTEC
[3]UMSNH
dmoctezuma@centrogeo.edu.mx, jortiz@umich.mx, eric.tellez@infotec.mx,
sabino.miranda@infotec.mx, mario.graff@infotec.mx

Resumen: El análisis de sentimientos sobre redes sociales consiste en analizar mensajes publicados por usuarios de dichas redes sociales y determinar la polaridad de dichos mensajes (p.e. positivos, negativos, o una gama similar pero más amplia de dichos sentimientos). Cada lenguaje tiene características que podrían dificultar el análisis de polaridad, como la ambigüedad natural en los pronombres, la sinónimia o la polisemía; adicionalmente, dado que las redes sociales suelen ser un medio de comunicación poco formal ya que los mensajes suele tener una gran cantidad de errores y variantes léxicas que dificultan el análisis mediante enfoques tradicionales. En este artículo se presenta la participación del equipo INGEOTEC en TASS'18. Esta solución propuesta está basada en varios subsistemas orquestados mediante nuestro sistema de programación genética EvoMSA.

Palabras clave: Categorización automática de texto, programación genética, análisis de sentimientos, clasificación de polaridad

Abstract: The sentiment analysis over social networks determines the polarity of messages published by users. In this sense, a message can be classified as *positive* or *negative*, or a similar scheme using more fine-grained labels. Each language has characteristics that difficult the correct determination of the sentiment, such as the natural ambiguity of pronouns, the synonymy, and the polysemy. Additionally, given that messages in social networks are quite informal, they tend to be plagued with lexical errors and lexical variations that make difficult to determine a sentiment using traditional approaches. This paper describes our participating system in TASS'18. Our solution is composed of several subsystems independently collected and trained, combined with our EvoMSA genetic programming system.

Keywords: text categorization, genetic programming, sentiment analysis, polarity classification

1 Introduction

Sentiment Analysis is an active research area that performs the computational analysis of people's feelings or beliefs expressed in texts such as emotions, opinions, attitudes, appraisals, among others (Liu y Zhang, 2012). In social media, people share their opinions and sentiments. In addition to the inherent polarity, these feelings also have an intensity. As in previous years, TASS'18 organizes a task related to four level polarity classification in tweets. In this year, the corpus InterTASS, has been expanded with two more subsets, namely, a dataset containing tweets from Costa Rica and another one coming from Peruvian tweeters. Therefore, there are three varieties of the Spanish language, namely, Spain (ES), Peru (PE), and Costa Rica (CR). Moreover, several subtasks are also introduced:

- Subtask-1: Monolingual ES: Training and test using the InterTASS ES dataset.

ISSN 1613-0073

- Subtask-2: Monolingual PE: Training and test using the InterTASS PE dataset.
- Subtask-3: Monolingual CR: Training and test using the InterTASS CR dataset.
- Subtask-4: Cross-lingual: Here, the training can be with a specific dataset and a different one is used to test.

These subtasks are mostly based on separating language variations in train and test datasets. Martínez-Cámara et al. (Martínez-Cámara et al., 2018) detail TASS'18 Task 1 and their associated datasets.

This paper details the Task 1 solution of our INGEOTEC team. Our approach consists of a number of subsystems combined using a non-linear expression over individual predictions using our EvoMSA genetic programming system. It is worth to mention that we tackle both Task 1 (this one) and Task 4 (good or bad news) using a similar scheme, that is, the same resources and the same portfolio of algorithms, we also applied the same hyper-parameters for the algorithms; of course, we use the given task's training set to learn and optimize for each task.

The manuscript is organized as follows. Section 2 details subsystems that compose our solution. Section 3 presents our results, and finally, Section 4 summarizes and concludes this report.

2 System Description

Our participating system is a combination of several sub-systems that tackles the polarity categorization of the tweets independently, and then all these independent predictions are combined using our EvoMSA genetic programming system. The rest of this section details the use of these sub-systems and resources.

2.1 EvoMSA

EvoMSA[1] is a multilingual sentiment analysis system based on genetic text classifiers, domain-specific resources, and a genetic programming combiner of the parts. The first one, namely B4MSA (Tellez et al., 2017), performs a hyper-parameter optimization over a large search space of possible models. It uses a meta-heuristics to solve a combinatorial optimization problem over the configuration space; the selected model is described in Table 1. On the second hand, EvoDAG (Graff et al., 2016; Graff et al., 2017) is a classifier based on Genetic Programming with semantic operators which makes the final prediction through a combination of all the decision function values. The domain-specific resources can be also added under the same scheme. Figure 1 shows the architecture of EvoMSA. In the first part, a set of different classifiers are trained with datasets provided by the contests and others resources as additional knowledge, i.e., the idea is to be able to integrate any other kind of related knowledge into the model. In this case, we used tailor-made lexicons for the aggressiveness task: aggressiveness words and affective words (positive and negative), see Section 2.2 for more details. The precise configuration of our benchmarked system is described in Section 3.

Table 1: Example of set of configurations for text modeling

Text transformation	**Value**
remove diacritics	yes
remove duplicates	yes
remove punctuation	yes
emoticons	group
lowercase	yes
numbers	group
urls	group
users	group
hashtags	none
entities	none
Term weighting	
TF-IDF	yes
Entropy	no
Tokenizers	
n-words	$\{1, 2\}$
q-grams	$\{2, 3, 4\}$
skip-grams	—

2.2 Lexicon-based models

To introduce extra knowledge into our approach, we used two lexicon-based models. The first, Up-Down model produces a counting of affective words, that is, it produces two indexes for a given text: one for positive words, and another for negative words. We created the positive-negative lex-

[1] https://github.com/INGEOTEC/EvoMSA

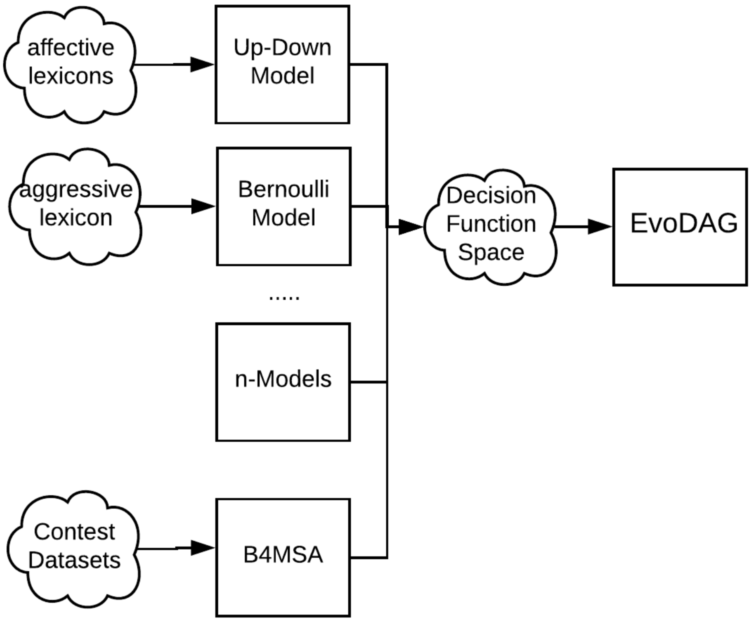

Figure 1: Architecture of our EvoMSA framework

icon based on the several Spanish affective lexicons (de Albornoz, Plaza, y Gervás, 2012; Sidorov et al., 2013; Perez-Rosas, Banea, y Mihalcea, 2012); we also enriched this lexicon with Spanish WordNet (Fernández-Montraveta, Vázquez, y Fellbaum, 2008). The other Bernoulli model was created to predict aggressiveness using a lexicon with aggressive words. We created this lexicon gathering common aggressive words for Spanish. These indexes and prediction along with B4MSA's (μTC) outputs are the input for EvoDAG system.

2.3 EvoDAG

EvoDAG[2] (Graff et al., 2016; Graff et al., 2017) is a Genetic Programming system specifically tailored to tackle classification problems on very large and high dimensional vector spaces. EvoDAG uses the principles of Darwinian evolution to create models represented as a directed acyclic graph (DAG). Due to lack of space, we refer the reader to (Graff et al., 2016) where EvoDAG is broadly described. It is important to mention that EvoDAG does not have information regarding whether input X_i comes from a particular class decision function, consequently from EvoDAG point of view all inputs are equivalent.

2.4 FastText

FastText (Joulin et al., 2017) is a tool to create text classifiers and learn a semantic vocabulary, learned from a given collection of documents; this vocabulary is represented with a collection of high dimensional vectors, one per word. It is worth to mention that FastText is robust to lexical errors since out-vocabulary words are represented as the combination of vectors of sub-words, that is, a kind of character q-grams limited in context to words. Nonetheless, the main reason of including FastText as part of our system is to overcome the small train set that comes with Task 4, which is fulfilled using the pre-trained vectors computed in the Spanish content of Wikipedia (Bojanowski et al., 2016). We use these vectors to create document vectors, one vector per document. A document vector is, roughly speaking, a linear combination of the word vectors that compose the document into a single vector of the same dimension. These document vectors were used as input to an SVM with a linear kernel, and we use the decision function as input to EvoMSA.

3 Experiments and results

The following tables show the performance of our system in the InterTASS dataset. We

[2] https://github.com/mgraffg/EvoDAG

also show the performance of a number of selected systems to provide a context for our solution. The following tables always show the top-k best results that include our system, i.e., we always show the best ones but sometimes we do not show all results below our system.

Please recall that the InterTASS dataset is split according to each sub-task. Table 2 shows the performance on monolingual datasets. For instance, the results of training with Spain-InterTASS and testing on tweets generated by people of Spain is shown in Table 2a where we reached seventh position from a total of nine participants teams. In the case training and test corpus of other Spanish varieties, in Table 2b and Table 2c show the result of training with CR and PE subsets, respectively. Our team achieved the fourth position among eight teams in CR, and the third one among eight participants. Notice that all our results are marked as bold to improve the readability.

In contrary, the results of training with the ES subset and test with subsets ES, CR, and PE are presented in Tables 3a, 3b, and 3c, respectively. Our team achieved the best result in cross-lingual task with Peruvian tweets, and also reached the second best results in ES (Spain) and CR (Costa Rica) subsets.

The performance of our method in cross lingual tasks 4 is shown in Table 3. For instance, Table 3a shows our performance on the ES subset; here, we achieved the second position among three teams. In general, the number of participants was smaller than the monolingual tasks. Table 3b show the rank of the four participant teams over the Peruvian subset of the test, here we reached the best position on the Macro-F1 score. Finally, we reached the second rank on the Costa Rica subset, just below of RETUYT-InCo.

4 Conclusions

It is worth to mention that we used the same scheme, explained in Section 2, to tackle all subtasks. Note that our EvoMSA allow to change the training set as specified for each subtasks, so we can optimize the pipeline for each particular objective.

Regarding the obtained results, our approach performs better when it is trained with tweets from Spain and test with other Spanish varieties. However, it is not clear if this performance is due to the data or a inherent feature of the Spanish variation.

Table 2: Monolingual subtasks

(a) Subtask-1, Spain dataset (ES)

Team's name	Macro-F1	Accuracy
ELiRF-UPV	0.503	0.612
RETUYT-InCo	0.499	0.549
Atalaya	0.476	0.544
UNSA_dajo	0.472	0.6
UNSA_UCSP_DaJo	0.472	0.6
MEFaMAF	0.46	0.55
INGEOTEC	**0.445**	**0.53**
ABBOT	0.409	0.482
ITAINNOVA	0.383	0.433

(b) Subtask-2, Costa Rica's dataset (CR)

Team's name	Macro-F1	Accuracy
RETUYT-InCo	0.504	0.537
ELiRF-UPV	0.482	0.561
Atalaya	0.475	0.582
INGEOTEC	**0.474**	**0.522**
MEFaMAF	0.418	0.512
ABBOT	0.408	0.46

(c) Subtask-3, Peruvian dataset (PE)

Team's name	Macro-F1	Accuracy
RETUYT-InCo	0.472	0.494
Atalaya	0.462	0.451
INGEOTEC	**0.439**	**0.447**
ELiRF-UPV	0.438	0.461
UNSA_dajo	0.413	0.319

Acknowledgements

The authors would like to thank *Laboratorio Nacional de GeoInteligencia* for partially funding this work.

References

Bojanowski, P., E. Grave, A. Joulin, y T. Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

de Albornoz, J. C., L. Plaza, y P. Gervás. 2012. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. En *Proceedings of LREC 2012*, páginas 3562–3567.

Fernández-Montraveta, A., G. Vázquez, y C. Fellbaum. 2008. The spanish version of

Table 3: Performance comparison of the cross-lingual (subtask-4) benchmark over three different test corpus.

(a) Spain's variation (ES).

Team's name	Macro-F1	Accuracy
RETUYT-InCo	0.471	0.555
INGEOTEC	**0.445**	**0.53**
Atalaya	0.441	0.485

(b) Peruvian variation (PE).

Team's name	Macro-F1	Accuracy
INGEOTEC	**0.447**	**0.506**
RETUYT-InCo	0.445	0.514
Atalaya	0.438	0.523
ITAINNOVA	0.367	0.382

(c) Costa Rica's variation (CR).

Team's name	Macro-F1	Accuracy
RETUYT-InCo	0.476	0.569
INGEOTEC	**0.454**	**0.538**
Atalaya	0.453	0.565
ITAINNOVA	0.409	0.440

wordnet 3.0. *Text Resources and Lexical Knowledge. Mouton de Gruyter*, páginas 175–182.

Graff, M., E. S. Tellez, S. Miranda-Jiménez, y H. J. Escalante. 2016. Evodag: A semantic genetic programming python library. En *2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, páginas 1–6, Nov.

Graff, M., E. S. Tellez, H. J. Escalante, y S. Miranda-Jiménez. 2017. Semantic Genetic Programming for Sentiment Analysis. En O. Schütze L. Trujillo P. Legrand, y Y. Maldonado, editores, *NEO 2015*, numero 663 en Studies in Computational Intelligence. Springer International Publishing, páginas 43–65. DOI: 10.1007/978-3-319-44003-3_2.

Joulin, A., E. Grave, P. Bojanowski, y T. Mikolov. 2017. Bag of tricks for efficient text classification. En *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, páginas 427–431. Association for Computational Linguistics, April.

Liu, B. y L. Zhang, 2012. *A Survey of Opinion Mining and Sentiment Analysis*, páginas 415–463. Springer US, Boston, MA.

Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, y J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. En E. Martínez-Cámara Y. Almeida-Cruz M. C. Díaz-Galiano S. Estévez-Velarde M. A. García-Cumbreras M. García-Vega Y. Gutiérrez A. Montejo Ráez A. Montoyo R. Muñoz A. Piad-Morffis, y J. Villena-Román, editores, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volumen 2172 de *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Perez-Rosas, V., C. Banea, y R. Mihalcea. 2012. Learning sentiment lexicons in spanish. En *LREC*, volumen 12, página 73.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, y J. Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. En *Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, páginas 1–14, Berlin, Heidelberg. Springer-Verlag.

Tellez, E. S., S. Miranda-Jiménez, M. Graff, D. Moctezuma, R. R. Suárez, y O. S. Siordia. 2017. A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters*, 94:68–74.