TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 111-115

INGEOTEC solution for Task 4 in TASS'18 competition

Solución del grupo INGEOTEC para la tarea 4 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: En este artículo se presenta un sistema de clasificación de noticias basado en clasificadores B4MSA, vectores de documentos calculados mediante vectores de palabras pre-entrenados, así como recursos especializados para detectar agresividad y afectividad en el texto. Todos estos recursos fueron entrenados de manera independiente para luego ser combinados en un modelo único mediante Programación Genética, utilizando nuestro clasificador EvoMSA. Utilizando este esquema, nuestro sistema alcanzó los mejores resultados de este año en la competencia en dos de los tres corpus de prueba.
Palabras clave: Categorización automática de texto, programación genética, clasificación automática de noticias seguras o inseguras.

Abstract: This paper describes a classification system based on our generic classifier B4MSA, sequence vectors computed with pre-trained Spanish word embeddings, and a list of specialized resources to detect aggressiveness and affectivity in text. These resources, along with the official training set, were independently trained and combined into a single model using Genetic Programming with our EvoMSA classifier. Using this approach, our system achieves the best performances, in two of three test corpus, of this year in Task 4.
Keywords: text categorization, genetic programming, safe-unsafe classification of news.

1 Introduction

News classification problem is closely related to traditional text classification applications such as topic classification (e.g., classifying a news-like text into sports, politics, or economy). Knowing any kind of categorization of news can reflect the problems of society in several domains. For instance, the discovery of negative news over time can be helpful to have a reference for leaders or decision-makers to do something about the current situation (Wang et al., 2018).

In this year, in TASS competition (Martínez-Cámara et al., 2018), a new task was proposed (Task 4), this task is related to an emotional categorization of news articles. With the purpose for knowing if a new article is SAFE or UNSAFE, a corpus was built from RSS feeds of a number of online newspapers in different varieties of Spanish (Argentina, Chile, Colombia, Cuba, Spain, USA, Mexico, Peru, and Venezuela). For the purpose of classifying these news, as SAFE or UNSAFE, the headlines were provided.

From Task 4, two sub-tasks were specified: Subtask-1 Monolingual classification and Subtask-2 Multilingual classification. The main difference between these two tasks is that, in the first case, the algorithm must be trained and tested with the same Spanish variety. In contrary, in the second case, the algorithm can be trained with a Spanish variety and tested with a different one. More information about details from Task 4 please see (Martínez-Cámara et al., 2018).

ISSN 1613-0073

In this paper, the solution proposed for the INGEOTEC team is presented. This solution is based on our B4MSA classifier and a number of specialized resources related to aggressiveness and affectivity detection. Finally, our EvoMSA classifier based on Genetic Programming is used to combine all the resources and the available training data. It is worth to mention that our scheme to create our systems for Task 1 (monolingual and cross-lingual polarity classification) and Task 4 (this one) is pretty similar; of course, we use the given task's training set to learn and optimize for each task.

The manuscript is organized as follows. In Section 2 the description of our solution is detailed. In Section 3 our results achieved in task 4 is presented. Finally, the conclusions are commented in Section 4.

2 System Description

As commented before, we use a combination of several sub-systems to tackle the *(un)safeness* categorization of the given news. Firstly, we use our generic text classifier B4MSA (Tellez et al., 2017) and a vocabulary of pre-trained vectors of FastText (Mikolov et al., 2013). Also, we use two different domain-specific lexicon resources, one of them designed to detect aggressiveness and the other one designed to detect emotions in text. All these sub-systems and resources are combined using our genetic programming scheme (EvoMSA) over the decision functions of several classifiers built on top of these resources. The rest of this section details the use of these sub-systems and resources.

2.1 EvoMSA

EvoMSA[1] has two stages. The first one, namely B4MSA (Tellez et al., 2017), uses SVMs to predict their decision function values of a given text. 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. Furthermore, EvoMSA is open to being fed with different models such as B4MSA (Tellez et al., 2018), and lexicon-based models, and EvoDAG. It is an architecture of two phases to solve classification tasks, see Figure 1. In the first part, a set of different classifiers are trained with datasets provided by the contests and others as additional knowledge, i.e., whatever knowledge could be integrated into EvoMSA. In this case, we used tailor-made lexicons for identifying aggressiveness, positiveness, and negativeness in texts, see Section 2.2 for more details. The precise configuration of our benchmarked system is described in Section 3.

2.1.1 B4MSA

B4MSA[2] (a.k.a. μTC) is a minimalistic system able to tackle general text classification tasks independently of domain and language. For complete details of the model see (Tellez et al., 2018). Roughly speaking, μTC creates text classifiers searching for the best models in given configuration space. A configuration consists of instructions to enable several preprocessing functions, a combination of tokenizers among the power set of several possible ones (character q-grams, n-word grams, and skip-grams), and a weighting scheme such as TF, TFIDF, or several distributional schemes. μTC uses an SVM (Support Vector Machine) classifier with a linear kernel. A text transformation feature could be binary options (yes/no) or ternary options (group/delete/none). Tokenizers denote how texts must be split after applying the process of each text transformation to texts, all tokens generated are part of the text representation. In Table 1, we can see details of the preprocessing, tokenizers, and term weighting scheme.

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 lexicon 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

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

[2] https://github.com/INGEOTEC/microTC

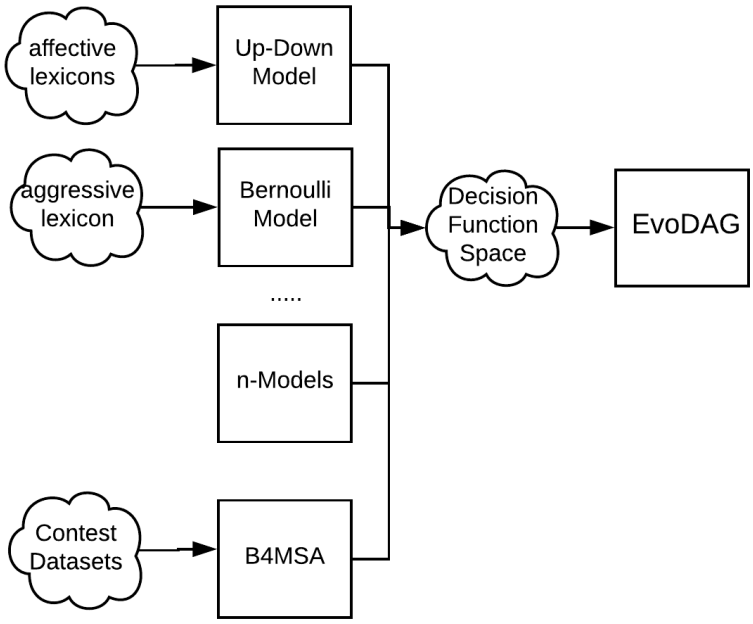

Figure 1: Architecture of our EvoMSA framework

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	—

Table 1: Example of set of configurations for text modeling

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[3] (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 in-

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

cluding 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

In order to test all the approaches in the Task-4, the SANSE (Spanish brANd Safe Emotion) corpus was established. The SANSE corpus is composed of 2,000 headlines of news written in the Spanish language along several Spanish speaker countries Spain, Mexico, Cuba, Chile, Colombia, Argentina, Venezuela, Peru, and U.S.A.

In the case of Subtask-1, Monolingual Classification, the goal was training with a Spanish variety, e.g., Mexico, and then testing with the same Spanish variety. In this case, our results and the results of the best five teams ranked by Macro-F1 metric, are presented in Table 2.

Team's name	Macro-F1	Accuracy
INGEOTEC	**0.795**	**0.802**
ELiRF-UPV	0.79	0.8
rbnUGR	0.774	0.786
MeaningCloud	0.767	0.776
SINAI	0.728	0.742
lone_wolf	0.700	0.718
TNT-UA-WFU	0.492	0.518

Table 2: Subtask-1: Monolingual Classification results: SANSE-TEST-500

Table 3 shows the best five teams in the SANSE-TEST-13152 corpus. With this corpus, our team reached the third position with a 0.866 and 0.871 of Macro-F1 and accuracy, respectively. Regarding Multilingual Classification subtask, in Table 4, all the results obtained by the best five teams, ranked by Macro-F1 metric, are reported.

In nutshell, from three datasets, our solution reached highest Macro-F1 in two corpus and middle position in the other corpus.

Team's name	Macro-F1	Accuracy
ELiRF	0.883	0.893
rbnUGR	0.873	0.888
INGEOTEC	**0.866**	**0.871**
MeaningCloud	0.793	0.801
SINAI	0.773	0.793
TNT-UA-WFU	0.544	0.552
lone_wolf	0	0

Table 3: Subtask-2: Multilingual Classification: SANSE-TEST-13152

Team's name	Macro-F1	Accuracy
INGEOTEC	**0.719**	**0.737**
ELiRF-UPV	0.699	0.722
rbnUGR	0.683	0.631
MeaningCloud	0.651	0.658
ITAINNOVA	0.617	0.575

Table 4: Subtask-2: Multilingual Classification: SANSE-408

4 Conclusions

Our solution based on Genetic Programming reached the best result in SubTask-1 Monolingual Classification SANSE-TEST-500 and SubTask-2 Multilingual Classification SANSE 408 corpus. In the largest corpus in SubTask-2 (SANSE-TEST-13152) our system reached the third best team solution.

Our approach, EvoMSA, is able to deal with several data sources through an ensemble of decision functions from each different bunch of data such as extra knowledge coded into lexicons for sentiment analysis and aggressiveness identification, and semantic information from word vectors.

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 senti-

ment analysis. En *Proceedings of LREC 2012*, páginas 3562–3567.

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

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.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

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.

Tellez, E. S., D. Moctezuma, S. Miranda-Jiménez, y M. Graff. 2018. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, 149:110–123.

Wang, B., L. Gao, T. An, M. Meng, y T. Zhang. 2018. A method of educational news classification based on emotional dictionary. En *2018 Chinese Control And Decision Conference (CCDC)*, páginas 3547–3551, June.