TASS 2017: Workshop on Semantic Analysis at SEPLN, septiembre 2017, págs. 23-28

A Genetic Programming Approach to Sentiment Analysis for Twitter: TASS'17

Un Enfoque de Programación Genética para el Análisis de Sentimiento en Twitter: TASS'17

Daniela Moctezuma[*], **Mario Graff**[†], **Sabino Miranda-Jiménez**[†], **Eric S. Tellez**[†], **Abel Coronado**[‡], **Claudia N. Sánchez**[+], **José Ortiz-Bejar**[°]
[*]dmoctezuma@centrogeo.edu.mx CONACyT-CentroGEO
[†]{mario.graff, sabino.miranda, eric.tellez}@infotec.mx CONACyT-INFOTEC;
[‡]abel@inegi.mx INFOTEC-INEGI;
[+]cnsanchez@up.edu.mx INFOTEC-Universidad Panamericana;
[⋆]job@dep.fie.umich.mx INFOTEC-UMSNH

Abstract: In this paper, we present the approach proposed by INGEOTEC team for global polarity classification at tweet level task of TASS'17 contest. We use B4MSA algorithm, a proposed entropy-based term-weighting scheme and, EvoDAG as an ensemble.
Keywords: Sentiment analysis, Opinion mining, Twitter.

Resumen: En este artículo se describe el enfoque propuesto por el equipo INGEOTEC para la tarea de clasificación global de polaridad a nivel de tweet de la competencia TASS'17. Utilizamos el algoritmo B4MSA, un nuevo esquema de pesado de términos basado en entropía y EvoDAG como un ensamble de clasificación.
Palabras clave: Análisis de sentimiento, Minería de opinión, Twitter.

1 Introduction

The rapid growth of social media such as review websites, microblogging sites, social networks, etc. has made both researchers and entrepreneurs interested in the analysis of that amount of information to create applications like sentimental analysis and opinion mining. Sentiment Analysis is used to analyze people's feelings or beliefs expressed in texts such as emotions, opinions, attitudes, etc. (Liu and Zhang, 2012). Thus, determining whether a text document has a positive, negative or neutral polarity is an essential tool for both public and private organizations.

Twitter is one of the most used social networking app, and as a result it has received a lot of attention. Twitter is considered as a huge and fast source of information (6,000 tweets each second).[1] Then, due to this important task, many international contests have been launched around the world in several languages. This is the case of the Spanish language handled in TASS (Taller de Análisis de Sentimientos en la SEPLN) contest. The TASS workshop is an event of SEPLN Conference, which is a conference in Natural Language Processing for the Spanish language.

Our participation in this contest, is mainly based on Genetic Programming (GP), which is an evolutionary algorithm that can be used to solve many different types of problems. Recently, the incorporation of semantic knowledge in GP has improved its performance. The use of efficient geometric semantic crossover operators as (Moraglio, Krawiec, and Johnson, 2012; Graff et al., 2015b; Graff et al., 2015a), combined with an optimized implementation as (Castelli et al., 2013; Graff et al., 2016) makes possible the use of GP to solve hard real problems like anti-coagulation level prediction in pharmacogenetics (Castelli et al., 2013) or sentimental analysis (Graff et al., 2017).

This paper describes the approach used in our participation in TASS'17 contest, and it is organized as follows, a brief overview of related works is shown in Section 2, the proposed methodology is described in Section 3. Section 4 shows the experimental results and analysis, and finally, Section 5 concludes.

[1]https://www.brandwatch.com/blog/44-twitter-stats-2016/

ISSN 1613-0073

2 Related work

As we know, several methods have been proposed in the community of opinion mining and sentiment analysis. A lot of these works employ Twitter as a primary source of data due to its easy and fast accessibility. Considering these advantages, most of the text polarity classification contests employed this source of data. Nevertheless, the way people write on Twitter provides a very complex task due to the tweets are full of slang and misspellings, new words are generated every day, etc. Therefore, Twitter is an easy source of acquisition, but it is very complex analyzing its content.

Tweet level and entity level polarity classification was treated in (Saif et al., 2016) using an approach based on lexicons, called SentiCircles, which creates a dynamic representation of words in order to capture their contextual semantics. Here, semantics refers to the co-occurrence patterns from each word in the text. Another approach is feature engineering, e.g. in (Ghiassi, Zimbra, and Lee, 2016) a feature engineering produced a final representation only of seven dimensions. This feature engineering was carried out in five analysis aspects: frequency, affinity, valence shifter, feature sentiment scoring and categorization. As can be seen, different types of representations or text models can be used or proposed, based on dictionaries and lexical aspects of text (Murillo and Raventós, 2016), word embeddings (Quirós, Segura-Bedmar, and Martínez, 2016), word and character n-gram (Cerón-Guzmán and de Cali, 2016), among others.

3 Proposed solution

Our participation in TASS'17 is based on an ensemble of SVM classifiers combined into a non-linear model created with Genetic Programming (GP). We used B4MSA (Tellez et al., 2017b), which is a baseline supervised learning system based on the SVM classifier, an entropy-based term-weighting scheme, and EvoDAG (Graff et al., 2016; Graff et al., 2017), a GP system that combines all decision values predicted by B4MSA systems. Figure 1 shows the architecture of our approach.

Furthermore, our approach uses two kinds of datasets; datasets labeled by human annotators provided by TASS contest, and also datasets generated by distant supervision approach.

Distant supervision has been used for tasks such as information extraction (Mintz et al., 2009), or sentiment analysis (Go, Bhayani, and Huang, 2009). In sentiment analysis, emoticons, some words, and hashtags are usually used as indicators of emotion to create automatically labeled dataset without human assistance. These new labeled datasets are expected to improve the performance of systems based on training data. We introduce a set of heuristics for distant supervision based on affective lexicons to generate labeled datasets for positive and negative sentiment.

Roughly speaking, our approach uses two layers. In the first layer, a set of B4MSA classifiers are trained with two kind of datasets; datasets labeled by human annotators: the InterTASS training set and the Spanish dataset of (Mozetič, Grčar, and Smailović, 2016), called HA dataset. We also used around 18 million tweets automatically generated by distant supervision approach, called DS dataset (for more detail see Section 4). In the case of HA datasets, each B4MSA classifier produces four real output values, one for each sentiment, that correspond to each class N, NEU, NONE, and P.

In the case of DS, around 18 million tweets are divided into chunks of 30K items (15K positive and 15K negative tweets). Each chunk produces a B4MSA model that predict a polarity level (from -1 to 1). To speed up the combination of partial results from DS, we rank the $\frac{18\times10^6}{30\times10^3}$ B4MSA classifiers with the affinity between each example and the vocabulary known by each classifier. Therefore, we select only the k classifiers with largest vocabulary intersection. The optimal k ($k = 30$) was experimentally determined.

Finally, EvoDAG's inputs are the concatenation of all the decision functions predicted by individual B4MSA classifiers. The following subsections detail the parts of our approach. The precise configuration of our benchmarked system is described in Section 4.

3.1 B4MSA

B4MSA[2] (Tellez et al., 2017b; Tellez et al., 2017a) system is our framework to create multilingual sentiment analysis systems; in particular, it produces sentiment classifiers

[2]https://github.com/INGEOTEC/b4msa

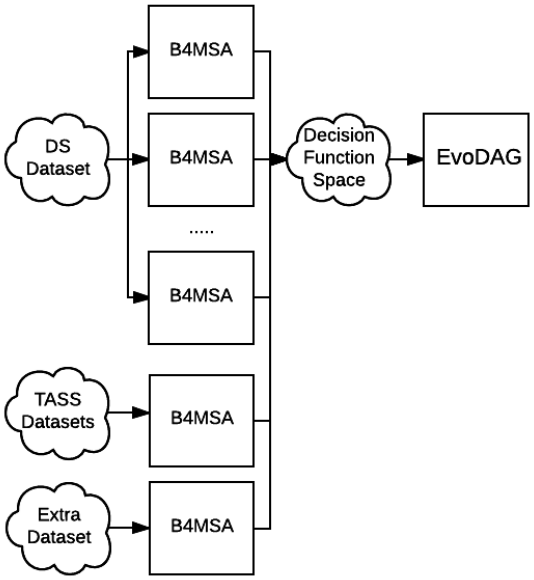

Figura 1: Prediction Scheme

that are weakly linked to language dependent methods.

In B4MSA, the whole process is stated as a combinatorial optimization problem, where the set of configurations is define by the possible solutions. In practice, finding the best text configuration for a particular problem has a high computational cost due to the large configuration space. However, a competitive solution can be found using hyper-heuristics.

Each configuration is composed by a set of text transformation functions, tokenizers (n-grams of words, q-grams of characters, and skip-grams), and a term weighting scheme. All of these parts are optimized to maximize the performance of the desired task.

In addition, we enriched the standard text transformation functions of B4MSA with the *complement of stemming*, i.e., we selected nearly word's inflections identified by a Spanish stemmer; in this sense, the clues of words (inflections) stand for the original text. The TASS datasets is transformed, tokenized, and then, the fully B4MSA pipeline is followed.

3.2 Entropy-based term-weighting

Instead of the term weighting found in B4MSA (TFIDF & TF), here we use the *entropy+b* term-weigthing scheme, firstly reported in (Eric S. Tellez, 2017).

In entropy+b each term is represented by a distribution over the available classes. Instead of using the raw probabilities per class, we weight each term with the entropy+b function, defined as follows:

$$\mathsf{entropy}_b(w) = \log|C| - \sum_{c\in C} p_c(w,b)\log\frac{1}{p_c(w,b)},$$

where C is the set of classes, and $p_c(w,b)$ is the probability of term w in class c parametrized with b. More detailed,

$$p_c(w,b) = \frac{\mathsf{freq}_c(w)}{b\cdot|C| + \sum_{c'\in C}\mathsf{freq}_{c'}(w)}.$$

Here, freq_c denotes the frequency of the given term in the class c. The idea behind $\mathsf{entropy}_b(w)$ is to weight each term using the entropy of the underlying distribution, that is, large entropy values (terms uniformly distributed along all classes) have a low weight while terms being skewed to some class are close to $\log|C|$. The parameter b is introduced to *absorb* the possible *noise* that occurs in low populated terms.

3.3 EvoDAG

EvoDAG[3] (Graff et al., 2016; Graff et al., 2017) is a Genetic Programming system specifically tailored to tackle supervised classification and regression problems on very high dimensional vector spaces and large datasets. In particular, EvoDAG uses the principles of Darwinian evolution to create models represented as a directed acyclic graph (DAG). EvoDAG evolves the solution using either steady-state or generational evolution, with a tournament selection of size two, the fitness function is set to be the balance error rate which is equivalent to macro-recall, and has as many outputs as classes (for a more detail description, we reader to (Graff et al., 2016)).

The models evolved by EvoDAG have three distinct node's types; the inputs nodes, that, as expected, received the independent variables, the output node that corresponds to the label, and the inner nodes are the different numerical functions such as: sum, product, sin, cos, max, and min, among others. In order to provide an idea of the type of models being evolved, Figure 2 depicts a model evolved for the polarity classification at global level task.

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

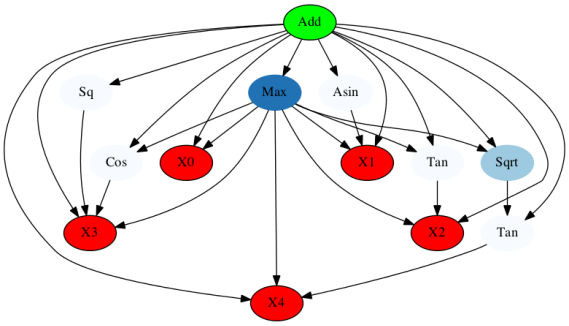

Figura 2: An evolved model for the polarity classification at global level task

As it can be seen, the model is represented using a directed acyclic graph (DAG) where direction of the edges indicate the dependency, e.g., cos depends on X_3, i.e., cosine function is applied to X_3. There are three types of nodes; the inputs nodes are colored in red, the inner nodes are blue (the intensity is related to the distance to the height, the darker the closer), and the green node is the output node. As mentioned previously, EvoDAG uses as inputs the decision functions of B4MSA, then first three inputs (i.e., X_0, X_1, and X_2) correspond to the decision function values of the negative, neutral, none, and positive polarity of B4MSA model trained with InterTASS dataset, and the later two (i.e., X_3 and X_4) correspond to the decision function values of two B4MSA systems each one trained with our distant supervision dataset.

It is important to mention that EvoDAG does not have information regarding whether input X_i comes from a particular polarity decision function, consequently from EvoDAG point of view all inputs are equivalent.

4 Experimental results

Table 1 shows the data distribution of training and test datasets used in our experiments. For training step, we used the datasets provided by the organizers (training set of TASS'16 and training set of InterTASS'17), extra data annotated by humans described in (Mozetič, Grčar, and Smailović, 2016)[4] In addition, we created a corpus using distant supervision approach using words from Spanish affective lexicons (Perez-Rosas, Banea, and Mihalcea, 2012; Sidorov et al., 2013).

In case of distant supervision (DS) dataset, around 18 million tweets were selected from more than 500 million tweets collected along one year. The tweets were classified into two classes positive or negative based on the words of the affective lexicons. Each tweet has no contradictions, i.e., the tweet has only positive or negative words, tweets with negative markers or some discourse markers (*no, aunque, sin embargo*) were avoided to ensure the class.

DataSet	Positive	Negative	Neutral	None	Total
train-TASS'16	2884	2182	670	1483	7219
train-TASS'17	473	635	202	201	1511
Extra-data	69,571	16,472	54,017	-	140,060
DS-dataset	9M	9M	-	-	18M

Tabla 1: Statistics of Spanish training data.

Regarding the parameters used by B4MSA and EvoDAG, it is important to mention that these parameters where optimized using random search an a hill climbing technique (only used by B4MSA) in the parameter search space. Specifically, we follow the instructions provided by these developments which consist in: firstly, optimize the algorithm's parameters using the training set, secondly, train the model with the parameters obtained in the previous step; and, finally, use the model to predict the data given. It is important to mention that B4MSA and EvoDAG can be used from command line and we decided to follow that path.

Table 2 shows our results on gold standard of TASS'17, namely, *InterTASS, General Corpus 1K* and *General Corpus 60K*. To solve the task with these three test sets we use two models, both using all datasets listed in Table 1. For the InterTASS subtask, EvoDAG creates a model that optimizes the combination of the internal classifiers. For this purpose, we use the provided train and validation partition, the first one to train and the later to compute the objective function; more detailed, EvoDAG uses the geometric mean per-class of F_1 scores as objective function. On the other hand, the Global corpus (both 1K and 60K) is optimized to maximize the geometric mean per-class of F_1 scores, over TASS'16.

5 Conclusions

In this paper, we presented the system used to tackle the task of global polarity classification at tweet level, in Spanish. From the results, it is observed that our system,

[4]The datasets are available at `http://hdl.handle.net/11356/1054`.

Corpus	Macro-P	Macro-R	Macro-F1	Accuracy
InterTASS	0.459	0.455	0.457	0.507
General Corpus 1K	0.501	0.553	0.526	0.595
General Corpus 60K	0.559	0.595	0.577	0.645

Tabla 2: Results of INGEOTEC participation in TASS'17

which combined B4MSA algorithm for text representation, a new entropy-based term-weighting scheme, and EvoDAG as an ensemble reaches good performance achieving high positions in the three corpora provided in TASS'17.

Acknowledgments

This research is partially supported by the Cátedras CONACyT project.

References

Castelli, Mauro, Davide Castaldi, Ilaria Giordani, Sara Silva, Leonardo Vanneschi, Francesco Archetti, and Daniele Maccagnola. 2013. An efficient implementation of geometric semantic genetic programming for anticoagulation level prediction in pharmacogenetics. In *Portuguese Conference on Artificial Intelligence*, pages 78–89. Springer.

Cerón-Guzmán, Jhon Adrián and Santiago de Cali. 2016. Jacerong at tass 2016: An ensemble classifier for sentiment analysis of spanish tweets at global level. In *TASS@ SEPLN*, pages 35–39.

Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff Daniela Moctezuma. 2017. Gender and language-variety identification with microtc. *Working Notes of CLEF 2017*, 1866.

Ghiassi, Manoochehr, David Zimbra, and Sean Lee. 2016. Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems*, 33(4):1034–1058.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

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

Graff, Mario, Eric S. Tellez, Hugo Jair Escalante, and Sabino Miranda-Jiménez. 2017. Semantic Genetic Programming for Sentiment Analysis. In Oliver Schütze, Leonardo Trujillo, Pierrick Legrand, and Yazmin Maldonado, editors, *NEO 2015*, number 663 in Studies in Computational Intelligence. Springer International Publishing, pages 43–65. DOI: 10.1007/978-3-319-44003-3_2.

Graff, Mario, Eric S Tellez, Hugo Jair Escalante, and Jose Ortiz-Bejar. 2015a. Memetic genetic programming based on orthogonal projections in the phenotype space. In *2015 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, pages 1–6. IEEE.

Graff, Mario, Eric Sadit Tellez, Elio Villasenor, and Sabino Miranda-Jiménez. 2015b. Semantic genetic programming operators based on projections in the phenotype space. *Research in Computing Science*, 94:73–85.

Liu, Bing and Lei Zhang, 2012. *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Moraglio, Alberto, Krzysztof Krawiec, and Colin G. Johnson, 2012. *Geometric Semantic Genetic Programming*, pages 21–31. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mozetič, Igor, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Murillo, Edgar Casasola and Gabriela Marín Raventós. 2016. Evaluación de modelos de representación del texto con vectores de dimensiónn reducida para análisis de sentimiento. In *TASS@ SEPLN*, pages 23–28.

Perez-Rosas, Veronica, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3077–3081, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1645.

Quirós, Antonio, Isabel Segura-Bedmar, and Paloma Martínez. 2016. Labda at the 2016 tass challenge task: Using word embeddings for the sentiment analysis task. In *TASS@ SEPLN*, pages 29–33.

Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. *Information Processing and Management*, 52(1):5 – 19.

Sidorov, Grigori, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, pages 1–14, Berlin, Heidelberg. Springer-Verlag.

Tellez, Eric S., Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, and Elio A. Villaseñor. 2017a. A case study of spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, 81:457 – 471.

Tellez, Eric S., Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart R. Suárez, and Oscar S. Siordia. 2017b. A simple approach to multilingual polarity classification in twitter. *Pattern Recognition Letters*, 94:68 – 74.