INGEOTEC at SemEval-2018 Task 1: EvoMSA and μTC for Sentiment Analysis

Mario Graff and **Sabino Miranda-Jiménez*** and **Eric S. Tellez**
CONACyT - INFOTEC, Aguascalientes, México
{mario.graff,sabino.miranda,eric.tellez}@infotec.mx

Daniela Moctezuma
CONACyT - CentroGEO, Aguascalientes, México
dmoctezuma@centrogeo.edu.mx

Abstract

This paper describes our participation in Affective Tweets task for emotional intensity and sentiment intensity subtasks for English, Spanish, and Arabic languages. We used two approaches, μTC and EvoMSA. The first one is a generic text categorization and regression system; and the second one is a two-stage architecture for Sentiment Analysis. Both approaches are multilingual and domain independent.

1 Introduction

Sentiment Analysis is a research area where does a computational analysis of people's feelings or beliefs expressed in texts such as emotions, opinions, attitudes, appraisals, etc. (Liu and Zhang, 2012). People communicate not only the emotion or sentiment they are feeling, but also the intensity, that is, the degree of emotion or sentiment. In this context, SemEval is one of the forums that conducts evaluations on semantics at different levels, for instance, it proposes tasks such as sentiment analysis, the intensity of emotion or sentiment (affective tweets) (Mohammad et al., 2018), irony detection, among others (SemEval, 2017).

In this work, we present the results of our participation in Affective Tweets task for four of the five subtasks in English, Spanish, and Arabic languages and for all emotions available: anger, fear, joy, and sadness.

The subtasks are A) emotion intensity regression (EI-REG): given a tweet and an emotion, determine the intensity of the emotion that best represents the mental state of the tweeter, a real-value score between 0 and 1.

B) Emotion intensity ordinal classification (EI-OC): given a tweet and an emotion E, classify the tweet into one of four ordinal classes of intensity of emotion: anger, fear, joy, and sadness, that best represents the mental state of the tweeter.

C) A sentiment intensity regression task (V-REG): given a tweet, determine the intensity of sentiment, a real-valued score between 0 (most negative) and 1 (most positive).

D) A sentiment analysis, ordinal classification (V-OC): given a tweet, classify it into one of seven ordinal classes, corresponding to several levels of positive and negative sentiment intensity.

In this context, one crucial step is the procedure used to transform the data (i.e., tweets) into the inputs (vectors) of the supervised learning techniques used. Typically, Natural Language Processing (NLP) approaches for data representation use n-grams of words, linguistic information such as dependency relations, syntactic information, lexical units (e.g., lemmas, stems), affective lexicons, error correction, etc. However, selecting the best configuration of those characteristics could be a cumbersome task, many times disregarded in favor of some *well-known* competitive setups. (Tellez et al., 2017b) studies the dependency between the performance and the proper selection of the text model. This selection can be seen as a combinatorial optimization problem where the objective is to maximize the performance metric of the classifier being used; this approach is implemented by μTC, (Tellez et al., 2018). Due to its combinatorial nature, and the kind of parameters that compose the configuration space, the resulting classifiers are multilingual and domain independent. Therefore, with a tight dependency on the training set, it is mandatory to provide additional information about the particular task to avoid overfitting. In this sense, the use of multiple knowledge sources is essential, and combining them simply and effectively is the idea be-

*corresponding author: sabino.miranda@infotec.mx

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 146–150
New Orleans, Louisiana, June 5–6, 2018.

hind EvoMSA. EvoMSA (§2.2) is a stacking system based on genetic programming, and particularly on the use of semantic genetic operators, that focus on sentiment analysis. The core of our contribution is to use both μTC and EvoMSA to learn from different annotated collections and then use that diverse knowledge to tackle the SemEval 2018 Task 1 challenge.

Looking at systems that obtained the best results in previous SemEval editions, it can be concluded that it is necessary to include more datasets, see for instance BB_twtr system (Cliche, 2017) for Sentiment Analysis in the Twitter task, which uses more datasets besides the one given in the competition. Here, it was decided to follow a similar approach by including an additional human-annotated dataset publicly available for English, Spanish, and Arabic to build robust models.

2 System Description

As commented, we use two systems to evaluate the Affective Tweets task: μTC and EvoMSA. On the one hand, μTC is used mainly to evaluate two tasks for the Arabic language because in our experiments it obtained the best performance in almost all subtask in this language both for regression and classification tasks. On the other hand, EvoMSA is used to evaluate English and Spanish languages, and ordinal sentiment classification (valence) task for Arabic. In the following paragraphs, we describe these approaches.

2.1 μTC

μTC[1] is a minimalistic and wide system able to tackle text classification and regression tasks independent of domain and language a detail. For complete details of the model see (Tellez et al., 2018). Essentially, μTC creates text classifiers (or a text regressors) searching for the best models in a 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, skip-grams, etc.), and a weighting scheme (application of frequency filters and the use of TF, TFIDF, or several distributional schemes). μTC seeks the best configurations optimizing a score which is evaluated through a classifier or a regressor; currently, it uses SVM for both tasks. In Table 1, we can see details of text transformations used in our solution for detecting *Anger* emotion for Arabic. This set of text transformations was selected among millions of possible configurations through the combinatorial optimization process implemented in μTC. In ordinal classification tasks the model is found out based on the training dataset provided for each emotion, if this is the case.

2.2 EvoMSA

EvoMSA[2] is a Sentiment Analysis System based on B4MSA and EvoDAG. It is an architecture of two phases to solve classification or regression tasks, see Figure 1. EvoMSA improves the performance of a global classifier combining the predictions of a set of classifiers with different models on the same text to be classified. Roughly speaking, in the first stage, a set of B4MSA classifiers (see Sec. 2.2.1) are trained with two kind of datasets; datasets provided by SemEval, and large datasets annotated by humans for sentiment analysis for English and Spanish languages (Mozetič et al., 2016), called HA datasets. In the case of HA datasets, it is split into balanced small datasets that feed each B4MSA classifier which produces three real output values, one for each sentiment (negative, neutral and positive). In the case of SemEval datasets, for instance, for EI-OC, the classifier produces one of four ordinal classes of intensity of emotion (0, 1, 2, 3). It creates a decision functions space with mixtures of values coming from different views of knowledge. Finally, EvoDAG's inputs are the concatenation of all the decision functions predicted by each B4MSA system, and EvoDAG produces a final value or prediction. The following subsections describe the internal parts of EvoMSA. The precise configuration of our benchmarked system is described in Sec. 4.

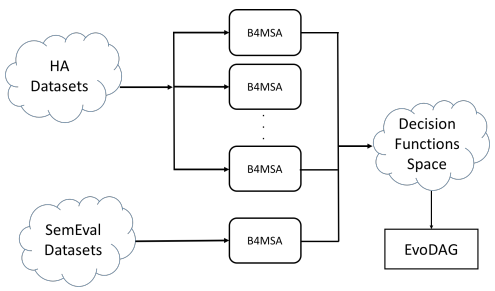

Figure 1: EvoMSA Architecture

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

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

2.2.1 B4MSA

B4MSA[3] is related to μTC, but this framework is mainly focused for multilingual sentiment analysis. For complete details of the model see (Tellez et al., 2017a,b).

The core idea behind B4MSA is similar to that of μTC, i.e., it tackles the sentiment analysis problem as a model selection problem, yet using a different view of the underlying combinatorial problem. Also, contrarily to μTC, B4MSA takes advantage of several domain-specific particularities like emojis and emoticons and makes explicit handling of negation statements expressed in texts. Nonetheless, EvoMSA avoids the sophisticated use of B4MSA fixing the model for each language in favor of performing an optimization process at the level of the decision functions of several models. Table 1 shows text transformation parameters used in our system for English and Spanish languages.

2.2.2 EvoDAG

EvoDAG[4] (Graff et al., 2016, 2017) is a Genetic Programming system specifically tailored to tackle 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). An EvoDAG model has 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. Due to lack of space, we refer the reader to (Graff et al., 2016) where EvoDAG is broadly described. In fact, in this research, we followed the steps explained there. In order to give an idea of the type of models being evolved, Figure 2 depicts a model evolved for the Arabic polarity classification at global message task. As can be seen, the model is represented using a DAG where direction of the edges indicates the dependency, e.g., cos depends on X_3, i.e., cosine function is applied to X_3. As commented above, 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 the first three inputs (i.e., X_0, X_1, and X_2) correspond to the decision functions values of the negative, neutral, and positive polarity of B4MSA model trained with SemEval Arabic 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 other dataset for two classes. 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.

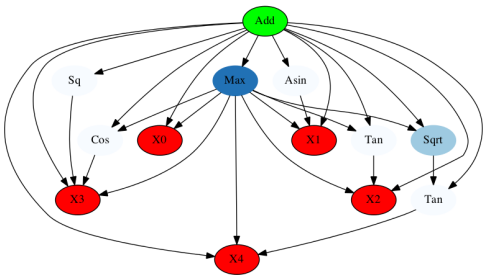

Figure 2: An evolved model for the Arabic task.

3 Experimental Settings

As we mentioned, to determine the best configuration of parameters for text modeling, μTC and B4MSA integrate a hyper-parameter optimization phase that ensures the performance of the classifier based on the training data. The text modeling parameters for B4MSA were set for all process as we show in Table 1 for English and Spanish language for classification and regression tasks. In the case of the Arabic language, the parameters were calculated by the optimization phase; an example is showed in Table 1. A text transformation feature could be binary (yes/no) or ternary (group/delete/none) option. Tokenizers denote how texts must be split after applying the process of each text transformation to texts. Tokenizers generate text chunks in a range of lengths, all tokens generated are part of the text representation. Both, B4MSA and μTC, allow selecting tokenizers based on n-words, $q-$grams, and skip-grams, in any combination. We call n-words to the well-known word n-grams; in particular, we allow to use any combination of unigrams, bigrams, and trigrams. Also, the configuration space allows selecting any combination of character q-grams (or just q-grams) for $q = 1$ to 9. Finally, we allow to

[3] https://github.com/INGEOTEC/b4msa
[4] https://github.com/mgraffg/EvoDAG

use $(2, 1)$ and $(3, 1)$ skip-grams (two words separated by one word, and three words separated by a gap).

Table 1 shows the final configurations for English and Spanish and an example for one emotion for Arabic. For example, *numbers* are deleted in Arabic, but it is grouped in English and Spanish. In the case of English, it is split in unigrams, bigrams, character q-grams of sizes 2, 3, and 4.

Text transformation	English	Spanish	Arabic
remove diacritics	yes	yes	yes
remove duplicates	yes	yes	yes
remove punctuation	yes	yes	yes
emoticons	group	group	group
lowercase	yes	yes	false
numbers	group	group	delete
urls	group	group	group
users	group	group	none
hashtags	none	none	none
entities	none	none	none
Term weighting			
TF-IDF	yes	yes	no
Entropy	no	no	yes
Tokenizers			
n-words	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$
q-grams	$\{2, 3, 4\}$	$\{2, 3, 4\}$	$\{2, 3, 7, 9\}$
skip-grams	—	—	—

Table 1: Example of set of configurations for text modeling

3.1 Datasets

SemEval provides datasets to train systems for each subtask. For instance, for emotion Anger in English, subtask emotion intensity ordinal classification, OC, the training data is distributed for four classes (class 0 = 445, class 1 = 322, class 2 = 507, class 3 = 427). The Arabic datasets for each emotion have around 800 samples each one, for English the sizes are between 1500 and 2200 samples, and for Spanish are between 1000 and 1150 samples, for more details of the data distribution and how the datasets were built we refer the reader to (Mohammad et al., 2018; Mohammad and Kiritchenko, 2018). In addition of SemEval data, we use extra datasets annotated by humans around 73 thousand tweets for English, 223 thousand for Spanish (Mozetič et al., 2016), and two thousand for Arabic (NRC, 2017). Table 2 shows the distribution of classes for datasets. Those datasets are mainly used for sentiment analysis; however, we use this extra information to improve the final decision in the approach we implemented (EvoMSA).

HA-DataSet	Positive	Neutral	Negative	Total
English	21,166	33,620	18,454	73,240
Spanish	107,252	89,782	26,272	223,306
Arabic	448	202	1,350	2,000

Table 2: Statistics of Human-Annotated training data. We used the labeled English and Spanish data from (Mozetič et al., 2016), and the Arabic data from (NRC, 2017).

4 Results

We present the results of our approaches in Table 3 and Table 4. All experiments were tested on the development dataset provided by SemEval. In the case of OC tasks, we use the macro-F1 score to measure the performance, and in the case of Reg tasks, we use the Pearson correlation coefficient. Table 3 shows the results of emotional intensity for ordinal classification (OC) and regression tasks (Reg) grouped by each emotion and language. Table 4 shows the results of sentiment analysis, ordinal classification task (V-OC) and sentiment intensity regression task (V-Reg) group by each emotion and language. We present three system configurations in Table 3 and Table 4. EvoMSA configuration uses only the training datasets provided by SemEval, and it is used as regressor or classification system. In addition of SemEval data, EvoMSA-HA uses extra information comes from sentiment analysis domain, and this information improves the performance as we can see. And μTC uses only the training data provided by the contest as the knowledge base to calculate the final class or real value. As we can see in Table 3, the best performance obtained are grouped by EvoMSA-HA configuration for both OC and Reg tasks for English and Spanish languages. For the Arabic language, μTC is quite good with OC and Reg task. According to the results we obtained, we decided to use for the evaluation phase the following configuration: EvoMSA-HA is used for OC, Reg, V-OC, and V-Reg tasks for English and Spanish; also for OC (Fear and Joy) and V-OC tasks for Arabic; and μTC is used for Arabic in OC (Anger and Sadness), Reg, and V-Reg tasks. In the table, the performance of our configuration systems, on gold standard, is labeled by subscripts; they stand for the rank in the general evaluation. For example, for Spanish in OC task, we were ranked for Anger emotion in position 4; Fear, position 2; Joy, position 3; and Sadness, position 2.

Configuration	Anger	Fear	Joy	Sadness
English				
(OC) EvoMSA	0.3938	0.3820	**0.3983**	0.4249
(OC) EvoMSA-HA	**0.4188**	**0.4187**	0.3977	**0.4389**
(OC) μTC	0.3300	0.4120	0.3167	0.3908
(Reg) EvoMSA	0.4948	0.4758	0.5371	0.5714
(Reg) EvoMSA-HA	**0.5756**	**0.5380**	**0.6249**	**0.6105**
(Reg) μTC	0.3301	0.5158	0.5042	0.5087
Performance on gold standard				
(OC) Our Approach	$0.560_{(14)}$	$0.489_{(15)}$	$0.643_{(9)}$	$0.584_{(13)}$
(Reg) Our Approach	$0.643_{(26)}$	$0.621_{(29)}$	$0.684_{(20)}$	$0.626_{(28)}$
Spanish				
(OC) EvoMSA	0.4210	**0.5013**	0.4811	0.4419
(OC) EvoMSA-HA	**0.4405**	0.5006	**0.5275**	**0.4835**
(OC) μTC	0.3741	0.4070	0.4353	0.3757
(Reg) EvoMSA	**0.5487**	**0.7338**	0.7051	**0.5965**
(Reg) EvoMSA-HA	0.4990	0.7265	**0.7129**	0.5941
(Reg) μTC	0.5241	0.6568	0.4897	0.5693
Performance on gold standard				
(OC) Our Approach	$0.468_{(4)}$	$0.634_{(2)}$	$0.655_{(3)}$	$0.628_{(2)}$
(Reg) Our Approach	$0.543_{(4)}$	$0.675_{(4)}$	$0.682_{(3)}$	$0.633_{(5)}$
Arabic				
(OC) EvoMSA	0.4062	**0.3721**	0.3688	0.4039
(OC) EvoMSA-HA	0.3805	0.3620	**0.3768**	0.3637
(OC) μTC	**0.4182**	0.3092	0.3347	**0.4689**
(Reg) EvoMSA	0.3661	0.2770	0.3782	0.5142
(Reg) EvoMSA-HA	0.2118	0.1117	**0.4279**	0.5952
(Reg) μTC	**0.4700**	**0.5011**	0.4090	**0.6191**
Performance on gold standard				
(OC) Our Approach	$0.387_{(4)}$	$0.440_{(4)}$	$0.498_{(4)}$	$0.425_{(6)}$
(Reg) Our Approach	$0.501_{(5)}$	$0.501_{(6)}$	$0.628_{(5)}$	$0.537_{(6)}$

Table 3: Results for Emotion Intensity: Ordinal Classification (OC) and Regression (Reg), in terms of macro-F1 (OC) and Pearson correlation coefficient (Reg).

Configuration	English	Spanish	Arabic
(V-OC) EvoMSA	0.3148	0.3367	**0.3304**
(V-OC) EvoMSA-HA	**0.3430**	**0.3902**	0.3251
(V-OC) μTC	0.2848	0.3418	0.2671
(V-Reg) EvoMSA	0.5993	0.6571	0.2977
(V-Reg) EvoMSA-HA	**0.6213**	**0.6693**	0.0045
(V-Reg) μTC	0.3440	0.5834	**0.6263**
Performance on gold standard			
(V-OC) Our Approach	$0.760_{(11)}$	$0.749_{(3)}$	$0.698_{(4)}$
(V-Reg) Our Approach	$0.761_{(24)}$	$0.701_{(5)}$	$0.746_{(5)}$

Table 4: Results for Valence: Ordinal Classification (OC) and Regression (Reg), in terms of macro-F1 (OC) and Pearson correlation coefficient (Reg).

5 Conclusions

In this paper was presented our solution for Affective Tweets task combining two approaches EvoMSA and μTC. Both systems are designed to be multilingual and language and domain independent as much as possible. For the training step, we use extra human annotated datasets out of any specific emotion, but related to sentiment-analysis information; our solution performs well in Spanish and Arabic languages; however, there is room for further improvements in performance for tasks in English language using another sort of knowledge such as semantic information (word embeddings) into EvoMSA architecture.

References

Mathieu Cliche. 2017. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 573–580.

M. Graff, 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.

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

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

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

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

NRC. 2017. Syrian tweets arabic sentiment analysis dataset. `http://saifmohammad.com/WebPages/ArabicSA.html`. Accessed 17-Feb-2017.

SemEval. 2017. Semeval-2017: Sentiment analysis task 4. `http://alt.qcri.org/semeval2017/task4/`. Accessed 17-Feb-2017.

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

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

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda-Jiménez, and Mario Graff. 2018. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, 149:110–123.