INGEOTEC at SemEval 2017 Task 4: A B4MSA Ensemble based on Genetic Programming for Twitter Sentiment Analysis

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

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

Abstract

This paper describes the system used in SemEval-2017 Task 4 (Subtask A): Message Polarity Classification for both English and Arabic languages. Our proposed system is an ensemble of two layers, the first one uses our generic framework for multilingual polarity classification (B4MSA) and the second layer combines all the decision function values predicted by B4MSA systems using a non-linear function evolved using a Genetic Programming system, EvoDAG. With this approach, the best performances reached by our system were macro-recall 0.68 (English) and 0.477 (Arabic) which set us in sixth and fourth positions in the results table, respectively.

1 Introduction

Sentiment Analysis is the computational analysis of people's feelings or beliefs expressed in texts such as emotions, opinions, attitudes, appraisals, etc. (Liu and Zhang, 2012). At the same time, with the growth of social media (review websites, microblogging sites, etc.) on the Web, Twitter has received particular attention because it is a huge source of opinionated information (6,000 tweets each second) [1], and has potential uses for decision-making tasks from business applications to political campaigns.

In this context, SemEval is one of the forums that conducts evaluations of Sentiment Analysis Systems on Twitter at different levels such as polarity classification at global or topic-based message, tweet quantifications, among other tasks (Nakov et al., 2016; SemEval, 2017).

In this research, the sentiment analysis task is faced as a classification problem, thus supervised learning techniques are used to tackle this problem. Particularly, we used Support Vector Machines (SVM) and a Genetic Programming system called EvoDAG (Graff et al., 2016, 2017).

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, etc.; however, selecting the best configuration of those characteristics could be a huge problem. In fact, this selection can be seen as a combinatorial optimization problem where the objective is to improve the accuracy (or any performance measure) of the classifier being used. The proposed system uses our generic framework for multilingual polarity classification (B4MSA) (Tellez et al., 2016) to transform the data into the inputs of an SVM. Furthermore, B4MSA uses random search and hill climbing to find a suitable text transformation pipeline among the possible ones.

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 SwissCheese system (Deriu et al., 2016), besides the one given in the competition. Here, it was decided to follow this approach by including an extra dataset for English, and a number of datasets automatically labeled using a distant supervision approach in both languages, English and Arabic. Regarding this point, it was observed that it is important to have a good balance between quality and amount of samples. We take care of this issue by removing the repeated samples in our training set and at the same time using a lot of sam-

*corresponding author: mario.graff@infotec.mx

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

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 771–776,
Vancouver, Canada, August 3 - 4, 2017.

ples.

In this paper, we describe our classification system used in SemEval-2017 contest for Task 4, subtask A: polarity classification at global message. This task consists in classifying given a tweet whether is positive, negative, or neutral sentiment according to its content. Our system was evaluated on the English and Arabic languages.

2 System Description

Our framework comprises two subsystems: B4MSA (Tellez et al., 2016), which is a supervised learning system based on SVM; and EvoDAG (Graff et al., 2016, 2017) that acts as integrator of agreements among the decision functions values predicted by a set of B4MSA systems. Figure 1 shows the architecture of our approach. The basic idea of this framework is to make maximum use of synergies between both approaches B4MSA and EvoDAG.

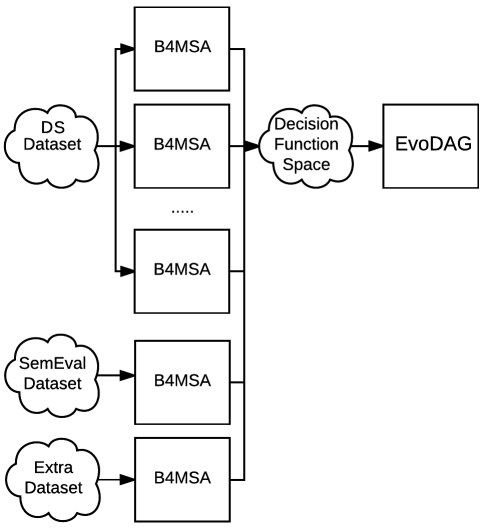

Figure 1: Prediction Scheme

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: SemEval datasets from 2013-2016 and the English dataset of (Mozetič et al., 2016), called HL dataset, and also datasets generated by distant supervision approach, called DS dataset, (see section 3.2). In case of HL datasets, each B4MSA classifier produces three real output values, one for each sentiment (negative, neutral and positive).

In the case of DS, the entire collection is divided into a number of disjoint parts to obtain a number of individually trained B4MSA classifiers. Each B4MSA is only trained to predict if a tweet is positive or negative, based on the distant supervision procedure described in §3.2. Since there are only two classes, then each classifier produces a real output. To improve the classification performance, we fix the size of the parts to contain 30K tweets (positive and negative) for large datasets. Due to the large number of parts for very large DS collections, we take into account just a few classifiers, k, in the decision process. To select the k best classifiers, we define a vocabulary-affinity measure that scores what a classifier knows about the vocabulary (content) of a tweet to be classified. All B4MSA classifiers compute its vocabulary affinity; then, the top k classifiers are selected dynamically for each tweet according to its content using the vocabulary-affinity measure. The optimal k should be experimentally determined.

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

2.1 B4MSA

B4MSA[2] (Tellez et al., 2016, 2017) is a framework to create multilingual sentiment analysis systems; in particular, it produces sentiment classifiers that are weakly linked to language dependent methods. For instance, B4MSA avoids the usage of computational expensive linguistic tasks such as lemmatization, stemming, part-of-speech tagging, etc., and take advantage of data representations, mostly based on simple text transformations and a number of text tokenizers.

The core idea of B4MSA is to determine automatically the best text transformation pipeline along with the best performing set of tokenizers, given a large set of possible configurations. In B4MSA, the whole process is stated as a combinatorial optimization problem, where the set of configurations define the possible solutions. In practice, the best text configuration for a particular problem has a high computational cost to evaluate each configuration, due to the large configuration space; however, a competitive solution can be found using hyper-heuristics.

We use a plain B4MSA setup, see Table 1 for details of text transformations used in our sys-

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

tem. This set of text transformations was selected among millions of possible configurations through the combinatorial optimization solution implemented in B4MSA.

2.2 EvoDAG

EvoDAG[3] (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 described, and, we followed, in this research, the steps mentioned there.

In order to provide 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. 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, 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 our DS 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.

3 Data Preparation

To determine the best configuration of parameters for text modeling, B4MSA integrates a hyper-parameter optimization phase that ensures the performance of the sentiment classifier based on the training data. The text modeling parameters determined for the English and Arabic languages related to text transformations, weighting scheme, and tokenizers are described 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. B4MSA allows 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 $q = 3, 5, 7$, and 9. Finally, we allow to use $(2, 1)$ and $(3, 1)$ skip-grams (two words separated by one word, and three words separated by a gap).

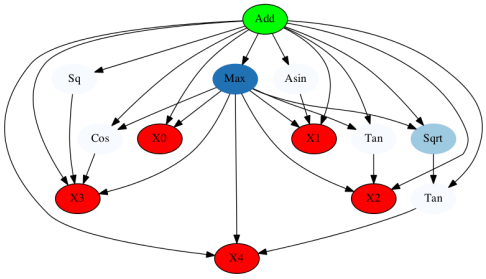

Figure 2: An evolved model for the arabic task.

Our parameter set contains five binary features, four ternaries ones, and nine individual tokenizers; thus the configuration space contains $5^2 \times 4^3 \times (2^9 - 1) = 817,600$ different items. For instance, for the English dataset, a commodity workstation needs close to ten minutes to evaluate each configuration,[4] such that an exhaustive evaluation of the configuration space will take 15 years. We use a number of hyper-heuristics to find a competitive model in a few hours, the interested reader on the optimization process is referenced to (Tellez et al., 2016, 2017).

Table 1 shows the final configurations for each language, for example, *remove diacritics* is not applied to English, but it is applied to Arabic. Although *lowercase* transformation is weird for

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

[4] Only human labeled data, this time does not apply to the distant supervision dataset.

Text transformation	English	Arabic
remove diacritics	no	yes
remove duplicates	no	yes
remove punctuation	no	yes
emoticons	none	delete
lowercase	yes	yes
numbers	group	group
urls	group	delete
users	delete	none
Term weighting		
TF-IDF	yes	yes
Tokenizers		
n-words	$\{2, 3\}$	$\{2\}$
q-grams	$\{3, 5, 9\}$	$\{3, 5\}$
skip-grams	–	$\{(3, 1)\}$

Table 1: Set of configurations for text modeling

Arabic since there is no such concept in Arabic text, it makes sense when text is not constrained to Arabic, e.g., tweets are full of text in other languages, it has URLs, user's names, or hashtags.

In case of English, the model selection procedure performed by B4MSA determined to use trigrams, bigrams, and character q-grams of sizes 3, 5, and 9. In case of Arabic, each tweet must be split into $(3, 1)$-skip-grams, bigrams and trigrams of words, and 3-grams, 5-grams of characters. TF-IDF term weighting is applied to both languages.

The processes associated to text modeling, shown in Table 1, are applied to all datasets as text representation model.

3.1 Training Data

For this year, SemEval provides training data from 2013 to 2016 (Nakov et al., 2016) evaluations to train systems. In addition, we use an extra dataset annotated by humans around 73 thousand tweets and 2,000 available for English (Mozetič et al., 2016) and Arabic (NRC, 2017) languages, respectively. Table 2 shows the distribution of classes for English and Arabic datasets. We consider, essentially, three kind of datasets as training data: all datasets provided from SemEval are as a cross-domain dataset for evaluation on the contest, Extra-data as out-of-domain dataset of SemEval evaluations, and DS dataset (distant supervision) as general domain dataset, mainly, for learning affective vocabulary and related words. In case of DS dataset, we obtained 11 million tweets for English after processing a huge amount of tweets, and 16 thousand tweets for Arabic (see section 3.2). For Arabic, due to the lack of data, all human labeled tweets are considered as a dataset.

DataSet	Positive	Neutral	Negative	Total
Statistics of English training data				
train2013	3,662	4,600	1,466	9,728
dev2013	575	739	340	1,654
test2013	1,572	601	1,640	3,813
test2014	982	669	202	1,853
test2015	1,040	987	365	2,392
train2016	3,094	863	2,043	6,000
dev2016	844	765	391	2,000
devtest2016	994	681	325	2,000
test2016	7,059	10,342	3,231	20,632
Extra-data	21,166	33,620	18,454	73,240
DS-dataset	5.5M	-	5.5M	11M
Statistics of Arabic training data				
train2017	743	1,470	1,142	3,355
Extra-data	448	202	1,350	2,000
DS-dataset	8,108	-	8,108	16,216

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

3.2 External Data

In addition of the training datasets provided by SemEval'17, and annotated Extra-datasets, we generate an affective dataset using distant supervision approach. Distant supervision has been used for tasks such as information extraction (Mintz et al., 2009), or sentiment analysis (Go et al., 2009). In sentiment analysis, emoticons, some words, and hashtags are usually used as indicators of emotion in order to create 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.

Our approach consists in filtering tweets considering the affective degree that each tweet contains based on its affective words. First, we have collected more than 220 million tweets for U.S. English according to their geolocation (from July to December 2016), and more than 130 thousand tweets for Arabic without restriction of geolocation (one week of January 2017). Later, tweets are filtered using a large affective lexicon built for this purpose. The tweets are selected based on its positive or negative words. Some heuristic rules are used, for example, if a tweet contains negative markers that could reverse the sentiment such as *no, not, although, however, but*, etc., question marks, or both positive and negative words, then

the tweet is discarded; if a tweet has only positive or negative words (no contradiction), then it is selected and labeled with the corresponding sentiment according to the affective words. Also, English and Arabic stemmers from NLTK (Bird et al., 2009) are used in order to maximize the matches between affective words and tweets.

Our distant supervision approach uses an affective lexicon that was created based on the SentiSense lexicon (de Albornoz et al., 2012) to extract affective tags (sadness, anger, love, etc.) related to WordNet synsets (Miller, 1995). A synset defines a group of words with semantic similarities, thus, a synset label defined in SentiSense is applied to all words in the WordNet synset, these words are part of our lexicon. Negative emotions in SentiSense (sadness, fear, anger, hate, disgust) are mapped to negative tag, and positive emotions (love, joy, like) are mapped to positive tag. In addition, opinion words from Bing Liu's lexicon (Liu, 2017) are also added to our lexicon. In case of Arabic language, the affective lexicon was created translating the affective lexicon from English into Arabic by means of python translation package for Bing translator service (LittleCoder, 2017). The same heuristic rules for English are applied to Arabic to create the labeled dataset for positive and negative emotions.

Finally, we remove near duplicated tweets to reduce the final dataset (DS-dataset); the idea is to select only the essential dataset while the vocabulary around affective words is maximized. The process of near duplicate removal is performed as follows. We performed a linear scan of the retrieved dataset, we transform the tweet with a number of coarsening text transformations (all those supported by B4MSA, see Table 1). Whether the transformed text has not be seen; that is, if the text was already generated, the tweet is discarded. In the end, from an initial collection of 220 million tweets, we obtained around 11 million exemplars for English, and, from an initial set of more than 130K of Arabic tweets, around 16 thousand exemplars were obtained (see Table 2 for more details).

4 Results

We present the results of our system in subtask A for both the English and Arabic languages. Table 3 shows the performance on some configurations for EvoDAG. 2-HL indicates the use of two human labeled datasets, SemEval and the presented in (Mozetič et al., 2016); 44-DS indicates the use of $k = 44$ for the 11 million DS-dataset. More detailed, the best 44 classifiers are chosen from 367; each classifier is trained over chunks of $30K$ tweets. The selection is made based on the vocabulary-affinity between an object and each classifier, see §2 for more details. In the end, this configuration produces 50 inputs for EvoDAG. Six inputs correspond to 2-HL since each dataset contributes with three inputs, i.e. the B4MSA's decision functions for positive, negative, and neutral classes. Also, the rest of the inputs correspond to 44 best B4MSA classifiers trained with our distant supervision process. Each value describes if a tweet is just positive or negative, as decided by the corresponding B4MSA classifier. We obtained 0.680 of macro-recall in our training stage, and achieve 0.649 in the SemEval's gold-standard.

In addition, we tested our system without using DS-dataset in order to show the improvement of our distant supervision approach (see 2-HL-train/4 configuration, Table 3). The training dataset was divided into 4 subsets to train our scheme with EvoDAG, this configuration, only with training dataset annotated by humans, is below nearly 3% of our best performance. Thus, we use the same DS approach for both English and Arabic languages.

In the case of Arabic, due to the lack of data, there are only five inputs for EvoDAG. As shown in Figure 2: three inputs come from a B4MSA trained with annotated datasets (1-HL), and two additional inputs come from trained classifiers with DS-dataset (16K tweets). The last dataset is partitioned into two subsets of around 8K tweets (2-DS), the only evaluation is shown in Table 4, we obtained 0.642 of macro-recall in our training stage and 0.477 in the SemEval's gold-standard.

configuration	macro-F1	macro-recall	accuracy
2-HL, 44-DS (11M)	0.649	**0.680**	0.667
2-HL, 44-DS (3.5M)	0.648	0.679	0.666
2-HL, train/4	0.632	0.652	0.655
Performance on gold standard of SemEval'17			
2-HL, 44-DS (11M)	0.645	**0.649**	0.633

Table 3: Results for substask A on English datasets. (HL) Human labeled, (DS) Distant supervision

configuration	macro-F1	macro-recall	accuracy
1-HL, 2-DS (16K)	0.642	**0.642**	0.662
Performance on gold standard of SemEval'17			
1-HL, 2-DS (16K)	0.455	**0.477**	0.499

Table 4: Results for substask A on Arabic datasets. (HL) Human labeled, (DS) Distant supervision

5 Conclusions

In this paper was presented the proposed approach combining a generic framework for multilingual polarity classification, B4MSA, with a genetic programming system, EvoDAG. For the training, we use several datasets: human annotated datasets, and our datasets generated with distant supervision approach. Our performance, macro-recall 0.649, brought us to the sixth position in the English language, and fourth position, macro-recall 0.477, for the Arabic language.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of LREC 2012*. pages 3562–3567.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1124–1128. http://www.aclweb.org/anthology/S16-1173.

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

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. https://doi.org/10.1109/ROPEC.2016.7830633.

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*, Springer International Publishing, number 663 in Studies in Computational Intelligence, pages 43–65. DOI: 10.1007/978-3-319-44003-3_2.

LittleCoder. 2017. Translation a python translation package based on website service. https://pypi.python.org/pypi/translation. Accessed 20-Jan-2017.

Bing Liu. 2017. English opinion lexicon. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. Accessed 20-Jan-2017.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, Springer US, Boston, MA, pages 415–463. https://doi.org/10.1007/978-1-4614-3223-4-13.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Mike Mintz, 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*. Association for Computational Linguistics, pages 1003–1011.

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

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, San Diego, California, SemEval '16.

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

Eric Sadit Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart Rodrigo Suárez, and Oscar Sánchez Siordia. 2016. A simple approach to multilingual polarity classification in twitter. *CoRR* abs/1612.05270. http://arxiv.org/abs/1612.05270.