INGEOTEC at SemEval-2019 Task 5 and Task 6: A Genetic Programming Approach for Text Classification

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 HatEval and OffensEval challenges for English and Spanish languages. We used several approaches, B4MSA, FastText, and EvoMSA. Best results were achieved with EvoMSA, which is a multilingual and domain-independent architecture that combines the prediction from different knowledge sources to solve text classification problems.

1 Introduction

Social media platforms, like Twitter and Facebook, are spaces where people interact with others and express themselves; while these platforms encourage free speech, other issues could emerge such as the usage of offensive language that could mock or insult individuals or groups of people. Thus, detecting offenses and misbehavior expressed in text form is interesting to measure the people's feelings and warn them about possible attacks on others such as abusive language, hate speech, cyberbullying, trolling, among others (Waseem et al., 2017).

In order to tackle these text classifications problems, SemEval-2019 proposed two tasks: multilingual detection of hate speech against immigrants and women in Twitter HatEval, task 5 (Basile et al., 2019), and identification and categorization of offensive language in social media OffensEval, task 6 (Zampieri et al., 2019b). In this paper, we present the results from our participating in these two tasks.

The HatEval challenge consists in detecting hate speech for two targets, immigrants and women, in Twitter for Spanish and English languages. There are two subtasks, subtask A is a binary classification where systems have to predict whether a tweet with a given target (immigrants or women) is hateful or not hateful; subtask B is about aggressive behavior and target classification, systems are asked to classify hateful tweets as aggressive or not aggressive, and identify the target harassed (individual or group).

On the other hand, OffensEval challenge consists in determining if a given message has offensive content. It is divided into three subtasks. Subtask A is dedicated to identifying the offensive language, i.e., determine if a message is offensive or not offensive. Subtask B is about categorizing offense types; that is, a tweet containing an insult or threat to someone, or a tweet containing non-targeted profanity and swearing. Finally, subtask C focus on identifying the target, i.e., whether the offensive post is about an individual, a group, or others.

Both HatEval and OffensEval are related tasks to abusive language, Waseem et al. (Waseem et al., 2017) describe tasks on this theme; authors focus their analysis on two primary factors that could guide the modeling of systems: i) language is directed towards a specific individual, entity, or generalized group; ii) the abusive content may be explicit or implicit.

For instance, Schmidt and Wiegand (Schmidt and Wiegand, 2017) present a collection of works on hate speech detection highlighting the features commonly used such as surface-level features. For instance, authors use bag of words (n-grams) and character-level n-grams to attenuate the spelling variation issue on informal text, frequency of URL mentions, punctuation, token lengths, capitalization, among others; word generalization such as topic identification (LDA) and word embeddings (Mikolov et al., 2013); outcomes from sentiment analysis classifiers (for example, samples predicted as negative polarity) as auxiliary evidence of hate for multi-step approaches; usage of lexical resources containing specific negative words (slurs, insults, etc.); linguistic aspects such as parts

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 639–644
Minneapolis, Minnesota, USA, June 6–7, 2019.

of speech and syntactic information; knowledge information such as ontologies and taxonomies (ConcepNet, WordNet, etc.).

For both tasks, we use the same approach for final runs. Our approach takes into account several features mentioned above. For example, the effects of character-level n-grams are broadly studied for related tasks in (Tellez et al., 2017b). In particular, text modeling is a crucial factor in our approach; therefore we used the approach presented in (Tellez et al., 2018) that selects the best configuration on the datasets concerned. We also use external knowledge to the given training set to support the classification task; in this sense, our approach named EvoMSA (§2.1) is a stacking system based on genetic programming, and particularly on the use of semantic genetic operators, that focus on sentiment analysis, and, in general, on text classification.

2 System Description

We used our framework based on genetic programming named EvoMSA to evaluate HatEval and OffensEval tasks. EvoMSA is composed of a stack of B4MSA classifiers to produce predictions, and EvoDAG combines the predictions into the final one.

2.1 EvoMSA

EvoMSA[1] (Graff et al., 2018a,b) is a Generic Sentiment Analysis System based on B4MSA and EvoDAG. It is an architecture of two phases to solve classification 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. 2.1.1) are trained from several views of the same datasets; datasets provided by SemEval. It creates a decision functions space with mixtures of values coming from different views of knowledge, one coming from B4MSA trained with the training set of the competition (it is used as generic classifier), a lexicon-based model (it only counts affective words: positive and negative, based on several lexicons (Liu, 2017; Albornoz et al., 2012; Sidorov et al., 2013; Perez-Rosas et al., 2012)), an emoji-based space (the sixty-four most probable emoticons for the message) (Graff et al., 2018b), and the output of FastText (Grave et al., 2018) (word embeddings of dimension of 100) trained with the training set. Finally, EvoDAG's inputs are the concatenation of all the decision functions predicted, 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.

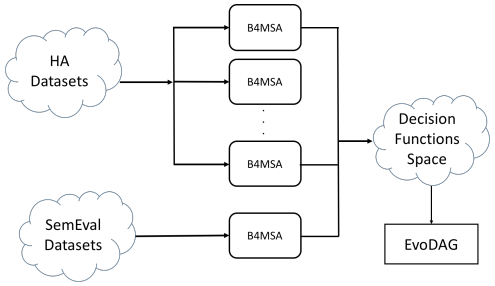

Figure 1: EvoMSA Architecture

2.1.1 B4MSA

B4MSA[2] focus on multilingual sentiment analysis. For complete details of the model see (Tellez et al., 2017a,b). The core idea behind B4MSA is to tackle the sentiment analysis problem as a model selection problem, using a different view of the underlying combinatorial problem, i.e., B4MSA combines a bunch of different text tokenization, text transformations, weighting methods, and internally uses an SVM with a linear kernel to classify. Also, 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 (Miranda-Jiménez et al., 2017). Table 1 shows text transformation parameters used in our system for English and Spanish languages.

2.1.2 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

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

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

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

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.

3 Experimental Settings

As we mentioned, to determine the best configuration of parameters for text modeling, B4MSA integrates 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 languages. 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 popular 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, for $q = 1$ to 9. Finally, we allow skip-grams such as $(3, 1)$ and $(2, 2)$, three words separated by one word (gap), and two words separated by two gaps.

We use two baselines B4MSA and the FastText's classifier (Bojanowski et al., 2016) for both contests. FastText represents sentences with a weighted bag of words, and each word is represented as a bag of character n-gram to create text vectors based on word embeddings. Our custom FastText searches automatically the best parameters, e.g., for OffensEval with parameters such as window $size = 9$, learning $rate = 0.01$, $epochs = 10$, size of word $vectors = 10$, minimum and maximum length of character n-grams, 2 and 5, respectively; and some other preprocessing steps such as group numbers and reduce duplicated characters.

3.1 Datasets

SemEval contests provide datasets to train systems for each task. Table 2 presents the data distribution of the HatEval dataset. *Hate* class (HATE) defines tweets that convey hate against immigrants or women; its complement correspond to these messages not having hate content (NO-H), aggressive (AGGR) and no aggressive (NO-A), and target harassed (TARG) as individual and group.

Table 3 shows the OffensEval data distribution. In Task A, class OFF defines tweets that have offenses or insults; while class NOT describes tweets with no offensive content. Messages with labeled as TIN contain an insult or threat to an entity; UNT defines the opposite. Group (GRP), individual (IND), and others (OTH) classes contain the target of the offensive messages. The OffensEval collection is described in detail in Zampieri et al. (2019a).

Text transformation	English (HE)	Spanish (HE)	English (OE)
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	delete	delete
urls	group	none	group
users	group	group	none
hashtags	none	none	none
entities	none	none	none
stemming	yes	yes	yes
Term weighting			
TF-IDF	yes	yes	yes
Entropy	no	no	no
Tokenizers			
n-words	$\{1, 3\}$	$\{1, 2\}$	$\{1, 2, 3\}$
q-grams	$\{3, 5, 9\}$	$\{2, 5, 7, 9\}$	$\{3, 4, 5, 9\}$
skip-grams	$\{(3, 1)\}$	$\{(3, 1)\}$	$\{(3, 1), (2, 2)\}$

Table 1: Example of set of configurations for text modeling, HatEval (HE), and OffensEval (OE)

DataSet	NO-H	HATE	NO-A	AGGR	NO-T	TARG
training (English)	5,217	3,783	7,441	1,559	7,659	1,341
development (English)	573	427	796	204	781	219
training (Spanish)	2,643	1,857	2,998	1,502	3,371	1,129
development (Spanish)	278	222	324	176	363	137

Table 2: Statistics of HatEval datasets.

DataSet	Task A		Task B		Task C		
	NOT	OFF	TIN	UNT	GRP	IND	OTH
training	8,840	4,400	3,876	524	1,074	2,407	395
development	243	77	34	39	4	30	2

Table 3: Statistics of OffensEval datasets for English language.

4 Results

We present the results of our approaches for HatEval contest in Table 4 and Table 5. We performed our experimentation on the development dataset

provided by HateEval. Table 4 shows the results of task A, given a tweet is hateful or not hateful for English and Spanish languages. In the case of task A, the macro-F1 score is used to measure the performance. Table 5 shows the results of task B, classify tweets as aggressive or not aggressive and the target harassed.

In the case of OffenseEval, Table 6 shows the results for the three task proposed offensive language identification (Task A), categorization of offense types (Task B), and offense target identification (Task C).

We present three system configurations for both tasks. B4MSA uses only the training data provided by the contest as the knowledge base to classify texts, i.e., B4MSA is our baseline, but it is also its outcome is an additional input for our more sophisticated classifier (EvoMSA). FastText generates word embeddings from the provided dataset. We do not use pre-training vectors, using pre-trained vectors did not provide any significant improvement in this case, but increased the complexity of the models and the processing pipeline. EvoMSA (Graff et al., 2018a) combines, using EvoDAG, the output of different text models such as B4MSA, a lexicon-based model, an emoji-space model, and FastText.

As we can see the performance in all results Tables, EvoMSA is systematically better than our other systems; under these circumstances, we decided to use EvoMSA firstly in the evaluation phase. Following the rules of HatEval, only the last run would be valid; therefore we used EvoMSA for this chance. In the case of OffensEval, up to three predictions were allowed on the test dataset, but only the best one was compared with other systems. As we can see, Table 6 shows the performance of our three systems on gold standards; EvoMSA stays ahead in all tasks including the baselines from the contest. The table also shows the performance of two baselines, "All NOT" and "ALL OFF", that correspond to labeling all tweets as NOT or OFF, respectively; similarly, the rest of the tasks have baselines for "All TIN", "All UNT", "All GRP", "ALL IND", and "ALL OTH" labeling strategies.

System	F1	Accuracy
English		
B4MSA	0.736	0.752
EvoMSA	0.736	0.733
FastText	0.728	0.756
Performance on gold standard		
EvoMSA	0.350	0.447
Spanish		
B4MSA	0.812	0.838
EvoMSA	0.821	0.834
FastText	0.822	0.801
Performance on gold standard		
EvoMSA	0.710	0.710

Table 4: Results of HateEval: Task A

5 Conclusions

In this paper was presented our solution for HatEval and OffensEval, two campaigns of SemEval 2019. We show the competitiveness of our approach in both training and test phases. EvoMSA and B4MSA are designed to be multilingual and language and domain independent as much as possible. For the training step, we used extra knowledge from datasets out of any specific emotion of the contests, but categories or emotions related to sentiment-analysis information. Our solution performs well in Spanish and some task for English languages; however, there is room for further improvements in performance for tasks in English language using another sort of knowledge for specific domains.

References

Jorge Carrillo De Albornoz, Laura Plaza, and Pablo Gerv. 2012. Language Resources and Evaluation SentiSense : an affective lexicon for sentiment analysis SentiSense : An easily scalable concept-based affective lexicon for sentiment analysis. In *International Conference on Language Resources and Evaluation*, pages 3562–3567, Istanbul, Turkey. European Language Resources Association (ELRA).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

M. Graff, E. S. Tellez, S. Miranda-Jiménez, and H. J. Escalante. 2016. Evodag: A semantic genetic programming python library. In *2016 IEEE Interna-*

System	Aggressiveness		Hate		Target		Avg-F1
	F1	Accuracy	F1	Accuracy	F1	Accuracy	
English							
B4MSA	0.495	0.814	0.736	0.752	0.659	0.858	0.630
EvoMSA	0.556	0.746	0.736	0.733	0.711	0.851	0.668
FastText	0.536	0.806	0.729	0.752	0.710	0.866	0.658
Performance on gold standard							
EvoMSA	0.515	0.542	0.348	0.445	0.653	0.699	0.506
Spanish							
B4MSA	0.820	0.726	0.785	0.810	0.767	0.880	0.791
EvoMSA	0.755	0.818	0.821	0.834	0.810	0.890	0.795
FastText	0.744	0.818	0.796	0.824	0.791	0.888	0.777
Performance on gold standard							
EvoMSA	0.737	0.765	0.71	0.71	0.816	0.862	0.754

Table 5: Results of HateEval: Task B

System	F1	Accuracy
Task A		
B4MSA	0.767	0.831
EvoMSA	0.774	0.828
FastText	0.741	0.803
Performance on gold standard.		
All NOT baseline	0.419	0.721
All OFF baseline	0.218	0.279
B4MSA	0.729	0.801
EvoMSA	0.731	0.791
FastText	0.697	0.797
Task B		
B4MSA	0.398	0.507
EvoMSA	0.694	0.699
FastText	0.618	0.644
Performance on gold standard.		
All TIN baseline	0.470	0.888
All UNT baseline	0.101	0.113
B4MSA	0.507	0.892
EvoMSA	0.671	0.871
FastText	0.634	0.896
Task C		
B4MSA	0.418	0.833
EvoMSA	0.392	0.611
FastText	0.550	0.806
Performance on gold standard.		
All GRP baseline	0.179	0.366
All IND baseline	0.213	0.470
All OTH baseline	0.094	0.164
B4MSA	0.486	0.653
EvoMSA	0.576	0.676
FastText	0.504	0.639

Table 6: Results of OffensEval: Offensive language identification (Task A), categorization of offense types (Task B), and offense target identification (Task C).

tional Autumn Meeting on Power, Electronics and Computing (ROPEC), pages 1–6.

Mario Graff, Sabino Miranda-Jiménez, Eric S Tellez, and Daniela Moctezuma. 2018a. Evomsa: A multilingual evolutionary approach for sentiment analysis. *arXiv preprint arXiv:1812.02307*.

Mario Graff, Sabino Miranda-Jiménez, Eric Sadit Tellez, and Daniela Moctezuma. 2018b. Evomsa: A multilingual evolutionary approach for sentiment analysis. *CoRR*, abs/1812.02307.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 3483–3487. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Bing Liu. 2017. English Opinion Lexicon.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sabino Miranda-Jiménez, Mario Graff, Eric S Tellez, and Daniela Moctezuma. 2017. INGEOTEC at SemEval 2017 Task 4: A B4MSA Ensemble based on Genetic Programming for Twitter Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 771–776. Association for Computational Linguistics.

Veronica Perez-Rosas, 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.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, No 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.

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.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.