INGEOTEC at SemEval-2020 Task 12: Multilingual Classification of Offensive Text

Sabino Miranda-Jiménez and **Eric S. Tellez** and **Mario Graff**
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 OffensEval challenges for English, Arabic, Danish, Turkish, and Greek languages. We used several approaches, such as μTC, TextCategorization, 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

Freedom of expression is one of the important factors that make social media platforms popular; there, people interact with others and express themselves freely. However, offensive content has become pervasive in social media that could mock or insult both individuals or groups of people. Thus, detecting offenses and misbehavior expressed in text form is useful to measure the people's feelings and warn them about possible attacks on others such as abusive language, hate speech, cyberbullying, trolling, among others social problems (Waseem et al., 2017; Zampieri et al., 2019a).

The scientific community organize several challenges periodically, like OffensEval, to tackle these text classification problems. In 2019, OffensEval (Zampieri et al., 2019c) organized three tasks for English, and this year (OffensEval-2020) (Zampieri et al., 2020) are included Arabic, Greek, Danish, and Turkish languages. The tasks consist in identifying the type of offensive language in short texts. In this paper, we present the results of our participation in all tasks and all languages previously mentioned.

OffensEval challenge consists in determining whether a given message has offensive content. It is divided into three tasks A, B, and C for English and only task A for Arabic, Greek, Danish, and Turkish. Task A is dedicated to identifying the offensive language, i.e., determine if a message is offensive or not offensive. Task 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, task C focuses on identifying the target, i.e., whether the offensive content is about an individual, a group, or others.

Abusive and offensive language identification problems includes aggression (Kumar et al., 2018; Aragón et al., 2019), hate speech (Basile et al., 2019), cyberbullying (Smith et al., 2008; Arroyo-Fernández et al., 2018), and offenses on targets (individuals or groups) (Waseem et al., 2017; Zampieri et al., 2019c). Typically, these problems are tackled as supervised learning problems, i.e., classification. For example, to identify the offensive language, the work by Liu et al. (2019) uses pre-trained BERT with fine-tuning on the training dataset. The approach described by Nikolov and Radivchev (2019) uses pre-trained BERT and GloVe vectors, along with techniques for overcoming unbalanced class distribution in the provided test data, authors show how these techniques increase the performance in general. Kebriaei et al. (2019) study the performance of combining TF-IDF weighting, lexicon-based approaches, and Support Vector Machines (SVM). A rule-based blacklist approach is used by Pedersen (2019).

For our participation in the contest, for all tasks and all languages, we use the same approach for final runs proposed by Graff et al. (2020). Our approach takes into account several features; for example, the effects of character-level n-grams, that is broadly studied for related tasks in the work of Tellez et al. (2017b). In particular, text modeling is a crucial factor in our multilingual approach; therefore, we used the approach presented by Tellez et al. (2018) that selects the best configuration on the datasets

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1992–1997
Barcelona, Spain (Online), December 12, 2020.

concerned. We also use external knowledge to the given training set to support the classification task; in this sense, our approach named EvoMSA (Section 2.1) is a stacking system that focuses on sentiment analysis, and, in general, on text classification.

2 System Description

We used our framework based on stacking generalization and genetic programming named EvoMSA to tackle the OffensEval tasks. EvoMSA is composed of a stack of several classifiers among them B4MSA, EmoSpace, lexicon-based to produce predictions, later a classifier, e.g. EvoDAG, combines the predictions into the final one.

2.1 EvoMSA

EvoMSA[1] (Graff et al., 2020) is a classifier that combines the output of different text classifiers to produce the final prediction. It is an architecture of two phases to solve classification tasks. 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 classifiers (e.g., B4MSA (Tellez et al., 2017a), SVM, Naive Bayes) are trained to produce several views of the same datasets. It creates a decision-value space with mixtures of predictions coming from different views of knowledge, one coming from B4MSA trained with the training set of the competition (it is used as a generic classifier), a lexicon-based model, an emoji-based space (the sixty-four most probable emoticons for the message), and the output of FastText (Grave et al., 2018) (100-dimensional word embeddings) trained with the training set. Finally, a classifier such as EvoDAG (Graff et al., 2016) produces a final prediction using the concatenation of all the decision functions predicted by the previous phase. The precise configuration of our benchmark system is described in Section 5. For a deep understanding of the architecture see the work presented by Graff et al. (2020).

2.2 MicroTC (μTC)

μTC (Tellez et al., 2018) is a minimalist and easy-to-use library that generates text models maximizing a performance measurement. μTC uses a SVM with a linear kernel as the classifier. The core idea behind μTC is to define a parameter space describing a massive number of text-classifiers. Parameters include transformations on text such as convert case, numbers, hashtags, n-grams of characters and words, skip grams, punctuation, and others, for more details see the work of Tellez et al. (2018). The problem is posed as a combinatorial optimization problem, and an efficient set of meta-heuristics are used to find very competitive solutions.

2.3 TextCategorization package

TextCategorization[2] is a Julia package inspired by μTC. The main difference with μTC is that it performs a full model selection, and this means the combinatorial problem represents the entire text-classification pipeline. That is, each configuration (model) describes all preprocessing functions, different tokenization schemes as μTC, several term weighting schemes, and several parts of the classifier used (a kernel-based and prototype-based classifier). The selection of both kernel and prototyping schemes are part of the combinatorial problem. The combinatorial problem is tackled using a local search algorithm (Beam Search); the configuration space is sampled randomly for the initial population, and then it is explored with a mutation and crossover strategy.

3 EvoMSA - BeamSelection / TextModels

One of the biggest problems with EvoMSA, and in general with most stacking generalization schemes, is what kind of models must be combined to produce well performance final models. While EvoMSA can handle several data complications with its genetic programming approach, it is only able to handle a relatively small number of models due to its computational cost. In order to handle more than 30

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

[2] https://github.com/sadit/TextClassification.jl

different models, that can produce more than 10^9 possible combinations, TextModels uses a Beam search approach to explore the combinatorial space, probing each combination with stacking generalization over the models and combine them with a Naïve Bayes classifier. Note that while we are looking for high-quality predictions with the proper selection of the combination, the stacking generalization process itself needs to be fast to be able to find those suitable models in a reasonable amount of time.

4 Experimental Settings

As we mentioned, to determine the best configuration of parameters for text modeling, μTC (aka B4MSA) integrates a hyper-parameter optimization phase that ensures the performance of the classifier based on the training data. The text modeling parameters for μTC and TextCategorization are computed for each language. A text transformation feature could be binary (yes/no) or ternary (group/delete/none) option and tokenizers generate text chunks in a range of lengths, all tokens generated are part of the text representation, see the approach proposed by Tellez et al. (2018).

4.1 Datasets

SemEval contests provide datasets to train systems for each task. Table 1 shows the OffensEval data distribution for English (Zampieri et al., 2019b) and Table 2 shows the data distribution for task A for Arabic (Mubarak et al., 2020), Greek (Pitenis et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), and Turkish (Çöltekin, 2020). In Task A, class OFF defines tweets that have offenses or insults, while class NOT describes messages with no offensive content. Tweets 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 for task C. Also, a large dataset for English is provided for training, the SOLID dataset (Rosenthal et al., 2020). It contains labeled tweets in a semi-supervised manner for task A (over 9 million tweets), task B (nearly 200 thousand), and task C (over 100 thousand).

DataSet	**Task A**		**Task B**		**Task C**		
	NOT	OFF	TIN	UNT	GRP	IND	OTH
training	9,083	4,477	3,910	563	1,078	2,437	397
2019-test	620	240	213	27	78	100	35

Table 1: Statistics of OffensEval-2019 datasets for English language.

DataSet	**Arabic**		**Greek**		**Danish**		**Turkish**	
	NOT	OFF	NOT	OFF	NOT	OFF	NOT	OFF
training	5,590	1,410	4,395	1,725	1,800	272	17,949	4,280
development	821	179	1,862	761	777	112	7,676	1,851

Table 2: Distribution of classes in OffensEval-2020 Task A datasets.

5 Results

We present the results of our approaches for the OffensEval contest in Table 3 for English and Table 4 for the languages mentioned. Table 3 shows the results on the three tasks proposed as offensive language identification (Task A), categorization of offense types (Task B), and offense target identification (Task C). Table 4 shows the results on only for task A, offensive language identification. We performed our experimentation on the 2019-test dataset released by OffensEval-2019 organizers. The rank obtained in the global ranking is indicated in parentheses.

We present three systems. μTC uses only the training data provided by the contest as the knowledge base to classify texts, i.e., μTC is our baseline, but it is also its outcome is an additional input for our more sophisticated classifier (EvoMSA). TextCategorization is a full model selection being able to adjust almost any part in its text classification pipeline.

EvoMSA combines, using EvoDAG, the output of different text models such as μTC (B4MSA), a lexicon-based model, an emoji-space model, and FastText. A complete study of the effects of wide range of models for EvoMSA, we refer the interested readers to the work presented by Graff et al. (2020).

As we can see the performance in all results tables, EvoMSA+BeamSelection is systematically better than our other systems; under these circumstances, we decided to use EvoMSA+BeamSelection in the evaluation phase. Table 3 shows the performance of our system on gold standards for English and Table 4 for the other languages.

System	English		
	TASK A	TASK B	TASK C
μTC	0.7356	0.6446	0.4916
EvoMSA+BeamSelection	0.8147	0.8383	0.7273
TextCategorization	0.7787	0.7124	0.5961
Performance on gold datasets			
Best System	0.9222	0.7461	0.7145
INGEOTEC	0.9061 (54)	0.6321 (12)	0.5626 (27)

Table 3: Results of task A (offensive language identification), task B (automatic categorization of offense types), and task C (offense target identification). The rank obtained is indicated in parentheses.

System	Arabic	Greek	Danish	Turkish
μTC	0.8353	0.7998	0.7208	0.7561
EvoMSA+BeamSelection	0.8568	0.8060	0.8270	0.8270
TextCategorization	0.8571	0.8037	0.8006	0.7561
Performance on gold datasets				
Best System	0.9017	0.852	0.8120	0.8257
INGEOTEC	0.8743 (8)	0.8200 (10)	0.7240 (21)	0.7757 (11)

Table 4: Results of Task A: Offensive language identification. The rank obtained is indicated in parentheses.

6 Conclusions

In this paper, we presented our solution for OffensEval 2020. We showed the competitiveness of our approach in both training and test phases. Our systems 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 English (task A), Arabic, and Greek languages; however, there is room for further improvements in performance using another sort of knowledge for specific domains and languages.

References

Mario Ezra Aragón, Miguel Ángel Álvarez Carmona, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, and Daniela Moctezuma. 2019. Overview of MEX-A3T at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 478–494. CEUR-WS.org.

Ignacio Arroyo-Fernández, Dominic Forest, Juan-Manuel Torres-Moreno, Mauricio Carrasco-Ruiz, Thomas Legeleux, and Karen Joannette. 2018. Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18

trac-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 140–149, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. ELRA.

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, Nov.

M. Graff, S. Miranda-Jiménez, E. S. Tellez, and D. Moctezuma. 2020. Evomsa: A multilingual evolutionary approach for sentiment analysis [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):76–88, Feb.

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), 2.

Emad Kebriaei, Samaneh Karimi, Nazanin Sabri, and Azadeh Shakery. 2019. Emad at SemEval-2019 task 6: Offensive language identification using traditional machine learning and deep learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 600–603, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Ted Pedersen. 2019. Duluth at SemEval-2019 task 6: Lexical approaches to identify and categorize offensive tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 593–599, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *arxiv*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

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, jul.

Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, and Elio A. Villaseñor. 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, jun.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420.

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

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.