Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets

Mario Ezra Aragón[1], Miguel Á. Álvarez-Carmona[4,5], Manuel Montes-y-Gómez[1], Hugo Jair Escalante[1], Luis Villaseñor-Pineda[1,2], and Daniela Moctezuma[3]

[1] Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
[2] Centre de Recherche en Linguistique Française GRAMMATICA (EA 4521),
Université d'Artois, France.
[3] Centro de Investigación en Ciencias de Información Geoespacial A.C., Mexico
[4] Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico
[5] Unidad de Transferencia Tecnológica,
Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE-UT3), Mexico
{mearagon,mmontesg,hugojair,villasen}@inaoep.mx,
malvarezc@cicese.mx, dmoctezuma@centrogeo.edu.mx

Abstract. This paper presents the framework and results from the MEX-A3T track at IberLEF 2019. This track considers two tasks, author profiling and aggressiveness detection, both of them using Mexican Spanish tweets. The author profiling task consists on determining the gender, occupation and place of residence of users from their tweets. As a novelty in this year's edition, it considers the use of text and images as information sources, with the aim of studying the relevance and complementarity of multimodal data for profiling social media users. On the other hand, the aggressiveness detection task follows the same design than the previous edition; it aims to discriminate between aggressive and non-aggressive tweets. For both tasks, we have built new corpora considering tweets from Mexican Twitter users. This paper compares and discusses the results of the participants.

1 Introduction

Twitter platform is constantly growing thanks to the information generated by a massive community of active users. The analysis of shared information has become very relevant for several applications in marketing, security, and forensics, among others.

One essential task for social media analysis is *author profiling* (AP), which consists in predicting general or demographic attributes of authors by examining the content of their posts [2, 4]. On the other hand, *detecting aggressive content* targeted to specific people or vulnerable groups is also a task of high relevance to preventing possible viral destructive behaviors through social networks.

The objective of the MEX-A3T is to encourage research on the analysis of social media content in Mexican Spanish. Mainly, it aims to push research into the treatment of a variety of Spanish that has cultural traits that make it significantly different from the peninsular Spanish. Accordingly, the 2019 edition of MEX-A3T consider two main tasks: author profiling, whose aim was to develop methods for profiling users according to non-standard dimensions (gender, occupation, and place of residence), and aggressiveness detection in tweets. Particularly, the main novelty for this edition is the use multimodal data (text and images) for AP, with the aim of exploring the relevance of multimodal information for profiling social media users.

To evaluate these tasks, we built two ad hoc collections. The first one is a multimodal author profiling corpus consisting of 5 thousand Mexican users, each one having eleven images, the profile image as well as ten random selected pictures. This corpus is labeled for the subtasks of gender, occupation and place of residence identification. Whereas the second corpus is oriented to the aggressiveness detection and contains more than 11 thousand tweets. In this case, each tweet is labeled as aggressive or not.

The remainder of this paper is organized as follows: Section 2 covers a brief description of the first edition of the MEX-A3T; Section 3 presents the evaluation framework used at MEX-A3T 2019; Section 4 shows an overview of the participating approaches; Section 5 reports and analyses the results obtained by the participating teams; finally, Section 6 draws the conclusions of this evaluation exercise.

2 MEX-A3T 2018

Last year, the first edition of the MEX-A3T shared task was carried out [1]. This represented the first attempt for organizing an evaluation forum for the analysis of social media content in Mexican Spanish. A variety of methods were proposed by participants, comprising content-based (bag of words, word n-grams, term vectors, dictionary words, and so on) and stylistic-based features (frequencies, punctuation, POS, Twitter-specific elements, slang words, and so forth) as well as approaches based on neural networks (CNN, LSTM and others). In both tasks, author profiling and aggressiveness identification, the baseline results were outperformed by most participants.

For author profiling, the approach proposed by the MXAA team [10] obtained the best results with an approach based on emphasizing the value of personal information for building the text representation. In the case of the aggressiveness identification, the top-ranked team was INGEOTEC [7], which proposed an approach based on MicroTC and EvoMSA. MicroTC is a minimalistic text clas-

sifier independent from domain and language. EvoMSA is another text classifier which combines models (as MicroTC) with Genetic Programming.

3 Evaluation framework

This section outlines the construction of the two used corpus, highlighting particular properties, challenges, and novelties. It also presents the evaluation measures used for both tasks.

3.1 A multimodal Mexican corpus for author profiling

This new corpus is based on previous year's collection. For the MEX-A3T 2018, we labeled 5 thousand Twitter users for occupation and place traits divided into 3500 users for training and 1500 for the test. For the occupation label, we considered the following eight classes: *arts, student, social, sciences, sports, administrative, health,* and *others*. For the place of residence trait, we considered the following six classes: *north, northwest, northeast, center, west, and southeast.* For more details, we recommend consulting [1].

For this year's edition, we added the gender trait to the corpus, and in this way, each user is characterized by three labels: gender, occupation, and location. Another important novel aspect of this new corpus is the addition of 11 images for user. We selected the profile picture for each user as well as 10 randomly selected images from their tweets[6].

Table 1 shows the distribution of the corpus according to the gender trait. For this corpus, the gender trait is balanced. Also, Table 2 shows the distribution of the corpus according to the place of residence trait. As it is possible to observe, the distributions of training and test sets are very similar. The majority class corresponds to the *center* region, with more than 36% of the profiles, whereas the minority class is the *north* region with only 3% of the instances. On the other hand, Table 3 shows the distribution of the occupation trait. It also shows similar distributions in the training and test partitions. The majority class are *students* with almost 50% of the profiles, whereas *sports* correspond to the minority class, with approximately 1% of the instances.

In the Tables 2 and 3, the class imbalance was calculated as proposed in [15]. The place of residence trait shows a value of 396.1, whereas the occupation trait has a value of 502.4. Considering that zero represents a perfect balance, these numbers indicate that the imbalance is bigger for the occupation trait and, therefore, that it could be more complex to be predicted that the place of residence.

Finally, Table 4 presents some additional statistics for the author profiling corpus. For computing these numbers, we have considered words, numbers, punctuation marks and emoticons as terms. We also applied a normalization over user

[6] For most users we collected 11 images, although there are a small number of users with fewer images since in total they did not shared 10 images. In this cases, we take all available images.

Table 1. Gender distribution

Class	Train Corpus (%)	Test Corpus (%)
Male	1750 (50)	750 (50)
Female	1750 (50)	750 (50)
Σ	3500	1500
Class imbalance	0	0

Table 2. Mexican author profiling corpus: distribution of the place of residence trait.

Class	Train Corpus (%)	Test Corpus (%)
North	106 (3.02)	34 (2.26)
Northwest	576 (16.45)	229 (15.26)
Northeast	914 (26.11)	389 (25.93)
Center	1266 (36.17)	554 (36.93)
West	322 (9.20)	144 (9.60)
southeast	316 (9.02)	150 (10.00)
Σ	3500	1500
Class imbalance	396.45	173.23

Table 3. Mexican author profiling corpus: distribution of the occupation trait.

Class	Train Corpus (%)	Test Corpus (%)
Arts	240 (6.85)	103 (6.86)
Student	1648 (47.08)	740 (49.33)
Social	570 (16.28)	234 (15.60)
Sciences	185 (5.28)	65 (4.33)
Sports	45 (1.28)	26 (1.73)
Administrative	632 (18.05)	264 (17.60)
Health	105 (3.00)	43 (2.86)
Others	75 (2.14)	25 (1.66)
Σ	3500	1500
Class imbalance	502.42	226.04

mentions, hashtags, and URLs. It is possible to observe that the lexical diversity is very close for the training and test partitions. Also, the same goes for the tweets per profile averages. Nevertheless, the standard deviation in training and test is quite large, implying that the length of the profiles is very variable. Finally, the last row in the table shows the number of images in the corpus.

Table 4. Statistics for the Mexican Author profiling corpus.

Measure	Train Corpus	Test Corpus	Full corpus
Tweets per profile	1354.21(±917.61)	1353.38(±905.58)	1353.96(±914.02)
Number of terms	78,542,124	34,032,819	112,574,943
Vocabulary size	2,540,580	1,274,902	3,506,826
Lexical diversity	0.0323	0.0374	0.0311
Images	38249	16354	54603

3.2 A Mexican corpus for aggressiveness identification

As the author profiling corpus, for the previous edition of MEX-A3T we also built an corpus of tweets for the task of aggressiveness detection. To build this corpus, we used rude words and controversial hashtags to narrow the search. The hashtags were related to topics of politics, sexism, homophobia, and discrimination. The collected tweets were labeled by two persons. At the end each tweet of the corpus was labeled as *aggressive* or *non-aggressive*. Table 5 shows some examples labeled as aggressive and non-aggressive. As can be intuited, the task of labeling aggressiveness is challenging, especially because in most of the cases it is necessary to interpret the message in a given context.

Table 5. Aggressive and Non-Aggresive Tweets.

Aggressive	Non-Aggressive
Tu novia la gata esa que usa hashtag hasta para poner hola, tu novia la acapulqueña esa	Aquí me juego la vida, o leo el libro o leo las diapos, porque nuestro capítulo es de mil putas hojas. *literal*
Deja de estar de calientagüevos, que te vas a ganar una madriza	Soy una enamoradiza sin remedio". -La emperatriz de todas las putas.
Es una tipa tan cagante que no tiene amigos	Atendiendote apartir de las 5 pm zona centro #SQUIRT #MILF #CULOS #NALGONA #HOTWIFE #SCORT #PUTAS

The collected corpus consists of 11 thousand tweets. For the evaluation exercise, the corpus was divided into two parts, one for training and the other for

test. Table 6 shows the distribution of this corpus. It is noticed that the non-aggressive class is the majority class in both partitions. For more details of the labeled methodology, please consult [1].

Table 6. Mexican aggressiveness corpus: distribution of the classes.

Class	Training Corpus (%)	Test Corpus (%)
Not Aggressive	4973 (65)	2372 (75)
Aggressive	2727 (35)	784(25)
Σ	7700	3156

3.3 Performance measures

Author profiling. For the author profiling task, we used as final score the average of the macro F_1 measures for gender, place of residence, and occupation traits, as shown in Formula 1.

$$F_{average} = \frac{F_{macro}(C_{gender}) + F_{macro}(C_{location}) + F_{macro}(C_{occupation})}{3} \quad (1)$$

The F_{macro} measures were computed using Formula 2, where C indicates the set of classes for a given trait[7], and $F_1(c)$ is the F_1-measure of each of the categories from that trait.

$$F_{macro}(C) = \frac{1}{|C|} \sum_{c \in C} F_1(c) \quad (2)$$

Aggressiveness identification. For this task, the final score corresponds to the F_1-measure for the aggressive class.

4 Overview of the Submitted Approaches

For this study, 8 teams have submitted one or more solutions, of which, 2 participated in the author profiling task and 6 participated in the aggressiveness identification task. By what they explained in their notebook papers, this section presents a summary of their approaches regarding preprocessing steps, features, and classification algorithms.

The participating methods are listed below:

[7] C_{gender} = {male, female}, $C_{location}$ = {north, northwest, northeast, center, west, southeast}, and $C_{occupation}$ = {arts, student, social, sciences, sports, administrative, health, others}

– *CerpamidUA at MexA3T 2019: Transition Point Proposal* [6]
  - **Task:** Author Profiling
  - **Team name:** Cerpamid
  - **Features:** Bag of Words.
  - **Classification:** Support Vector Machine.
  - **Summary:** In this paper, the authors proposed an approach that follows the traditional pipeline of a non-thematic text classification system, where they employed a BOW representation and a SVM classifier. The authors focused on determining a reduced subset of features that represent frequent words for each profile, and propose using the theory of Transition Points for the selection of these features.

– *Author profiling from images using 3D Convolutional Neural Networks* [16]
  - **Task:** Author Profiling
  - **Team name:** CIC-VCR
  - **Features:** Hierarchical features obtained with CNN.
  - **Classification:** CNN.
  - **Summary:** In this paper, the authors focused on determining the profile of an author using images only. They proposed a 3D Convolutional Neural Network for extracting features from the images and classifying them in the different classes. They concluded that predicting the AP of a Twitter user using only images is a difficult task due to the generality of purpose of the images on this platform.

– *Aggressive analysis in Twitter using a combination of models* [13]
  - **Task:** Aggressiveness Detection
  - **Team name:** PRHLT
  - **Features:** Bag of Words with TF-IDF weights, hierarchical features obtained with CNN.
  - **Classification:** CNN, LSTM and Multi-layer Perceptron.
  - **Summary:** In this work, the authors proposed a method that combines different classification strategies: a Convolutional Neuronal Networks whose outputs feed a LSTM Neural Network; a pre-trained Universal Sentence Encoder for encoding sentences into embedding vectors; and a simple Multi-layer Perceptron which gets the TF-IDF representation of the tweet. The best results were obtained with the simplest model (the multi-layer perceptron), which can be explained by the lack of data to train deep learning models.

– *Aggressiveness Identification in Twitter at IberLEF2019: Frequency Analysis Interpolation for Aggressiveness Identification* [12]
  - **Task:** Aggressiveness Detection
  - **Team name:** OscarGaribo

- **Features:** Statistical descriptors.
- **Classification:** Support Vector Machine.
- **Summary:** In this paper, the authors proposed a new text representation that reduces the dimensionality of the information for each author or text to 6 characteristics per class. The proposed representation aims to capture the level of association of each word to each one of the classes and, therefore, to model the probability distribution of the presence or evidence of each class in the texts. This representation, named as Frequency Analysis Interpolation, is used to codify the texts for each user, and then this codified information is used to feed a Support Vector Machines classifier.

– *Attribute selection techniques for classification of aggressive tweets. LyR-UAMC participation at MexA3T 2019 Task* [14]
  - **Task:** Aggressiveness Detection
  - **Team name:** LyR
  - **Features:** Document frequency, mutual information, and lexical Availability.
  - **Classification:** Naïve Bayes.
  - **Summary:** In this work, the authors proposed an approach that follows the traditional pipeline of a non-thematic text classification system. They employed a BOW representation and evaluated the impact of distinct features selection strategies. Their goal was to test if a condensed set of words can be indicative of the aggressiveness of a tweet. They proposed a new criterion to select relevant words: the lexical availability, and reach the following conclusion: different feature selection techniques favor different aspects of the aggressiveness in a short text.

– *Detection of Aggressive Tweets in Mexican Spanish Using Multiple Features with Parameter Optimization* [11]
  - **Task:** Aggressiveness Detection
  - **Team name:** mineriaUNAM
  - **Features:** Linguistically features and different types of n-grams.
  - **Classification:** Support Vector Machine
  - **Summary:** In this work, the authors approached the problem using linguistically motivated features and several types of n-grams (words, characters, functional words, punctuation symbols, among others). They trained a Support Vector Machine using a combinatorial framework that optimizes the results of the classifier.

– *UACh at MEX-A3T 2019: Preliminary results on detecting aggressive tweets by adding author information via an unsupervised strategy* [5]
  - **Task:** Aggressiveness Detection

- **Team name:** UACh
- **Features:** Character n-grams and word embeddings.
- **Classification:** Support Vector Machine and a multilayer perceptron.
- **Summary:** In this paper, the authors considered the application of a traditional classification method to the problem of aggressiveness detection in Spanish tweets. They used two main kinds of features: character n-grams and word embeddings. Then employed two different classifiers, a SVM and a multilayer perceptron. The main idea of their participation was the inclusion of features to try to give context to the text messages and explore if people verbally attack differently depending on their traits and overall environment. The obtained results indicated that adding context features produce almost unnoticeable changes in the performance.

– *Aggressiveness Detection through Deep Learning Approaches* [9]
  - **Task:** Aggressiveness Detection
  - **Team name:** VRAIN
  - **Features:** Hierarchical features obtained by a CNN.
  - **Classification:** CNN, LSTM, GRU.
  - **Summary:** In this paper, the authors explore three deep learning approaches to the task: a convolutional network, a recurrent network and a self-attention network. They did not obtain good results in the test set. They assumed that that was due to the fact that the content of the test data is too different from the training set.

– *Ensemble learning to detect aggressiveness in Mexican Spanish tweets* [8]
  - **Task:** Aggressiveness Detection
  - **Team name:** CEATIC
  - **Features:** Bag of words with term frequency.
  - **Classification:** Support Vector Machine, Logistic Regression, Multinomial Naïve Bayes.
  - **Summary:** In this work, the authors used a traditional BOW representation, considering unigrams and bigrams, and applied a TF weighting. They evaluated multiple classification algorithms, among them Logistic Regression, Multinomial NB and SVM, and proposed an ensemble classifier combining the three best individual algorithms by majority vote.

5 Experimental evaluation and analysis of results

This section summarizes the results obtained by the participants, comparing and analyzing in detail the performance of their submitted solutions. For the final phase of the challenge, participants sent their predictions for the test partition,

Table 7. General description for the different approaches; with the preprocessing, representation (features) and classification of each participant. In the classification section, the C is to indicate a classifier and A to indicate a general approach.

Type	Approach	PRHLT	OGaribo	LyR	mineriaUNAM	UACh	VRAIN	CEATIC	Cerpamid	CIC-VCR
Preprocessing	Lowercase			✓	✓	✓		✓		
	Normalize tweets	✓		✓	✓	✓	✓	✓		
Representation *(features)*	Characters n-grams				✓	✓				
	Words n-grams	✓						✓	✓	
	Aggressive words	✓								
	Word embeddings	✓				✓	✓			
	Statistical descriptors		✓							
	LIWC				✓					
	Hierarchical (over texts)	✓					✓			
	Hierarchical (over images)									✓
Classification	**C** Logistic regression							✓		
	C Naïve Bayes			✓				✓		
	C SVM		✓		✓	✓		✓	✓	
	A Deep-learning	✓					✓			✓
	A Model selection/Ensembles	✓					✓	✓		

the performance on this data was used to rank them. The macro average F1 was used as the main evaluation measure.

For computing the evaluation scores we relied on the EvALL platform [3]. EvALL is an online evaluation service targeting information retrieval and natural language processing tasks. It is a complete evaluation framework that receives as input the ground truth and the predictive outputs of systems and returns a complete performance evaluation. In the following, we report the results obtained by participants as evaluated by EvALL.

As baseline methods, we implemented two popular approaches that have shown to be hard to beat in both tasks: *i*) a classification model trained on the bag of words (BoW) representation, and *ii*) a classifier trained on a character 3-grams representation. Also, we compared the systems' results versus the best results for both tasks in the previous year edition. For author profiling we consider the results from the MXAA approach [10], and for aggressiveness detection we use the results from the INGEOTEC system [7],

For the BOW approach, all the corpus vocabulary was used, but stopwords and special characters were removed. The size of the representation of each text was 14,913. For the 3-grams representation, all 3-grams were used. As in BOW, stopwords and special characters were removed. The size of the representation of each text was 5,212. A SVM with linear kernel and $C = 1$ was applied for classification in both tasks.

5.1 Author profiling results

Table 8 shows a summary of the results obtained by each team in the three AP subtasks as well as their average performance. The average macro F1 was used to rank participants. The approach of the CerpamidTeam (run 1) team obtained the best performance. Nevertheless, this system do not overcome all baselines. In particular, it is noticeable that for the two traits considered in the 2018 edition, i.e., occupation and place of residence, any of the two participant teams was able to improve the results from winner team (MXAA) of the last year's edition.

Table 8. Average Macro $F_1 performance for the three traits in the author profiling task$

Team	Source	Gender	Occupation	Location	$Average - F_1$
Baseline (*MXAA*)	Text	-	**0.51**	**0.83**	-
Baseline (*BoW*)	Text	0.72	0.48	0.63	**0.61**
CerpamidTeam run 1	Text	**0.84**	0.40	0.50	0.58
Baseline (*3-grams*)	Text	0.68	0.42	0.60	0.57
CerpamidTeam run 2	Text	0.83	0.38	0.48	0.56
CIC-VCR run 1	Image	0.52	0.12	0.15	0.26
CIC-VCR run 2	Image	0.47	0.10	0.11	0.23

Tables 9 and 10 show the results obtained by each team for the location and occupation traits respectively. Although we used the macro average of F_1

to rank the participants, we also show the accuracy results as well as the F_1 for each class. For these two particular traits the two participant systems could not outperform any of the proposed baselines.

Table 9. Results for the location trait in the author profiling task.

	Global		Per class performance					
Team	F_{macro}	Accuracy	center	southeast	northwest	north	northeast	west
Baseline (*MXAA*)	*0.83*	*0.86*	*0.87*	*0.81*	*0.86*	*0.78*	*0.90*	*0.75*
Baseline (*BoW*)	*0.63*	*0.75*	*0.79*	*0.60*	*0.78*	*0.32*	*0.83*	*0.45*
Baseline (*3-grams*)	*0.60*	*0.72*	*0.75*	*0.50*	*0.77*	*0.31*	*0.80*	*0.47*
CerpamidTeam run 1	0.50	0.63	0.69	0.39	0.67	0.17	0.71	0.29
CerpamidTeam run 2	0.48	0.62	0.70	0.40	0.67	0.25	0.73	0.26
CIC-VCR run 1	0.15	0.24	0.42	0.04	0.10	0.03	0.03	0.10
CIC-VCR run 2	0.11	0.17	0.36	0.0	0.087	0.05	0.14	0.03

Table 10. Results for the occupation trait in the author profiling task

	Global		Per class performance							
Team	F_{macro}	Accuracy	others	arts	student	social	sciences	sports	admin	health
Baseline (*MXAA*)	*0.51*	*0.74*	*0.04*	*0.51*	*0.91*	*0.69*	*0.47*	*0.49*	*0.59*	*0.38*
Baseline (*BoW*)	*0.48*	*0.71*	*0.15*	*0.48*	*0.90*	*0.61*	*0.37*	*0.52*	*0.54*	*0.23*
Baseline (*Trigrams*)	*0.42*	*0.69*	*0.13*	*0.32*	*0.90*	*0.62*	*0.26*	*0.28*	*0.52*	*0.32*
CerpamidTeam run 1	0.40	0.66	0.10	0.25	0.86	0.57	0.20	0.31	0.51	0.26
CerpamidTeam run 2	0.38	0.66	0.13	0.33	0.86	0.55	0.21	0.35	0.48	0.25
CIC-VCR run 1	0.12	0.27	0.0	0.09	0.45	0.14	0.06	0.0	0.21	0.0
CIC-VCR run 2	0.09	0.23	0.0	0.12	0.44	0.07	0.04	0.0	0.09	0.0

From an overall analysis, it was possible to notice that for all traits the best results corresponded to textual-based solutions. In spite of this general behaviour, we could identify 110 users out of 1500 from the the test set that were correctly classified only by the image-based systems (runs 1 and 2 from the CIC-VCR team). We hypothesize that this result could be caused by the lower number of tweets from these users in comparison to the rest. They have on average 1218 tweets, whereas the average from the complete test set is of 1353 tweets per user.

To analyze the complementarity of the predictions by the two participants, we built a theoretically perfect ensemble from their four runs. That is, we considered that a test instance was correctly classified if at least one of the participating teams (i.e., one of their runs) classified it correctly. Additionally, we considered a majority vote approach; for this we choose the class with the greatest number of predictions among the four runs.

Table 11 shows the results of the perfect ensemble and the majority vote approach, and compares them with the best result obtained for each trait by a

single participant system. From these results, it is possible to observe that the perfect ensemble performance is considerably greater than the best approach for the three traits, suggesting that the two participant systems are complementary to each other. Nevertheless, the bad results from the majority vote approach indicate that the intersection of correctly classified instances by the two systems is quite small, and therefore, that automatically taking advantage of these complementarity is a complex task

Table 11. Combining AP results from the different systems: perfect ensemble and the majority vote approach

Trait	Best approach	Vote	Perfect ensemble
Gender	0.83	0.76	0.97
Location	0.50	0.45	0.71
Occupation	0.39	0.35	0.57

5.2 Aggressiveness identification results

Table 12 shows the results obtained by the participating teams in the aggressiveness detection task. For this task, we sort the teams by their F_1 results in the aggressive class, but for completion we also report the accuracy, the macro F_1 and the F_1 in the non-aggressive class. The approach submitted by the University of Chihuahua (UACh) obtained the best performance, outperforming all proposed baselines, except the results from INGEOTEC, which the winner team in the 2018 edition. Nevertheless, it is important to point out that the UACh approach is considerably much simpler than the one from INGEOTEC.

As previously done with the profiling task, we also built a theoretically perfect ensemble and a majority vote approach from all submitted submissions to the aggressiveness task. Table 13 shows these results. Again, the perfect ensemble shows a very good result, in this case a $F_1 = 0.99$, but also the majority vote approach obtained a low performance, indicating that also for this task it is difficult to find a way to merge the information from the different approaches, even when they are complementary to each other.

As a result of the perfect theoretical ensemble, it was possible to identify those common errors across all the systems. In fact, there are only 9 tweets that no system could classify correctly. All of them are aggressive tweets that were classified as non-aggressive. Below we show some of these tweets, where we can identify ironic comments, the use of out of the training vocabulary words, such as some named entities, as well as offenses with no vulgar or profane words.

- *Y hablando de cosas feas, ¿cómo está tu novia?*
- *A mí más real se me hace Carla Morrison porque está super gorda*
- *Ponete a correr gorda, está bien que las puertas del gimnasio de abren*
- *pero va a llegar el momento en que no vas a pasar el mejor profff????*
- *Tu siempre tan tonta Viviana ????*

Table 12. Results for the aggressiveness identification task

Team	Accuracy	F_{macro}	Aggresive	Non aggressive
UACh	**0.73**	**0.65**	**0.48**	**0.82**
Baseline (*INGEOTEC*)	*0.73*	*0.65*	*0.48*	*0.81*
PRHLT-run2	0.70	0.63	0.47	0.79
PRHLT-run4	0.69	0.62	0.46	0.78
mineriaUNAM-run2	0.71	0.63	0.45	0.80
mineriaUNAM-run1	0.71	0.63	0.45	0.80
PRHLT-run3	0.65	0.59	0.44	0.74
PRHLT-run1	0.65	0.59	0.44	0.74
Baseline (*Trigrams*)	*0.69*	*0.60*	*0.43*	*0.79*
LyR-run3	0.69	0.61	0.43	0.79
LyR-run6	0.68	0.59	0.42	0.77
VRAIN-run1	0.50	0.49	0.41	0.57
OscarGaribo-run1	0.68	0.59	0.40	0.79
LyR-run5	0.67	0.59	0.38	0.79
LyR-run2	0.70	0.55	0.38	0.77
LyR-run1	0.65	0.57	0.38	0.76
Baseline (*BoW*)	*0.68*	*0.58*	*0.37*	*0.78*
OscarGaribo-run2	0.67	0.57	0.37	0.77
LASTUS-UPF-run2	0.60	0.52	0.32	0.72
LASTUS-UPF-run1	0.58	0.50	0.30	0.70
CEATIC	0.72	0.56	0.30	0.82
VRAIN-run2	0.61	0.51	0.29	0.73
Aspie96-run2	0.63	0.52	0.29	0.75
LyR-run4	0.66	0.57	0.28	0.81
hzegheru	0.60	0.50	0.28	0.73
Aspie96-run1	0.68	0.53	0.27	0.79

Table 13. Results of majority vote and perfect ensemble for aggressiveness identification

Best approach	Vote	Perfect ensemble
0.47	0.40	0.99

6 Conclusions

This paper described the design and results of the MEX-A3T shared task collocated with IberLef 2019. MEX-A3T stands for *Authorship and Aggressiveness Analysis in Mexican Spanish Tweets*. Two tasks were proposed, one targeting author profiling and the other focused on aggressiveness detection. Mainly, given a set of tweets in Mexican Spanish, the participants had to identify the gender, location and occupation of their authors as well as the aggressive messages. For these tasks we employed the same data sets than from the previous MEX-A3T edition, but we extended the author profiling collection by including eleven images for each user, with the aim of evaluating multimodal profiling approaches. The shared task lasted more than two months and attracted the participation of nine teams from three different countries, Mexico, Spain and Cuba.

A variety of methodologies were proposed by the participants, from traditional supervised methods to deep learning approaches. For author profiling, the approach proposed by the Cerpamid team obtained the best results with an approach based on dimentionality reduction in text. However; their results did not overcome the best results from the previous year edition.For aggressiveness identification, the top-ranked team was UACh with an apporach based on two main kinds of features: character n-grams and word embeddings. Their results were equal to the previous year winner but employing a simpler approach.

In general terms, the competition was a success: the solutions proposed by nine participants were diverse regarding methodologies and performances, and new insights on how to deal with tweets on Mexican Spanish were obtained. Among the most interesting findings was the complementarity of the predictions from the different participants, a phenomenon that was also observed in the previous edition. This opens the possibility to study how to take advantage of the different information extracted by the teams in such a way that results reach those from the perfect ensemble.

Acknowledgements Our special thanks go to all of MEX-A3T's participants. We would like to thank CONACyT for partially supporting this work under grants CB-2015-01-257383, FC-2016-2410 and the Thematic Networks program (Language Technologies Thematic Network). The first author thanks for doctoral scholarship CONACyT-Mexico 654803 and the second for doctoral scholarship CONACyT-Mexico 401887.

References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September (2018)

2. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. pp. 151–162. Springer (2016)
3. Amigó, E., Carrillo-de Albornoz, J., Almagro-Cádiz, M., Gonzalo, J., Rodríguez-Vidal, J., Verdejo, F.: Evall: Open access evaluation for information access systems. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1301–1304. ACM (2017)
4. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN- **23**(3), 321–346 (2003)
5. Casavantes, M., López, R., González, L.C.: Uach at mex-a3t 2019: Preliminary results on detecting aggressive tweets by adding author information via an unsupervised strategy. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
6. Castro Castro, D., Artigas Herold, M.F., Ortega Bueno, R., Muñoz, R.: Cerpamidua at mexa3t 2019: Transition point proposal. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
7. Graff, M., Miranda-Jiménez, S., Tellez, E.S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., Sánchez, C.N.: Ingeotec at mex-a3t: Author profiling and aggressiveness analysis in twitter using μtc and evomsa. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
8. Molina-González, M.D., Plaza-del Arco, F.M., Martín-Valdivia, M.T., Ureña López, L.A.: Ensemble learning to detect aggressiveness in mexican spanish tweets. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
9. Nina-Alcocer, V., González, J.Á., Hurtado, L.F., Pla, F.: Aggressiveness detection through deep learning approaches. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
10. Ortega-Mendoza, R.M., López-Monroy, A.P.: The winning approach for author profiling of mexican users in twitter at mex.a3t@ibereval-2018. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
11. Ortiz, G., Gómez-Adorno, H., Reyes-Magaña, J., Bel-Enguix, G., Sierra, G.: Detection of aggressive tweets in mexican spanish using multiple features with parameter optimization. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
12. Garibo i Orts, O.: Aggressiveness identification in twitter at iberlef2019: Frequency analysis interpolation for aggressiveness identification. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
13. De la Peña Sarracén, G.L., Rosso, P.: Aggressive analysis in twitter using a combination of models. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)
14. Ramírez-de-la Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H.: Attribute selection techniques for classification of aggressive tweets lyr-uamc participation at mexa3t 2019 task. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)

15. Tellez, F.P., Pinto, D., Cardiff, J., Rosso, P.: Defining and evaluating blog characteristics. In: Artificial Intelligence, 2009. MICAI 2009. Eighth Mexican International Conference on. pp. 97–102. IEEE (2009)
16. Valdez-Rodríuez, J.E., Calvo, H., Felipe-Riverón, E.M.: Author profiling from images using 3d convolutional neural networks. In: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings (2019)