Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets

Miguel Á. Álvarez-Carmona[1], Estefanía Guzmán-Falcón[1], Manuel Montes-y-Gómez[1], Hugo Jair Escalante[1], Luis Villaseñor-Pineda[1,4] Verónica Reyes-Meza[2], and Antonio Rico-Sulayes[3]

[1] Language Technologies Lab., Computational Sciences Department, Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Mexico.
[2] Centro Tlaxcala de Biología de la Conducta, Universidad Autónoma de Tlaxcala, Mexico
[3] Universidad de las Américas Puebla, Mexico
[4] Centre de Recherche en Linguistique Française GRAMMATICA (EA 4521), Université d'Artois, France
{miguelangel.alvarezcarmona,fany.guzman,mmontesg,hugojair,villasen}@inaoep.mx, veronica.reyesm@uatx.mx, antonio.rico@udlap.mx

Abstract. This paper presents the framework and results from the MEX-A3T track at IberEval 2018. This track considers two tasks, author profiling and aggressiveness detection, both of them using Mexican Spanish tweets. The author profiling task aims to identify the place of residence and occupation of Twitter users. On the other hand, the aggressiveness detection task aims to discriminate between aggressive and non-aggressive tweets. For these two tasks we have built new corpora considering tweets from Mexican Twitter users. This paper compares and discusses the results from the participants.

1 Introduction

Nowadays there is a tremendous amount of information available on the Internet. Specifically, social media platforms such as Twitter are constantly growing thanks to the information generated by a huge community of active users. The analysis of shared information has became very relevant for several applications in security, marketing and forensics, among others.

One key task for social media analysis is *author profiling* (AP), which consists in predicting general or demographic attributes of authors such as gender, age, personality and mother tongue, by examining the content of their posts [1, 5]. On the other hand, *detecting aggressive content* targeted to people or vulnerable groups is also a task of great relevance to prevent possible viral destructive behaviors through social networks.

These two tasks have been recently studied in different academic forums, most of them focused on English language [21, 26, 27]. Although some have included

Spanish data, they have mainly included profiles from the peninsular Spanish variant [23].

The objective of the MEX-A3T is to encourage research on the analysis of social media content in Mexican Spanish. Particularly, it aims to push research in the treatment of a variety of Spanish that has cultural traits that make it significantly different from peninsular Spanish. In addition, it considers two dimensions of author profiling that have not been studied deeply by the community: occupation and place of residence. Most research so far has focused on age and gender, although useful, the considered dimensions are more challenging and could have a greater applicability.

To evaluate these tasks, we have built two ad hoc collections. The first one is an author profiling corpus consisting of 5 thousand Mexican users. This corpus is labeled for the subtasks of 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 some previous evaluation forums for the tasks of author profiling and aggressiveness identification; Section 3 presents the evaluation framework used at MEX-A3T; Section 4 shows an overview of the participating approaches; Section 5 reports and analyses the results obtained by participating teams. Finally, Section 6 draws the conclusions of this evaluation exercise.

2 Related Work

This section presents a brief description of some previous evaluation forums for the author profiling and aggressiveness detection tasks.

2.1 Author profiling task

The PAN lab at CLEF[5] is the most well know evaluation forum for the author profiling task. At its first edition, in 2013, it considered the identification of age and gender of users from English and Spanish blogs [20]. Then, in the following editions, it included other aspects such as different social media [19], additional traits as personality [22], and more languages [22, 23]. The PAN 2017 edition is of particular relevance for our track since it considered the task of language variety identification including the Mexican variant from the Spanish. Nevertheless, it did not approach the profiling of other traits for the considered Mexican users.

The RepLab at CLEF has also considered the author profiling task. In its 2014 edition[6], the participants were asked to categorize English Twitter profiles by type of author (Company, Professional, Celebrity, Employee, Stockholder, Investor, Journalist, Sportsman, Public Institution, and Non-Governmental Organization) as well as rank them according to their influence [3].

[5] https://pan.webis.de/clef18/pan18-web/index.html
[6] http://nlp.uned.es/replab2014/

Among other efforts to evaluate the author profiling task, there is a forum oriented to mother language identification. In its first edition in 2013 it considered the analysis of essays written during college-entrance test [25]; then, in a subsequent edition it also considered information from the students records [13].

2.2 Aggressiveness detection task

The detection of aggressive language is considered a central task to tackle more complex problems such as cyberbulling, hate speech and harassment, among others. The growing interest in these issues has prompted the recent organization of workshops and forums aimed at finding relevant solutions for their detection.

In 2017, it was organized the first Workshop on Abusive Language Online (ALW1) [27], where were presented different approaches for the detection of abusive language in social media. In the context of this workshop it was also carried out a task focused on the detection of abusive language in English and German messages.

Another forum related to these kind of tasks was the Workshop of Discourses of Aggression and Violence in Greek Digital Communication [12]. This forum explored the multifaceted relationship between language and aggression/violence, with a particular focus on the discourse of Greek users of social media.

It was also recently organized the Workshop on Online Harassment at the Conference on Human Factors in Computing Systems (CHI) [6]. The purpose of this workshop was to build an online harassment corpus. The participants could share vocabulary, hashtags, and features of harassment.

Finally, there was a Workshop on Trolling, Aggression and Cyberbullying at COLING 2018 [15]. This workshop focused on the phenomena of online aggression, trolling, cyberbullying and other related aspects, in text as well as in speech.

It is important to highlight that non of these workshops were especially focused on the analysis of aggressive language in Spanish.

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 Mexican corpus for author profiling

In order to study the characteristics of the different Mexican Twitter profiles, we built a Mexican corpus for author profiling. Each of the authors (social media users) was labeled with occupation and place of residence information. For the occupation label, we considered the following 8 classes: *arts, student, social, sciences, sports, administrative, health,* and *others*. For the place of residence trait, we considered the following 6 classes: *north, northwest, northeast, center, west, and southeast.*

Construction of the corpus. Two labelers, working three months each, were needed for building this corpus. They applied the following methodology: (i) to find a set of Twitter accounts corresponding to famous persons and/or organizations from each region of interest. These accounts usually were from local civil authorities, known restaurants, and universities; (ii) to search for followers of the initial accounts, assuming that most of them belong to the same region than the initial accounts; (iii) to select only those followers that explicitly mention, in Twitter or in other social network (as Facebook and Instagram) their place of residence and occupation. Table 1 shows some examples of tweets where users reveal information from their place of residence and occupation.

Table 1. Example of tweets mentioning information related to the place of residence and/or occupation of users.

Trait detected	Original text	Translation
Residence	La pura carnita asada en **Monterrey**	Roast beef in **Monterrey**
Residence	Nunca me canso de pasear en el zócalo de **Puebla**	I never get tired of walking in the **Puebla** Zocalo
Occupation	Porque los **arquitectos** nunca descansamos	Because we, the **architects** never rest
Occupation	**Programando** en el trabajo ando	**Programming** at work

Statistics. The corpus consists of 5 thousand profiles from Mexican Twitter users. Each profile is labeled with information about the occupation and place of residence of the user. For the MEX-A3T evaluation exercise, the corpus was divided in two parts, one for training and other for test. 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 partitions are very similar. The majority class correspond 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 for 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 both tables, 2 and 3, the class imbalance was calculated as proposed in [24]. The place of residence trait shows a value of 396.1, while the occupation trait has a value of 502.42. Considering that 0 represents a perfect balance,

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

these numbers indicate that the imbalance is bigger for the occupation trait, and therefore, that it could be more complex to be predicted than 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 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.

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

3.2 A Mexican corpus for aggressiveness identification

As mentioned in Section 2 the study of aggressive language in social media has been done mainly in English language; hence, there is a lack of resources to study this kind of language in Spanish. Motivated by this situation we built a Mexican corpus for aggressiveness detection in Twitter. We considered Twitter as the source media since it is open and its anonymity allows people to write judgments or assessments about other people, including offenses or aggressions.

To build the corpus, we collected tweets for three months. We used rude words and controversial hashtags to narrow the search. The hashtags were related to topics of politics, sexism, homophobia, and discrimination. To ensure their origin, the tweets were collected taking into account their geolocation. We considered Mexico City as the center and extracted all tweets that were within a radius of 500km. Finally, the collected tweets were labeled by two persons. At the end each tweet of the corpus was labeled as *aggressive* or *non-aggressive.*

Labeling methodology. It does not exist a formal definition of aggressive language, but there have been reported several of its characteristics [7, 17] that help us to build one base definition. Mainly, we considered that *a message is aggressive if the purpose is to humiliate, belittle, discredit a person or group of people using rude words or pejorative language.* Supported on this definition, we created the next rules to help our labelers to classify the tweets:

- **Rule 1.** An aggressive text can contain one or all of the following elements:

- **Nicknames** Nickname that the author will assign to the person who is directed the message, alluding to a disability or defect.
- **Jokes** Jokes are messages that will always be considered offensive, when the intention is to humiliate or attack the recipient.
- **Derogatory adjectives** Words that describe a person with the intention to humiliate and belittle.
- **Rudeness** As long as they are not used in an informal context.

– **Rule 2.** Tweets that are aggressive to objects are not considered as aggressive texts.
– **Rule 3.** Tweets where the author is self-attacking are not considered as aggressive texts.
– **Rule 4.** Tweets where there are ads for pornography and prostitution are not considered as aggressive texts.
– **Rule 5.** Tweets that tell what a person said in any situation, recounting an conversation or citing an aggression, are not considered as aggressive texts.

Table 5 shows some examples labeled by using these rules. As can be intuited, the task of labeling aggressiveness is very difficult, specially because in most of the cases it is necessary to interpret the message in a given context. Considering this, we applied a pilot labeling to train our labelers as well as clarify and verify the initial rules. Table 6 shows the kappa in the pilot and final labeling phases.

Table 5. Tweets extracted using the rules.

Aggressive	Not 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

Table 6. Labeling agreement before and after the training phase.

Sessions	Kappa
Pilot labeling	0.4240
Final labeling	0.5867

Statistics. The collected corpus consists of 11 thousand tweets. For the MEX-A3T evaluation exercise, the corpus was divided in 2 parts, one for training and other for test. Table 7 shows the distribution of this corpus. It is noticed that the non-aggressive class is the majority class in both partitions.

Table 7. 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 both traits, place of residence and occupation, as shown in Formula 1.

$$F_{average} = \frac{F_{macro}(C_{location}) + F_{macro}(C_{occupation})}{2} \tag{1}$$

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 the trait.

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

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, eight teams have submitted their solutions, of which, 1 participated in the author profiling task, 4 participated in the aggressiveness identification task and 3 participated in both tasks. On the basis of what they explained in their notebook papers, this section presents a summary of their approaches in terms of preprocessing steps, features, and classification algorithms.

The participating methods are listed below:

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

- *CIC-GIL Approach to Author Profiling in Spanish Tweets: Location and Occupation* [14]
  - **Task:** Author profiling.
  - **Team name:** CIC-GIL
  - **Preprocessing:** All letters converted to lowercase, normalize digits, user mentions, hashtags, picture links and urls; replace slang words by their standardized version.
  - **Features:** Typed character n-grams, function-word n-grams, and regionalisms, with tf weighting.
  - **Classification:** logistic regression algorithm (but also SVM and Bayes)
  - **Summary:** This paper presents the CIC-GIL approach for the identification of location and occupation of Twitter users from Mexico. This approach follows the traditional supervised methodology for a multi-class classification task. On the one hand, it considers a set of handcrafted features to represent the tweets from each user. These features include typed character n-grams, as well as function word n-grams and regionalisms for the location identification subtask. Then, based on this representation, it trains a logistic regression algorithm. The results are encouraging, 73.63 F1-macro score for location and 48.94 for occupation; they corroborate the appropriateness of (typed) character n-grams for authorship related tasks, given their capability to capture different levels of information.

- *Machine Learning Approach for Detecting Aggressive tweets in Spanish* [10]
  - **Task:** Aggressiveness identification.
  - **Team Name:** Nochebuena
  - **Preprocessing:** All letters converted to lowercase, digits were replaced by a single symbol, the "@" was removed from mentions, links to pictures were replaced by a specific character. Slang words were replaced by their standardized version.
  - **Features:** Language patterns, character, and word n-grams, POS tags of n-grams and aggressive word n-grams.
  - **Classification:** Logistic regression + SMOTE (oversampling method)
  - **Summary:** The authors implemented a standard pipeline comprising preprocessing, feature engineering and classification. Two runs were submitted, one using SMOTE and another one without any oversampling strategy. In their cross-validation experiments authors achieved quite high performance (above 70 %), however, test set performance was lower than the baseline. We hypothesize that authors overfitted the training set.

- *Deep analysis in aggressive Mexican tweets* [9]
  - **Task:** Aggressiveness identification.
  - **Team name:** SimSom
  - **Features:** Textual features, the density of written knowledge, style, writing density, bag of words, list of aggressive words, syntactic patterns,

affective features for Spanish (SEL). Deep learning was used for feature extraction.
  - **Classification:** Deep neural network model is used for classification and feature extraction.
  - **Summary:** The authors describe a deep learning based method for detecting aggressive tweets. Two runs were submitted by the authors, in the first one, handcrafted features are combined with the deep learning model. In another one, only learned features were considered for classification. The performance of both methods was very similar. Yet, performance was extremely low.

– *Attention mechanism for aggressive detection* [16]
  - **Tasks:** Aggressiveness identification.
  - **Team name:** CGP
  - **Preprocessing:** The authors clean the tweets, eliminating links and urls. The emoticons are replaced with words expressing the corresponding sentiment. Each tweet is analyzed with FreeLing, and each word is replaced with its corresponding lemma.
  - **Features:** The tweets are represented as vectors using a word embedding model. The model was generated by using the word2vec algorithm from the Wikipedia collection in Spanish.
  - **Classification:** The authors use a model consisting of a Bidirectional LSTM neural network at the word level. Later, an attention layer is applied over each hidden state. The attention weights are learned using the concatenation of the current hidden state of the Bi-LSTM and the past hidden state of a second LSTM (a Post-Attention LSTM). Finally, the target aggressiveness of the tweet is predicted by this Post-Attention-LSTM network.
  - **Summary:** The authors proposed a method using an Attention-based Long Short-Term Memory Recurrent Neural Network with an attention mechanism. The authors submitted two runs. The difference between them is that the second run includes a linguistic characteristic indicating the occurrence of vulgar or obscene phrases. This second run demonstrated the relevance of this type of information, achieving high results.

– *Linguistic generalization of slang used in Mexican tweets, applied in aggressiveness detection* [8]
  - **Tasks:** Aggressiveness identification.
  - **Team name:** SEAL-UPV
  - **Preprocessing:** Before tokenization, the authors replace all the different ways to express the same insult (i.e. *hijo de puta*, *hijo de su reputa madre*, etc.) for a simplified form (v.gr. **hdp**). The same was done for other insults expressions. Also, during preprocessing, a normalization process was applied for laughs and question marks.

- **Features:** The authors applied the PCA technique reducing the original dimensions (word forms).
- **Classification:** The authors experiment with different classification algorithms including SVM, NN, Decision Trees, Nave Bayes and K-NN.
- **Summary:** The authors applied the Principal Component Analysis (PCA) to reduce the feature space, eliminating the non-informative words. They experimented with different automatic learning algorithms and showed results by applying or not applying the normalization of insults and other expressions. The results submitted were calculated by performing normalization and training with SVM. These results were superior to the traditional BOW approach and demonstrate the importance of dimensionality reduction, even though the use of simplified forms of insults is not conclusive.

– *INGEOTEC at MEX-A3T: Author profiling and aggressiveness analysis in Twitter using μTC and EvoMSA* [11]

- **Tasks:** Author profiling and aggressiveness identification.
- **Team name:** INGEOTEC
- **Preprocessing:** Stemming.
- **Features:** For author profiling the author used: character n-grams, word n-grams, skip-grams, with tf and tfidf weights. On the other hand, for aggressiveness identification they used: character n-grams, word n-grams, but also word embeddings and tailor-made lexicons.
- **Classification:** For author profiling the authors applied SVM classifier with a linear kernel. Nevertheless, for aggressiveness identification they applied an ensemble of different classifiers.
- **Summary:** This paper presents two different systems to tackle the author profiling and the aggressive text detection tasks: microTC and EvoMSA, respectively. MicroTC is a text classification approach supported on model selection techniques. It mainly builds text classifiers searching for the best models in a given configuration space, consisting of several preprocessing functions, different tokenizers (i.e., kind of features, such as word and character n-grams) and weighting schemes. In all the cases, it uses a SVM with linear kernel as classifier. On the other hand, the EvoMSA is an ensemble approach that combines the decisions from different models to produce a final prediction. In particular, for the aggressiveness detection subtask, EvoMSA considers the decisions from MicroTC, from a lexicon-based model that takes into account the presence of aggressive and affective words, and from a model based on the fastText representation of texts. Results show to be very competitive for both tasks, indicating that learning specific models for the recognition of each user category is a good idea.

- *Author Profiling and Aggressiveness Detection in Spanish Tweets: MEX-A3T 2018* [4]
  - **Tasks:** Author profiling and aggressiveness identification.
  - **Team name:** Aragon-Lopez
  - **Features:** Bag of Terms, Second Order Attributes, words and Characters N-Grams. They selected the most important features with the χ^2 distribution.
  - **Classification:** CNN Models as CNN-Rand, CNN-Static and CNN-NonStatic.
  - **Summary:** The authors used some different representations that have been useful in the author profiling task for others forums evaluation. The used the bag of terms and the second order attributes (SOA). SOA has obtained the best result throughout 3 editions of the PAN. Nevertheless, the best results are obtained by the n-grams ensemble. The authors separated the training corpus in 70 % and 30 % for training and test respectively. They used a n-gram representation and it can observe that this representation gets the best results in the three different tasks. The authors conclude that this representation captures important words for the classification especially in the aggressive class where the words show a clear aggressiveness.

- *The Winning Approach for Author Profiling of Mexican Users in Twitter at MEX.A3T@IBEREVAL-2018* [18]
  - **Tasks:** Author profiling and aggressiveness identification.
  - **Team name:** MXAA
  - **Features:** The authors used a technique called discriminative personal purity (DPP). DPP consists of two components: first, a descriptive factor, defined as the maximum value of the function categorical personal purity, that captures the capability of a term to describe personal information of authors belonging to the category; and second, a discriminative factor, based on the *gini* coefficient for scoring the ability of the term to discriminate among the different profiles.
  - **Classification:** Support Vector Machine with L2 normalization.
  - **Summary:** The aim of the authors is using feature selection and term weighting strategies that emphasize the value of personal information for building the text representation which feeds the machine learning algorithms. The base of these strategies is a measure called Personal Expression Intensity (PEI), which determines the amount of personal information revealed by each term. In general, they used a combination of content and style attributes, which include unigrams of content words, punctuation marks, slang words and out-of-dictionary terms like emoticons. They also considered the occurrences of function words. By means of the n top terms according to DPP, they built a standard BoW representation where the weights of the terms are estimated with the DPP scheme. The results indicate that the approach appears to be effective

in AP for Spanish supporting the idea that personal phrases (sentences having a first-person pronoun) integrate the essence of texts for the AP task. On the other hand, for the aggressiveness identification task, the results showed that the proposed approach, configured with word unigrams, has lower performance than the baseline which considers word sequences.

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 partitions, the performance in these data was used to rank participants. Average of macro F-measure performance was used as the main evaluation measure to rank participants.

For computing the evaluation scores we relied in the EvALL platform [2]. EvALL is an online evaluation service targeting information retrieval and natural language processing tasks. It is a very complete evaluation framework that receives as input the ground truth and 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 systems we implemented two popular approaches that have proved to be hard to beat for both tasks: (i) a classification model trained on the bag of words (BoW) representation and another classifier trained on 3-grams of characters (Trigrams) representation.

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

5.1 Author profiling results

First we analyze the author profiling performance. Table 8 shows a summary of results obtained by each team and for both tasks, as well as the average between location and occupation traits. The latter is evaluation measure used to rank participants. The approach of the Aragon-Lopez (run 1) team obtained the best performance for the location trait, while the method of the MXAA team was the best for the occupation trait. In average, the MXAA team was the top ranked team for the author profiling task. In general terms all systems but Aragon-Lopez (run 2) outperformed the baselines, evidencing the success of participants and the feasibility of the proposed task.

Table 9 shows the results obtained by each team for the location trait of the author profiling task. Although we used F_{macro} for ranking participants, we also show accuracy and micro F-measure for each class.

Table 8. Average Macro F-measure performance for both traits in the author profiling task

Team	Occupation	Location	$F_{average}$
MXAA	**0.5122**	0.8301	**0.6711**
Aragon-Lopez (run 1)	0.4910	**0.8388**	0.6649
INGEOTEC	0.4470	0.8155	0.6312
CIC-GIL (run 2)	0.4894	0.7363	0.6128
CIC-GIL (run 1)	0.4727	0.7310	0.6018
BoW	0.47675	0.6295	0.5531
Trigrams	0.41875	0.6004	0.5095
Aragon-Lopez (run 2)	0.3824	0.619	0.5007

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
Aragon-Lopez (run 1)	**0.838**	**0.879**	**0.884**	**0.821**	**0.889**	0.727	**0.932**	**0.776**
MXAA	0.830	0.858	0.874	0.812	0.862	**0.782**	0.900	0.748
INGEOTEC	0.815	0.856	0.867	0.811	0.8826	0.736	0.904	0.690
CIC-GIL (run 2)	0.736	0.798	0.835	0.703	0.807	0.620	0.853	0.598
CIC-GIL (run 1)	0.731	0.798	0.833	0.686	0.800	0.607	0.859	0.599
Baseline (*BoW*)	0.629	0.746	0.788	0.605	0.783	0.325	0.827	0.449
Aragon-lopez (run 2)	0.619	0.709	0.752	0.518	0.778	0.542	0.808	0.314
Baseline (*3-grams*)	0.601	0.718	0.750	0.504	0.769	0.308	0.805	0.466

The approach of the Aragon-Lopez team (run 1) obtained the best overall performance, with a F_{macro} higher than 0.83. This submission consistently outperformed every other submitted run in all but the north location trait, where the best performance was obtained by the MXAA team. In fact, MXAA obtained a very similar performance as the top ranked team. All teams outperformed the baselines (except run 2 from the top ranked team), showing the feasibility of the proposed task.

In terms of location traits, it can be seen that the class with higher performance was *northeast*, where three teams obtained performance higher than 0.9. On the other hand, the class with lowest performance was *north*, where no team reached a F-measure of 0.80. The latter can be due to the fact that this was the class with the smallest number of samples.

Table 10 shows the results for the occupation trait in the author profiling task. In this trait, the approach of the MXAA team obtained the best F_{macro} performance (0.51). This run obtained the best results for the *arts, social, sciences* and *administrative* classes, whereas the Aragon-Lopez run achieved the best performance for the *student, sports* and *health* classes; the best results of the *others* class was obtained by the 2 runs of the CIC-GIL team.

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
MXAA	**0.512**	**0.744**	0.045	**0.507**	0.915	**0.689**	**0.474**	0.488	**0.590**	0.385
Aragon-Lopez (run 1)	0.491	0.737	0.000	0.451	**0.921**	0.664	0.372	**0.555**	0.568	**0.393**
CIC-GIL (run 2)	0.489	0.726	**0.153**	0.486	0.904	0.636	0.370	0.476	0.584	0.303
Baseline (*BoW*)	0.476	0.709	0.150	0.485	0.905	0.611	0.373	0.522	0.536	0.232
CIC-GIL (run 1)	0.472	0.718	**0.153**	0.469	0.905	0.624	0.333	0.4091	0.5613	0.3235
INGEOTEC	0.447	0.717	0.069	0.444	0.891	0.630	0.326	0.322	0.558	0.333
Baseline (*Trigrams*)	0.418	0.692	0.130	0.316	0.902	0.622	0.264	0.278	0.521	0.317
Aragon-Lopez (run 2)	0.382	0.669	0.095	0.298	0.902	0.640	0.263	0.243	0.444	0.170

From Table 10 it can be seen that this is problem is harder than the location task. In fact the performance across classes is quite diverse. The class with highest performance was *student* with all but one team above 0.9 F-measure, this is not surprising as this is the majority class in the data set with almost 50% of the profiles. The class with lower performance was the *others* class with 0.15, this can be due to the fact that it is one of the minority classes with a little more than 2 % of the profiles and also it could be that this is a heterogeneous class, as it comprises profiles from any occupation no considered in the other classes.

Unlike the location trait, for occupation only the approaches of MXAA, Aragon-Lopez (run 1) and CIC-GIL (run 2) outperformed the baselines. CIC-GIL (run 1) and the INGEOTEC teams overcome only the trigrams baselines.

Finally, the approach of Aragon-Lopez (run 2) was outperformed by both baselines.

In order to further analyze the results obtained by the participants, Figure 1 shows the distribution of F_{macro} performance across all submitted runs associated to the location trait. It can be seen that participants obtained results between 0.30 and 0.93. Also, it is possible to confirm that the highest deviations were for the *north* and *west* classes, which are of the classes less represented in the data set. The categories in which most teams performed well were *center, northwest* and *northeast* which were the 3 classes with more samples. Hence, the sample size was the main factor that determined the success of evaluated methods.

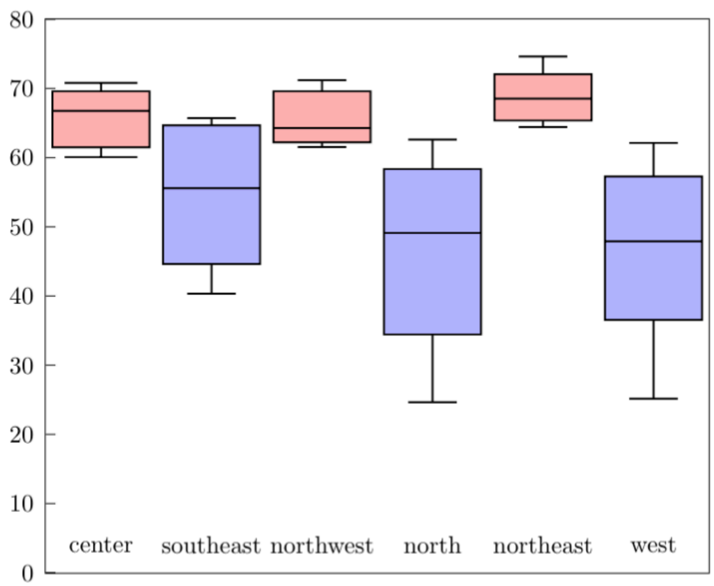

Fig. 1. F_{macro} distribution of results for the location class.

On the other hand, Figure 2 shows the distribution of results from participants for the occupation class. It can be seen that performance was quite varied across different occupation traits, the results rage 0 and 0.92. As previously mentioned, *others* was the most difficult class for all teams, whereas student the simplest: all teams succeeded. The highest deviation in performance was obtained for the *sport* class.

Figure 3 left shows the average confusion matrix over all participating teams for the location trait in author profiling. Each (i, j) position represents the percentage of the instances of the class i classified as the class j. In the heat map is possible to see that the most confusion appears in tweets from the *west* class, which are confused, mainly with the *center* trait with more than 25%. Also the

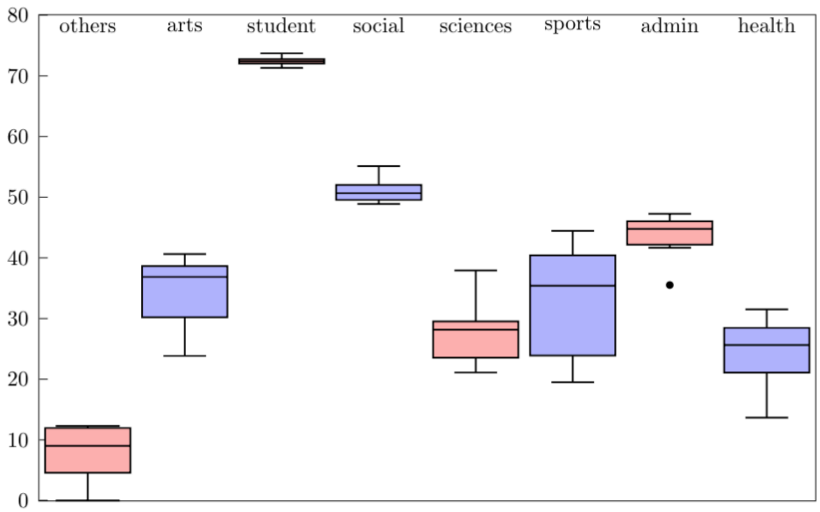

Fig. 2. F_{macro} distributions of teams performance for the occupation results

center class is confused with *north* and *southeast* with 22.55% and 19.78% respectively. In the main diagonal the best performance was obtained by the center class with more than 87%.

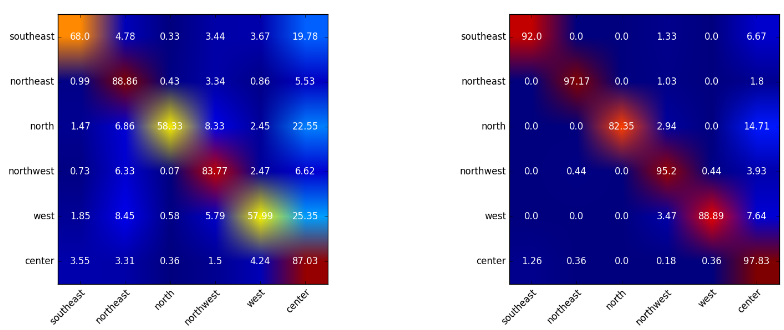

Fig. 3. Heat map of the confusion matrix average for the location results

In order to analyze the complementariness of predictions by participants, we built a theoretically perfect ensemble from the predictions of all participants. Basically, we say a test instance is correctly classified if at least one the participating teams classified it correctly. If an instance is not correctly classified by any team, that instance is assigned to the class with more predictions among the teams. The right plot in Figure 3 shows the corresponding confusion matrix. It can be seen that in the diagonal, only the *west* and *north* classes did not make it over 90%. The perfect ensemble would get a F_{macro} of 0.94, which is considerably higher than that achieved by the top ranked team, (0.83). This result confirm there is a considerable complimentary among predictions of participant teams, and that it is possible to further push performance in order to solve the task.

On the other hand, Figure 4 shows the corresponding confusion matrices for the occupation trait in the author profiling task. In the main diagonal of the left plot the best performance is obtained by the student class with more than 90%. The matrix corresponding to the perfect ensemble (right) shows a

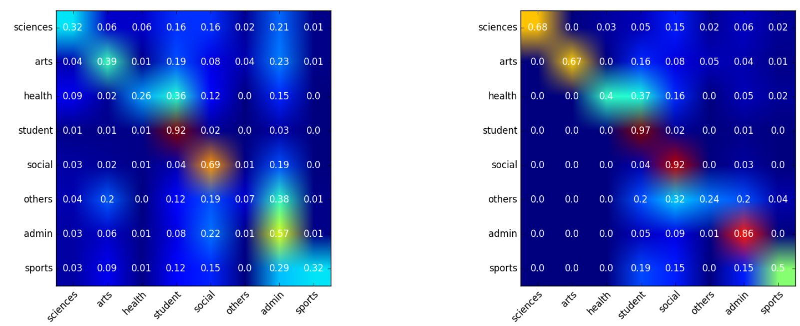

Fig. 4. Heat map of the confusion matrix average for the occupation results

considerable improvement in its main diagonal where the 3 majority classes obtained a performance above 85%. Recall the top ranked team in this task achieved a F_{macro} of 0.51, whereas this artificial ensemble could obtain up to 0.70. Hence, it is worth studying ensemble construction methods for further boosting performance in this task.

5.2 Aggressiveness detection results

We now analyze the results obtained in the aggressiveness detection task. Table 11 shows the results obtained by the participating teams for this problem sorted by the F-measure results of the aggressive class (leading evaluation measure). For completion, we also report accuracy, F_{macro} and performance on the non-aggressive class.

It can be seen that the best performance was obtained by the INGEOTEC team, overcoming the other approaches with an F-measure in the aggressive class higher than 0.48 and a F_{macro} higher than 0.62. Overall, the performance in reported in Table 11 show that detecting aggressiveness in tweets is a very complicated task. Only three systems, INGEOTEC, CGP (run 2) and Aragon-Lopez, outperformed both baselines, while Nochebuena, MXAA and SEAL-UPV teams overcome the BoW baseline and both baselines outperformed the SimSom and CGP (run 1) teams. The F_{macro} is slightly higher than 0.5, with about half of the teams obtaining performance higher than 0.6. In particular, the aggressive

Table 11. Results for the aggressiveness identification task

Team	Accuracy	F_{macro}	Aggresive	Non aggressive
INGEOTEC	0.667	**0.620**	**0.488**	0.753
CGP (run 2)	0.667	0.605	0.450	0.761
Aragon-Lopez	0.711	0.619	0.431	0.806
Baseline (*Trigrams*)	0.688	0.608	0.430	0.786
Nochebuena (run 2)	0.664	0.595	0.428	0.762
MXAA	0.7136	0.614	0.419	0.809
SEAL-UPV	0.694	0.598	0.402	0.794
Nochebuena (run 1)	0.691	0.596	0.401	0.792
Baseline (*BoW*)	0.677	0.576	0.369	0.783
SimSom (run 2)	0.670	0.558	0.336	0.780
SimSom (run 1)	0.586	0.509	0.315	0.703
CGP (run 1)	**0.764**	0.583	0.309	**0.858**

class was much more complicated than the non aggressive one. Regarding the latter class, the CGP team (run 1) obtained the best performance, although they also obtained the lowest score for the aggressive class.

As previously done with the profiling task, we also build a theoretically perfect ensemble with all submissions submitted to the aggressiveness task. We found that it is possible to obtain an F-measure in the aggressive class of up to 0.92 with such artificial ensemble, that is, it is almost as twice the performance obtained by the top ranked team (0.48). This is a very interesting result suggesting the outputs of the system are complimentary to each other and motivating further research on ensemble learning for aggressiveness detection.

As a result of the theoretical perfect ensemble, it was possible to identify those recurrent errors in all the runs. Based on a short analysis, some of the errors are due to:

- Offenses where vulgar or rude language is not used. In this type of messages, the offense is not identified, since rude or vulgar words are not used explicitly. These messages were wrongly classified as non-offensive by all systems:
  - *Hable español vieja payasa! La que es payasa cae gorda!*
  - *Esa chava está bien prieta y fea y su foto de perfil está bien blanca.*
  - *No le crean nada a esa gorda sólo miente para convivir y qué le pongan la atención qué su familia no le da.*
- Use of vernacular expressions as intensifiers. This type of messages uses a colloquial, vulgar or rude language to amplify discontent or liking for a situation or event. These messages were wrongly classified as offensive by all systems:
  - *Que pedo con el puto olfato de los gatos, al menos el de ella esta de la versh.*
  - *La liguilla es otro puto torneo alv paren de mamar*

- *Somos bien hdp con ella, la verdad. A veces hasta me siento mal de ser así, pero todos dicen que se lo gana.*

– Use of indirect discourse. In this case, the message refers to another message expressed at another time or by a third person. For example:

- *No puedes escribir que se vayan a la verga, que dejen de mamar, que están bien pendejos o pendejas, en fin, con cualquier Tweet se ofenden.*
- *Disculpen que esté tuiteando letras ardidas, pendejas o súper cursis pero tengo actitud "cantinerona".*

On the other hand, it should be mentioned that, due to the context of interpretation of the message, the labeling was not consistent. Apparently the most difficult case was to identify an offense if the message was addressed to an *undefined* group of people. Some examples of this type of messages are:

– *Dicen que no hay preguntas tontas, pero si hay gente tontita (pendej....) que pregunta pura chingadera*
– *Todas estan feas menos la que me gusta.*
– *Por unas cuantas mujeres tontas no dejemos que se nos olvide porque marchamos las demás q no expulsamos a nadie*

Finally, with the goal of getting further insights into the complementariness and redundancy of evaluated systems we show in Table 12 the number of instances correctly classified by at least one team, the number of instances wrongly classified by all teams, and the number of instances correctly classified by all teams.

Table 12. Instances statistics

Task	Well by some team	Wrong by all teams	Well by all teams
Location	1432 (95.46 %)	68 (4.53 %)	874 (58.26 %)
Occupation	1308 (87.20 %)	192 (12.80 %)	809 (53.93 %)
Aggressiveness	3048 (96.57 %)	108 (3.42 %)	759 (24.04 %)

It can be confirmed that the hardness of the two problems in the author profiling task was comparable (55% of instances correctly classified by all teams), although for the occupation trait there were more instances that were not correctly classified by any team (12.8%). The latter is in part due to the number of classes involved in the problem (8 vs. 6 in the location problem). Interestingly, it is confirmed that for the aggressiveness identification task there is a substantial complementariness among solutions that could be exploited: only 24% if instances were correctly classified by all teams, but more than 96% were classified by any of them. This result is encouraging.

6 Conclusions

This paper described the design and results of the MEX-A3T shared task collocated with IberEval 2018. MEX-A3T stands for *Authorship and Aggressiveness Analysis in Mexican Spanish Tweets.* Two tasks were proposed targeting author profiling (location and occupation) and aggressiveness detection. Given a set of tweets in Mexican Spanish for training, the participants had to identify location, occupation, and aggressiveness. Two novel data sets associated to the two tasks were introduced, together with an evaluation protocol and baselines. The competition lasted more than two months, and attracted 8 teams.

A variety of methodologies were proposed by participants, comprising content-based (bag of words, word n-grams, term vectors, dictionary words, slang words, and so on) and stylistic-based features (frequencies, punctuations, POS, Twitter-specific elements, and so forth) as well as approaches based on neural networks (CNN, LSTM and others). In all tasks, the baselines were outperformed by most participants.

For author profiling, the approach proposed by the MXAA team obtained the best results with an approach based on emphasizing the value of personal information for building the text representation [18]. For aggressiveness identification the top ranked team was INGEOTEC [11], This approach is based on MicroT, which one is a text classification approach supported on model selection techniques, from a lexicon-based model that takes into account the presence of aggressive and affective words, and from a model based on the fastText representation of texts.

In general terms, the competition was a success: performance was considerably improved with respect to the baseline, solutions proposed by participants were diverse in terms of methodologies and performances, and new insights on how to deal with tweets on Mexican Spanish. Among the most interesting findings of the task was the fact that predictions from participants resulted complimentary. For aggressiveness detection, the best team obtained a performance of 0.48 whereas an artificial ensemble yielded up to 0.92 of F_{macro}. This result is encouraging as it motivates research on ensemble generation for further boosting performance.

Acknowledgements Our special thanks go to all of MEX-A3Ts 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). Álvarez-Carmona thanks for doctoral scholarship CONACyT-Mexico 401887. Also, Guzmán-Falcón thanks for scholarship CONACyT-Mexico 776534.

References

1. 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)

2. 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)
3. Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 307–322. Springer (2014)
4. Aragón, E.M., López-Monroy, A.P.: Author profiling and aggressiveness detection in spanish tweets: Mex-a3t 2018. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
5. 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)
6. Ashlund, S., et al.: Conference on human factors in computing systems. Interchi
7. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). pp. 71–80. IEEE (2012)
8. Correa, S., Martin, A.: Linguistic generalization of slang used in mexican tweets, applied in aggressiveness detection. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
9. Frenda, S., Banerjee, S.: Deep analysis in aggressive mexican tweets. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
10. Gómez-Adorno, H., Bel-Enguix, G., Sierra, G., Sánchez, O., Quezada, D.: A machine learning approach for detecting aggressive tweets in spanish. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
11. 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)
12. Hatzidaki, O.: Fyi: Discourses of aggression in greek cmc
13. Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., Qian, Y.: A report on the 2017 native language identification shared task. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 62–75 (2017)
14. Markov, I., Gómez-Adorno, H., Mónica, J.R., Grigori, S.: Cic-gil approach to author profiling in spanish tweets: Location and occupation. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
15. Mey, J.: International conference on computational linguistics (2018)

16. Muñiz Cuza, C.E., De la Pea Sarracén, G.L., Rosso, P.: Attention mechanism for aggressive detection. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
17. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. pp. 145–153. International World Wide Web Conferences Steering Committee (2016)
18. 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)
19. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014. pp. 1–30 (2014)
20. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
21. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF (2017)
22. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF. p. 2015. sn (2015)
23. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. pp. 750–784 (2016)
24. 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)
25. Tetreault, J., Blanchard, D., Cahill, A.: A report on the first native language identification shared task. In: Proceedings of the eighth workshop on innovative use of NLP for building educational applications. pp. 48–57 (2013)
26. Villegas, M.P., Garciarena Ucelay, M.J., Fernández, J.P., Álvarez Carmona, M.A., Errecalde, M.L., Cagnina, L.: Vector-based word representations for sentiment analysis: a comparative study. In: XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). (2016)
27. Waseem, Z., Chung, W.H.K., Hovy, D., Tetreault, J.: Proceedings of the first workshop on abusive language online. In: Proceedings of the First Workshop on Abusive Language Online (2017)