Author Profiling in Social Media with Multimodal Information

By:

Miguel Ángel Álvarez Carmona

A dissertation submitted in partial fulfillment
of the requirements for the degree of:

PH.D. IN COMPUTER SCIENCE

at

Instituto Nacional de Astrofísica, Óptica y Electrónica

March, 2019
Tonantzintla, Puebla

Supervised by:

PhD. Luis Villaseñor Pineda, INAOE
PhD. Esaú Villatoro Tello, UAM-C

Abstract

Determine aspects of a person as gender, age, residency, occupation, among others, through his/her texts is a task that is part of the *natural language processing* and is known as *author profiling*.

In this thesis work, we propose a solution for the task of profiling authors in social networks. Our solution uses a multimodal approach to extracting information from written messages and images shared by users. Previous work has shown the existence of useful information for this task in these modalities; however, our proposal goes further demonstrating the complementarity of the modalities when merging these two sources of information. To do this, we propose to map images in a text, and with that, to have the same framework of representation through which to achieve the fusion of information.

Our work explores different methods for extracting information either from the text or from the images. To represent the textual information, different distributional term representations approaches were explored in order to identify the topics addressed by the user. For this purpose, an evaluation framework was proposed in order to identify the most appropriate method for this task. To represent visual information, approaches were explored to convert an image into a set of descriptive terms.

The results show that the textual descriptions of the images contain information for the author profiling task, and the fusion of textual information with information extracted from the images increases the accuracy of this task.

Resumen

Determinar aspectos de una persona como su género, edad, lugar de origen, ocupación, entre otros, a través de sus textos es una tarea que se enmarca dentro del *procesamiento del lenguaje natural* y se le conoce como *perfilado del autor*.

En este trabajo de tesis se propone una solución para la tarea de perfilado de autores en redes sociales. Nuestra solución utiliza un enfoque multimodal extrayendo información tanto de los mensajes escritos como de las imágenes compartidos por los usuarios. Trabajos previos han demostrado la existencia de información útil para esta tarea en estas modalidades, sin embargo, nuestra propuesta va más allá demostrando la complementariedad de las modalidades al fusionar estas dos fuentes de información. Para ello, se optó por llevar a las imágenes a una representación textual, y con ello contar con un mismo marco de representación a través del cual lograr la fusión de información.

Nuestro trabajo explora diferentes métodos para la extracción de información ya sea a partir del texto o de las imágenes. Para representar la información textual, se exploraron diferentes representaciones distribucionales con la finalidad de identificar los temas abordados por el usuario. Para ello se propuso un marco de evaluación de manera a identificar el método más adecuado para esta tarea. Para representar la información visual, se exploraron enfoques para convertir una imagen en conjuntos de términos descriptivos.

Los resultados muestran que las descripciones textuales de las imágenes contienen información para la tarea de perfilado de autor; y la fusión de la información textual con información extraída de las imágenes incrementa la exactitud en esta tarea.

Agradecimientos

Mamá, no hay palabras para poder agradecerle todo tu apoyo a lo largo de toda una vida y más. ¡Simplemente Gracias!

Mi agradecimiento al pueblo mexicano y al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca no. 401887. También agradezco el apoyo que se dio a esta tesis a través de los proyectos CB-2015-01-258588, CB-2015-01-257383 y FC-2410.

A mis asesores, los Dres. Esaú Villatoro Tello y Luis Villaseñor Pineda, mi más sincero agradecimiento por su apoyo constate, sus comentarios acertados y sus consejos que me acompañaron a lo largo de mi doctorado.

Al Dr. Manuel Montes y Gómez que sin ser oficialmente mi asesor, me dio su apoyo y se convirtió en una fuerte guía que, a pesar de todo, nunca se rindió en su afan por mi desarrollo académico, y sobre todo, mostró siempre un gran entusiasmo para que pudiera llegar más lejos en esta investigación.

A Alan, Alfredo, Gibrán, Gustavo y Julio. No importa de cúantas fiestas, eventos, reuniones o pláticas me he perdido, siempre me han apoyado y comprendido.

A Lau, Luis, Jona, Samuel, Shender y Mau; Fuimos una gran generación

A esos amigos, camaradas y compañeros todos cazadores de jirafas, recuerden que la vida es un safari interminable donde el que no caza es cazado. ¡A cazar!

Soy tan afortunado que he tenido muchos profesores que me han inspirado a seguir adelante, más de los que podrían caber en esta hoja de agradecimientos. A todos ellos: ¡Gracias!

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

A different language is a different vision of life.

Federico Fellini

The Internet has been consolidated as an interactive and massive means of communication to allow the exchange of information among people from different geographical areas, gender, age, socio-economic level, etc.

Recently, social media has gained an important popularity thanks to some services that invite to easily share information such as messaging, chats, blogs, among others. This impact has become evident in recent years. According to the site Qmee[1], every minute more than 350 GB of data are generated in Facebook[2], more than 278 thousand tweets are written[3], there are more than 11 thousand users uploading photos in Pinterest[4] and there are more than 347 new posts in WordPress[5], just to mention some data.

These numbers show us that, per minute there is a significant number of new texts and images shared by authors of which, most of the time we do not know anything about them.

There are several reasons why, it is essential to know some relevant data of the social networks users. For example, from marketing, there is interest in knowing the identity and demographic characteristics of the various users, with the intention of directing the advertising for exploiting in a better way (Bentolila et al., 2015). In the human-computer interaction area, it is essential to know the specific characteristics of people to be able to show an interface according to the features and personality of each (De Andrés et al., 2015). In addition to that significant impact and especially to

[1] http://blog.qmee.com
[2] https://www.facebook.com/
[3] https://twitter.com/
[4] https://pinterest.com/
[5] https://wordpress.com/

the facility of exchanging information hiding the profile of people, the web has been used to perform illegal or deceptive acts such as sexual harassment and extortion (Hall and Hall, 2007; Escalante et al., 2015). To detect and prevent this type of illicit acts, the discipline is known as forensic linguistics (Aronoff, 2017), makes use of linguistic knowledge to study texts that evidence this type of bad behavior.

That is why, the need to determine the profile of users on social networks has emerged. Given that, to carry out an analysis of this type manually on social networks is unthinkable, the need arises to carry out this analysis automatically using computational technologies.

In natural language processing, the task entrusted to study issues related to the author of a text is known as authorship analysis (Indurkhya and Damerau, 2010). Authorship analysis is the process of examining the characteristics of a text with the intention of obtaining conclusions from its author (El and Kassou, 2014). Several papers divide the authorship analysis into two mains areas. (Zheng et al., 2003; Abbasi and Chen, 2005; Zheng et al., 2006): **Authorship attribution** and **Author profiling**.

The authorship attribution consists in determining the probability that text belongs to a given author (Stamatatos, 2009). On the other hand, the author profiling consists of extracting as much information as possible from the author through what he/she writes (Argamon et al., 2009). The hypothesis behind this area is that the way we write reveals our behavior and personality. It is in this area where this thesis work is centered.

The author profiling task (AP) is to extract *demographic aspects* of a person from their texts. For example gender, age, location, occupation, socio-economic level or native language[6](Corney et al., 2002; Koppel et al., 2005; Schler et al., 2006). Efforts have also been made to determine other aspects such as the level of well-being (Schwartz et al., 2013b), personality traits such as extraversion or neuroticism (Argamon et al., 2005; Mairesse and Walker, 2006; Rangel et al., 2015) as well as political ideology (Koppel et al., 2009), affinity for some products (Argamon et al., 2005), among others.

At the beginning of the AP task, formal texts such as books, newspapers or magazines were analyzed to determine the features of their authors (Argamon et al., 2003). However, determining the profile of people through their social network accounts is a task that has taken force in recent years (Stamatatos et al., 2015; Rangel et al., 2016, 2017).

Traditionally, there are two types of approaches that have proven to be effective in addressing the problem of AP in social networks: style-based approaches and

[6]If the text is written in a different language of the native language

content-based approaches (Argamon et al., 2003; Álvarez-Carmona et al., 2016). The approaches based on style refer to the fact of analyzing how the author expresses himself when writing, on the other hand, in the content-based approaches the thematic area of the text is analyzed. The main contribution of various works, is based on the selection of attributes that can measure the style and content of the author (Schler et al., 2006; Mairesse and Walker, 2006; Rangel et al., 2017).

Beyond the relevance and advantages that these approaches can have in this type of tasks, they also begin to identify particular problems and challenging aspects that, they require more elaborate approaches and techniques than those that have been used up to now. Among these more advanced approaches, we should mention those that, incorporate information from another available modality, beyond which, it can be derived from the style and content of the text. This information may be of different nature, such as that provided by users of social networks in the images they share, information about their contact networks, interaction behavior in social networks, among others. To these approaches that take advantage of different sources of information are known as multimodal information approaches.

In the AP context, it can be seen that most of the recent works, in the field of social networks, have focused, mainly on the definition of thematic attributes and style-metrics appropriate for this task. However, there is a sign of progress towards the description of multimodal representations that, for example, integrate different types of information or that, due to the nature of social networks, information about the images shared by users or their social environment is also incorporated. This thesis work is part of the author profiling in social networks with multimodal information.

1.1 Problem

Most of the works that have tried to solve the task of AP are based solely on the textual information that users share in social networks. Utilize only text generates that, much of the information available by the nature of social networks is not exploited. Most approaches do not take advantage of images, videos, contact list, activity schedules or another information. For this reason, it is not known which of these different information modalities is more valuable for the AP task. This is why it is essential to analyze how the multimodal information impacts the AP task.

Another aspect to highlight is that the works in AP have given evidence of the importance of the content of the texts. Nevertheless, the most common approach that has been used is the Bag of Words (BoW). The problem with this approach when

working on social networks is the lack of information because regularly short texts are analyzed. Besides that, the texts are not formal which causes that there are words out-of-the-dictionary and spelling mistakes.

A set of approaches that have not been deepened enough to represent the content of the texts and, that can be useful for the AP task are the **distributional term representations** (**DTRs**). The basic intuition behind the DTR's is the called distributional hypothesis (Skalmowski, 2016), which states that terms with similar distributional patterns tend to have the same meaning (Lavelli et al., 2004a; Levy et al., 2015). This distributional hypothesis could capture the content of the users' text in a better way than the traditional content approaches used for AP. In this thesis work we compare these representations experimentally to know their impact on the AP task.

On the other hand, there are few works that have taken advantage of the information extracted from the images shared by users, this despite the fact that various works in psychology have concluded that the photos that are shared on social networks can tell a lot of the people (Hum et al., 2011; Grimshaw, 2013; Eftekhar et al., 2014; Wu et al., 2014; Kharroub and Bas, 2015). Some works have applied the color histogram of the images to determine the gender of the users, but no studies have been done for other traits of the authors. Others works have converted the images to texts with automatic labelers of images, through **automatic images annotation** techniques, that assign a list of labels from a previously established set, and from there, infer the user's profile.

These approaches commonly are supervised and with a **closed vocabulary**. This means that the labelers select from a limit list of labels the elements in each image. The problem is that a limited vocabulary could be insufficient to represent the interest of the profiles in a collection. In this thesis work we propose to apply an approach based on **open vocabulary** to the AP task, under the idea that, it describes in a better way the social media profiles. The automatic images annotation based on open vocabulary approaches does not select the labels set from a limit list, but they select from a vocabulary from a large collection, normally, extracted from Internet pages. With this idea, we can represent each image in the collection as a text and it is possible apply text approaches to classify the profile of each user.

1.2 Research Questions

Throughout this thesis, we intend to answer the following research questions:

1. What information is being captured by distributional-based methods, and how

effective is for representing user's information when facing the problem of author profiling?

2. How to extract information from images shared by users through an open vocabulary approach in such a way that it is possible to determine their author profile?

3. How to take advantage of the information obtained from texts and images for solving the author profiling task?

1.3 Objectives

1.3.1 General Objective

Proposing an automatic multimodal method for author profiling that combines information obtained from images and texts produced by users, both in English and Spanish, and that demonstrate being more effective than uni-modal approaches.

1.3.2 Particular Objectives

1. Evaluate and analyze distributional-based representations for representing the textual information in the author profiling task for users from different social media sources.

2. Design a method that captures the interests of users through their shared images and represent them using an open vocabulary approach, in such a way that they can be classified for the author profiling task in Twitter collections.

3. Design and implement a method for determining the profiles of authors, in such a way that it takes advantage of textual and images information from users' tweets.

1.4 Description of the Proposed Work and its Scientific Relevance

The proposed solution for the Multimodal Author Profiling problem is divides into 4 steps:

1. Construct corpora for Multimodal Author Profiling. Two novels corpora were built. These collections contain text and image information. One of them is an extension of the Pan 14 corpus. This corpus has gender and age labels for English users. Until now, this would be the first multimodal corpus labeled for age. On the other hand, we built a collection with only Mexican users. The labels of the corpus are gender, occupation, and location of the users. Also, this is the first multimodal corpus exclusively with information from Mexicans. The importance of this step is that the collection will be accessible for the community.

2. Propose a framework to apply Distributional Term Representations (DTR's) in the author profiling task. In previous works, some authors have applied DTRs to determine some traits of the authors. Nevertheless, we do not know which are the best DTR's for the task and their advantages. The importance of this step is that we can discover the best DTR's to represent users on different traits and social media domains.

3. Apply an open vocabulary approach to transform each image into a list of words, and then, represent it for the author profiling task. Until now, the authors have proposed closed vocabulary methods to solve the task. Nevertheless, we think that with open vocabulary methods it is possible to represent the users in a better way. The importance of this step is the experimentation of open vocabulary on the task to be able to understand if there are advantages.

4. Propose fusion schemes to join text and image information. For this, we will use the best DTR's found in step 2, and the best image representations approach found in step 3. The importance of this step is to understand if it is possible to fuse different modalities represented with DTR's and open vocabulary approaches and overcome the individual results.

1.5 Document Outline

This thesis is structured in three parts besides the introductory one. These are listed below.

- In Part II, three chapters describe the background concepts, required to make this document as self-contained as possible. This part encompasses the following chapters:

  - In chapter 2, we describe basic concepts related to machine learning and text classification.
  - In chapter 3, the general concepts related to automatic image annotation are described.
  - In chapter 4, the concepts and related work of the author profiling task are described.

- Part III presents the contributions of this thesis, and is organized in the following chapters:
  - In chapter 5 the different collections built and used in this thesis are presented.
  - In chapter 6, we describe the general framework of distributional based approaches for representing users in the author profiling task.
  - In chapter 7, the author profiling method with the images shared by the users is described. Also, the fusion schemes for the text and images information are analyzed in this chapter.
- Part IV outlines the general conclusions of this thesis.

Part II

Theoretical Background Review

Chapter 2

Supervised Machine Learning

A vida é uma aprendizagem diária. Afasto-me do caos e sigo um simples pensamento: Quanto mais simples, melhor!

José Saramago

The principal aim of supervised machine learning is to create a function, capable of predicting the corresponding value to any valid input object, after having seen a series of examples, i.e., the training data. For this, a the supervised machine learning algorithm has to generalize from the data presented to the situations not previously seen (Raschka and Mirjalili, 2017).

Classification is a common task in supervised learning, in which the output is a categorical value that represents a class label. Its popularity relies on its application in a wide range of fields. As a result, many learning algorithms to construct a classification model have been proposed, including support vectors machines, decisions trees, the k-nearest neighbors, among others (Olson and Wu, 2017). Despite the variety of algorithms currently available, to date there is not a universal "best" one; this is sometimes referred to as the *No Free Lunch Theorem* (Wolpert and Macready, 1997; Kang et al., 2018).

In this Section, we focus on describing some concepts related to supervised learning and metrics commonly used to assess the performance of the constructed models.

2.1 Classification

Classification is the task of estimating the output value of a data sample, where the output is characterized for being a categorical value (Lotte et al., 2007). For constructing a classification model, we usually require a learning algorithm and a training dataset. A dataset used for training consists of a set of data samples, where each sample,

$\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$, is described by a set of d attributes. It also has a target attribute $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ with k classes, which is called the class attribute. This dataset records a set of samples, also called instances, which represent the examples of the task to be learned (Aha et al., 1991).

Given a dataset, the objective of the learning algorithm is to produce a function to map a sample from the attribute space to a class label, i.e., $f(\mathbf{x}) : \mathbf{x} \rightarrow c \in \mathcal{C}$, where $\mathbf{x}$ is the sample and $\mathcal{C}$ is the set of class labels. This function can be used to predict the class labels of the future samples. This function is also called classification model or simply classifier.

There exist many learning algorithms that can be used for constructing the classification model. We describe some of the most popular approaches next.

2.2 Instance-Based Learners

Instance-based learners (IBL) are a kind of algorithms belonging to the lazy learning family. This means that, instead of performing a training phase in order to make abstraction from data, IBL represents each training sample as a point in a multi-dimensional feature space, these points are stored in memory, and new samples are classified based on the labels assigned to their closest samples kept in memory (de Haro-García et al., 2018).

The principal idea behind such algorithms is that similar instances will have similar classifications, i.e., the label result should be the same for nearby instances. The issue is how to define the similarity between a pair of instances. For this, a similarity function is used, which typically computes distance among instances. IBL also has a classification function which specifies how instance similarities yield a final classification. Examples of instance-based learners are the nearest neighbor algorithm, k^* algorithm, among others (Witten et al., 2016).

Nearest neighbor algorithm (Cover and Hart, 1967; Fadaei-Kermani et al., 2017) is an instance-based learner, which uses a specific distance function to determine the single most similar instance from the training set. The class label of the most similar instance is given as the classification for the new instance. A generalization of the nearest neighbor rule is the k-nearest neighbor algorithm. The k-nearest neighbors of the new instance are found, and the new instance's classification is based on the predominant class label among them (Zhang et al., 2018).

2.3 Decision Trees

Decision trees are approaches that allow constructing a model following a divide and conquer strategy (Huang and Siu, 2017). Generally, a decision tree is a graphical representation in which each internal node is associated with a decision, and the terminal nodes are usually associated with a class label. Each internal node is associated to test an attribute to decide what path should be taken. The path between two nodes is represented by a link, which contains the value of the decision (Jin et al., 2009).

The construction of decision trees generally involves a splitting process in order to choose an attribute at each internal node to make a decision. In the beginning, the most important attribute is selected for being used in the root (Friedl and Brodley, 1997; Sadr et al., 2018). At each node, the dataset is split, and the outcome is used to construct a new decision tree. One of the main issues when building a decision tree is to determine what attribute should be chosen next, which is approached by selecting, at each level, the most discriminative one. A useful attribute should be able of separating (as much as possible) the samples among the different classes. This attribute is the one that decreases (a set of samples is pure when all samples belong to the same class. A set of samples is impure when it has more than one class) in a set of samples as much as possible. There are several measures to determine the impurity, such as entropy-based, as it is used by the ID3 and C4.5 algorithms (Quinlan, 1986; Quinlan and Cameron-Jones, 1993; Salzberg, 1994), Gini index, χ^2, or G-square, as they are used in CART (Breiman, 2017).

Figure 2.1 shows an example of a decision tree. In the root node, the most discriminative feature is located, according to the adopted criterion. Links represent the path to be taken, based on the value of the feature. This is constructed recursively. The terminal nodes have the class labels, and one of them is reached when an instance is classified.

2.4 Support Vector Machines

Support vector machines (SVM) are supervised learning algorithms that can be used for both classification and regression (Cortes and Vapnik, 1995; Joachims, 2002). SVM was initially proposed for linearly separable classification problems. For a two-class classification problem, SVM finds the hyperplane that maximizes the margin separation between two classes. A particular characteristic of SVMs is that the solution to the classification problem is represented by the support vectors that determine the maximum margin hyper-plane (Hsu et al., 2003). The optimum separation hyper-plane

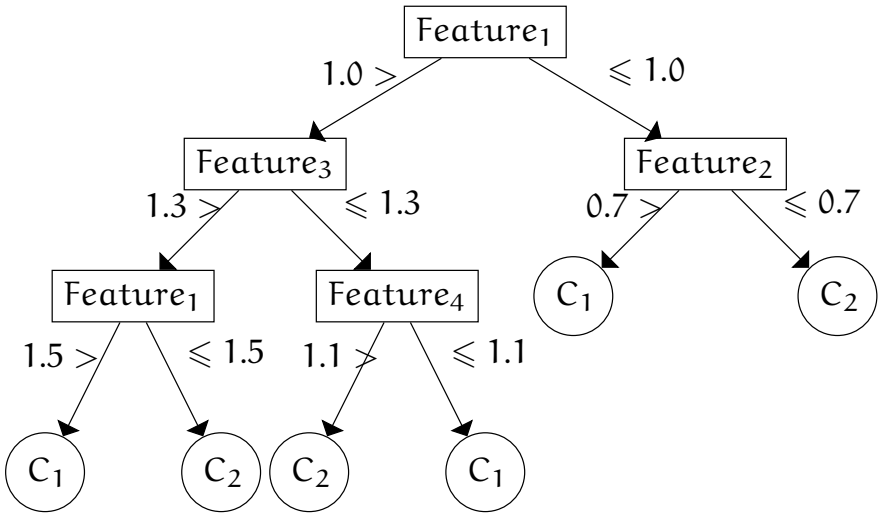


Figure 2.1: Example of a decision tree.

is the linear classifier with the maximum margin for the given training set. The Figure 2.2 shows an example of the optimum separation hyper-plane.

SVMs can also be used to non-linearly separable classification problems. In such cases, the data is mapped to a higher dimensional feature space using a kernel function, where the classes can be linearly separable. This is usually known as the *kernel trick* (Schölkopf, 2001).

2.5 Naïve Bayes

Naïve Bayes classifier is a statistical classifier. It can predict class membership probabilities, such as the probability that a given sample belongs to a particular class (Murphy, 2006).

A Bayesian classifier is based on Bayes' theorem (Rish, 2001). Nevertheless, the Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered "naïve" (Lewis, 1998).

Abstractly, naïve Bayes is a conditional probability model. Given a problem instance to be classified, represented by a vector $X = \{x_1, \cdots, x_n\}$ representing some n features (independent variables), it assigns to this instance probabilities $p(\mathcal{C}_i|\{x_1, \cdots, x_n\})$ for each of K possible classes $\mathcal{C}_k$ (Murty and Devi, 2011).

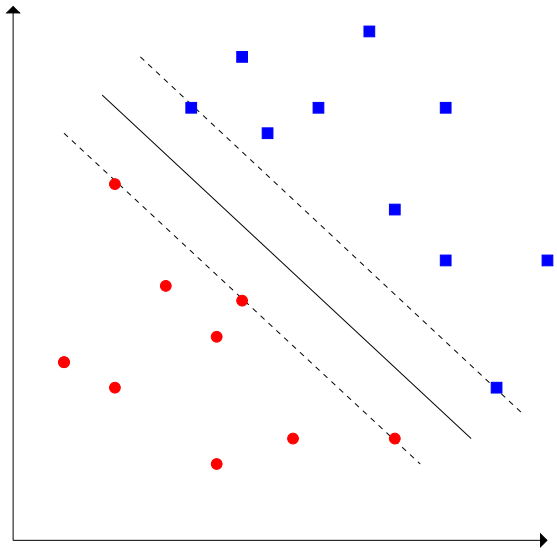


Figure 2.2: Example of a linear classifier with optimum separation hyper-plane

Using Bayes' theorem, the conditional probability can be decomposed as:

$$p(\mathcal{C}_i|X) = \frac{p(\mathcal{C}_i)p(X|\mathcal{C}_i)}{p(X)} \tag{2.1}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on $\mathcal{C}$ and the values of the features x_i are given so that the denominator is effectively constant. Now, it applies the conditional independence assumptions. It assumes that each feature x_i is conditionally independent of every other feature x_j if $j \neq i$, given the category $\mathcal{C}_k$. This means that:

$$p(\mathcal{C}_k|x_1, \cdots, x_n) = p(\mathcal{C}_k)p(x_1|\mathcal{C}_k)p(x_2|\mathcal{C}_k)\cdots p(x_n|\mathcal{C}_k) \tag{2.2}$$

This is possible to express as:

$$p(\mathcal{C}_k|x_1, \cdots, x_n) = p(\mathcal{C}_k)\prod_{i=1}^{n} p(x_i|\mathcal{C}_k) \tag{2.3}$$

Finally, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \mathrm{argmax}_{k\in\{1,2\cdots,K\}}p(\mathcal{C}_k)\prod_{i=1}^{n} p(x_i|\mathcal{C}_k) \tag{2.4}$$

2.6 Fusion of Information

The extraction of several kinds of features raises new issues about the way to properly use them inside a classification system. In most works, authors have combined the extracted features in order to improve the performance of their methods, then it is interesting to exploit this kind of information into the final representation. Nonetheless, the most used ways to combine heterogeneous attributes are simple fusion approaches; early fusion and late fusion (Bekkerman and Allan, 2004; Tan et al., 2002; Rokach, 2009).

2.6.1 Early Fusion

The main idea of early fusion is to concatenate the different feature spaces (e.g., words and n-grams) into single vectors, which are fed to a learning method (Rokach, 2009; Kuncheva, 2004). Given a vector $v_1 = \{v_{1,1}, v_{1,2}, \ldots, v_{1,n}\}$ with n elements and a vector $v_2 = \{v_{2,1}, v_{2,2}, \ldots, v_{2,m}\}$ with m elements, the results of the early fusion is a vector $v_{EF} = \{v_{1,1}, v_{1,2}, \ldots, v_{1,n}, v_{2,1}, v_{2,2}, \ldots, v_{2,m}\}$ of $n + m$ elements.

The Support Vector Machine (SVM) has shown to be effective using the early representation (Boiman et al., 2008; Cruz-Roa et al., 2011).

The problem of early fusion approaches is that they can be affected if the feature spaces are not diverse enough (Rokach, 2009; Kuncheva, 2004).

2.6.2 Late Fusion

The underlying late fusion idea is to represent instances using vectors corresponding to each feature space, to provide different perspectives/views of each instance.

Late fusion strategies consider each feature space independently and build an ensemble learning system to combine the outputs of classifiers trained on different inputs for instance a weighting vote ensemble classifier (Rokach, 2009; Breiman et al., 1996).

Except for voting, stacking (Kotsiantis et al., 2006) aims to improve efficiency and scalability by executing a number of learning processes and combining the collective results. If for a test instance V_i, R_1 is the result of a classifier C_1, R_2 is the result of a classifier C_2 and so on, R_n is the result of a classifier C_n, then, the stacking representation is $V_i = \{R_1, R_2, \ldots, R_n\}$

The main difference between voting and stacking is that the latter combines base classifiers in a non-linear fashion (Kotsiantis et al., 2007). The combining task, called a meta-learner, integrates the independently computed base classifiers into a higher level

		Predicted Class $\hat{C}^+$	Predicted Class $\hat{C}^-$
Correct Class	C^+	TP	FN
Correct Class	C^-	FP	TN

Figure 2.3: Confusion matrix for a binary classification problem.

classifier, a meta-classifier, by relearning the meta-level training set. This meta-level training set is created by using the base classifiers' predictions on the validation set as attribute values and the true class as the target (Sikora, 2015).

2.7 Performance Assessment

Once a model is constructed, one crucial question is how to assess its predictive performance on unknown samples. The ability to correctly classify these unknown samples is called *generalization capability* (Schmidhuber, 1997; Bottou, 2010). One usually wants to find a model with a good generalization capability. Therefore, the evaluation is crucial, since it can tell us how good a particular model or classifier is for a particular problem. In this section, we describe some evaluation methods used to assess the expected performance of a model.

2.7.1 Measurements of Model Efficacy

Before explaining the techniques for assessing the model efficacy, some measurements are introduced.

Figure 2.3 shows an example of a confusion matrix for a binary classification problem. It shows the positive samples that are correctly classified (TP), the positive samples that are incorrectly classified (FN), the negative samples that are incorrectly classified (FP), and the negative samples that are correctly classified (TN). From it, several scores or measurements can be computed. The list of scores in supervised classification may be large, including standard scores and those designed for specific classification problems. Here, we revisit the best well-known scores.

Among the existing scores, some of the most popular are the following (Powers, 2011):

- **Accuracy** measures the portion of samples that are correctly classified, i.e.,

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

  Accuracy ranges from 0 to 1, in which 0 means all samples are incorrectly classified and 1 that all samples are correctly classified.

- The **error rate** is the complement of accuracy and is computed as:

$$Err = \frac{FP + FN}{TP + TN + FP + FN} \tag{2.6}$$

- **Precision** (also called positive predictive value) is the fraction of positive instances among the total of the instances, and is computed as:

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

- **Recall** is the fraction of positive instances that have been correctly classified over the total amount of positive instances, and is computed as:

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

- **F-measure** is approximately the average of precision and recall measures, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean. It is computed as:

$$F_\beta = (1 + \beta^2)\frac{Precision * Recall}{\beta * Precision + Recall} \tag{2.9}$$

  β is a weight factor. Commonly, this measures is named F_1 measure, because recall and precision are evenly weighted, i. e., $\beta = 1$.

2.8 Text Classification

The problem of classification has been widely studied in machine learning and information retrieval communities with applications in many diverse domains, such as target marketing, medical diagnosis, newsgroup filtering, and document organization (Aggarwal, 2015).

It is possible to observe that the text classification is a subtask of the classification as it was described at Section 2.1. A dataset used for the classification process consists of a set of text documents samples (books, magazines, chats, journals, tweets, posts,

emails, etc), where each sample, $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$, is described by a set of d attributes extracted from the text. The training data is used to construct a classification model, which relates the features in the underlying record to one of the class labels. For a given test instance for which the class is unknown, the training model is used to predict a class label for this instance (Aggarwal, 2015).

This model assumes that only information about the presence or absence of words is used in a document. The frequency of words plays a helpful role in the classification process, and the typical domain size of text data is much higher than a typical classification problem (Sebastiani, 2002).

Commonly, the text classification problem consists of three steps (Forman, 2003): i) the features selection, ii) the text representation and iii) the application of a learning algorithm to classify.

In this Section, we describe the features selection process and describe the most popular text representation methods for the classification process.

2.8.1 Feature Selection

While feature selection is also desirable in other classification tasks, it is especially important in text classification due to the high dimensionality of text features and the existence of irrelevant (noisy) features (Rogati and Yang, 2002).

The most common feature selection method is that of stop-word removal and stemming. In stop-word removal, it determines the common words in the documents which are not specific or discriminatory to the different classes. On the other hand, for stemming, different forms of the same word are consolidated into a single word.

Nevertheless, several approaches have been proposed in order to select the most discriminative words in the training set. This kind of selection process ensures that, those features which are highly skewed, towards the presence of a particular class label are picked for the learning process (Yang and Pedersen, 1997).

We will discuss some feature selection methods in this section.

Information Gain

Information gain measures how much information a feature gives us about the class (Koller and Sahami, 1996; Bu et al., 2018). Formally, let P_i be the global probability of class i ,and $p_i(w)$ be the probability of class i, given that the document contains the word w. Let $F(w)$ be the documents that contain the word w. The information gain

measure $I(w)$ for a given word w is defined as follows:

$$\begin{aligned} I(w) = &- \sum_{i=1}^{k} P_i * \log P_i + F(w) * \sum_{i=1}^{k} p_i(w) * \log p_i(w) \\ &+ (1 - F(w)) * - \sum_{i=1}^{k} (1 - p_i(w)) * \log (1 - p_i(w)) \end{aligned} \tag{2.10}$$

The higher the value of the information gain $I(w)$, the higher the discriminatory power of the word w.

Mutual Information

Mutual information measure provides a formal way to model the mutual information between the features and the classes (Bi et al., 2018). The mutual information $M_i(w)$ between the word w and the class i is defined on the basis of the level of co-occurrence between the class i and word w (Peng et al., 2005).

Note that the expected co-occurrence of class i and word w on the basis of mutual independence is given by $P_i * F(w)$. Nevertheless, the co-occurrence is given by $F(w) * p_i(w)$. In practice, the value of $F(w) * p_i(w)$ is probably much larger or smaller than $P_i * F(w)$, depending upon the level of correlation between the class i and word w.

The mutual information is defined in terms of the ratio between these two values. Specifically:

$$M_i(w) = \log \frac{F(w) * p_i(w)}{F(w) * P_i} = \log \frac{p_i(w)}{P_i} \tag{2.11}$$

The word w is positively correlated to the class i, when $M_i(w) > 0$, and the word w is negatively correlated to class i, when $M_i(w) < 0$.

χ^2-statistic

The χ^2-statistic is a different way to compute the lack of independence between the word w and a particular class i (Jin et al., 2006). Let n be the total number of documents in the collection, $p_i(w)$ be the conditional probability of class i for documents which contain w, P_i be the global fraction of documents containing the class i, and $F(w)$ be the global fraction of documents which contain the word w.The χ^2-statistic of the word between word w and class i is defined as follows:

$$\chi_i^2 = \frac{n * F(w)^2 * (p_i(w) - P_i)^2}{F(w) * (1 - F(w)) * P_i * (1 - P_i))} \tag{2.12}$$

We note that the χ^2-statistic and mutual information are different ways of measuring the correlation between terms and categories. One significant advantage of the χ^2-statistic over the mutual information measure is that it is a normalized value, and therefore these values are more comparable across terms in the same category (Forman, 2003).

2.8.2 Text Representation

Although there are different models to represent the text, the Vector Space Model (VSM) (Sidorov et al., 2014) is the principal model for many textual tasks, which is used to represent the text documents and define the similarity among them (Liu et al., 2005; Han et al., 2018).

Bag of Word (BOW) (Wallach, 2006) is an approach used to represent each document as a histogram of words under the VSM. In the BOW representation, each document is encoded as a feature vector, with each element in the vector indicating the presence or absence of a word in the document.

For a document d_i the BoW representation is given by $d_i = \{w_1, w_2, \ldots, w_n\}$, where w_j is the $j-$th word in the corpus collection and represents the weigh of the word j in the document i. It is a way of extracting features from a text for use in modeling, such as with machine learning algorithms (Schmitt and Schuller, 2017).

Others attempts have been made to incorporate the word-order knowledge with the vector space representation. N-gram statistical language model (Stolcke, 2002; Bakhtin et al., 2018) is a well-known one among them. The entries of the document vector by N-gram representation are strings of n consecutive words extracted from the collections. They are effective approximations, and they not only keep the word-order information but also style of the author (Gómez-Adorno and Sidorov, 2017). However, the high-dimensional feature vectors of them is a clear disadvantage of the approach.

These motivate us to seek for others models, for instance, the **Latent Semantic Analysis (LSA)**. LSA is a method to extract and represent the meaning of the words and documents. LSA is built from a matrix $\mathbf{M}$. LSA uses the Singular Value Decomposition (SVD) to decompose $\mathbf{M}$ as follows.

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Where The $\mathbf{\Sigma}$ values are called the singular values and $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively. $\mathbf{U}$ and $\mathbf{V}$ contains a reduced dimensional representation of words and documents respectively. $\mathbf{U}$ and $\mathbf{V}$ emphasizes the most influential relationships and throws away the noise (Landauer et al., 1998). In other

words, it makes the best possible reconstruction of the **M** matrix with the less possible information (Landauer et al., 2013). Using **U** and **V** computed only from the training documents, words and documents are represented for training and test. For this is necessary provide a k parameter to choose the first k dimensions for making the lower-dimensional approximation reconstruction M_k of the M matrix. In this way, M_k is the semantic space representation for the train and test documents (Soundar and Ponesakki, 2016).

Another attempt to represent the text with a topic-based approach is the *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003). LDA find the topics in a text with a statistical approach. Latent Dirichlet Allocation (LDA) is a Bayesian probabilistic model of text documents. It assumes a collection of k topics. Each topic defines a multinomial distribution over the vocabulary and is assumed to have been drawn from a Dirichlet distribution (Hoffman et al., 2010).

Chapter 3

Automatic Image Annotation

Pinto autorretratos porque estoy mucho tiempo sola. Me pinto a mí misma porque soy a quien mejor conozco.

Frida Kahlo

Nowadays, there is a significant amount of images available on the Internet. This amount causes that the searching of the images becomes a crucial task. For this reason, a large amount of research has been carried out on image retrieval (IR). The IR is the task that is responsible for browsing, searching and retrieving images from a database. Since in this thesis work, we are going to browse for the user's images, we propose to apply specific approaches IR based. In this chapter, we present the main approaches inside the task.

In general, IR research efforts can be divided into three types of approaches (Zhang et al., 2012). The first approach is the traditional text-based annotation. In this approach, images are annotated manually by humans and images are then retrieved in the same way as text documents (Russakovsky et al., 2015). The second type of approach focuses on content-based image retrieval (CBIR), where images are automatically indexed and retrieved with low-level content features like color, shape and texture (Chang et al., 2003). The last approach tries to capture the content of the image but in a higher level, i.e., these approaches fond the objects in the images (Murthy et al., 2015). With this information, a IR is built.

The manual annotation approach has several apparent disadvantages as the subjectivity of the people, without mentioning that it is time-consuming and costly. Also, it is impractical for general users to use a CBIR system because users are required to provide query images Zhang et al. (2012).

From these disadvantages, in recent years, a third approach called Automatic Image

Annotation emerged. The Automatic Image Annotation (AIA) is the process by which a system automatically assigns keywords to an image (Jeon et al., 2003; Murthy et al., 2015).

Once images are annotated with semantic labels, images can be retrieved by keywords, which is similar to text document retrieval. The critical characteristic of AIA is that it offers keyword searching based on image content and it employs the advantages of both the text-based annotation and CBIR (Zhang et al., 2012; Uricchio et al., 2017). In this thesis, we propose to apply this approach for representing the images in our corpora.

There are generally three types of AIA approaches. The first approach is the single labeling annotation using conventional classification methods. The second approach is the closed vocabulary or multi-labeling annotation which annotates an image with multiple concepts defined previously. The third approach is the open vocabulary approach or web-based image annotation which uses metadata to annotate images (Zhang et al., 2012).

In the next sections, we present these three different approaches for the AIA task.

3.1 Single Labelling Annotation

In this approach, low-level features are extracted from image content, and the features are fed directly into a conventional binary classifier which gives a yes or no vote. The output of the classifier is the semantic concept(s) which is used for image annotation (Wei et al., 2014). The idea of single labeling is equivalent to collective labeling, that is, instead of labeling images individually, images are first clustered and then labeled collectively. The conventional machine learning tools include support vector machines, artificial neural network, and decision tree (Zhang et al., 2012).

The advantage of this type of approach is that the retrieval is efficient, as there is no need to do image indexing and expensive online matching, as in other IR approaches (Zhang and Zhang, 2010). The disadvantage is that it does not consider the fact that many images belong to multiple categories (Boutell et al., 2004). As a result, many relevant images can be missed from the retrieval list if a user does not type the right keyword exactly. One way to alleviate this problem is to label each category with multiple keywords reflecting different themes within the category. Another issue with the single labeling annotation is that images within each category are not ranked, leading to reduced retrieval accuracy (Zhang et al., 2012; Huang et al., 2018).

3.2 Closed Vocabulary Approaches

Also known as multi-labeling annotation. Different from the binary classification approaches, multiple labeling methods annotate an image with multiple semantic concepts or categories (Darwish, 2016; Masoud, 2018). The idea of these approaches is related to multi-instance learning, or, multi-instance multi-label (MIML) learning (Andrews et al., 2003).

In MIML, an image is represented by a bag of features or a bag of regions. The image is annotated with a concept label if any of the regions in the bag is associated with the label. As a result, an image is annotated with multiple labels. A typical MIML is achieved using probabilistic tools such as the Bayesian methods (Zhang and Zhou, 2014). The Bayesian methods try to find the probability that an image belongs to any particular concept, given the observation of certain features from the image or region (Zhang et al., 2012). This makes it possible to assign an image to multiple concepts and rank images with the same concept according to the probabilities. Given a set of images $I = \{I_1, I_2, \cdots, I_N\}$ from a set of given semantic classes $C = \{c_1, c_2, \cdots, c_n\}$, Bayesian models determine the probability from the conditional probabilities and the priors. An image I_k is represented by a vector x. Given the probabilities $p(c_i)$ and conditional probability densities $p(x|c_i)$, the probability of an unknown image I_k belonging to class c_i is determined as in 3.1.

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)} \tag{3.1}$$

From Expression 3.1, it can be seen that a Bayesian approach has four components: one output component $p(c_i|x)$ and three input components: $p(c_i)$, $p(x|c_i)$, and $p(x)$. Because the distribution $p(x)$ is usually uniform for all classes, the class of image I can be decided using the "maximizing a posterior" (MAP) criterion as indicates the expression 3.2.

$$\hat{c} = argmax_{c_i} p(x|c_i)p(c_i) \tag{3.2}$$

3.3 Open Vocabulary Approaches

The web is a rich source of images and text information. The images in the web, often come with descriptors elements as text, URL, HTML, etc. This information can be used for image annotation (Zhang et al., 2012). A number of techniques have been

developed for annotating web images, most of them integrating both metadata and visual features for accurate image annotation (Pellegrin et al., 2016).

In (Cai et al., 2004), the author proposed a two level annotation and clustering mechanism: textual clustering for semantic annotation and visual clustering for reorganization of images within each semantic category. In this process, images from web pages are first represented using three types of features: textual features (derived from surrounding text), link graph (derived from three complex hyperlink matrices) and visual features (derived from color moments on local Fourier transform). The textual features and link graph are used to cluster images into semantic category which is equivalent to annotation. However, images within each of the semantic categories may not be perceptually similar. Therefor, they apply a second level of clustering on each of the semantic categories to reorganize the images into clusters based on visual features. The major issue with this method is that the textual features especially the link graph features are not reliable, as shown in existing image search engines (Zhang et al., 2012).

Wang et al. (2006b) also propose an automatic system that annotates images using both web description and content features. The system needs at least one correct initial keyword and one example image to initiate the process. The keyword is used to search the web to find images and their web descriptions. Thereafter, a 36 dimensional color correlogram is used to select a number of top ranked images similar to the example image. The web descriptions of the selected images are clustered using a text clustering algorithm. Each cluster is scored either by its size or by the average number of words. The words in the top scored clusters are used for annotations. The advantage in this approach is that it does not need any training samples. However, the performance is subject to the quality of the description of the images (Uricchio et al., 2017).

The annotations from text description could be noisy, therefore these annotations need refinement. This is especially needed for web based image annotation, because each image is usually annotated with multiple words which may not be related to each other (Wu et al., 2015). In a refinement stage, it preserves the annotations which are strongly correlated and rejects those which are not so strongly related to each other.

(Wang et al., 2006a) calculate the similarity between two words as the normalized frequency of images annotated by both words. On the other hand, (Wang et al., 2007) calculate this similarity as the normalized sum of content-similarities between the candidate image and the images annotated by both words. The similarity values are used to find strongly correlated annotation words.

Chapter 4

The Author Profiling Task

Tonantlajtol kemej toyoltlajtol.

Nahuatl quote

Author Profiling (AP) is the task of determining demographic features of authors like native language, education, gender, age, personality traits, location, occupation, among others, by analysing understanding his writing styles (Reddy et al., 2016).

For its nature, AP is an essential technique in the present information era which has applications in marketing, security, and forensic analysis (Kanellis, 2006; Lakkaraju et al., 2018).

In the case of the marketing, text is analyzed to classify the consumers based on their age, gender, occupation, native language, nationality and personality traits. The classification results of these traits help to direct advertising better (Kumar and Reinartz, 2018). AP is also beneficial in the education domain. It helps in revealing the exceptional talent of students. It also helps in estimating the suitable level of knowledge of each student or a student group in the educational forum (Vinokur, 2015). Also AP helps in crime investigation to identify the perpetrator of a crime by considering the characteristics of writing styles (Layton, 2016). Social websites are an integral part of our lives through which, crimes are cropping up like public embarrassment, fake profiles, defamation, blackmailing, stalking, among others. To identify the perpetrator, it is useful to sketch the writing style of a perpetrator using Author Profiling (Douglas et al., 2013; Schilling and Marsters, 2015).

In general, every human being has his/her style of writing and each one continues to write the same style in tweets, blogs, reviews, social media and also in documents (Reddy et al., 2016). Nevertheless, the content of texts, is usually more important for some traits as gender or occupation Stamatatos et al. (2015). Exploiting the content of texts to find the authors' profile is the primary focus of this thesis work.

Author profiling commonly is faced by two approaches: i) the style based approaches and ii) the content based approaches (Álvarez-Carmona et al., 2016). The approaches based on style captures the way authors write their texts, and the approaches based on content captures the thematic of the texts. Both approaches consist in constructing a vector from the text characteristics to feed a machine learning algorithm to determine the profile of the author.

In the next sections, we describe the different methods proposed in the literature based on style, syntax and content for AP.

4.1 Style Based Approaches

In the literature, the research on Author Profiling proposed a set of stylistic features to enumerate the writing style of authors. Every category of features has their importance to predict demographic features of authors. The combinations of these features were also used to discriminate authors' writing style.

Typically, the style of the author can be captured from three different ways: the style based on characters, the style based on lexicon and style based on syntax.

In this section, it is going to describe the features used to capture the author's style.

4.1.1 Features Based on Characters

A text is a sequence of characters, and various character-based features were defined by researchers to differentiate the texts. For example, Goswami et al. (2009) and Weren et al. (2014b) used the total number of characters. Tam and Martell (2009) the number of capital letters and Gilad Gressel et al. (2014) the frequency of special characters. With the same idea, Baker (2014) extracted the ratio of different features as: capital letters, the white-space characters, the tab spaces, the white spaces, the capital letters, and the numeric data.

A more complicated approach, for the stylistic purposes, consists in extract the most frequent n-grams. The n-grams require no special tools and is language-independent. Nevertheless, the dimensionality of these representations increases, significantly, when it is compared with the word-based approach. Pham et al. (2009), Hernández et al. (2013) and Daneshvar and Inkpen (2018) used frequencies of the most common character 4-grams of the considered documents. Rao et al. (2010) considered character unigrams, trigrams and 5-grams for text characterization. Romania (2015) observed that the best tf-idf features were at character-level where n-gram ranges from 2 to 6.

Table 4.1: Summary of the most common features based on characters

Features description
The total number of characters
Character n-grams
The number of capitalized letters
The frequency of special characters
The ratio of capital letters to the total number of characters
The ratio of white-space characters to the total number of characters
The ratio of tab spaces to the total number of characters
The ratio of white spaces to non-white spaces
The ratio of capital letters to the lower case letters
The ratio of numeric data in the text

Table 4.1 enlists the most common characters-based features for the Author Profiling task.

4.1.2 Lexical Features

Several functions in the grammar enumerate the variety of vocabulary of the text. For instance, a function which finds the ratio between the total number of different stems and the total number of words after applying stemming (Nowson et al., 2015; HaCohen-Kerner et al., 2018).

Mechti et al. (2014) and Castillo et al. (2018) used hapax legomena (i.e., words occurring once) and Hapax dislegomena (the number of words that occur twice) to represent the vocabulary to determine the gender of the author.

As we mentioned before, the most straightforward approach for the researchers is to represent text as vectors of word frequencies. The studies which were focused on Author Profiling were based on the features of word combinations for representing the style. This approach is similar to the conventional Bag Of Words (BoW) representation.

The size of the feature terms also places a predominant role in the document representation. Some works used the top 200 frequent terms as features and incremented up to 50000 frequent terms (Álvarez-Carmona et al., 2015; Mechti et al., 2013; Lopez-Monroy et al., 2013).

To take a benefit of the contextual information, n-grams of words (n consecutive words) were proposed as a textual feature in several works (Nowson and Oberlander, 2006; Gómez-Adorno and Sidorov, 2017; Sanchez-Perez et al., 2017; Martinc et al.,

Table 4.2: Summary of the most common lexical based features

Features description
Number of words
Word n-grams
The number of positive/ negative emotional words
Number of acronyms
The number of Hapax legomena
The number of Hapax dislegomena
List of foreign words
Average word length
The number of capitalized words
The number of words with repetitive letters
The maximum length of a word
The number of words with digits
The ratio of words with length greater than k words to total words
The ratio of words shorter than m letters to total words
The ratio of to the total number of words in the text

2017).

An acronym is an abbreviation, used as a word, which is formed from the initial components in a phrase or a word. The occurrences of acronym words were used by Company and Wanner (2007) as a feature set.

Flekova and Gurevych (2013), for example, used features as:

- The number of words with repetitive characters.
- The number of words with digits.
- The ratio of five letter words to the total words.
- The ratio of three letter words to the total words.
- The ratio of distinct words to the total words.

Table 4.3 enlists the most common lexical based features for the Author Profiling task.

4.1.3 Features Based on Syntax Analysis

The syntax is the set of rules, principles, and processes that govern the structure of sentences in a given language, usually including word order (Carnap, 2014). These aspects have been important for the Author profiling task.

In syntax, the function words have little lexical meaning or have ambiguous meaning and express grammatical relationships among other words within a sentence, or specify the attitude or mood of the speaker (Klammer, 2007; Siirtola et al., 2017). These are considered as structured grammatical words which have a structural relationship with other words in a sentence. These function words include part of speeches, such as pronouns (she, they), determiners (the, that), prepositions (in, of), auxiliary verbs (be, have), modal verbs (may, could), conjunctions(and, but) and quantifiers (some, both).

Some authors have used function words as features and proved that male authors tend to use more prepositions in their writings when compared to female authors (Argamon et al., 2005; Koppel et al., 2005; Ortega-Mendoza et al., 2018). For instance, Gilad Gressel et al. (2014) extracted features from text, which include adjectives, nouns, determiners, pronouns, adverbs and foreign words.

The morpho-syntactic information tags that are assigned to every word token based on the contextual information is a process carried out by a Part Of Speech (POS) tagger (Belinkov et al., 2018). With these POS taggers is possible to identify the style of authors by using POS tags n-grams frequencies or POS tag frequencies (Corney et al., 2002; Pham et al., 2009; Gilad Gressel et al., 2014; Tschuggnall et al., 2017) from a text. Also the proportion of plural and singular nouns, pronouns and proper nouns, the ratio of past and future verb tenses, ratios of comparative and superlative adjectives and adverbs (Flekova and Gurevych, 2013).

Several authors have used the frequency of punctuations in the context of AP (Maharjan et al., 2014; Aleman et al., 2013; Lim et al., 2013; Simaki et al., 2017). The ratio of punctuations to the text was used by Baker (2014), whereas, spelling and grammatical errors were used by Marquardt et al. (2014).

Table 4.3 enlists the most common syntactic based features for the Author Profiling task.

4.2 Content Based Approaches

Some works have shown the importance of the content approaches on the AP task (Ortega-Mendoza et al., 2018), affirming that, the content based features are more discriminative than style features (Reddy et al., 2016).

Table 4.3: Summary of the most common syntactic based features

Features description
POS n-grams
Syntactic n-grams
Frequency of function words
The number of contraction words
Frequency of punctuations
Stop-words
The ratio of singular to plural nouns
The ratio of singular to plural proper nouns
The ratio of singular to plural pronouns
The ratio of punctuations to text
Spelling and grammatical errors
The ratio of future and past verb tenses

The principal aim of the content features is to capture the topics that users talk about and share on social networks. This is the key of the importance of these features, since people tend to talk and write about the same issues if they are from the same group (same age or gender, or location, etc).

There exist several methods for capturing the content of texts in a corpus. These methods have been applied to the AP task. For instance, the bag of words representation (Zhang et al., 2010).

Marquardt et al. (2014) extracted fourteen terms as features from the MRC psycholinguistic database and 68 terms as features from Linguistic Inquiry and Word Count (LIWC) dictionary and the features concerned with negative, positive or neutral sentiment based expressed sentences. MRC data features capture the information about the word frequencies that predict the concepts of psycholinguistic features such as imagery, concreteness and familiarity (Liu et al., 2014). On the other hand, LIWC is able to calculate how people use different categories of words through a wide variety of texts. LIWC allows to determine the degree to which users use words that connote positive or negative emotions, self-references, extended words or words that refer to different categories as sex, eating or religion. LIWC was designed to analyze more than 70 language dimensions (Pennebaker et al., 2001).

Arroju et al. (2015) categorized motion, anger and religion based on frequency of words that are helpful while classifying the age and gender in hotel reviews. Also,

Table 4.4: Summary of the most common content based features

Features description
Frequency of content specific words
Frequency of content specific words categories
Topic specific features
Dictionary topics features
Sentiment words

the authors mentioned that the writing style, word choice, and grammar rule solely depends on the topic of interest and the differences were found with topic variations. It is observed that the gender-specific topic will have an impact in their writing styles. Also, it is observed that female authors tend to write more about wedding styles and fashions, whereas males stress more on technology and politics. These phenomena also occur with reference to the age. People of 20's write more about their college life and the people of 30's write more about marriage, job and politics and more so the teenagers tend to write about their friends and mood swings. With these statistics, it is evident that the content-based features have a dominant role while distinguishing between the authors of different groups (Reddy et al., 2016).

Pavan et al. (2013); Seroussi et al. (2011); Chen and Ren (2017) considered the topic distribution model and used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in order to get the topics in the documents using the probabilistic distribution function.

Table 4.4 enlists the most common content based features for the Author Profiling task.

4.3 Multimodal Based Approaches

As we mentioned before, in the case of social media, the majority of the works have focused on using *content* and *stylistic* text features (Rangel et al., 2013, 2014; López-Monroy et al., 2014; Ortega-Mendoza et al., 2018).

However, in recent years, some works have started to explore others modalities of information in the social media platforms for the author profiling task. One of the most popular has been the images information.

In (You et al., 2014), the authors analyze the behavior of the users posting pictures on Pinterest. For the prediction, they group all images in 34 categories. The authors also extract Scale-invariant feature transform (SIFT) to discover local features for each image in the dataset. Visual words are discovered by clustering all the SIFT features

and defining the center of each cluster as a visual word. In this way, each image can be represented as the frequencies of visual words discovered. With this method, they achieved 71% of accuracy in the task.

In Merler et al. (2015), the authors proposed a union of several information modalities for gender identification in Twitter. They used a Face Based Gender Predictor. Also, they chose 25 categories to label the images. Besides, they represented each images using one thousand labels, extracted from ImageNet, using a convolutional neural network. Finally, they extracted the image color. With all this information the authors achieved 88 % of accuracy.

In (Farseev et al., 2015) the authors used a Facebook corpus with the location, images, and text information. For text features, the authors selected topic representations as LIWC and LDA, and for the visual information, the authors used color and concept from ImageNet. With this information, they achieve 87 % of accuracy for gender and 50 % for age prediction.

In (Estruch et al., 2017) the authors used a dataset with information on three social media platforms: Foursquare, Instagram and Twitter for gender identification. The authors proposed represent the text using LDA (Latent Dirichlet Allocation) (Blei et al., 2003) and LIWC. For the images, they used a representation from ImageNet (Jia et al., 2018). The architecture proposed fuses the information with Deep Learning. With this, they achieved 91 % of accuracy for the gender prediction task.

In (Segalin et al., 2017) the authors analyzed the importance of the images to predict personality. They used a Flickr dataset and extracted the images characteristics with AlexNet (Iandola et al., 2016). With this, they achieved 80 % of accuracy for personality traits.

In (Wendlandt et al., 2017), the authors use a corpus with images and text from students. The work was about classified gender and personality. For the text features, the authors used a representation as the bag of words, n-grams, word2vec (Mikolov et al., 2013a) and LIWC. On the other hand, for images, the authors used several features like color, texture, face detection, circles, and objects. For the objects detection, the authors used the AlexNet representation. With all this information the authors achieve 71 % of accuracy for gender and over of 60 % for personality.

In (Farnadi et al., 2018) the authors analyze the gender, age and personality identification on Facebook. They took advantage of the text representing it through LIWC Tausczik and Pennebaker (2010), the images of the profile represented as 64 facial features using the Oxford Face API (Cao et al., 2010) and the pages that the users liked. Also, they trained an unsupervised deep neural network approach called Node2Vec on

their relational graph. The authors join the information with their architecture named UDMF (User Profiling Through Deep Multimodal Fusion). With all this, the authors achieved more than 90 % for gender and age recognition.

Takahashi et al. (2018) approached the gender identification task in Twitter with RCNN for texts and ImageNet-based CNN for images. They, apply deep learning approach to join text and image information. This approach achieve 85 % of accuracy.

Most of the works use a closed vocabulary for user representation as in (You et al., 2014; Estruch et al., 2017; Segalin et al., 2017; Takahashi et al., 2018). No one tries to apply a method with an open vocabulary. The majority uses representations based on ImageNet, AlexNet or RCNN for their analysis. We propose to use an open vocabulary method under the hypothesis that with an open vocabulary, the representation is better for the user identification task.

4.4 Summary

The Author Profiling task has attracted the attention of the scientific community in recent years. This has caused that numerous approaches have been proposed with the intention of solving the task.

Until recently, most of these works have applied a selection of features, the majority of which are content and style-textual features. Nevertheless, the content-textual features seem to be those who provide more information for the AP task. This is somehow intuitive since AP is not focused on distinguishing a particular author through modeling its writing style, but on characterizing a group of authors. Typically, the principal content features have been: BoW, the frequency of content specific words, the rate of content particular words categories, the topic-specific features, dictionary topics features, and sentiment words. The disadvantage of these features is that some of them are manually selected by experts, as in the case of dictionaries. Also, another problem, in the topic-specific features is that they are parameterized approaches, and it means that to find the best configuration could be difficult.

On the other hand, recent approaches for author profiling tend to use more information in addition to the textual modality. Usually, the main information comes from the images shared by the users. Some works have used low-level information of the images; nevertheless the information extracted by the objects into the images seem to be most valuable. For this, most of the works use closed vocabulary approaches as AlexNet, ImageNet or RNN. These approaches, having a limited vocabulary, may not provide adequate information to represent the user's interests.

Part III

Proposal and Experiments

Chapter 5

Corpora

Non c'è certezza nella scienza se la matematica non può esservi applicata, o se non vi è comunque in relazione.

Leonardo da Vinci

There exist several collections for the evaluation of AP task approaches. Nevertheless, most of them provide just labels for gender, which does not allow to analyze other types of traits. Also, these collections have only textual information. This causes that information that can be valuable for the task is not exploited. For example, images, videos, date or behavior information. On the other hand, there exist corpora which were built from social networks whose main feature is the publication of images as Pinterest or Flickr but, there is no collections for AP that have both text and image information labeled with different traits.

Also, several of the collections for the AP task have information on English-speaking user accounts. Although there are other collections in Spanish, the tagged accounts are from Spain. Although that Mexico is the country with the most Spanish speakers in the world, there is not an exclusive corpus for the evaluation and analysis of mexican social media accounts.

As a contribution of this thesis, we present two novel corpora that have been designed for the Author Profiling task evaluation with text and image information.

First, we present an extension of the well known PAN 14 Twitter corpus (Rangel et al., 2014), with aiming to use a well-known corpus enriching it with image information.

Also, we present a mexican Twitter corpus for the AP task. The specific application of this corpus is in the the analysis of several traits of mexicans Twitter users by text and image information. The data contains for each account the activity schedule on

Twitter, its tweets and its images divided into two categories: i) the images that the user uploads (personal) and ii) the images that the user shares through another person with a re-tweet (extern). This corpus is labeled for gender, place where he/she lives and occupation. The annotation of the data has been accomplished manually.

This chapter is organized as follows. Section 5.1 describes the PAN 14 corpus for the text experiments. Section 5.2 shows the description of the images extension for the PAN 14 Twitter corpus. Finally, Section 5.3 describe the new Mexican Twitter corpus for the author profiling task.

5.1 Pan 14 Corpus

For our experiments, we employed the English dataset from the PAN 14 AP track. This corpus was specially built for studying AP in social media. It is labeled by gender (i.e., female and male), and five non-overlapping age categories (18-24, 25-34, 35-49, 50-64, 65+). Although all documents are from social media domains, four distinct genres were provided: blogs, social media, hotel reviews, and Twitter posts. A more detailed description of how these datasets were collected can be found in (Rangel et al., 2014). Table 5.1 provides some basic statistics regarding the distribution of profiles across the different domains (i.e., genres). It can be noticed that gender classes are balanced, whereas for the age classification task the classes are highly unbalanced. Notably, there are very few instances for the 65+ category.

Table 5.1: Distribution of the gender and age classes across the different social media domains.

Classes	**Genres**			
	Blogs	*Reviews*	*Social-media*	*Twitter*
Female	73	2080	3873	153
Male	74	2080	3873	153
Total:	147	4160	7746	306
18-24	6	360	1550	20
25-34	60	1000	2098	88
35-49	54	1000	2246	130
50-64	23	1000	1838	60
65+	4	800	14	8
Total:	147	4160	7746	306

5.2 Extended PAN 14 Corpus

Images shared by social media users tend to be strongly correlated with their thematic interests as well as to their style preferences. Motivated by these facts, we tackled the task of assembling a corpus considering text and images from Twitter users. Mainly, we extended the PAN-2014 (Rangel et al., 2014) dataset by obtaining images from the already existing Twitter users.

The PAN-2014 dataset includes tweets (only textual information) from English users. Based on this dataset, we obtained more than 42,000 images, corresponding to a subset of 279 profiles in English[1]. The images associated with all of the users were downloaded to existing user profiles, resulting in a new multimodal Twitter corpus for the AP task. Each profile has an average of 304 images.

Tables 5.2 and 5.3 present additional statistics on the values that both variables, gender and age can take, respectively. On the one hand, Table 5.2 divides profiles by age ranges, i.e., 18-24, 25-34, 35-49, 50-64 and 65+. It shows an important level of imbalance, being the 35-49 class the one having the greatest number of users. Nonetheless, the users from the 65+ range are the ones with the greatest number of posted images as well as the lower standard deviation values. It is also important to notice that the users belonging to the 50-64 range share in average a lot of images, but show a large standard deviation, indicating the presence of some users with too many and very few images.

Table 5.2: Statistics of images shared by each age category.

Ages	Profiles	Average images (α)	Average tweets (α)
18-24	17	246.45 ($\pm$80.34)	706.18($\pm$361.76)
25-34	78	286.42 ($\pm$202.65)	796.01($\pm$291.18)
35-49	123	301.74 ($\pm$253.83)	640.41($\pm$362.28)
50-64	54	334.19 ($\pm$238.24)	527.68($\pm$354.24)
65+	7	441.65 ($\pm$102.52)	651.85($\pm$432.28)

On the other hand, Table 5.3 reports some statistics for each gender profile. It is observed a balanced number of male and females users in both corpora as well as a similar number of shared images.

[1]Note that the PAN-2014 corpus includes more profiles, however, for some Twitter users, it was impossible to download their associated images.

Table 5.3: Statistics of images shared by each gender category.

Ages	Profiles	Average images (α)	Average tweets (α)
Female	140	162.21 ($\pm$294.13)	543.53($\pm$395.93)
Male	139	141.76 ($\pm$274.98)	784.88($\pm$265.86)

Figure 5.1: Regional division for Mexico

5.3 Mex-A3T-500 Corpus

To study the characteristics of the different Mexican Twitter profiles, we built a Mexican corpus for author profiling named Mex-A3T-500. Each of the Twitter users was labeled with gender, occupation, and place of residence information. 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 (norte)*, *northwest (noroeste)*, *northeast (noreste)*, *center (centro)*, *west (occidente)*, *and southeast (sureste)*. Figure 5.1 shows the division in the Mexico's map.

5.3.1 Construction of the Corpus

Two human annotators, working three months each, were needed for building this corpus. They applied the following methodology: (i) to find a set of Twitter accounts

Table 5.4: 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

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 with 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 5.4 shows some examples of tweets where users reveal information from their place of residence and occupation.

5.3.2 Statistics

The corpus consists of 500 profiles from Mexican Twitter users. Each profile is labeled with information about the gender, occupation, and place of residence of the user. Tables 5.5, 5.6 and 5.7 present additional statistics on the distribution of user accounts on gender, occupation and location. Table 5.6 divides profiles into the different Mexican regions on the corpus, i.e., north, northeast, northwest, center, west, and southeast. Also, it shows an important level of imbalance, being the center class the one having the greatest number of users, while the north is the class with the lowest. On the other hand, Table 5.7 divides profiles on the eight different occupations on the corpus. It is possible to see that the majority class is the center region whereas the classes with the least instances are the others and sports.

5.3.3 Mexican Important Words

For both traits, we retrieve the top mutual information words for each class.

Figure 5.2 shows the clouds of the most representative words of each region. We

Table 5.5: Mexican author profiling corpus: distribution of the gender trait.

Class	Profiles	Average images (α)	Average tweets (α)
Female	250	715.46 ($\pm$722.89)	1225.00($\pm$868.17)
Male	250	480.90($\pm$459.36)	1500.01($\pm$946.66)

Table 5.6: Mexican author profiling corpus: distribution of the place of residence trait.

Class	Profiles	Average images (α)	Average tweets (α)
North	13	625.23($\pm$442.49)	1594.23($\pm$855.17)
Northwest	80	385.92($\pm$345.95)	1162.17($\pm$866.14)
Northeast	123	460.54($\pm$482.02)	1071.60($\pm$800.66)
Center	191	755.58($\pm$732.74)	1597.83($\pm$922.49)
West	46	611.91($\pm$488.10)	1525.80($\pm$990.62)
Southeast	47	659.12($\pm$732.35)	1284.51($\pm$916.36)

Table 5.7: Mexican author profiling corpus: distribution of the occupation trait.

Class	Profiles	Average images (α)	Average tweets(α)
Arts	38	826.21($\pm$754.71)	1828.23($\pm$834.09)
Student	253	336.57($\pm$259.81)	1184.66($\pm$838.81)
Social	64	1158.15($\pm$867.03)	1362.62($\pm$921.89)
Sciences	25	474.28($\pm$461.97)	1549.64($\pm$947.44)
Sports	12	682.41($\pm$652.27)	1113.00($\pm$892.95)
Administrative	82	894.59($\pm$651.72)	1597.52($\pm$965.65)
Health	15	248.20($\pm$275.05)	1410.20($\pm$1127.04)
Others	11	1026.90($\pm$747.28)	1873.27($\pm$965.63)

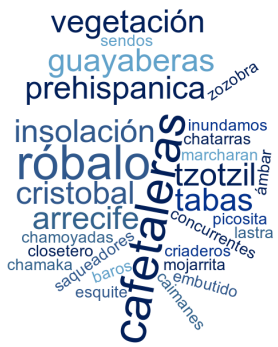
(a) North

(b) Northwest

(c) Northest

(d) Center

(e) West

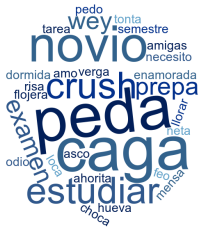


(f) Southeast

Figure 5.2: Clouds of the most representative words of each region

can see that most of the essential words are places and some regionalisms. On the other hand, figure 5.3 shows the most important words for each occupation. Unlike the region traits, the top words are more related with each occupation, for example, Figure 5.3c shows words as violence, politics, rights, etc. for representing the Social class, or Figure 5.3e has words as tournament, sport, team, among others to represent the Sports class.

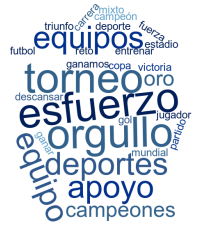
(a) Arts



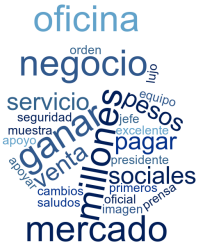
(b) Students

(c) Social

(d) Sciences



(e) Sports



(f) Administrative

(g) Health

Figure 5.3: Clouds of the most representative words of each occupation

Chapter 6

Analysis of Distributional Term Representations

Es ist nicht das Wissen, sondern das Lernen,
nicht das Besitzen, sondern das Erwerben,
nicht das Dasein, sondern das Hinkommen,
was den größten Genuß gewährt.

Carl Friedrich Gauss

This Chapter describes a general framework for Author Profiling using distributional term representations (DTRs). By exploiting DTRs, we aim to represent documents from social media users in a low dimensional and non-sparse space, which captures more discriminative information.

Our goal is to overcome, to some extent, the issues naturally inherited by the BoW representation and build instead a more semantically related representation. Intuitively, DTRs can capture the semantics of a term t_i by exploiting the distributional hypothesis: "words with similar meanings appear in similar contexts". Thus, different DTRs can capture the semantics through the context in different ways and at different levels.

As we mentioned in Chapter 4, traditionally, the Author Profiling task has been approached as a single-labeled classification problem, where the different categories (e.g., *male* vs. *female*, or *teenager* vs. *young* vs. *old*) stand for the target classes. The common pipeline is as follows: *i*) extracting textual features from the documents; *ii*) building the documents' representation using the extracted features, and *iii*) learning a classification model from the built representations. As it is possible to imagine, extracting the relevant features is a key aspect for learning the textual patterns of the different profiles. Accordingly, previous research has evaluated the importance of thematic (content-based) features (Koppel et al., 2002; Poulston et al., 2017) and stylistic characteristics (Bergsma et al., 2012). More recently, some works have also considered learning such representations utilizing Convolutional and Recurrent Neural Networks

(Sierra and González, 2018; Kodiyan et al., 2017; Takahashi et al., 2018).

Although many textual features have been used and proposed, a common conclusion among previous research is that content-based features are the most relevant for this task. The latter can be confirmed by reviewing the results from the PAN[1] competitions (Rangel et al., 2018), where the best-performing systems employed content-based features for representing documents regardless of their genre. This result is somehow intuitive since AP is not focused on distinguishing a particular author through modeling his/her writing style, but on characterizing a group of authors. For example, in (Schler et al., 2006) authors performed an exhaustive study of non-formal documents in order to determine the pertinence of content-based features. They found that stylistic features do not provide any additional information to the learning algorithm. In contrast, content words such as *linux* and *office*, and *love* and *shopping*, showed to be highly discriminant for males and females respectively.

In line with these findings, previous research has focused on evaluating the pertinence of distinct content-based representations for solving the AP task. Mainly, we went beyond the traditional bag of words by considering distributional and topic-based representations. The idea behind both approaches was to develop enriched representations that help to overcome the small-length and high-sparsity issues of social media documents by considering contextual information computed from document occurrence and term co-occurrence statistics. Mainly, we proposed a family of distributional representations based on second order attributes which allow capturing the relationships between terms and profiles and subprofiles (López-Monroy et al., 2015). These representations obtained the best results in the AP tasks at PAN 2013 and PAN 2014 (López-Monroy et al., 2014). Also, we evaluated topic-based representations such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) in the AP task (Álvarez-Carmona et al., 2015), obtaining the best performance at the PAN 2015 as well as showing its superiority against a representation based on manually defined topics utilizing LIWC (Álvarez-Carmona et al., 2016).

Motivated by the good results of our subprofile-specific representation (SSR) (López-Monroy et al., 2015), as well as by the recent use of word embeddings in the AP task (López-Santillán et al., 2018), in this chapter, we present a thorough analysis on the pertinence of *distributional term representations* (DTRs) for solving the problem of AP in social media. We aim to highlight the advantages and disadvantages of this type of representations in comparison with traditional topic-based representations such as LSA and LDA.

[1] A set of shared tasks on digital text forensics: http://pan.webis.de/

In summary, the main contributions of this chapter are:

- We introduce a framework for supervised author profiling in social media domains using DTRs. This framework encompasses the extraction of distributional representation of terms as well as the construction of the authors' representation by aggregating the representations of the terms from their documents.
- We evaluate for the first time the document-occurrence representation (DOR) and the term co-occurrence representation (TCOR) in the AP task. These are two simple and well-known term representations from distributional semantics (Lavelli et al., 2004b).
- We present a comparative analysis of several distributional representations, namely DOR, TCOR, SSR, and word2vec, using the proposed framework for AP. Additionally, we compare their performance against the results from classic bag-of-words and topic-based representations.

For evaluating the proposed framework and performing the analysis of the distinct DTRs, we employed the PAN 14 dataset described at Section 5.1. This corpus was specially built for studying AP in social media domains, as it contains data from blogs, Twitter, and reviews. We performed several experiments aiming at determining the suitability of DTRs for solving the AP task in social media domains. Our initial intuitions suggest that through the use of these representations it will be possible to obtain richer content-based features as well as –for some of them– easily interpretable results. Thus, we carry out an analysis of the obtained results and their relation to different characteristics of the considered text collections such as their lexical complexity, shortness, and class imbalance.

6.1 Distributional Framework for AP

This Section describes a general framework for Author Profiling using distributional term representations (DTRs). Intuitively, DTRs can capture the semantics of a term t_i by exploiting the distributional hypothesis; "words with similar meanings appear in similar contexts". Thus, different DTRs can capture the semantics through the context in different ways and at different levels.

The proposed framework is shown in Figure 6.1 and it comprises two main stages: *i*) to determine the terms' vector representations, and *ii*) to build the document[2]

[2]Hereafter we are going to use the term *document* as a synonym of *user*, under the consideration that all the posts from a user form a document.

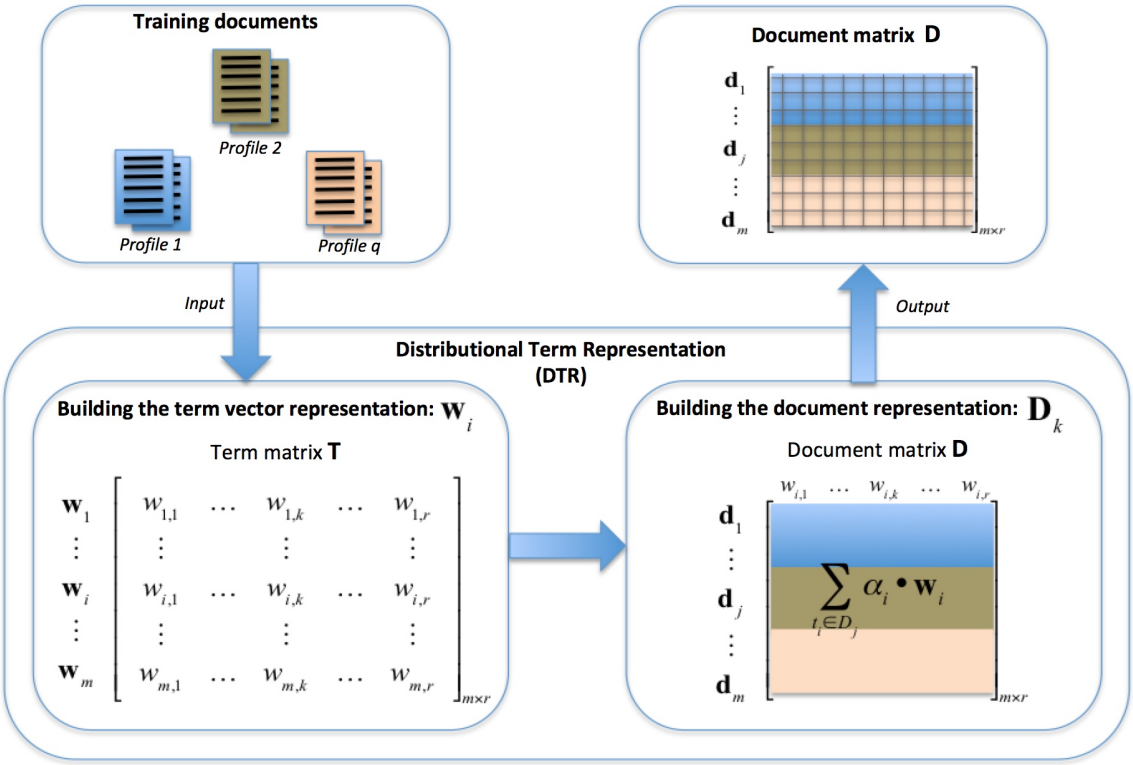


Figure 6.1: General diagram of the proposed framework for building the distributional term representation of documents.

representations. Notice that term representations account for discriminative semantic relationships between terms. Then, document representations are obtained by aggregating the representations of terms that occur in each document, leading to a distributional-based representation. The resultant document representations are non-sparse and capture useful profile information. The way in which terms and documents are represented in each stage is the same. The only difference is on how the semantics of each term is determined, i.e., how the DTR is computed. Once documents are represented, a standard classifier is considered to build an AP model. The rest of the section details the way in which terms and documents are represented, and how the distinct DTRs are obtained.

6.1.1 Distributional Term Representations

As mentioned, our proposed framework requires two steps (Figure 6.1) to represent a document using DTRs. To put things simple, lets consider words in the vocabulary as the base terms for building the DTR. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_n, y_n)\}$ be a training set of $n-$pairs of documents (d_j) and labels/categories $y_i \in \mathcal{C} = \{C_1, \ldots, C_q\}$. Also let $\mathcal{V} = \{t_1, \ldots, t_m\}$ be the collection vocabulary. In this context, DTRs associates each term $t_i \in \mathcal{V}$ with a term vector $\vec{w_i} \in R^r$, i.e., $\vec{w_i} = \langle w_{i,1}, \ldots, w_{i,r} \rangle$. In this notation $w_{i,j}$ indicates the contribution of distributional feature j to the representation of term t_i. This contribution is particular of each DTR and can be computed in a number of ways. In the following sections we describe in detail each of the DTRs that we selected for this study. The second step consists in building the document representations by using the term vectors. More formally, the representation of document a d_j, the vector $\vec{d_j}$, is obtained by using the expression 6.1, where the scalar α_i weighs the relevance of term t_i in the document d_j. Although there are several ways to define this weighting, the most widely used approach is the average of the distribution (i.e., α_i is proportional to the number of terms in the document).

$$\vec{d_j} = \sum_{t_i \in d_j} \alpha_i \cdot \mathbf{w_i} \tag{6.1}$$

Document Occurrence Representation

The document occurrence representation (DOR) can be considered the dual of the TF-IDF representation widely used in the Information Retrieval field (Lavelli et al., 2004a). DOR is based on the hypothesis that the semantics of a term can be revealed by its distribution of occurrence-statistics over the documents in the corpus. A term t_i

that belongs to the vocabulary $\mathcal{V}$ is represented by a vector of weights associated to documents $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,N} \rangle$ where N is the number of documents in the collection and $0 \leqslant w_{i,j} \leqslant 1$ represents the contribution of document d_j to the specification of the semantics of t_i:

$$w_{i,j} = df(t_i, d_j) \log \frac{|\mathcal{V}|}{N_j} \tag{6.2}$$

where N_j is the number of different terms from the dictionary $\mathcal{V}$ that appear in document d_j, $|\mathcal{V}|$ is the number of terms in the vocabulary, and

$$df(t_i, d_j) = \begin{cases} 1 + \log(\#(t_i, d_j)) & \text{if } \#(t_i, d_j) \geqslant 0 \\ 0 & otherwise \end{cases} \tag{6.3}$$

where $\#(t_i, d_j)$ denotes the number of times the term t_i occurs in the document d_j. Intuitively, the more frequent the term t_i is in the document d_j, more important is d_j to characterize the semantics of t_i. In the same way, the more terms contain d_j, the less its contribution in characterizing the semantics of t_i.

Term Co-Occurrence Representation

Term Co-Occurrence Representation (TCOR) is based on co-occurrence statistics (Lavelli et al., 2004a). The underlying idea is that the semantics of a term t_i can be revealed by the terms that co-occur with it across the documents collection. Here, each term $t_i \in \mathcal{V}$ is represented by a vector of weights $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,|\mathcal{V}|} \rangle$ where $0 \leqslant w_{i,j} \leqslant 1$ represents the contribution of term t_j to the semantic description of t_i, and is computed as follows:

$$w_{i,j} = tf(t_i, t_j) \log \frac{|\mathcal{V}|}{\mathcal{V}_k} \tag{6.4}$$

where $\mathcal{V}_k$ is the number of different terms in the dictionary $\mathcal{V}$ that co-occur with t_i in at least one document, and:

$$tf(t_i, t_j) = \begin{cases} 1 + \log(\#(t_i, t_j)) & \text{if } \#(t_i, t_j) \geqslant 0 \\ 0 & otherwise \end{cases} \tag{6.5}$$

where $\#(t_i, t_j))$ denotes the number of documents in which term t_j co-occurs with the term t_i. Intuitively, the more frequent the co-occurrence among the terms t_i and t_j is in a document d, more important is d to characterize the semantics of t_i and t_j. In the same way, the more terms contain d, the less its contribution in characterizing the semantics of t_i and t_j.

Word Embeddings: Word2vec

Recently, a very popular group of related models for producing word embeddings is word2vec (Mikolov et al., 2013b). These models are shallow, two-layer neural networks trained to reconstruct the linguistic contexts of words. Word2vec takes as its input a large corpus of texts and produces a vector space, typically of a few hundreds of dimensions, where each term in the corpus is assigned to a corresponding vector $\vec{w}_i$ in the space. Thus, once the word vectors have been computed and positioned in the vector space, words that share common contexts in the corpus are located close to each another in the space (Mikolov et al., 2013a).

Word2vec employs either one of two model architectures to produce the distributed representation of words: *i*) continuous bag-of-words (CBOW), or *ii*) continuous skip-gram (Mikolov et al., 2013a). In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence the prediction. In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture assigns a higher weight to those nearby context words while more distant context words are considered less important (Mikolov et al., 2013b). The CBOW model is faster than the skip-gram model. However, the later does a better job for handling infrequent words (Zhuang et al., 2017).

In our experiments we built the word embeddings (i.e., vectors $\vec{w}_i$) using the skip-gram model. It mainly considers the conditional probabilities $p(c|t)$ for all terms t and their respective contexts c. Thus, given a corpus D, it aims to set the parameters θ of $p(c|t;\theta)$ so as to maximize the corpus probability:

$$D_{parameters} = \operatorname{argmax}_{\theta} \prod_{(t,c)\in D} p(c|t;\theta) \tag{6.6}$$

The purpose of word2vec is to build an accurate representation of words in a space R^r, where similar vectors correspond to semantically related words (Mikolov et al., 2013a). For example, the difference of vectors for words **France** and **Paris** will be similar to the difference between **Germany** and **Berlin**, since both are nations and capitals, so as the vectors from *elephant* and *dog* since both are animals.

Subprofile Specific Representation

The intuitive idea of the second order attributes consists in representing the terms by their relation with each target class (Li et al., 2011; López-Monroy et al., 2015). This can be done by exploiting occurrence-statistics over the set of documents in each

one of the target classes. In this way, we represent each term $t_i \in \mathcal{V}$ with a vector $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,q} \rangle$, where the scalar $w_{i,k}$ is the degree of association between word t_i and class C_k. Under this DTR, the weight $w_{i,k}$ is directly related to the number of occurrences of term t_i in documents that are labeled with class C_k. The relationship between the i^{th} word and the k^{th} class can be defined according to:

$$w_{i,k} = \sum_{\forall d_j : y_j == C_k} \log_2 \left(1 + \frac{tf(t_i, d_j)}{len(d_j)}\right) \tag{6.7}$$

where $tf(t_i, d_j)$ is the occurrence frequency of the word t_i in the document d_j, and $len(d_j)$ indicates the number of words in d_j. The $\log_2$ function aims to soften the relevance of highly frequent words. In our case, the classes are the different profiles that we aim to identify. Thus, d_j represents documents that were produced by users with the same profile, e.g., same gender or same age rate.

The computed raw weights $w_{i,k}$ from Equation 6.7 can be directly used to build the term vectors. However, a term representation based on raw $w_{i,k}$ weights is sensitive to highly unbalanced data. Thus, in order to produce the final $\vec{w_i}$ representation, we consider applying two normalizations: a kind of row-based normalization to consider the proportion of the $|V|$ terms in each class, and then a kind of column-based normalization to take into account the weights computed for the $|C|$ classes, making weights $w_{i,k}$, comparable among classes. Finally, $\vec{w_i}$ can be seen as a probability distribution of t_i over the distinct k author profiles.

In López-Monroy et al. (2015), second order attributes were modeled at sub-profile level; mainly, it was proposed to cluster the instances from each target in order to generate several subclasses. The idea was to consider the high heterogeneity of social media users. Utilizing this process, the set of target classes C will now correspond to the set of all subgroups from the original target classes. This new representation is called *Subprofile-based Representation* (SSR), and is considered one of the state-of-the-art representations for AP.

6.2 Experiments and Results

This section explains the experiments that were carried out using the proposed framework. As we have previously mentioned, we aim at determining the pertinence of distributional term representations (DTRs) to the AP task in distinct social media domains. Accordingly, this section is organized as follows: first, Subsection 6.2.1 explains the experimental settings for all the experiments, then, Subsection 6.2.2 describes the results obtained by each DTR in the four different social media domains.

6.2.1 Experimental Setup

Preprocessing: For computing the DTRs of each social media domain we considered the 10,000 most frequent terms. We did not remove any term, i.e., we preserved all content words, stop words, emoticons, punctuations marks, etc. In one previous work López-Monroy et al. (2015) demonstrated that preserving only the 10,000 most frequent words is enough for achieving a good representation of the documents.

Text representation: The different DTRs were computed as described in Section 6.1. For the particular case of the word2vec representation, we employed two distinct configurations: *w2v-wiki*, where the model was trained using a Wikipedia dataset, and *w2v-sm*, where we trained a model for each one of the domains using the available training documents[3]. In both cases, we used the word2vec skip-gram architecture as suggested in (Zhuang et al., 2017). Regarding the representation of the documents, in all cases, for all DTRs, we built their vectors by averaging the vectors from their words.

Classification: Following the same configuration as in previous works (please refer to (Álvarez-Carmona et al., 2016)), in all the experiments we used the linear Support Vector Machine (SVM) from the LIBLINEAR library with default parameters Fan et al. (2008).

Baseline: As baseline we employed the traditional bag-of-words (BoW) representation. We also compared the results from the different DTRs to those obtained by topic modeling representations such as LSA and LDA as well as to those from the top systems from the PAN@2014 AP track.

Evaluation: We performed a stratified 10 cross-fold validation (10-CFV) strategy. For comparison purposes, and following the PAN guidelines, we employed the accuracy as the main evaluation measure. Finally, we evaluated the statistical significance of the obtained results using a 0.05 significance level utilizing the Wilcoxon Signed-Ranks test since is recommended for these cases by Demšar (2006).

6.2.2 Results

This section is organized as follows: first, we show the results from different DTRs for the age and gender classification tasks; then, we compare them against some topic-based representations and the best approaches from PAN 2014.

[3]with the default parameters of gensim for python 2.7 for both cases

Table 6.1: Accuracy results obtained by the DTRs for the *age* classification problem. Last column depicts the average performance of each approach across the distinct genres.

Approach	Text genres				*Average*
	Blogs	Reviews	Social Media	Twitter	
DOR	**0.49***	**0.36***	**0.38***	0.47*	**0.42**
TCOR	0.38*	0.31	0.32	0.35	0.34
w2v-wiki	0.37	0.31	0.36*	0.43	0.36
w2v-sm	0.38*	0.30	0.36*	0.41	0.36
SSR	0.48*	0.34*	0.37*	**0.48***	0.41
Baseline	0.34	0.28	0.32	0.42	0.34

Age and Gender Identification Using DTRs

Tables 6.1 and 6.2 show the accuracy and F-measures results respectively for *age*. Also, Table 6.3 shows the obtained results for the *gender*[4] classification problems respectively. Each row represents one of the described DTRs, i.e., DOR, TCOR, word2vec, and SSR, while the last row represents the baseline results. Every column refers to a distinct social media genre, and the last column (i.e., *Average*) represents the average performance for each method across all genres. In these tables, the best results are highlighted using boldface, and the star symbol ($\star$) indicates the differences that are statistically significant concerning the baseline results (in accordance to the used test; for details refer to Section 6.2.1).

Obtained results indicate that all DTRs, except for TCOR, outperformed the baseline method. In particular, DOR and SSR show statistically significant differences. These two methods obtained comparable results, being DOR slightly better than SSR in 5 out of 8 classification problems, which is an interesting result since SSR was among the winning approaches at PAN 2014. On the other hand, we attribute the low accuracy results showed by TCOR to the strong expansion that it imposes to the document representations. Considering direct term co-occurrences causes the inclusion of many unrelated and unimportant terms in the document vectors, and, therefore, it complexities the extraction of profiling patterns.

Finally, another essential aspect to notice is the fact that both *w2v-wiki* and *w2v-sm* obtained similar results in each of the classifications problems, although the former

[4]Since the gender trait is balanced it is not necessary show the F-measure table because the results are very similar with the accuracy

Table 6.2: F-measure results obtained by the DTRs for the *age* classification problem. Last column depicts the average performance of each approach across the distinct genres.

Approach	**Text genres**				***Average***
	Blogs	Reviews	Social Media	Twitter	
DOR	**0.38**	**0.30**	**0.29**	**0.35**	**0.33**
TCOR	0.22	0.21	0.23	0.31	0.24
w2v-wiki	0.21	0.21	0.23	0.30	0.23
w2v-sm	0.20	0.20	0.24	0.28	0.23
SSR	0.36	0.27	0.26	0.33	0.30
Baseline	0.21	0.19	0.23	0.21	0.30

Table 6.3: Accuracy results obtained by the employed DTRs for the *gender* classification task. Last column depicts the average performance of each approach across the distinct genres.

App.	**Text genres**				***Average***
	Blogs	Reviews	Social Media	Twitter	
DOR	**0.78***	**0.69***	0.52	0.70	0.66
TCOR	0.56	0.62	0.41	0.54	0.53
w2v-wiki	0.75*	0.64	0.52	0.69	0.65
w2v-sm	0.74	0.64	0.54	0.66	0.64
SSR	**0.78***	**0.69***	**0.55***	**0.71**	**0.68**
Baseline	0.72	0.62	0.52	0.70	0.64

Table 6.4: Comparison of the best DTRs against topic-based methods in the *age* classification task. The last column shows the average performance of each approach across the different genres.

Approach	Text genres				*Average*
	Blogs	Reviews	Social Media	Twitter	
DOR	**0.49**†	0.36†	**0.38**†‡	0.47‡	**0.42**
SSR	0.48†	0.34†	0.37‡	**0.48**†‡	0.41
LDA	0.44	0.27	0.37	0.47	0.38
LSA	**0.49**	**0.37**	0.36	0.45	**0.42**
Maharjan et al. (2014)	0.38	0.33	0.36	0.44	0.37
Villena Román and González Cristóbal (2014)	0.39	0.31	0.35	0.41	0.36
Weren et al. (2014a)	0.45	0.37	0.42	0.52	0.44

learned the embeddings from a corpus that is not thematically and neither stylistically similar to the social media content. We presume these results could be explained by the relatively small size of the social media training collections, and, at the same time, by the large size and broad coverage of the used Wikipedia dataset, which has a vocabulary of 1,033,013 words.

DTRs vs. Topic-Based Representations

Tables 6.4 and 6.5 compare the results from DOR and SSR, the best DTRs according to the previous results, against the results from two well-known topic-based representations, namely LDA and LSA. For both topic modeling representations, the tables only show their best result in each domain obtained after evaluating a varying number of topics. The results marked with a ‡ indicate that they are significantly better than LSA, whereas results marked with † indicate that they are significantly better than LDA. Details on the test of statistical significance are given in Section 6.2.1.

The obtained results clearly show that LDA was unable to obtain good results in both classification problems. This performance is in line with our previous findings reported in Álvarez-Carmona et al. (2016). We hypothesize this poor performance is due to the dataset sizes; bigger corpora are needed for extracting relevant and discriminative topics.

Regarding the LSA results, it is possible to observe, on the one hand, that for *age* classification (refer to Table 6.4), its average performance is similar to the one from DOR, i.e., 42%. However, the only domain in which LSA outperforms DOR is in the reviews dataset. Nonetheless, there is no significant difference between these results.

Table 6.5: Comparison of best DTRs against topic-based methods in the *gender* classification task. The last column depicts the average performance of each approach across the different genres.

Approach	Text genres				*Average*
	Blogs	Reviews	Social Media	Twitter	
DOR	**0.78**†	**0.69**†	0.52	0.70†	0.66
SSR	**0.78**†	**0.69**†	**0.55**†‡	**0.71**†	**0.68**
LDA	0.61	0.55	0.52	0.64	0.58
LSA	**0.78**	**0.69**	0.53	0.70	0.67
Maharjan et al. (2014)	0.57	0.66	0.53	0.66	0.60
Villena Román and González Cristóbal (2014)	0.64	0.68	0.54	0.51	0.59
Weren et al. (2014a)	0.82	0.71	0.57	0.78	0.72

On the other hand, for *gender* classification (Table 6.5), LSA was not able to improve any result from DOR and SSR. It is important to mention that, although their results are comparable, LSA is a parametric method, and, therefore, tunning is required.

Comparison Against Other Approaches

This section presents a comparison of the proposed framework, employing the DOR and SSR distributional term representations, against the works from PAN@2014 which reported results in the training partition. We mainly consider the following three works: Maharjan et al. (2014), based in a combination of term n-grams with different n values using the MapReduce programming paradigm; Villena Román and González Cristóbal (2014), which used a two-level classifier composed of a document-oriented classifier with a term vector model representation in combination with a voting strategy; Weren et al. (2014b), which considered a method based on information retrieval ideas.

The bottom rows from Tables 6.4 and 6.5 show the results for the age and gender classification tasks. As it is possible to observe, the employed DTRs outperform the results from Maharjan et al. (2014) and Villena Román and González Cristóbal (2014) in both tasks and for all genres. Nevertheless, they could not improve the results from Weren et al. (2014b). It is important to consider that in Weren et al. (2014b) authors reported the best results obtained after an exhaustive tunning stage, i.e., the selection of the best classification method from a broad family of algorithms, as well as the selection of the best subset of features for each social media genre. Also, this approach obtained an erratic performance during the test phase of PAN@1014 (Rangel et al., 2014), especially for the Twitter domain, where it achieved an accuracy 15-points lower

than SSR-based approach (López-Monroy et al., 2014). Hence, the overall outlook seems to indicate that the proposed framework is more robust than most previous approaches for AP, as some DTRs are nonparametric and therefore they do not require for a tunning stage.

6.2.3 Getting to Know the Learned Concepts: a Qualitative Analysis

Previous experiments indicate that the DOR (Lavelli et al., 2004a) representation has several advantages in comparison to other approaches, for example, it does not require tunning any parameter, it allows building relatively compact non-sparse representations, and it obtains very competitive results. Moving a step forward, we performed an analysis of the *interpretability* of DOR. As explained in Section 6.1.1, in the DOR representation each document is represented by its relation to other documents. Thus, in the context of AP, it means that each user is represented by its relation to or similarity with other users from the corpus. Accordingly, the features with greater IG are the more discriminative users among the classes (i.e., target profiles).

To exemplify this, Table 6.6 shows the top ten words from the three most representative male and female profiles. Words were selected according to their TF-IDF values. As it is shown, each one of these users tends to write about different topics, nevertheless, show interesting and important content aspects of their classes. For example, Male 1 writes about books and pictures, Male 2 writes about technology, and Male 3 writes about exercises and diets. In the case of females, notice that Female 1 writes about networks, accessories, and shopping, Female 2 writes about food and drinks ingredients, and Female 3 write about social media management. Given that these profiles are the most representative "features" for each class, it is possible to say that, DOR representation allows the classifier to assign an unknown profile to the class where are those users with similar distributional use of words.

6.2.4 On the Role of the Collection Characteristics

In order to enrich the performed analysis, we carried out some initial experiments for analyzing the role or influence of different characteristics from the collections over the performance of the considered DTRs. In particular, we measured the correlation between the value of these characteristics and the improvement in accuracy of the DTRs over the baseline result.

- Type Token Ratio (**TTR**). It measures the vocabulary richness of the collection as the ratio of different terms to the total number of terms in the collection (Laufer

Table 6.6: The three most representative *male* and *female* users from the blogs genre (used as features in the DOR representation). Showed words correspond to the top ten words according to their TF-IDF value for each user.

Male 1	Male 2	Male 3	Female 1	Female 2	Female 3
photo	google	game	bloglovin'	smooth	knowledge
book	width	exercise	pinterest	acidity	media
draw	sms	breakfast	style	palate	management
sketches	windows	baseball	instagram	alcohol	social
teaching	smile	salad	twitter	cherry	culture
learn	keyboard	dinner	accessories	licorice	change
learning	success	eating	facebook	vanilla	content
spent	delete	running	necklace	cheese	conversation
anxiety	forget	protein	shoes	chocolate	meeting
brothers	funny	training	shopping	aromatic	blogs

and Nation, 1995).

- Lexical Density (**LD**). This is another vocabulary richness measure. It is calculated as the ratio of content terms (nouns, verbs, adjectives, adverbs) to the total number of terms in the collection (Laufer and Nation, 1995).

- Sophistication (**SX**). It indicates the percentage of sophisticated terms compared with the number of terms in the collection. A term is considered as sophisticated if its length is greater than the average terms length with a standard deviation Lu (2012);

- Shortness (**S**). It is calculated as the arithmetic mean of the lengths of the document from the collection (Tellez et al., 2009). Thus, the higher its value, the longer the documents.

- Class imbalance (**In**). It is calculated as the standard deviation of the differences between the current and the ideal number of documents for each category. The ideal number of documents per category is defined as the ratio of the number of documents in the collection and the number of categories. The higher the value of class imbalance, the more unbalanced the collection is (Tellez et al., 2009).

- Hardness (**H**). It measures the vocabulary overlap among the texts from the different categories (profiles) in the collection. A collection is harder to classify if users from different profiles share much vocabulary. As well as if users from the same class write about very different things. For its computation, we considered all the combinations of two categories from the collection, and for each of them,

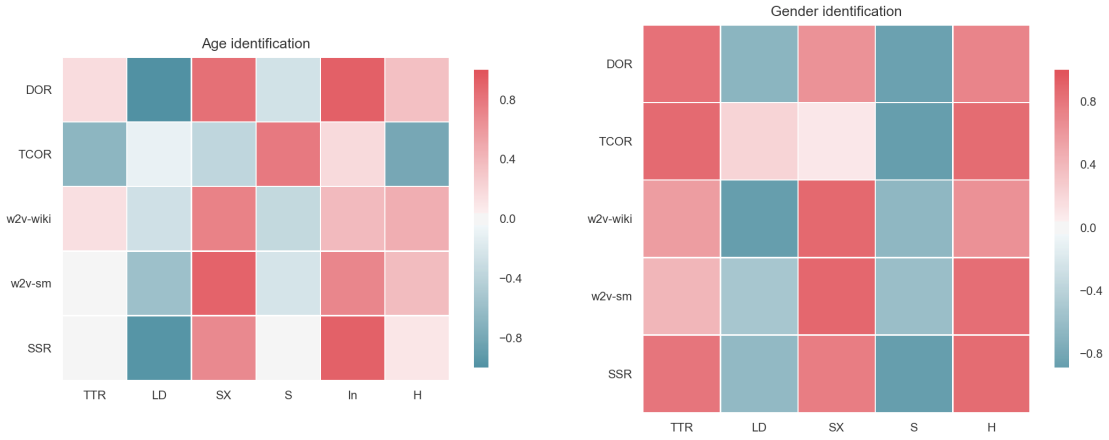


Figure 6.2: Correlation map between the obtained improvement from all considered DTRs and several collection characteristics. The accuracy improvement is obtained by comparing the result of each DTR against the BoW method.

we calculated the text overlap average. The text overlap is calculated using the Jaccard coefficient (Pinto and Rosso, 2007).

Figure 6.2 shows two heat maps that indicate the level of correlation between the evaluated characteristics and the improvement in accuracy of the DTRs over the BoW approach. Although this analysis was basic and straightforward, it helped in discovering that all DTRs, except TCOR, show some common patterns. On the one hand, the figures show a positive correlation concerning the vocabulary sophistication (SX) and the collection hardness (H). That means that the DTRs tend to obtain better results than BOW for collections having more strange, unusual words as well as for collections showing a considerable overlap among the vocabulary of the different profiles. On the other hand, they show a negative correlation with the shortness (S) and lexical density (LD) characteristics, indicating that DTRs tend to obtain better results than BoW for collections of short texts containing few content terms. Also, this analysis shows that, for the age classification problem, DTRs results positive correlate with the class imbalance (In). All these identified characteristics of the DTRs are important since it is quite common to have imbalanced training sets, short texts and a lot of unusual words in most social media applications.

6.3 Conclusions

Through the years, the majority of research work has focused on identifying and extracting relevant features for building a suitable representation that learns the textual

patterns of distinct profiles. In this same line, a common conclusion among previous research is that content-based features provide more important information than style-based features.

By previous findings, this work aimed to determine the pertinence of using distributional term representations (DTRs) for solving the author profiling task. Our intuition was that utilizing the semantics of the documents by exploiting the distributional hypothesis, it is possible to identify more discriminative content-based information. Thus, we proposed a novel framework for supervised author profiling in social media domains using DTRs, mainly, we studied for the first time the DOR and TCOR representations and compared their performance against other popular DTRs as well as against two well-known topic-based approaches, namely LDA and LSA.

For our experiments, we considered the PAN 14 dataset, which was specially designed for evaluating the AP problem in social media domains. The obtained results indicate that DTRs are suitable for the AP task in social media domains. Mainly, DOR representation achieved the best accuracy results, while enabling interpretability of results. Moreover, a detailed analysis of these results shows that they are robust to class imbalance as well as able to take advantage of the different characteristics of social media data such as their shortness and lexical richness.

A significant advantage of the proposed framework using DTRs is its robustness across different social media genres. Contrary to LDA and LSA, our proposed approach using the DOR term representation does not require any tuning phase, which is a tremendous benefit for social media applications.

Finally, this work represents the first attempt for carefully determining the pertinence of distinct DTRs approaches for the AP task as well as for analyzing the impact of topic-based methods by its own, which has not been done before. In concordance with previous research, our obtained results allow us to assert that content-based features are attributes for the AP task. The performed analysis in this work will help future research in AP since it represents a thoughtful and broad study on popular forms of representations.

Chapter 7

Multimodal Author profiling approach

La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. Ici, nous avons réuni théorie et pratique: Rien ne fonctionne... et personne ne sait pourquoi!

Albert Einstein

Currently, the vast majority of the research that has been done focuses in only text for detecting the author profile. Nevertheless, the nature of social media is multimodal, i.e., users share images, videos, audios, and texts, representing all of these valuable sources of information as well. For example the tweets' polarity (Rangel and Rosso, 2016; Rosso and Rangel, 2017), the pages that are liked by the user (Cao et al., 2010), emoticons (op Vollenbroek et al., 2016), ratios of links, hashtags, user mentions (Rangel et al., 2017; Cunha et al., 2014), the user's name (Liu and Ruths, 2013). Nevertheless, one of the most popular types of information for the AP task is the images that the users share. Some authors have used this information obtaining good results (You et al., 2014; Schwartz et al., 2013a; Wendlandt et al., 2017).

Traditionally, images are represented through a vector of semantic categories, where each component of the vector indicates, with a real value, the level of confidence whether that category is (or not) present in the analyzed image. The main hypothesis of these approaches indicates that similar users with analogous profiles will share related semantic information by means of their posted images.

Thus, a supervised learning approach is trained to distinguish among a set of pre-established semantic categories (labels), i.e., a supervised image annotation method. Although this type of strategies obtain an acceptable performance in the AP problem, they have a major disadvantage, they work under a closed-vocabulary configuration,

meaning that they are able to identify only those categories that were present in the training dataset. Consequently, if an image contains unknown objects, this type of methods could provide a set of erroneous tags. Hence, supervised image annotation methods are not suitable for describing the vast amount and highly diverse type of interests reflected in the posted images in current social networks environments.

Considering the latter scenario, in this chapter, we propose a novel and effective framework to identify user's profiles by means of employing an unsupervised image annotation approach. To face the AP problem, the proposed approach considers an open vocabulary strategy for labelling images, i.e., our method do not depend on a set of pre-established categories. Thereby, images are transformed into a list of textual categories describing the objects contained inside an image. Our main hypothesis establishes that authors from the same group, tend to share similar images which can be accurately labeled using an open vocabulary annotation approach.

Once all the information from the images is extracted, the proposed framework allows to employ distinct text-based representations for training a supervised approach to distinguish between profiles. We foresee this work will represent an important contribution for the development of novel methodologies for the multimodal author profiling problem, as well as motivate further research from the intelligent systems and text mining research communities.

The principal contributions of this chapter are as follows:

1. We show the pertinence of image annotation methods with open vocabulary for obtaining relevant information from posted images, which can later be used for improving the classification results in the AP task.

2. We evaluate the complementarity between the information obtained by the proposed image annotation method and the given textual information in the original post, i.e., a multimodal setting which combines visual and textual information.

3. We evaluate the cross-lingual capacity of the proposed AP framework employing the unsupervised image annotation method. We hypothesize that even if users are separated by the language gap, if they have a similar profile, they will post semantically related images as well.

The rest of the chapter is organized as follows. Section 7.1 explains the open vocabulary approach and how it is included in the proposed AP framework. Section 7.2 depicts the experimental setup as well as the obtained results. Section 7.4 shows

the fusion schemes results among text and images approaches. Also, Section 7.5 shows the results of the cross-lingual experiments. finally, in Section 7.6, we draw some conclusions.

7.1 Open Vocabulary Method for Images Representation

In order to exploit the semantic information of images for the AP task, we use a methodology based on the Automatic Image Annotation (AIA) task. AIA has the goal of assigning labels to images aiming at describing their visual content. AIA methods can be defined under supervised and unsupervised scenarios. In a supervised scenario, images are annotated according to a set of previously known labels, which were learned from training pairs (image, labels). In this case, the annotation process is usually defined as a classification task with a closed vocabulary. On the contrary, the unsupervised scenario uses a reference image collection where each instance is seen as a document. Every document in the collection consists of a pair of image and text, where the text is associated to the image. In this case, labels are derived from the associated texts allowing to unsupervised AIA (UAIA) methods use large vocabularies for annotating images.

Supervised and unsupervised scenarios have advantages that can be beneficial for the AP task. Supervised scenarios provide AIA methods that are quite competitive in the assignation of labels to images. However, they are limited to a closed number of labels, only those that are defined in the ground truth, e.g., ImageNet, one of the most popular AIA, has only a set of 1000 labels for annotating every possible image. In contrast, UAIA methods are capable of using a more significant number of labels. In this way, each image of each profile is transformed into text with the list of labels returned. For this reason, we are using a non-supervised approach and therefore an open vocabulary UAIA.

In order to obtain the corresponding labels from the objects contained in a picture, to use them as features in the machine learning process, we propose to use the labeler proposed in Pellegrin et al. (2016). In the next section, we explain this method.

7.1.1 Unsupervised Automatic Images Annotation

The proposed method is based on an Unsupervised Automatic Images Annotation (UAIA) of Pellegrin et al. (2016). The idea of the method is to relate an image with the words when this appears in the same context.

The UAIA method aims at taking advantage of the visual-textual relationships to

label images. In an offline step, it discovers associations between textual and visual terms for later using them to find relevant images to describe the content of the image. UAIA takes advantage of the interactions of textual t_i and visual v_i representations of images and their texts, respectively. It considers every pair of image and its associated text in the reference collection as a multimodal object that can be described under two different views: a visual view, v_i, and a textual one, t_i. The main idea behind this approach is that both, textual and visual views, have salience in the same objects when represented by the two different features.

Using these views of the multimodal objects (image+text), it performs a multimodal indexing Escalante et al. (2012). That is a representation that merges the two modalities, in this case, it represents text utilizing visual descriptors: each word in the vocabulary is associated with a visual representation. Hence, the multimodal indexing can be seen as a bunch of visual prototypes, one per word, where each prototype gathers the main characteristics of images sharing the corresponding word. Then, comparing the image to be annotated with the visual prototypes, we can retrieve words that can be used to describe the content of the image.

1. Multimodal indexing. The hypothesis behind this is that each word can be associated with a visual prototype. In this way, any query image described by visual features can be readily compared with prototypes, and we can determine what concepts (words) are most related to the query. For the construction of the multimodal indexing, we rely on multimodal term co-occurrence statistics, where the terms are both textual and visual features. The multimodal indexing associates each word with a distribution of weights over the visual features forming a visual prototype for each word. The multimodal indexing is built offline and is obtained as follows:

$$M = T^t \cdot V$$

Where M is the multimodal indexing obtained by the product of textual T and visual features V of multimodal objects. Therefore, we can see that $M_{i,j} = \sum_{l=1}^{m} w_{i,l} \cdot v_{l,j}$ is a scalar value that expresses the degree of association between word i and visual-feature j, across the whole collection of m images. In this way, each row of the matrix M can be seen as a visual prototype that is associated with one word. Thus, M is a matrix of size $|X| \times |Y|$, where $|X|$ is vocabulary in the corpus and $|Y|$ is the length of the visual features. that is, the dimension is determined by the sizes of the features that represent both textual and visual features respectively.

2. Content-based image retrieval (CBIR). To determine the labels that have to be associated with an image, a CBIR stage is performed, taking the image to be labeled a query and matrix M as the reference collection. The similarity between visual representations can be estimated with any measure, for instance, L_1-based similarity function, (1), or the cosine similarity (depending on the visual representation):

$$sim(I_q, W_i) = cosine(v_q, M_i) = \frac{v_q * M_i}{|vq| * |Mi|}$$

where v_q is the visual representation of the query image I_q and M_i is the visual prototype for word W_i, i.e., the $i-th$ row of matrix M. The query image is compared with each row of M and a score is generated for each word as follows:

$$score(W_i) = cosine(v_q, M_i)$$

The output of the global UAIA method is the set formed by the n-words with the highest score; these words are used to label the image.

3. Particularization problem. This approach could have a problem with the particularization of the labels. Figure 7.1 shows some labels for two images. For the car image, we can see that the method chooses car brands, while for the dog images the method selects dog breeds. If the labeler is not wrong at all, it is possible that using such a particular language could affect the results. For the car images maybe it is better to have the car label several times than to have several car names. Thus, for facing this problem we propose two approaches:

   - To use WordNet to extract the hyperonyms of the labels extracted from the images. For this, we add the hyperonyms of each label. We add three different levels in the WordNet tree. This variant will be called *Hyper*.
   - To use LSA for grouping the labels in the same contexts. We experiment with several values for the number of topics. This variant will be called *LSA-I*.

With the labels, we propose to use a bag-of-labels (BoL) representation.

For the experiments we use the collections described at Section 5.2, which is an extension of the PAN 14 corpus. This collection includes the original textual information and the images that users share in their Twitter accounts were also included. On the other hand, we also use the collection described at Section 5.3. This corpus is a collection of Mexican user Twitter accounts with the gender, location and occupation traits.

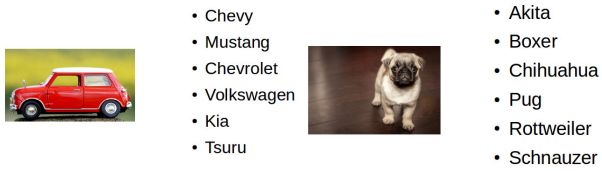


Figure 7.1: Some labels for two images with the UAIA method

7.2 Experimental Settings

In this section, we describe the configuration used in the experiments.

First, we apply the UAIA method. Each image in each profile is transformed into a list of labels. We take the 50[1] greatest confidence value labels for each image. Then, we apply a *bag-of-labels (BoL)*.

As baseline, we are going to compare the images BoL results with a group of textual representations with good results in the textual corpus. The representations are i) text BoW with 2000 and 10000 most frequent words and ii) Latent Semantic Analysis (LSA). Also, the BoL result is compared with the average of the colors of the images per user.

To compare the effectiveness of the open-vocabulary approach, we also compared the result of the BoL-UAIA method with the results obtained by AlexNet and RCNN as they are used in Girshick et al. (2014); Merler et al. (2015); Segalin et al. (2017). Also, we compare the result of ImageNet + color average as it is used in Farseev et al. (2015).

For the classification in all the experiments, we used the LibLINEAR classifier Fan et al. (2008) and performed a stratified ten cross-fold validation (10-CFV). Finally, as measures, we use accuracy and F1-measure.

7.3 Open vs. Closed Vocabulary Approaches Results

In this section, we present the results obtained by the BoL representation compared with the baseline and with closed vocabulary methods.

7.3.1 Results for the Extended PAN 14 Corpus

Table 7.1 shows the accuracy comparison of the labelers approaches with the baselines. BoW obtains his best results with 10k vocabulary. Nevertheless, the best result with textual information is obtained by LSA. Also, the color average obtains the worst

[1] Derived of empiric experiments variants several numbers from 10 to 200 where 50 was the best result

Table 7.1: Accuracy and F1-measure for age trait

Approach	Accuracy	F1-measure
BoW(2k)	0.39(0.12)	0.20
BoW(10k)	0.40(0.10)	0.21
LSA(k=100)	0.46(0.07)	0.20
color Avg	0.35(0.08)	0.20
BoL-AlexNet(Krizhevsky et al., 2012)	0.39(0.07)	0.22
BoL-RCNN (Girshick et al., 2014)	0.39(0.07)	**0.23**
ImageNet (Krizhevsky et al., 2012)	0.39(0.08)	0.21
ImageNet + color Avg Farseev et al. (2015)	0.38(0.07)	0.20
BoL-UAIA	0.40(0.06)	0.23
LSA-I(k=20)*	**0.49(0.06)***	0.24
Hyper(level=1)*	0.42(0.06)*	**0.30**

results. On the other hand, the BoL-UAIA representation obtains a better F1-measure result compared with the LSA result.

For these results, we can see how the visual information could be as important as the text information for the age trait. Finally, we notice that the fusion with ImageNet and color gets worse results than only ImageNet. Also, we can see that the open vocabulary method overcomes the closed vocabulary based labelers. For this trait, the best accuracy results were obtained by LSA-I method whereas the best F1-measure result is obtained by the Hyper method.

Like the previous section, the goal is to observe the results of the labelers for the author profiling task for the gender trait compared the open vocabulary method with the closed vocabulary approaches.

In Table 7.2 we show the results of the labelers and the baselines. As we can see, the best results are the BoW with 10 k vocabulary. Nevertheless, the BoL-UAIA method result overcomes the others labelers and the color average. Even, the LSA-I method overcomes the BoL-UAIA result. These results could be comparable with the BoW result.

These results, give evidence again, that the open vocabulary approach can represent better the users than the closed vocabulary methods for the AP task.

Table 7.2: Accuracy and F1-measure for gender trait

Approach	Accuracy	F1-measure
Bow(2k)	0.74(0.07)	0.74
Bow(10k)*	**0.75(0.07)**	**0.75**
LSA(k=200)	0.72(0.12)	0.72
color Avg	0.53(0.08)	0.53
BoL-AlexNet (Krizhevsky et al., 2012)	0.58(0.08)	0.58
BoL-RCNN (Girshick et al., 2014)	0.56(0.06)	0.56
ImageNet (Krizhevsky et al., 2012)	0.69(0.08)	0.69
ImageNet + color Avg Farseev et al. (2015)	0.63(0.09)	0.63
BoL-UAIA	0.70(0.08)	0.70
LSA-I(k=100)*	0.74(0.09)	0.74
Hyper(level = 1)	0.68(0.09)	0.68

7.3.2 Results for the MEX-A3T-500 Collection

In table 7.3 the gender results are shown for the MEX-A3T-500 collection. As we can see, as the gender trait in the PAN 14 corpus, the best result is obtained by the BoW representation. Nevertheless, the BoL-UAIA result overcomes the rest of the methods for images, including, the closed vocabulary approaches. The best methods for the images is the LSA-I reaching the LSA result.

In table 7.4 we show the results for the occupation trait in the MEX-A3T collection. Here, it is possible to observe as the open vocabulary methods overcome the closed vocabulary methods. Also, the best result is obtained by the LSA-I method like the LSA; nevertheless, the best F1-measure result is achieved by the Hyper method.

Finally, in Table 7.5 the results for the location trait are shown. As in the previous results, the open vocabulary methods overcome the closed vocabulary methods as much for accuracy as for F1-measure. Again, the best result is obtained by the LSA method.

As we can see, for all trait the open vocabulary methods overcome the closed vocabulary results. This can provide evidence that to apply an open vocabulary approach for author profiling could be a good option for the task to represent the images for AP.

Table 7.3: Accuracy and F1-measure for gender trait on the MEX-A3T-500 corpus

Approach	Accuracy	F1-measure
Bow(2k)	0.72(0.06)	0.72
Bow(10k)*	**0.80(0.04)**	**0.80**
LSA(k=400)*	0.79(0.03)	0.79
color Avg	0.61(0.05)	0.61
BoL-AlexNet (Krizhevsky et al., 2012)	0.65(0.04)	0.65
BoL-RCNN (Girshick et al., 2014)	0.64(0.03)	0.64
ImageNet (Krizhevsky et al., 2012)	0.65(0.07)	0.65
ImageNet + color Avg Farseev et al. (2015)	0.64(0.08)	0.64
BoL-UAIA*	0.74(0.05)	0.74
LSA-I(k=20)*	0.79(0.09)	0.79
Hyper(level = 1)*	0.73(0.04)	0.73

Table 7.4: Accuracy and F1-measure for occupation trait on the MEX-A3T-500 corpus

Approach	Accuracy	F1-measure
Bow(2k)	0.59(0.05)	0.30
Bow(10k)*	0.64(0.04)	0.34
LSA(k=50)*	**0.65(0.06)**	0.25
color Avg	0.48(0.06)	0.20
BoL-AlexNet (Krizhevsky et al., 2012)	0.52(0.02)	0.23
BoL-RCNN (Girshick et al., 2014)	0.54(0.04)	0.24
ImageNet (Krizhevsky et al., 2012)	0.56(0.04)	0.26
ImageNet + color Avg (Farseev et al., 2015)	0.53(0.03)	0.24
BoL-UAIA*	0.63(0.04)	0.34
LSA-I(k=50)*	**0.65(0.05)**	0.34
Hyper(level = 1)*	0.64(0.04)	**0.36**

Table 7.5: Accuracy and F1-measure for the location trait on the MEX-A3T-500 corpus

Approach	Accuracy	F1-measure
Bow(2k)*	0.51(0.03)	0.34
Bow(10k*)	0.52(0.05)	0.37
LSA(k=300)*	**0.71(0.07)**	**0.57**
color Avg	0.35(0.02)	0.21
BoL-AlexNet (Krizhevsky et al., 2012)	0.35(0.04)	0.24
BoL-RCNN (Girshick et al., 2014)	0.35(0.05)	0.23
ImageNet (Krizhevsky et al., 2012)	0.36(0.04)	0.26
ImageNet + color Avg (Farseev et al., 2015)	0.35(0.03)	0.25
BoL-UAIA*	0.44(0.06)	0.28
LSA-I(k=100)*	0.50(0.06)	0.31
Hyper(level = 1)*	0.44(0.05)	0.27

7.4 Complementary Information: Open vs Closed Vocabulary

Since we have two sources of information for the author profiling task, it is possible to take advantage of the images and text results. From here, the next question arises: What method works best merging image and text information?

To try to answer this question we fuse the information of BoW and DOR (the best text representations options according to the study of the previous chapter) methods with the different closed vocabulary methods (AlexNet, RCNN, and ImageNet) and with UAIA methods to observe if the fusion with the open vocabulary method overcomes the closed vocabulary.

For this, we apply two approaches to mix the two types of information:

- *Early fusion*. Given two vector spaces the early fusion scheme consists in combining both spaces obtaining a single space. For this fusion, we are concatenating the different spaces, whose dimension will be $m + n$ where m represents the dimension of the first space and n the dimension of the second space(Snoek et al., 2005).

- *Late fusion*. For this work, we apply a stacking scheme. We use the predicted label of the classification results of each approach to the fusion. The representation will be k predicted labels where k is the number of representations to merge and the $i-th$ label where $0 < i \leqslant k$ represents the output of the $i-th$ approach (Yang et al., 2008).

7.4.1 Fusion Results

Table 7.6 shows the results with all combinations for the early and late fusions for all traits in both corpora. For these combinations we use BoW and DOR. In Table, it is possible to observe that for each trait and with both fusion schemes, the best result is obtained by the methods based on open vocabulary approach. It is evidence that these approaches overcome the closed vocabulary results ones when fusion is applied.

LSA-I obtains the best individual accuracy result for the age trait in the Pan 14 collection (0.49) and Hyper F-measure (0.30). Nevertheless, these results are overcome by the combination Late(DOR+LSA-I) achieved 0.50 of accuracy and 0.42 of F-Measure.

For the gender trait in the Pan 14 collection, the best individual result is obtained by BoW with 0.75 accuracy and F-Measure. Nevertheless, with the Late(DOR+BoL) combination is possible to achieve 0.79.

In the case of the gender trait in the Mex-A3T-500 corpus, the best result is 0.80 but, once again, this result is overcome by the Late(BoW+LSA-I) with 0.82.

For the occupation trait, the combination Late(DOR+LSA-I) with 0.68 of accuracy and Late(DOR+Hyper) with 0.41 of F-Measure overcome the 0.65 of accuracy obtained by the LSA-I method and the 0.36 of F-Measure obtained by Hyper.

Finally, for the location trait, the best individual result was achieved by LSA with 0.71 of accuracy and 0.57 of F-Measure. In this case, the best combination cannot overcome the accuracy result (0.69). Nevertheless, the F-Measure is exceeded by the Late(DOR+Hyper) combination with 0.59.

7.4.2 Combining the Principal Approaches

Since it is possible to observe the advantage of combining text representations and open vocabulary approaches for image representation, we apply the fusion schemes for the best text representations (BoW, LSA, and DOR) and the image representations (BoL, LSA-I, and Hyper).

Table 7.7 shows the results of these fusion approaches for all traits. The first two columns represent the accuracy and F-Measure obtained by the early fusion scheme, and the last two represent the accuracy and F-Measure for the late fusion.

It is possible to see that late fusion obtains the best results in all cases. It is, possibly, because the difference of the dimensionality of each space in early fusion causes that the learning algorithm do not capture all valuable information correctly.

Tables 7.8 and 7.9 show the results of the text and images representation late fusion compared with the best combination results in order to observe if more information is

Text app.	Fusion	Method	Pan 14				MEX-A3T-500					
			Age		Gender		Gender		Occupation		Location	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BoW	Early	AlexNet	0.39(0.05)	0.21	0.60(0.03)	0.60	0.66(0.05)	0.66	0.60(0.03)	0.26	0.39(0.04)	0.23
		RCNN	0.39(0.06)	0.22	0.59(0.09)	0.59	0.67(0.04)	0.67	0.60(0.04)	0.27	0.40(0.05)	0.27
		ImageNet	0.40(0.03)	0.22	0.66(0.03)	0.66	0.69(0.04)	0.69	0.62(0.04)	0.30	0.44(0.03)	0.30
		BoL	0.42(0.10)*	0.23	**0.77(0.08)***	**0.77**	0.73(0.03)*	0.73	0.64(0.05)*	0.32	**0.47(0.06)***	0.32
		LSA-I	**0.47(0.07)***	0.26	0.75(0.04)*	0.75	**0.79(0.04)***	**0.79**	**0.65(0.03)***	**0.33**	**0.47(0.05)***	**0.33**
		Hyper	0.44(0.06)*	**0.30**	0.72(0.07)*	0.72	0.74(0.07)*	0.74	0.64(0.04)*	0.31	0.45(0.03)	0.32
	Late	AlexNet	0.43(0.07)	0.25	0.62(0.04)	0.62	0.70(0.04)	0.70	0.62(0.05)	0.32	0.43(0.04)	0.35
		RCNN	0.42(0.07)	0.26	0.62(0.05)	0.62	0.70(0.05)	0.70	0.62(0.03)	0.32	0.43(0.07)	0.38
		ImageNet	0.45(0.07)	0.25	0.68(0.07)	0.68	0.75(0.07)	0.75	0.63(0.07)	0.34	0.57(0.08)	0.40
		BoL	0.49(0.06)*	0.40	0.75(0.07)*	0.75	0.80(0.09)*	0.80	0.65(0.03)*	0.38	0.65(0.03)*	0.55
		LSA-I	**0.50(0.08)***	**0.41**	**0.78(0.05)***	**0.78**	**0.82(0.06)***	**0.82**	**0.67(0.04)***	**0.40**	0.66(0.05)*	0.57
		Hyper	0.44(0.03)	0.40	0.74(0.04)*	0.74	0.80(0.05)*	0.80	0.66(0.03)*	**0.40**	**0.69(0.08)***	**0.57**
DOR	Early	AlexNet	0.40(0.07)	0.24	0.60(0.04)	0.60	0.67(0.08)	0.63	0.61(0.06)	0.31	0.34(0.04)	0.30
		RCNN	0.41(0.05)	0.25	0.60(0.07)	0.60	0.66(0.07)	0.66	0.62(0.05)	0.30	0.41(0.09)	0.30
		ImageNet	0.40(0.05)	0.23	0.65(0.07)	0.66	0.69(0.05)	0.69	0.63(0.03)	0.29	0.45(0.07)	0.31
		BoL	0.43(0.08)*	0.24	**0.79(0.04)***	**0.79**	0.74(0.05)*	0.74	0.65(0.03)*	0.34	0.49(0.03)*	0.35
		LSA-I	0.48(0.07)*	0.25	0.75(0.09)*	0.75	0.76(0.09)*	0.76	**0.66(0.09)***	0.36	0.66(0.03)*	0.37
		Hyper	0.47(0.06)*	0.25	0.73(0.05)*	0.73	**0.79(0.05)***	**0.79**	**0.66(0.05)***	**0.38**	**0.67(0.04)***	**0.40**
	Late	AlexNet	0.42(0.05)	0.26	0.64(0.08)	0.64	0.71(0.06)	0.71	0.61(0.06)	0.32	0.45(0.07)	0.33
		RCNN	0.41(0.03)	0.27	0.64(0.04)	0.64	0.71(0.04)	0.71	0.63(0.06)	0.33	0.44(0.08)	0.38
		ImageNet	0.45(0.03)	0.26	0.69(0.04)	0.62	0.72(0.04)	0.75	0.64(0.05)	0.35	0.59(0.07)	0.41
		BoL	0.48(0.04)*	0.40	0.73(0.07)*	0.73	0.79(0.05)*	0.79	0.65(0.07)	0.39	0.66(0.08)*	0.57
		LSA-I	**0.50(0.04)***	**0.42**	**0.74(0.09)***	**0.74**	0.79(0.08)*	0.79	**0.68(0.05)***	0.39	**0.68(0.06)***	0.58
		Hyper	0.48(0.06)*	0.41	**0.74(0.07)***	**0.74**	**0.80(0.04)***	**0.80**	0.66(0.05)*	**0.41**	**0.69(0.03)***	**0.59**

Table 7.6: Fusion schemes of the different images methods with BoW and DOR

Table 7.7: Fusion schemes for author profiling traits

Trait	Early Accuracy	Early F-measure	Late Accuracy	Late F-measure
Age	0.46(0.09)	0.24	**0.55(0.06)**	**0.46**
Gender (Pan 14)	0.78(0.05)	0.78	**0.81(0.05)**	**0.81**
Gender (Mex-A3T-500)	0.77(0.06)	0.77	**0.86(0.03)**	**0.86**
Occupation	0.67(0.04)	0.36	**0.71(0.04)**	**0.47**
Location	0.54(0.08)	0.37	**0.73(0.03)**	**0.70**

Table 7.8: Late scheme compared with the best accuracy results

Trait	Late	Best	Improvement
Age	0.55(0.06)	0.50(0.08)	10.0 %
Gender (Pan 14)	0.81(0.05)	0.79(0.04)	2.5 %
Gender (Mex-A3T-500)	0.86(0.03)	0.82(0.06)	4.8 %
Occupation	0.71(0.04)	0.68(0.04)	4.4 %
Location	0.73(0.03)	0.71(0.07)	-2.8 %

captured with the principal representations than with all the combinations presented in Table 7.6. The tables show how this configuration overcomes all results.

These results provide evidence that among the textual and images information there is complementarity.

Figure 7.2 shows the decision tree obtained from the Pan 14 for the gender trait. The input of the algorithm was a predicted labels of each method (text and images). This tree shows that the most important feature for the model is Hyper feature. From now on, it is possible to follow the ways to choose the classes.

Table 7.9: Late scheme compared with the best F-measure results

Trait	Late	Best	Improvement
Age	0.46	0.42	9.5 %
Gender (Pan 14)	0.81	0.79	3.7 %
Gender (Mex-A3T-500)	0.86	0.82	4.8 %
Occupation	0.47	0.41	14.6 %
Location	0.70	0.59	18.6 %

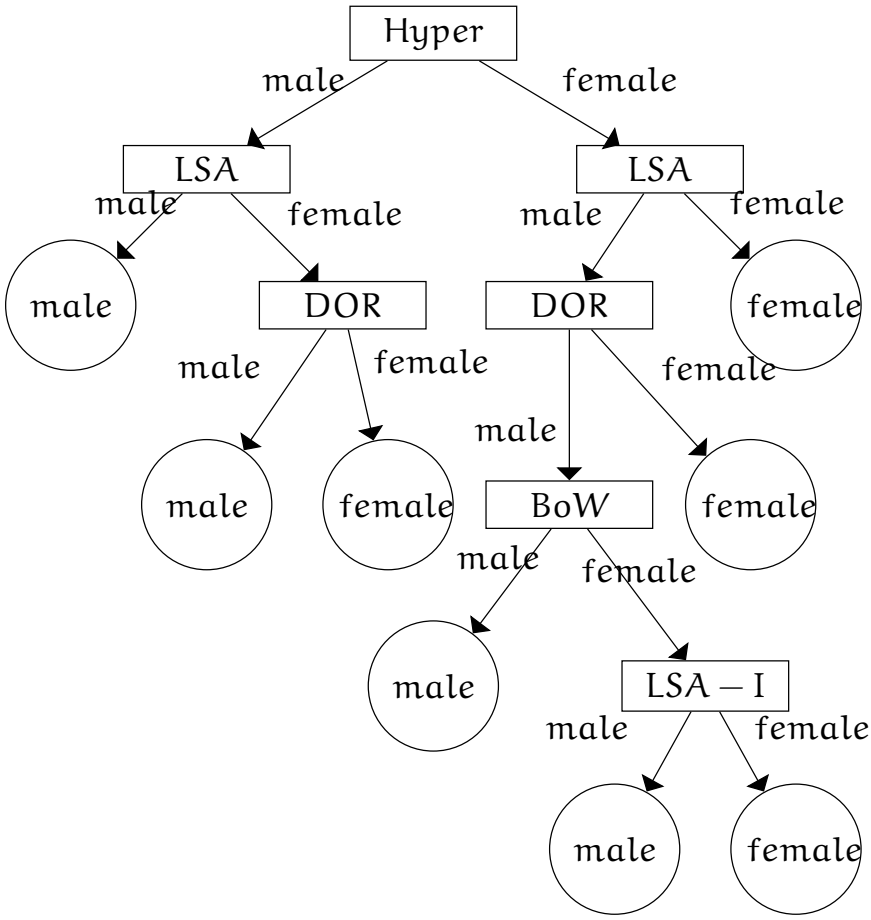


Figure 7.2: Decision Tree for gender for English corpus.

7.4.3 Important Images for Author Profiling

In order to determine the most representative images per class, we built an image retrieval system. First, we took the list of labels of each image. In these lists, we apply the mutual information method. This method will give more value to the most discriminative labels of each class. Now, for each class, the labels with positive mutual information value are taken. This new list of labels would represent the ideal image for each class. We compare this list with all images list of labels with the Jaccard coefficient. The images with the greatest Jaccard value are retrieved. The Jaccard coefficient is computed as the expression 7.1. where $Image_{MI}$ is the set of the top mutual information labels for some class, and $Image_i$ is the $i-th$ image compared with the $Image_{MI}$.

$$Jaccard_i = \frac{|Image_{MI} \cap Image_i|}{|Image_{MI} \cup Image_i|} \tag{7.1}$$

Figure 7.3 shows the most relevant images that men share in the corpus. As we can see, men mostly share images about sporting events and cars. On the other hand, Figure 7.4 also shows the important images for women. Unlike men, women share more flowers, puppies, and women.

Figure 7.3: The most relevant images for men

Figure 7.4: The most relevant images for women

7.5 Cross-Language Gender Prediction Through Images

Since we have an English corpus (PAN 14) and a Spanish corpus (MEX-A3T-500) we can observe if it is possible to carry on a cross-language method under the hypothesis

Table 7.10: Cross language results with LSA-I

Train	Test	Accuracy	F1-measure	Male	Female
English	English(k=100)	0.72(0.12)	0.72	0.71	0.72
Spanish	English (k=100)	0.60(0.7)	0.55	0.70	0.40
Spanish-English	English (k=300)	**0.96(0.03)**	**0.96**	0.96	0.96
Spanish	Spanish (k=20)	0.79(0.09)	0.79	0.79	0.79
English	Spanish (k=50)	0.64(0.06)	0.59	0.46	0.72
Spanish-English	Spanish (k=50)	**0.80(0.03)**	**0.80**	0.80	0.80

Table 7.11: Cross language results with Hyper

Train	Test	Accuracy	F1-measure	Male	Female
English	English	0.68(0.09)	0.68	0.68	0.68
Spanish	English	0.59(0.5)	0.59	0.60	0.58
Spanish-English	English	**0.80(0.04)**	**0.80**	0.80	0.80
Spanish	Spanish	0.73(0.04)	0.73	0.72	0.74
English	Spanish	0.62(0.03)	0.62	0.62	0.62
Spanish-English	Spanish	**0.78(0.06)**	**0.78**	0.78	0.78

that images are language independent.

For this, we experiment with both corpora combining the information. Since the corpora only share the gender trait, we test with this trait only.

Table 7.10 shows the result for the combination of train and test with the LSA-I approach. As we can see, the best result when we test with the English corpus is for the train with both corpora. Also, the same thing happens when we test for the Spanish corpus. On the other hand, Table 7.11 shows the results if we apply the Hyper approach. Like the previous approach, the best results are obtained when we combine both corpora.

These results give evidence that it is possible to use the images in corpora of different sources. This seems intuitive since the images are language independent.

7.6 Conclusions

Recently, some works have made efforts to solve the author profiling task with multimodal information. Commonly, the approaches proposed transform the images in a set of labels of the objects in each image. With this, it is possible to combine textual

and image information. Traditionally, the closed vocabulary approaches were used in this way.

In this chapter, we presented an open vocabulary-based approach to represent images information for the author profiling task.

The results obtained show the advantage of the open vocabulary over the closed vocabulary approaches. The proposed approach overcomes the traditional closed vocabulary approach. This indicates that to have more options to describe each image improves the quality of the representation.

Also, we can conclude that, it is possible to combine textual and images information for the task. The open vocabulary approaches show best results compared with the closed vocabulary approaches. This behavior is constant with each trait in both corpora. This indicates that the open vocabulary approaches are robust for data from different sources.

Finally, we show that it is possible to use information from another corpus, even if the corpus is in another language. This seems reasonable, taking into account that images are language independent.

Part IV

Conclusions

Chapter 8

Conclusions and Future Work

I eftychía synístatai sto na boreís na enoseis tin archí me to telos.

Pythagóras

In this thesis work, we faced the author profiling problem with multimodal information, particularity with text and images.

For this work, we presented an extension of the well-known PAN 14 text corpus including image information. Also, we introduced a novel multimodal collection from Mexican Twitter accounts named MEX-A3t. This helps to encourage the participation of the scientific community in the author profiling task.

For the text modality, we presented a framework based on the distributional term representation family (DTR's) showing their advantages and utility for the author profiling task compared with other content-based methods. Derived from the framework, we participated at the *PAN 2015* evaluation forum for the AP task, one of the most important forums worldwide for the AP. As a result of our participation, we obtained the 1st place in one of the most important competitions worldwide applying DTR's representations[1]. For more details of our participation, see the appendix A.

For the image modality, we presented a scheme based on an open vocabulary approach, comparing it with closed vocabulary approaches and showing its effectiveness. Besides, we apply the early and late fusion schemes for both modalities showing that it is possible to combine the text and image information and overcome their results.

Besides, from the Mexican corpus, we organized the evaluation forum named *"MEX-A3T"* at IberEval 2018. This is the first forum organized exclusively for the Mexican Spanish. The highlights of the forum are described at the appendix B. Also, we built a word2vec model trained with the Mexican user's tweets. This is the first

[1] https://pan.webis.de/clef15/pan15-web/author-profiling.html

model trained with Mexican Spanish labeled tweets. For more details about this model, see the appendix C.

In the following sections, we list the derived contributions, conclusions, and proposed future work from this thesis.

8.1 Contributions

This thesis has contributed with the following:

- A language-independent framework for author profiling based on the distributional hypothesis. Specifically, we provided empirical evidence regarding the pertinence of the DOR representation for solving the posed task across several social media domains.
- A multi-modal approach based on an open vocabulary image annotation technique. We showed that using this type of unsupervised image annotation techniques outperforms currently supervised image annotation strategies.
- We evidenced the level of complementariness among textual information and images' information. Particularly, we showed that open-vocabulary image annotation strategies provide more relevant information than those obtained from closed-vocabulary image annotation techniques
- We were able to demonstrate that a late-fusion strategy (stacking) allows learning algorithms to benefit the most from the DTR's representation and the information obtained from the open vocabulary image annotation method.

Additionally, we present two novel collections for the author profiling task with multimodal information. First, an extension of the well-known PAN 14 Twitter English corpus. As far as we know, this multimodal collection is the first one that is labeled for gender and age for English tweets. Also, we introduce a multimodal Mexican corpus for author profiling with labels for gender, location and occupation traits. This collection is the first one with only Mexican Spanish tweets.

8.2 Conclusions

As a result of this thesis, the following conclusions were obtained:.

- DTR's have advantages in the author profiling task compared with other approaches to capture the content of the texts. In particular, DOR presents the

best behavior, besides that DOR is not a parameterized approach, which causes it to be a simpler and more efficient approach to this task. Also, a significant advantage of DOR is its robustness across different social media genres, contrary to others approaches.

- Automatic image annotation based on open vocabulary approaches is better to represent the images than the closed vocabulary approaches for the Author profiling task. With this approach, it is possible to determine the profiles only with the images information. For gender, the results are over the 70 % of accuracy, for age almost to 50 % of accuracy, for occupation on 65 % and for location on 50 % of accuracy. Some results are comparative with the textual modality.
- To apply a generalization step seems to work well to represent the image information of the profiles. For most traits, these approaches overcome the baselines approaches. To group with LSA was better than the approach that generalize the labels with WordNet.
- There is complementarity among the textual and image modalities since it is possible to overcome the individual results with fusion schemes. Also, the best results are obtained with open vocabulary approaches. The best scheme to fuse this information is the stacking approach. With these approaches it is possible to classify gender, age, occupation, and location more effective than the rest of available methods, improving up to 10 % of accuracy and up to 18 % of F-measure, compared to the best individual results.
- It is possible to use image information from another corpus, even if the corpus is in another language. This seems reasonable, taking into account that images are language independent.

8.3 Future Work

We propose the following list of possible future work.

- Until now, we use only content information from the images. We propose to observe the behavior of the style information, for instance, the size of the images, color, quality or how often the images are uploaded.
- Deepen the cross-lingual study for image corpora. This study would have the idea that image information could break the language barrier and therefore, it could serve to train in some language and test in another.

- Apply different alternatives to combine information. For instance, explore Deep Learning approaches to combine information. It makes sense feed Deep Learning architectures specialized to mix information the the input would be the text and image information.
- Analyze the competence of applying the approaches described in this thesis for other essential traits. Some recent works have faced the author profiling problem for traits like depression, bulimia, anorexia or other mental disorders. The idea is to determine from the images shared on social media if a user has some mental disorder.
- Use the author profiling prediction for other machine learning tasks where some demographics traits are relevant for a classification process as the sentiment analysis.

8.4 Publications

As a result of this thesis work, the following list the papers derived from this research:

- **Journal**
  - **Álvarez-Carmona, M. Á.**, Pellegrin, L., Montes-y-Gómez, M., Sánchez-Vega, F., Escalante, H. J., López-Monroy, A. P., Villaseñor-Pineda, L. & Villatoro-Tello, E. (2018). A visual approach for age and gender identification on Twitter. Journal of Intelligent & Fuzzy Systems, 34(5), 3133-3145. (2 cites)
  - **Álvarez-Carmona, M. Á.**, Villatoro-Tello, E., Montes-y-Gómez M.,, & Villaseñor-Pineda, L. (2019). A Comparative Analysis of Distributional Term Representations for Author Profiling in Social Media. Journal of Intelligent & Fuzzy Systems. (Accepted).
- **Congress**
  - **Álvarez-Carmona, M. Á.**, López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Escalante, H. J. (2015). INAOE's participation at PAN'15: Author profiling task. Working Notes Papers of the CLEF. (37 cites)
  - **Álvarez-Carmona, M. Á.**, López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Meza, I. (2016, November). Evaluating Topic-Based Representations for Author Profiling in Social Media. In Ibero-American Conference

on Artificial Intelligence (pp. 151-162). Springer International Publishing. (6 cites)

- **Organized evaluation forum**
  - **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) (11 cites)
- **Divulgation**
  - **Álvarez-Carmona, M. Á.**, VIllaseñor Pineda, L., Villatoro-Tello, E., (2016). Determinación del perfil de autores en redes sociales con información multimodal. Tech. Rep. CCC-16-007, Instituto Nacional de Atrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Santa María Tonantzintla, México, CP 72840.
  - Montes-y-Gómez, M., Villaseñor-Pineda, L., Escalante, H. J., and **Álvarez-Carmona M. Á.**:"Dime qué posteas y te diré quién eres". Saberes y ciencias (2017).
  - Carrera-Trejo, J. V., **Álvarez-Carmona, M. Á.**. & Villaseñor-Pineda, L.: Identificación del perfil de usuario en Twitter utilizando recursos semánticos. Comia, Mérida Yucatán. (2018). 57–69.
- **Secondary papers**
  - Villegas, M. P., Garciarena Ucelay, M. J., Fernández, J. P., **Álvarez Carmona, M. Á.**, Errecalde, M. L., & Cagnina, L. (2016). Vector-based word representations for sentiment analysis: a comparative study. In XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). (7 cites)
  - **Álvarez-Carmona, M. Á.**, Franco-Salvador, M., Villatoro-Tello, E., Montes-y-Gómez, M., Rosso, P., & Villaseñor-Pineda, L. (2018). Semantically-informed distance and similarity measures for paraphrase plagiarism identification. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-8.

References

Abbasi, A., Chen, H., 2005. Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE 20 (5), 67–75.

Aggarwal, C. C., 2015. Mining text data. In: Data Mining. Springer, pp. 429–455.

Aha, D. W., Kibler, D., Albert, M. K., 1991. Instance-based learning algorithms. Machine learning 6 (1), 37–66.

Aleman, Y., Loya, N., Ayala, D. V., Pinto, D., 2013. Two methodologies applied to the author profiling task. In: CLEF (Working Notes). Citeseer, pp. 1–8.

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., Escalante, H. J., 2015. Inaoe's participation at pan'15: Author profiling task. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum 1391.

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I., 2016. Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. Springer, pp. 151–162.

Amigó, E., Carrillo-de Albornoz, J., Almagro-Cádiz, M., Gonzalo, J., Rodríguez-Vidal, J., Verdejo, F., 2017. Evall: Open access evaluation for information access systems. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 1301–1304.

Andrews, S., Tsochantaridis, I., Hofmann, T., 2003. Support vector machines for multiple-instance learning. In: Advances in neural information processing systems. pp. 577–584.

Aragón, E. M., López-Monroy, A. P., 2018. 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.

Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J. W., 2005. Lexical predictors of personality type. In: Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America.

Argamon, S., Koppel, M., Fine, J., Shimoni, A. R., 2003. Gender, genre, and writing style in formal written texts. Text-The Hague then Amsterdam then Berlin 23 (3), 321–346.

Argamon, S., Koppel, M., Pennebaker, J. W., Schler, J., 2009. Automatically profiling the author of an anonymous text. Communications of the ACM 52 (2), 119–123.

Aronoff, M., 2017. The handbook of linguistics. John Wiley & Sons.

Arroju, M., Hassan, A., Farnadi, G., 2015. Age, gender and personality recognition using tweets in a multilingual setting. In: 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction. pp. 23–31.

Baker, C., 2014. Proof of concept framework for prediction. In: CLEF (Working Notes). pp. 1110–1115.

Bakhtin, A., Szlam, A., Ranzato, M., Grave, E., 2018. Lightweight adaptive mixture of neural and n-gram language models. arXiv preprint arXiv:1804.07705.

Bekkerman, R., Allan, J., 2004. Using bigrams in text categorization. Tech. rep., Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., Glass, J., 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. arXiv preprint arXiv:1801.07772.

Bentolila, I., Zhou, Y., Ismail, L. K., Humpleman, R., May 21 2015. System, method, and software application for targeted advertising via behavioral model clustering, and preference programming based on behavioral model clusters. US Patent 20,150,143,414.

Bergsma, S., Post, M., Yarowsky, D., 2012. Stylometric analysis of scientific articles. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 327–337.

Bi, N., Suen, C. Y., Nobile, N., Tan, J., 2018. A multi-feature selection approach for gender identification of handwriting based on kernel mutual information. Pattern Recognition Letters, 1–10.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022.

Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, pp. 177–186.

Boutell, M. R., Luo, J., Shen, X., Brown, C. M., 2004. Learning multi-label scene classification. Pattern recognition 37 (9), 1757–1771.

Breiman, L., 2017. Classification and regression trees. Routledge.

Breiman, L., et al., 1996. Heuristics of instability and stabilization in model selection. The annals of statistics 24 (6), 2350–2383.

Bu, Y., Zou, S., Liang, Y., Veeravalli, V. V., 2018. Estimation of kl divergence: optimal minimax rate. IEEE Transactions on Information Theory 64 (4), 2648–2674.

Cai, D., He, X., Li, Z., Ma, W.-Y., Wen, J.-R., 2004. Hierarchical clustering of www image search results using visual, textual and link information. In: Proceedings of the 12th annual ACM international conference on Multimedia. ACM, pp. 952–959.

Cao, Z., Yin, Q., Tang, X., Sun, J., 2010. Face recognition with learning-based descriptor. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 2707–2714.

Carnap, R., 2014. Logical syntax of language. Routledge.

Castillo, E., Cervantes, O., Vilariño, D., 2018. Author profiling using a graph enrichment approach. Journal of Intelligent & Fuzzy Systems 34 (5), 3003–3014.

Chang, E., Goh, K., Sychay, G., Wu, G., 2003. Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology 13 (1), 26–38.

Chen, C., Ren, J., 2017. Forum latent dirichlet allocation for user interest discovery. Knowledge-Based Systems 126, 1–7.

Company, J. S., Wanner, L., 2007. How to use less features and reach better performance in author gender identification. In: The 9th edition of the Language Resources and Evaluation Conference (LREC). pp. 26–31.

Corney, M., De Vel, O., Anderson, A., Mohay, G., 2002. Gender-preferential text mining of e-mail discourse. In: Computer Security Applications Conference, 2002. Proceedings. 18th Annual. IEEE, pp. 282–289.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20 (3), 273–297.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory 13 (1), 21–27.

Cruz-Roa, A., Caicedo, J. C., González, F. A., 2011. Visual pattern mining in histology image collections using bag of features. Artificial intelligence in medicine 52 (2), 91–106.

Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., Almeida, V., 2014. He votes or she votes? female and male discursive strategies in twitter political hashtags. PloS one 9 (1), e87041.

Daneshvar, S., Inkpen, D., 2018. Gender identification in twitter using n-grams and lsa. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018).

Darwish, S. M., 2016. Combining firefly algorithm and bayesian classifier: new direction for automatic multilabel image annotation. IET Image Processing 10 (10), 763–772.

De Andrés, J., Pariente, B., Gonzalez-Rodriguez, M., Fernandez Lanvin, D., 2015. Towards an automatic user profiling system for online information sites: Identifying demographic determining factors. Online Information Review 39 (1), 61–80.

de Haro-García, A., Pérez-Rodríguez, J., García-Pedrajas, N., 2018. Combining three strategies for evolutionary instance selection for instance-based learning. Swarm and Evolutionary Computation 42, 160–172.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research 7 (Jan), 1–30.

Douglas, J., Burgess, A. W., Burgess, A. G., Ressler, R. K., 2013. Crime classification manual: A standard system for investigating and classifying violent crime. John Wiley & Sons.

Eftekhar, A., Fullwood, C., Morris, N., 2014. Capturing personality from facebook photos and photo-related activities: How much exposure do you need? Computers in Human Behavior 37, 162–170.

El, S. E. M., Kassou, I., 2014. Authorship analysis studies: A survey. International Journal of Computer Applications 86 (12), 22–29.

Escalante, H. J., Montes, M., Sucar, E., 2012. Multimodal indexing based on semantic cohesion for image retrieval. Information retrieval 15 (1), 1–32.

Escalante, H. J., Montes-y Gómez, M., Villaseñor-Pineda, L., Errecalde, M. L., 2015. Early text classification: a naive solution. arXiv preprint arXiv:1509.06053.

Estruch, C. P., Paredes, R., Rosso, P., 2017. Learning multimodal gender profile using neural networks. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 577–582.

Fadaei-Kermani, E., Barani, G., Ghaeini-Hessaroeyeh, M., 2017. Drought monitoring and prediction using k-nearest neighbor algorithm. Journal of AI and data mining 5 (2), 319–325.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874.

Farnadi, G., Tang, J., De Cock, M., Moens, M.-F., 2018. User profiling through deep multimodal fusion. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, pp. 171–179.

Farseev, A., Nie, L., Akbari, M., Chua, T.-S., 2015. Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, pp. 235–242.

Flekova, L., Gurevych, I., 2013. Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In: CLEF 2012 Labs and Workshop, Notebook Papers. Citeseer.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research 3 (Mar), 1289–1305.

Friedl, M. A., Brodley, C. E., 1997. Decision tree classification of land cover from remotely sensed data. Remote sensing of environment 61 (3), 399–409.

Gilad Gressel, H. P., Surendran K, T. S., Aravind A, P. P., 2014. Ensemble learning approach for author profiling. In: Proceedings of CLEF.

Girshick, R., Donahue, J., Darrell, T., Malik, J., June 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587.

Gómez-Adorno, H., Sidorov, G., 2017. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings. Vol. 10456. Springer, p. 145.

Goswami, S., Sarkar, S., Rustagi, M., 2009. Stylometric analysis of bloggers' age and gender. In: Third International AAAI Conference on Weblogs and Social Media. pp. 214–217.

Graff, M., Miranda-Jiménez, S., Tellez, E. S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., Sánchez, C. N., 2018. 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.

Grimshaw, M., 2013. The Oxford handbook of virtuality. Oxford University Press.

HaCohen-Kerner, Y., Yigal, Y., Elyashiv Shayovitz, D. M., Breckon, T., 2018. Author profiling: Gender prediction from tweets and images, 11–25.

Hall, R. C., Hall, R. C., 2007. A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. In: Mayo Clinic Proceedings. Vol. 82. Elsevier, pp. 457–471.

Han, D., Wang, w., Luo, S., Fan, W., Wang, S., 2018. The uncertainty mapping of ontologies based on three-dimensional combination weight vector space model. Information Discovery and Delivery (just-accepted), 00–00.

Hernández, D.-I., Guzmán-Cabrera, R., Reyes, A., Rocha, M.-A., 2013. Semantic-based features for author profiling identification: First insights. In: Proceedings of CLEF. pp. 123–126.

Hoffman, M., Bach, F. R., Blei, D. M., 2010. Online learning for latent dirichlet allocation. In: Advances in Neural Information Processing Systems. pp. 856–864.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al., 2003. A practical guide to support vector classification, 1–16.

Huang, J.-J., Siu, W.-C., 2017. Learning hierarchical decision trees for single-image super-resolution. IEEE Transactions on Circuits and Systems for Video Technology 27 (5), 937–950.

Huang, S.-J., Gao, W., Zhou, Z.-H., 2018. Fast multi-instance multi-label learning. IEEE transactions on pattern analysis and machine intelligence, 1868–1874.

Hum, N. J., Chamberlin, P. E., Hambright, B. L., Portwood, A. C., Schat, A. C., Bevan, J. L., 2011. A picture is worth a thousand words: A content analysis of facebook profile photographs. Computers in Human Behavior 27 (5), 1828–1833.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360.

Indurkhya, N., Damerau, F. J., 2010. Handbook of natural language processing. Vol. 2. CRC Press.

Jeon, J., Lavrenko, V., Manmatha, R., 2003. Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, pp. 119–126.

Jia, X., Song, S., He, W., Wang, Y., Rong, H., Zhou, F., Xie, L., Guo, Z., Yang, Y., Yu, L., et al., 2018. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. arXiv preprint arXiv:1807.11205.

Jin, C., De-Lin, L., Fen-Xiang, M., 2009. An improved id3 decision tree algorithm. In: Computer Science & Education, 2009. ICCSE'09. 4th International Conference on. IEEE, pp. 127–130.

Jin, X., Xu, A., Bie, R., Guo, P., 2006. Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In: International Workshop on Data Mining for Biomedical Applications. Springer, pp. 106–115.

Joachims, T., 2002. Learning to classify text using support vector machines: Methods, theory and algorithms. Vol. 186. Kluwer Academic Publishers Norwell.

Kanellis, P., 2006. Digital crime and forensic science in cyberspace. IGI Global.

Kang, J.-W., Park, H.-J., Ro, J.-S., Jung, H.-K., 2018. A strategy-selecting hybrid optimization algorithm to overcome the problems of the no free lunch theorem. IEEE Transactions on Magnetics 54 (3), 1–4.

Kharroub, T., Bas, O., 2015. Social media and protests: An examination of twitter images of the 2011 egyptian revolution. New Media & Society, 1461444815571914.

Klammer, T. P., 2007. Analyzing English Grammar, 6/e. Pearson Education India.

Kodiyan, D., Hardegger, F., Neuhaus, S., Cieliebak, M., 2017. Author profiling with bidirectional rnns using attention with grus, 1–10.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. Tech. rep., Stanford InfoLab.

Koppel, M., Akiva, N., Alshech, E., Bar, K., 2009. Automatically classifying documents by ideological and organizational affiliation. In: Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on. IEEE, pp. 176–178.

Koppel, M., Argamon, S., Shimoni, A. R., 2002. Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17 (4), 401–412.

Koppel, M., Schler, J., Zigdon, K., 2005. Determining an author's native language by mining a text for errors. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, pp. 624–628.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160, 3–24.

Kotsiantis, S. B., Zaharakis, I. D., Pintelas, P. E., 2006. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26 (3), 159–190.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Kumar, V., Reinartz, W., 2018. Customer relationship management: Concept, strategy, and tools. Springer.

Kuncheva, L. I., 2004. Classifier ensembles for changing environments. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.

Lakkaraju, S. K., Tech, D., Deng, S., 2018. A framework for profiling prospective students in higher education. In: Encyclopedia of Information Science and Technology, Fourth Edition. IGI Global, pp. 3861–3869.

Landauer, T. K., Foltz, P. W., Laham, D., 1998. An introduction to latent semantic analysis. Discourse processes 25 (2-3), 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W., 2013. Handbook of latent semantic analysis. Psychology Press.

Laufer, B., Nation, P., 1995. Vocabulary size and use: Lexical richness in l2 written production. Applied linguistics 16 (3), 307–322.

Lavelli, A., Sebastiani, F., Zanoli, R., 2004a. Distributional term representations: an experimental comparison. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, pp. 615–624.

Lavelli, A., Sebastiani, F., Zanoli, R., 2004b. Distributional term representations: An experimental comparison. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. CIKM '04. ACM, New York, NY, USA, pp. 615–624.
URL http://doi.acm.org/10.1145/1031171.1031284

Layton, R., 2016. Relative cyberattack attribution. In: Automating Open Source Intelligence. Elsevier, pp. 37–60.

Levy, O., Goldberg, Y., Dagan, I., 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3, 211–225.

Lewis, D. D., 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In: European conference on machine learning. Springer, pp. 4–15.

Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K., 2011. Fast text categorization using concise semantic analysis. Pattern Recognition Letters 32 (3), 441–448.

Lim, W.-Y., Goh, J., Thing, V. L., 2013. Content-centric age and gender profiling. Proceedings of the Notebook for PAN at CLEF, 1–8.

Liu, N., Zhang, B., Yan, J., Chen, Z., Liu, W., Bai, F., Chien, L., 2005. Text representation: From vector to tensor. In: Data Mining, Fifth IEEE International Conference on. IEEE, pp. 4–7.

Liu, T., Cho, K., Broadwell, G. A., Shaikh, S., Strzalkowski, T., Lien, J., Taylor, S. M., Feldman, L., Yamrom, B., Webb, N., et al., 2014. Automatic expansion of the mrc psycholinguistic database imageability ratings. In: LREC. pp. 2800–2805.

Liu, W., Ruths, D., 2013. What's in a name? using first names as features for gender inference in twitter. In: AAAI spring symposium: Analyzing microtext. Vol. 13. pp. 10–16.

Lopez-Monroy, A. P., Gomez, M. M.-y., Escalante, H. J., Villasenor-Pineda, L., Villatoro-Tello, E., 2013. Inaoe's participation at pan'13: Author profiling task. In: CLEF 2013 Evaluation Labs and Workshop.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., 2014. Using intra-profile information for author profiling. In: CLEF 2014 Working Notes. pp. 1116–1120.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Stamatatos, E., 2015. Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89, 134 – 147.

López-Santillán, R., González-Gurrola, L. C., Ramírez-Alonso, G., 2018. Custom document embeddings via the centroids method: Gender classification in an author profiling task. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). pp. 1121–1132.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for eeg-based brain–computer interfaces. Journal of neural engineering 4 (2), R1.

Lu, X., 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. The Modern Language Journal 96 (2), 190–208.

Maharjan, S., Shrestha, P., Solorio, T., 2014. A simple approach to author profiling in mapreduce. In: CLEF (Working Notes). pp. 1121–1128.

Mairesse, F., Walker, M., 2006. Words mark the nerds: Computational models of personality recognition through language. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society. pp. 543–548.

Markov, I., Gómez-Adorno, H., Mónica, J.-R., Grigori, S., 2018. 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.

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.-F., Davalos, S., Teredesai, A., De Cock, M., 2014. Age and gender identification in social media. In: Proceedings of CLEF 2014 Evaluation Labs. pp. 1129–1136.

Martinc, M., Škrjanec, I., Zupan, K., Pollak, S., 2017. Pan 2017: Author profiling-gender and language variety prediction, 1–10.

Masoud, M., 2018. Enhancing automatic annotation for optimal image retrieval. Ph.D. thesis, Georgia State University.
URL https://scholarworks.gsu.edu/cs_diss/140

Mechti, S., Jaoua, M., Belguith, L. H., Faiz, R., 2013. Author profiling using style-based features. In: Proceedings of CLEF. Citeseer, pp. 1107–1114.

Mechti, S., Jaoua, M., Belguith, L. H., Faiz, R., 2014. Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media. Proceedings of the PAN@ CLEF, Sheffield, England, 1137–1142.

Merler, M., Cao, L., Smith, J. R., 2015. You are what you tweet... pic! gender prediction based on semantic analysis of social media images. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on. IEEE, pp. 1–6.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119.

Murphy, K. P., 2006. Naive bayes classifiers. University of British Columbia 18.

Murthy, V. N., Maji, S., Manmatha, R., 2015. Automatic image annotation using deep learning representations. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, pp. 603–606.

Murty, M. N., Devi, V. S., 2011. Pattern recognition: An algorithmic approach. Springer Science & Business Media.

Nowson, S., Oberlander, J., 2006. The identity of bloggers: Openness and gender in personal weblogs. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 163–167.

Nowson, S., Perez, J., Brun, C., Mirkin, S., Roux, C., 2015. Xrce personal language analytics engine for multilingual author profiling. Working Notes Papers of the CLEF, 1412–1424.

Olson, D. L., Wu, D. D., 2017. Data mining models and enterprise risk management. In: Enterprise Risk Management Models. Springer, pp. 119–132.

op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M., 2016. Gronup: Groningen user profiling, 1412–1424.

Ortega-Mendoza, R. M., López-Monroy, A. P., 2018. 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.

Ortega-Mendoza, R. M., López-Monroy, A. P., Franco-Arcega, A., Montes-y Gómez, M., 2018. Emphasizing personal information for author profiling: New approaches for term selection and weighting. Knowledge-Based Systems 145, 169–181.

Pavan, A., Mogadala, A., Varma, V., 2013. Author profiling using lda and maximum entropy. Notebook for PAN at CLEF, 1–4.

Pellegrin, L., Escalante, H. J., Montes-y Gómez, M., González, F. A., 2016. Local and global approaches for unsupervised image annotation. Multimedia Tools and Applications 76 (15), 16389–16414.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 27 (8), 1226–1238.

Pennebaker, J. W., Francis, M. E., Booth, R. J., 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71, 2001.

Pham, D. D., Tran, G. B., Pham, S. B., 2009. Author profiling for vietnamese blogs. In: Asian Language Processing, 2009. IALP'09. International Conference on. IEEE, pp. 190–194.

Pinto, D., Rosso, P., 2007. On the relative hardness of clustering corpora. In: Text, Speech and Dialogue. Springer, pp. 155–161.

Poulston, A., Waseem, Z., Stevenson, M., 2017. Using tf-idf n-gram and word embedding cluster ensembles for author profiling, 1–6.

Powers, D. M., 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation 2 (1), 37–63.

Quinlan, J. R., 1986. Induction of decision trees. Machine learning 1 (1), 81–106.

Quinlan, J. R., Cameron-Jones, R. M., 1993. Foil: A midterm report. In: European conference on machine learning. Springer, pp. 1–20.

Rangel, F., Rosso, P., 2016. On the impact of emotions on author profiling. Information processing & management 52 (1), 73–92.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W., 2014. Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes). pp. 1–30.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G., 2013. Overview of the author profiling task at pan 2013. Notebook Papers of CLEF, 23–26.

Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B., 2018. Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. Working Notes Papers of the CLEF, 1–38.

Rangel, F., Rosso, P., Potthast, M., Stein, B., 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF, 1–26.

Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W., 2015. Overview of the 3rd author profiling task at pan 2015. In: CLEF. sn, p. 2015.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B., 2016. 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.

Rao, D., Yarowsky, D., Shreevats, A., Gupta, M., 2010. Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. ACM, pp. 37–44.

Raschka, S., Mirjalili, V., 2017. Python Machine Learning. Packt Publishing Ltd.

Reddy, T. R., Vardhan, B. V., Reddy, P. V., 2016. A survey on authorship profiling techniques. International Journal of Applied Engineering Research 11 (5), 3092–3102.

Rish, I., 2001. An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. IBM, pp. 41–46.

Rogati, M., Yang, Y., 2002. High-performing feature selection for text classification. In: Proceedings of the eleventh international conference on Information and knowledge management. ACM, pp. 659–661.

Rokach, L., 2009. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. Computational Statistics & Data Analysis 53 (12), 4046–4072.

Romania, B., 2015. Automatic profiling of twitter users based on their tweets, 10–18.

Rosso, P., Rangel, F., 2017. Author profiling in social media: The impact of emotions on discourse analysis. In: International Conference on Statistical Language and Speech Processing. Springer, pp. 3–18.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115 (3), 211–252.

Sadr, A. V., Farhang, M., Movahed, S., Bassett, B., Kunz, M., 2018. Cosmic string detection with tree-based machine learning. arXiv preprint arXiv:1801.04140.

Salzberg, S. L., 1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. Machine Learning 16 (3), 235–240.

Sanchez-Perez, M. A., Markov, I., Gómez-Adorno, H., Sidorov, G., 2017. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pp. 145–151.

Schilling, N., Marsters, A., 2015. Unmasking identity: speaker profiling for forensic linguistic purposes. Annual Review of Applied Linguistics 35, 195–214.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. W., 2006. Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. Vol. 6. pp. 199–205.

Schmidhuber, J., 1997. Discovering neural nets with low kolmogorov complexity and high generalization capability. Neural Networks 10 (5), 857–873.

Schmitt, M., Schuller, B., 2017. Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit. The Journal of Machine Learning Research 18 (1), 3370–3374.

Schölkopf, B., 2001. The kernel trick for distances. In: Advances in neural information processing systems. pp. 301–307.

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E., Ungar, L. H., 2013a. Toward personality insights from language exploration in social media. In: AAAI Spring Symposium: Analyzing Microtext. pp. 72–79.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E., et al., 2013b. Characterizing geographic variation in well-being using tweets. In: ICWSM. pp. 583–591.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR) 34 (1), 1–47.

Segalin, C., Cheng, D. S., Cristani, M., 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. Computer Vision and Image Understanding 156, 34–50.

Seroussi, Y., Zukerman, I., Bohnert, F., 2011. Authorship attribution with latent dirichlet allocation. In: Proceedings of the fifteenth conference on computational natural language learning. Association for Computational Linguistics, pp. 181–189.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D., 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas 18 (3), 491–504.

Sierra, S., González, F. A., 2018. Combining textual and visual representations for multimodal author profiling. Working Notes Papers of the CLEF 2125, 219–228.

Siirtola, H., Saily, T., Nevalainen, T., 2017. Interactive principal component analysis. In: 2017 21st International Conference Information Visualisation (IV). IEEE, pp. 416–421.

Sikora, R., 2015. A modified stacking ensemble machine learning algorithm using genetic algorithms. In: Handbook of Research on Organizational Transformations through Big Data Analytics. IGi Global, pp. 43–53.

Simaki, V., Simakis, P., Paradis, C., Kerren, A., 2017. Identifying the authors' national variety of english in social media text. Association for Computational Linguistics, pp. 1–8.

Skalmowski, W., 2016. Review of harris, zellig (1968) mathematical structures of language. ITL-International Journal of Applied Linguistics 4 (1), 56–61.

Snoek, C. G., Worring, M., Smeulders, A. W., 2005. Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia. ACM, pp. 399–402.

Soundar, K. R., Ponesakki, P., 2016. Cyberbullying detection based on text representation. International Journal of Engineering Science 6 (10), 2776–2785.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60 (3), 538–556.

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B., 2015. Overview of the pan/clef 2015 evaluation lab. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer, pp. 518–538.

Stolcke, A., 2002. Srilm-an extensible language modeling toolkit. In: Seventh international conference on spoken language processing. pp. 901–904.

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T., 2018. Text and image synergy with feature cross technique for gender identification. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Vol. 2125. pp. 10–22.

Tam, J., Martell, C. H., 2009. Age detection in chat. In: Semantic Computing, 2009. ICSC'09. IEEE International Conference on. IEEE, pp. 33–39.

Tan, J., Town, T., Mori, T., Obregon, D., Wu, Y., DelleDonne, A., Rojiani, A., Crawford, F., Flavell, R. A., Mullan, M., 2002. Cd40 is expressed and functional on neuronal cells. The EMBO journal 21 (4), 643–652.

Tausczik, Y. R., Pennebaker, J. W., 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29 (1), 24–54.

Tellez, F. P., Pinto, D., Cardiff, J., Rosso, P., 2009. Defining and evaluating blog characteristics. In: Artificial Intelligence, 2009. MICAI 2009. Eighth Mexican International Conference on. IEEE, pp. 97–102.

Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M., 2017. Overview of the author identification task at pan-2017: style breach detection and author clustering. In: Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al. pp. 1–22.

Uricchio, T., Ballan, L., Seidenari, L., Del Bimbo, A., 2017. Automatic image annotation via label transfer in the semantic space. Pattern Recognition 71, 144–157.

Villena Román, J., González Cristóbal, J. C., 2014. Daedalus at pan 2014: Guessing tweet author's gender and age, 1157–1163.

Vinokur, A. I., 2015. Information technologies in culture and education: Image processing issues. Modern Applied Science 9 (5), 314.

Wallach, H. M., 2006. Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 977–984.

Wang, C., Jing, F., Zhang, L., Zhang, H.-J., 2006a. Image annotation refinement using random walk with restarts. In: Proceedings of the 14th ACM international conference on Multimedia. ACM, pp. 647–650.

Wang, C., Jing, F., Zhang, L., Zhang, H.-J., 2007. Content-based image annotation refinement. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, pp. 1–8.

Wang, X.-J., Zhang, L., Jing, F., Ma, W.-Y., 2006b. Annosearch: Image auto-annotation by search. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, pp. 1483–1490.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S., 2014. Cnn: single-label to multi-label. arXiv preprint arXiv:1406.5726.

Wendlandt, L., Mihalcea, R., Boyd, R. L., Pennebaker, J. W., 2017. Multimodal analysis and prediction of latent user dimensions. In: International Conference on Social Informatics. Springer, pp. 323–340.

Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., Wives, L. K., 2014a. Examining multiple features for author profiling. Journal of Information and Data Management 5 (3), 266.

Weren, E. R., Moreira, V. P., de Oliveira, J. P. M., 2014b. Exploring information retrieval features for author profiling. In: CLEF (Working Notes). pp. 1164–1171.

Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A., 2004. Latent semantic analysis. In: Proceedings of the 16th international joint conference on Artificial intelligence. Citeseer, pp. 1–14.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wolpert, D. H., Macready, W. G., 1997. No free lunch theorems for optimization. IEEE transactions on evolutionary computation 1 (1), 67–82.

Wu, J., Yu, Y., Huang, C., Yu, K., 2015. Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3460–3469.

Wu, Y.-C. J., Chang, W.-H., Yuan, C.-H., 2014. Do facebook profile pictures reflect user's personality? Computers in Human Behavior 51, 880–889.

Yang, Y., Pedersen, J. O., 1997. A comparative study on feature selection in text categorization. In: International Conference on Machine Learning. Vol. 97. pp. 412–420.

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H. H., 2008. Toward multi-modal music emotion classification. In: Pacific-Rim Conference on Multimedia. Springer, pp. 70–79.

You, Q., Bhatia, S., Sun, T., Luo, J., 2014. The eyes of the beholder: Gender prediction using images posted in online social networks. In: Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, pp. 1026–1030.

Zhang, C., Zhang, P., 2010. Predicting gender from blog posts. Tech. rep., Technical Report. University of Massachusetts Amherst, USA.

Zhang, D., Islam, M. M., Lu, G., 2012. A review on automatic image annotation techniques. Pattern Recognition 45 (1), 346–362.

Zhang, M.-L., Zhou, Z.-H., 2014. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering 26 (8), 1819–1837.

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2018. Efficient knn classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems 29 (5), 1774–1785.

Zhang, Y., Jin, R., Zhou, Z.-H., 2010. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1 (1-4), 43–52.

Zheng, R., Li, J., Chen, H., Huang, Z., 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57 (3), 378–393.

Zheng, R., Qin, Y., Huang, Z., Chen, H., 2003. Authorship analysis in cybercrime investigation. In: Intelligence and Security Informatics. Springer, pp. 59–73.

Zhuang, Y.-t., Wu, F., Chen, C., Pan, Y.-h., 2017. Challenges and opportunities: from big data to knowledge in ai 2.0. Frontiers of Information Technology & Electronic Engineering 18 (1), 3–14.

Appendices

Appendix A

INAOE's Participation at PAN'15: Author Profiling Task

For this competition, we focus on the representation of the documents, to improve the representation of tweets for the Author Profiling task (Rangel et al., 2015; Álvarez-Carmona et al., 2015). The main goal of our approach is to compute high-quality *discriminative* and *descriptive* features built on the top of the state-of-the-art typical textual features (e.g., content words, function words, punctuation marks, etc.). For this, we proposed to combine two state-of-the-art dimensionality reduction techniques that best contribute to automatically stress the contribution of the *discriminative* and *descriptive* textual features. According to the literature, the most frequent textual features (e.g., function words, stopwords, punctuation marks) provide important clues about the discrimination of the authors. For this, we need a representation highly based on term frequencies, which stresses the contribution of such discriminative attributes and produces highly discriminative document representations. To capture this information contained among textual features we use Second Order Attributes (SOA) computed as in Lopez-Monroy et al. (2013). On the other hand, relevant thematic information usually are in *descriptive* terms, terms that are frequent only in some specific documents or classes. In this way, to represent documents, we bring ideas from the information retrieval field exploiting the Latent Semantic Analysis (LSA) Wiemer-Hastings et al. (2004). LSA represents terms and documents into a new semantic space. This is done by performing a singular value decomposition using a Term Frequency-Inverse Document Frequency (TFIDF) matrix. The descriptive terms and documents representation are stressed under the LSA formulation throwing out the noise but emphasizing strong patterns and trends. To the best of our knowledge, the idea of representing documents using the combination *discriminative* and the *descriptive* high-level features through dimensionality reduction techniques have never been explored before in AP task. Thus, it is promising to bring together two of the best document representations to better improve the AP; that is precisely the propose

of this work for this competition.

A.1 Exploiting the Jointly Use of Discriminative-Descriptive Features

The idea is to use the representations built under the whole feature space to highlight the *discriminative* automatically and *descriptive* properties in documents. The intuitive idea is to take advantage of both approaches in a representation using early fusion. Let $\mathbf{x}_j$ be the $j-th$ training instance-profile under LSA representation with k dimensions and $\mathbf{y}_j$ be the same instance-profile under the SOA representation with m dimensions, the final representation is shown in Expression A.1.

$$\mathbf{z_j} = \langle x_{j1}, \ldots, x_{jk}, y_{j1}, \ldots, y_{jm} \rangle \tag{A.1}$$

The collection of training documents are finally represented as:

$$\mathbf{Z} = \bigcup_{d_j \in D} \langle \mathbf{z_j}, c_j \rangle \tag{A.2}$$

Where c_j is the class of the $j-th$ training instance-profile.

A.2 Data Collection

We have approached the PAN 2015 AP task as a classification problem. PAN 2015 corpora are composed of 4 datasets in different languages (Spanish, English, Italian and Dutch). Each dataset has labels of gender (male, female), age [1] (18-24, 25-34, 35-49, 50-xx) and five personality traits values (extroverted, stable, agreeable, conscientious, open) between -0.5 and 0.5. Table A.1 we show the number of Author-Profiles per language.

For personality identification Table A.2 shows the relevant information (in terms of classes). For each language, it shows the range and the number of the classes for each trait[2]. For personality, we consider each trait value in the training corpus as a class. For example, if only two values (e.g., 0.2 and 0.3) are observed in the training corpus, then we built a two-class classifier (e.g., 0.2 and 0.3) [3].

[1] Age data for Italian and Dutch languages are not available.

[2] The ranges with an asterisk indicate that value between the range is missing. For example, in Spanish (extroverted and conscientious) the -0.1 is missing.

[3] For each personality trait in each language the number of the classes are variables between them, see Table A.2

Table A.1: Description of the dataset

Language	Author-Profiles
English	152
Spanish	100
Italian	38
Dutch	34

Table A.2: The personality traits information by language

	English		Spanish		Italian		Dutch	
Trait	Range	Classes	Range	Classes	Range	Classes	Range	Classes
Extroverted	[-0.3,0.5]	9	[-0.3,0.5]*	8	[0.0,0.5]*	5	[0.0,0.5]	6
Stable	[-0.3,0.5]	9	[-0.3,0.5]	9	[-0.1,0.5]	7	[-0.2,0.5]	8
Agreeable	[-0.3,0.5]	9	[-0.2,0.5]	8	[-0.1,0.5]*	6	[-0.1,0.4]	6
Conscientious	[-0.2,0.5]	8	[-0.2,0.5]*	7	[0.0,0.4]	5	[-0.1,0.4]	6
Open	[-0.1,0.5]	7	[-0.1,0.5]	7	[0.0,054]	6	[0.1,0.5]	5

A.3 Experimental Evaluation

A.3.1 Experimental Settings

We use for each experiment the following configuration: i) for terms we use words, contractions, words with hyphens, punctuation marks and a set of common emoticons, ii) we consider the terms with at least 5 occurrences in the corpus, iii) the number of concepts for LSA is set to $k = 100$. We perform a stratified ten cross-fold validation (CFV) using the training PAN15 corpus and a LibLINEAR classifier Fan et al. (2008). In order to determine the full profile of a document (gender, age, and the five personality traits) we built one classifier to predict each target profile for each language.

A.3.2 Experimental Results

This first experiment aims to analyze the performance of LSA, SOA and the BOW approach in the AP tasks. We experiment with LSA and SOA separately and finally with the two approaches together. We are interested in observing the contribution of discriminative-stylistic (captured by SOA) and descriptive-thematic (captured by LSA) information in the AP task. For gender prediction, in Table A.3 we can see that considering the individual representations, LSA obtains the best results, which outperforms the BOW approach in every language. When LSA and SOA are together, the result only improves in English, which is an important remark since the English language is the bigger-robust collection (see Table A.1). The following conclusions can

Table A.3: Detailed classification accuracy to gender

Language	BOW	SOA	LSA	LSA+SOA
English	74.00	70.86	74.34	**78.28**
Spanish	84.00	74.00	**91.00**	**91.00**
Italian	76.31	73.68	**86.84**	**86.84**
Dutch	82.35	91.07	**91.17**	**91.17**

Table A.4: Detailed classification accuracy to age

Language	BOW	SOA	LSA	LSA+SOA
English	74.83	68.21	78.94	**79.60**
Spanish	80.00	74.00	81.00	**82.00**

be outlined from Table A.3:

- The descriptive information captured by LSA is the most relevant information for gender prediction in PAN 2015 AP dataset. This is because LSA obtained the best average individual performance.
- The pure discriminative information captured by SOA only outperforms BOW in Dutch documents. However, the combination of LSA and SOA obtained an improvement of around 4% in accuracy for English gender detection. We think SOA could improve the results if more documents are available [4].

For age prediction, Table A.4 shows the experimental results. Recall that the age data is available only for English and Spanish languages. As in the last experiment, LSA obtains the best individual performance, but in this experiment, the combination of LSA and SOA obtains an improvement in both collections. It is worth noting that despite the small datasets, for age prediction SOA could contribute to improve the classification[5].

Finally for personality prediction Table A.5 shows the performance of BOW and LSA plus SOA performance by language in the personality detection task. For this experiment, although the results seem promising, they should be taken with caution. This is due to the lack of data, and the number of classes that we consider (one class for each observed value) one correct/wrong predicted instance is enough to change the results considerably. For this specific experiment in personality, we built

[4]SOA has proven outstanding results in recent years in the PAN AP tracks Rangel et al. (2014, 2013).
[5]The best results for SOA in previous PAN AP editions have been for age prediction

Table A.5: Detailed classification accuracy for personality

	English		Spanish		Italian		Dutch	
Trait	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA
Extroverted	64	**87**	62	**87**	65	**94**	64	**91**
Stable	56	**85**	69	**91**	52	**94**	61	**94**
Agreeable	60	**80**	62	**84**	71	**92**	61	**88**
Conscientious	61	**78**	62	**86**	57	**94**	67	**91**
Open	65	**86**	62	**74**	55	**84**	64	**97**

a representation on the entire dataset. Then we evaluate using a 10CFV. In general, the results suggest that the combination of LSA plus SOA gets similar or better results than the typical BOW approach. Given evidence of the usefulness of the *discriminative* features and the *descriptive* features.

A.4 Official Results

For participating in The PAN 2015 workshop, it was necessary to upload the training model to the organizers' platform[6]. In this point, the organizers tested the models of all the competitors with a secret test corpus for publishing the results.

For each language, the quadratic error between the output of each system was computed (f_{sal}) for personality and the ground truth result (f_{gt}) as follows:

$$RMSE = \sqrt{\frac{\sum_i^n (f_{gti} - f_{sali})^2}{n}}$$

The join accuracy for gender and age and personality is obtained for each language as follows:

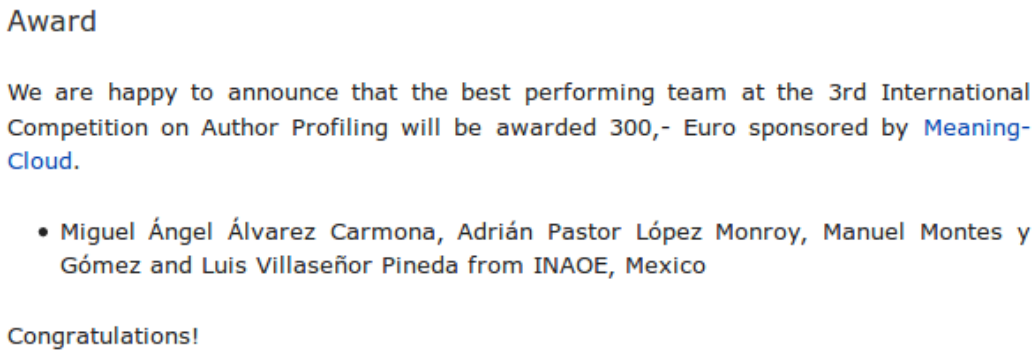
$$rank = \frac{(1 - RMSE) + jointAccuracy}{2}$$

The final rank result is obtained from the average of the four languages results. In the Figure A.1 it shows the official results published by the organizers. In this Figure it is possible to observe that the result of the INAOE team (alvarezcarmona15) obtains the best results for the English, Spanish and Dutch languages, making that the INAOE team obtain the best average in the overall result.

[6] www.tira.io/task/author-profiling/

Ranking	Team	Global	English	Spanish	Italian	Dutch
1	alvarezcarmona15	**0.8404**	**0.7906**	**0.8215**	0.8089	**0.9406**
2	gonzalesgallardo15	0.8346	0.7740	0.7745	**0.8658**	0.9242
3	grivas15	0.8078	0.7487	0.7471	0.8295	0.9058
4	kocher15	0.7875	0.7037	0.7735	0.8260	0.8469
5	sulea15	0.7755	0.7378	0.7496	0.7509	0.8637
6	miculicich15	0.7584	0.7115	0.7302	0.7442	0.8475
7	nowson15	0.7338	0.6039	0.6644	0.8270	0.8399
8	weren15	0.7223	0.6856	0.7449	0.7051	0.7536
9	poulston15	0.7130	0.6743	0.6918	0.8061	0.6796
10	maharjan15	0.7061	0.6623	0.6547	0.7411	0.7662
11	mccollister15	0.6960	0.6746	0.5727	0.7015	0.8353
12	arroju15	0.6875	0.6996	0.6535	0.7126	0.6843
13	gimenez15	0.6857	0.5917	0.6129	0.7590	0.7790
14	bartoli15	0.6809	0.6557	0.5867	0.6797	0.8016
15	ameer15	0.6685	0.6379	0.6044	0.7055	0.7260
16	cheema15	0.6495	0.6130	0.6353	0.6774	0.6723
17	teisseyre15	0.6401	0.7489	0.5049	0.6024	0.7042
18	mezaruiz15	0.6204	0.5217	0.6215	0.6682	0.6703
19	bayot15	0.6178	0.5253	0.5932	0.6644	0.6881
	ashraf15	-	0.5854	-	-	-
	kiprov15	-	0.7211	0.7889	-	-
	markov15	-	0.5890	0.5874	-	0.6798

Figure A.1: Table with the final results of the PAN 2015 for author profiling task



Award

We are happy to announce that the best performing team at the 3rd International Competition on Author Profiling will be awarded 300,- Euro sponsored by MeaningCloud.

- Miguel Ángel Álvarez Carmona, Adrián Pastor López Monroy, Manuel Montes y Gómez and Luis Villaseñor Pineda from INAOE, Mexico

Congratulations!

Figure A.2: Image extracted from the official PAN site

Appendix B

Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets

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 massive community of active users. The analysis of shared information has become very relevant for several applications in security, marketing, 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 such as gender, age, personality and native language, by examining the content of their posts (Álvarez-Carmona et al., 2016; Argamon et al., 2003). On the other hand, *detecting aggressive content* targeted to 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 peninsular Spanish. Also, 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 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.

Table B.1: 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

B.1 Evaluation Framework

B.1.1 A Mexican Corpus for Author Profiling

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 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.*

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 into two parts, one for training and the other for the test. Table B.1 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 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 B.2 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 both tables, B.1 and B.2, the class imbalance was calculated as proposed in (Tellez

Table B.2: 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

Table B.3: 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

et al., 2009). 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, 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 B.3 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.

B.1.2 Performance Measures

For the 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 B.1.

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

The F_{macro} measures were computed using Formula B.2, where C indicates the set of classes for a given trait[1], and $F_1(c)$ is the F_1-measure of each of the categories from the trait.

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

B.2 Overview of the Submitted Approaches

For this study, four teams have submitted their solutions. From 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:

- *CIC-GIL Approach to Author Profiling in Spanish Tweets: Location and Occupation* (Markov et al., 2018)
  - **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

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

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.

- *INGEOTEC at MEX-A3T: Author profiling and aggressiveness analysis in Twitter using μTC and EvoMSA* (Graff et al., 2018)
  - **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 the 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 an SVM with the linear kernel as the 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* (Aragón and López-Monroy, 2018)

  - **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 three 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 an evident aggressiveness.

- *The Winning Approach for Author Profiling of Mexican Users in Twitter at MEX.A3T@IBEREVAL-2018* (Ortega-Mendoza and López-Monroy, 2018)

  - **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 of 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. Utilizing 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 useful 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.

B.3 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 of 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 on the EvALL platform (Amigó et al., 2017). 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 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. For the case Trigrams, all 3-grams were used. As in BOW, stop words and special characters were removed. SVM with linear kernel and C = 1 was applied for classification of both tasks.

Table B.4: 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 B.5: 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

B.3.1 Results

First we analyze the author profiling performance. Table B.4 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 B.5 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.

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

Table B.6: 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

every other submitted to 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 to the top-ranked team. All teams outperformed the baselines (except run two from the top-ranked team), showing the feasibility of the proposed task.

Regarding location traits, it can be seen that the class with the higher performance was *northeast*, where three teams obtained performance higher than 0.9. On the other hand,

Table B.6 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.

From Table B.6 it can be seen that the problem is harder than the location task. In fact, the performance across classes is quite diverse. The class with the 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 dataset 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 B.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

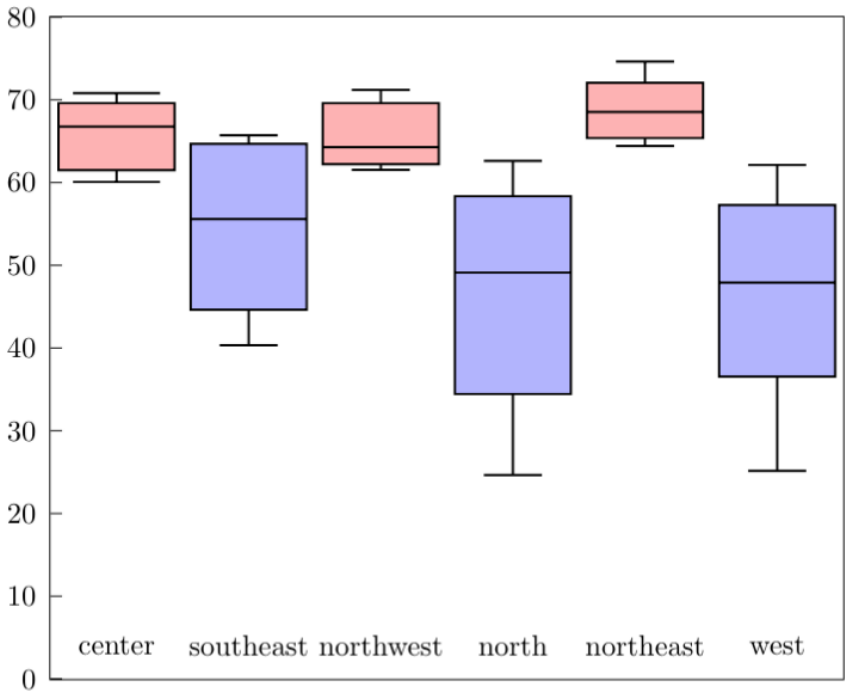


Figure B.1: F_{macro} distribution of results for the location class.

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.

On the other hand, Figure B.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 B.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 *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%.

In order to analyze the complementariness of predictions by participants, we built

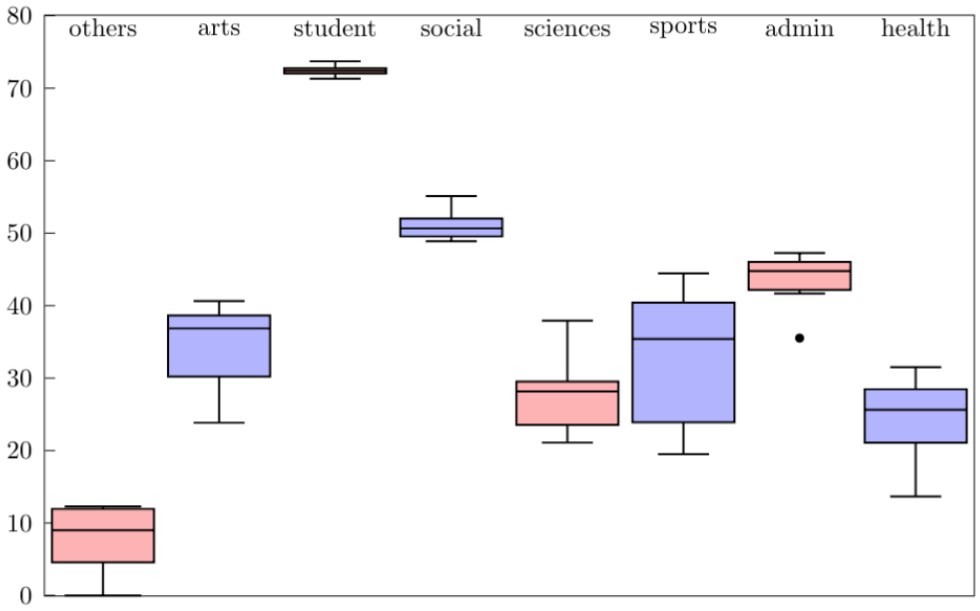


Figure B.2: F_{macro} distributions of teams performance for the occupation results

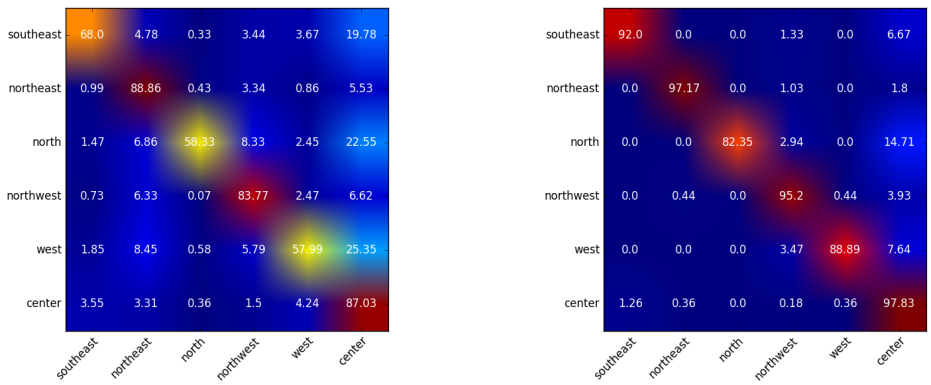


Figure B.3: Heat map of the confusion matrix average for the location results

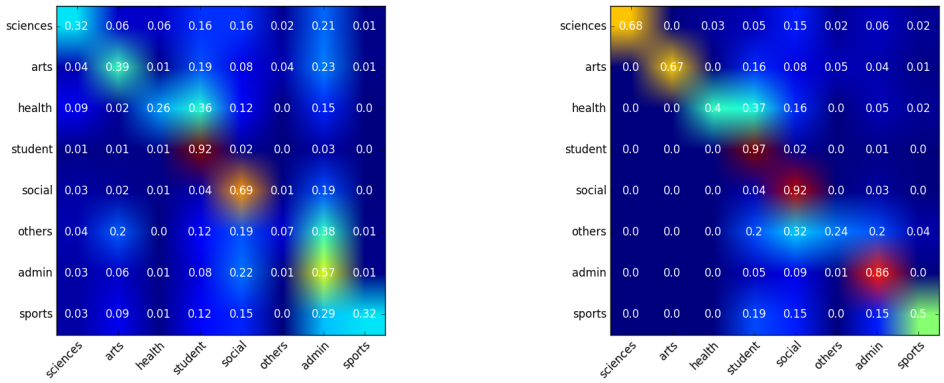


Figure B.4: Heat map of the confusion matrix average for the occupation results

a theoretically perfect ensemble from the predictions of all participants. 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 B.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 confirms there is a considerable complimentary among predictions of participant teams, and that it is possible to push performance further to solve the task.

On the other hand, Figure B.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 considerable improvement in its main diagonal where the three 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.

Finally, with the goal of getting further insights into the complementariness and redundancy of evaluated systems we show in Table B.7 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.

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 more instances 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

Table B.7: 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 %)

the location problem).

B.4 Conclusions

This chapter 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 with the two tasks were introduced, together with an evaluation protocol and baselines. The competition lasted more than two months and attracted eight 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 (Ortega-Mendoza and López-Monroy, 2018).

In general terms, the competition was a success: performance was considerably improved concerning the baseline, solutions proposed by participants were diverse regarding 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 complimentarily. 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.

Appendix C

Word2vec Model Trained from MEX-A3T

Today, there are several pre-trained word2vec models. Many of them in several languages. Nevertheless, although there are some models trained for the Spanish language, these models are trained with the peninsular Spanish. This causes that some aspects, slang, culture, and expressions of Spanish-speaking regions outside of Spain are not captured by these models. Moreover, regardless of whether Mexico is one of the largest Spanish-speaking countries, it still does not have a specialized word2vec model for Mexican Spanish. Taking advantage of the information collected for the MEX-A3t corpus, we trained a word2vec model in order to capture aspects of Mexican culture.

We use the skip-gram model architecture, and each vector has 200 dimensions. This model was made with the Gensim library[1] for Python 2.7 [2]. To test the efficiency of the model, we did some tests with the vectors. First, we wanted to check if the semantic information of the vectors make sense. Table C.1 presents some examples of relation tested for the model. The first row represents the most classic example in word2vec (King - man + woman = queen) but this time for the Spanish. We also note that other examples of this type are met. In the same way, we did some tests that only made sense for the Mexican language, for instance, with the relationship Puebla - Cholula + Celaya where result is effectively Guanajuato.

Later, we test with words that represent places. The idea was to observe if the most related word coincided with its geographical proximity. In Table C.2 we present some places, for example, UNAM is an university in Mexico City, and we can see that the most related words are other universities around the country, nevertheless, BUAP is a University of Puebla, and all related universities in the model are from the same state (upapep, udlap and upp). Also, for places as Puebla, Muchoacán or even Houston the results meet their geographical proximity.

[1] https://radimrehurek.com/gensim/

[2] https://www.python.org/download/releases/2.7/

Table C.1: Some interest relations extracted from the Mexican word2vec model

Relationship	1st Result	2nd Result	3th Result
Rey - hombre + mujer (king - man + woman)	reina (queen)	reyna (queen)	princesa (princess)
Investigador - hombre + mujer (Researcher man - man + woman)	investigadora (researcher woman)	colaboradora (collaborator)	egresada (graduated)
Investigadora - mujer + hombre (Researcher woman - woman + man)	investigador (researcher)	agente (agent)	promotor (promoter)
Adolescente - secundaria + universidad (Teen - secundary school + university)	adulto (adult)	estudiante (student)	empleado (employee)
Puebla - Cholula + Celaya	Guanajuato	Irapuato	Querétaro
Mole - Puebla + Veracruz	Pescado	Pozole	Chicharrón
Poblano - Puebla + Oaxaca	Oaxaqueño	Michoacano	Zacatecano
Pascua - abril + diciembre (Easter - april + december)	navidad (christmas)	noche buena (christmas eve)	halloween (halloween)

Table C.2: Relations among places in the word2vec model.

Place	1st Result	2nd Result	3th Result
UNAM	BUAP	UAEM	UJAT
BUAP	UPAEP	UDLAP	UPP
Puebla	Tlaxcala	Cholula	Tepeaca
Michoacán	Nayarit	Pátzcuaro	Uruapan
Monterrey	Tamaulipas	Guadalajara	Tijuana
Sinaloa	Culiacán	Mochis	Sonora
Cdmx	Iztapalapa	Xochimilco	Puebla
Francia	Alemania	Italia	España
Houston	Austin	Mcallen	Texas

Table C.3: Words related with Mexican context concepts in the word2vec model.

Concept	1st Result	2nd Result	3th Result
PRI	PAN	PVEM	PRD
Tlatlaya	Nochixtlán	Atenco	Ayotzinapa
Normalistas (Normalists)	militares (military)	desaparecidos (missing)	manifestantes (protesters)
neta (Slang for truth)	en serio (seriously)	verdad (truth)	vdd (abbreviation for truth)
Wey (Slang for guy)	wei (Slang for guy)	morro (Slang for guy)	vato (Slang for guy)
Escuincle (Slang for kid)	huerco (Slang for kid)	mocoso (Slang for kid)	niño (kid)
Gabriela	Leticia	Luisa	Claudia
Gaby	Fer	Sofi	Dany

Finally, we can see examples related to the Mexican context. In Table C.3 we show some examples of words related to the situations of Mexican culture. For instance, in the first row, we present the Word PRI, which is a Mexican political party, and we can observe that the most related word are PAN, PVEM, and PRD, which are also Mexican political parties. In the second row, we show the result of the Tlatlaya word. Tlataya is a small town and municipality located in the southeast of the State of Mexico. The importance of Tlatlaya is that there was a massacre at the hands of the Mexican government. Therefore, the most related words in the model are others towns where similar events occurred (as Ayotzinapa). The third row shows how the disappearance of Ayotzinapa normalists also is captured by the model. The rest rows show some slang own of the Mexican vocabulary.