Expert Systems With Applications 140 (2020) 112909

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Early author profiling on Twitter using profile features with multi-resolution

A. Pastor López-Monroy[a,*], Fabio A. González[b], Thamar Solorio[c]

[a] Department of Computer Science, Mathematics Research Center (CIMAT), Jalisco S/N, Col. Valenciana. Guanajuato, GTO 36023, México
[b] MindLab, Computing Systems and Industrial Engineering Department, Universidad Nacional de Colombia, Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia
[c] Department of Computer Science, University of Houston, 4800 Calhoun Rd, Houston, TX 77004, USA

ARTICLE INFO

*Article history:*
Received 29 March 2019
Revised 15 August 2019
Accepted 31 August 2019
Available online 2 September 2019

*Keywords:*
Early text classification
Author profiling
Social media analysis
Text mining

ABSTRACT

The Author Profiling (AP) task aims to predict demographic characteristics about the authors from documents (e.g., age, gender, native language). The research so far has focused only on forensic scenarios by performing post-analysis using all the available text evidence. This paper introduces the task of Early Author Profiling (EAP) in Twitter. The goal is to effectively recognize profiles using as few tweets as possible from the user history. The task is highly relevant to support social media analysis and different problems related to security and marketing, where prevention and anticipation is crucial. This work proposes a novel strategy that combines a state of the art representation for early text classification and specialized word-vectors for author profiling tasks. In this strategy we build prototypical features called Profile based Meta-Words, which allow us to model AP information at different levels of granularity. Our evaluation shows that the proposed methodology is well suited for profiling little text evidence (e.g., a handful of tweets) in early stages, but as more tweets become available other granularities better encode larger amounts of text in late stages. We evaluated the proposed ideas on gender and language variety identification for English and Spanish, and showed that the proposal outperforms state of the art methodologies.



## 1. Introduction

The Author Profiling (AP) task is a text classification problem that aims to predict demographic and socio-linguistic aspects from author's documents, for example: age, gender, native language, and personality traits (Schler, Koppel, Argamon, & Pennebaker, 2006). The AP dimensions are highly relevant for social media analysis and a wide range of problems related to security, marketing, business intelligence, etc. Nowadays most of the AP applications have been studied in the forensic scenario, where the methods perform a post-analysis by exploiting all the available information generated by the users. Although such analysis is important, many scenarios can be complemented by profiling the information as soon as it is generated. In this context, virtually any classification system devoted to operate on groups of online users could take advantage of early profiling analyses. For example in Sexual Predator detection, systems could pay more attention to conversations where at least one child (age profiling) is involved, and ignore adult conversations. In Cyberbullying detection, it would be worth to focus the analysis in highly vulnerable groups of people, for example; teenagers with specific personality traits. In other words, in the same way that companies exploit the author profiling systems to target users (e.g., using adds), security entities could enhance the effectiveness of detection systems focusing on subjects/profiles of interests. In other words, Early Author Profiling (EAP) systems can help to drive the analysis by exploiting the profiling information that highlights relevant groups of authors.

The Early Author Profiling (EAP) task can be seen as a particular formulation of Early Text Classification (ETC) (Dulac-Arnold, Denoyer, & Gallinari, 2011). ETC has the goal of identifying the target categories by using as few text as possible and with as much anticipation as possible (see Fig. 1). In Fig. 1 we show the current standard framework for evaluating this task. So far this emerging field in Natural Language Processing (NLP) has scant work, but recently the problem has been taking relevance in specialized forums and works, for example: early depression detection (Losada, Crestani, & Parapar, 2017), early anorexia detection (Losada, Crestani, & Parapar, 2018), early sexual predator identification (Escalante et al., 2015), and in general early text classification (Dulac-Arnold et al., 2011), etc. Notwithstanding the fact that those scenarios are im-

* Corresponding author.
*E-mail addresses:* pastor.lopez@cimat.mx (A.P. López-Monroy), fagonzalezo@unal.edu.co (F.A. González), solorio@cs.uh.edu (T. Solorio).

https://doi.org/10.1016/j.eswa.2019.112909
0957-4174/

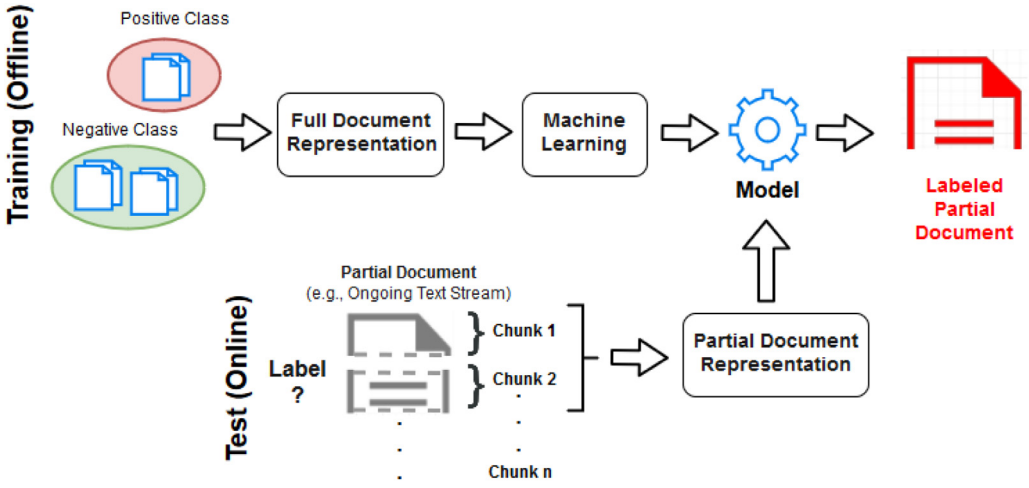



**Fig. 1.** Standard Early Text Classification Framework. Usually the performance of proposed model is analyzed with different amounts of information. One of the goals is to improve the classification performance in early chunks.

portant, it is somewhat surprising that there are no previous works addressing the EAP task despite its wide applicability. The EAP demands very specific research, similar to particular cases of text classification such as authorship attribution, author profiling and sentiment analysis. For example, in author profiling thematic features have proven to be effective but highly complemented by stylistic features such as the personal pronouns and its contextual words (Ortega-Mendoza, López-Monroy, Franco-Arcega, & y Gómez, 2018). In contrast, other tasks such as in Depression Detection there are different complementary features such as the time and day of posting (Farias-Anzaldua, Montes-Y-Gómez, Lopez-Monroy, & Gonzalez-Gurrola, 2017).

In this paper, we present for the first time the Early Author Profiling (EAP) task. For this, we focus in early detection of gender and language variety. In the proposed EAP framework, one aims to anticipate the user profile using as few text as possible. This scenario includes multiple types of problems and data, which could be useful to monitor text streams and perform a timely detection. In this work, we propose to extend the concept of meta-word (López Monroy, González, Montes, Escalante, & Solorio, 2018) by adapting specialized word-vectors (López-Monroy, y Gómez, Escalante, Villaseñor-Pineda, & Stamatatos, 2015) and weighting schemes for capturing discriminative information (Lavelli, Sebastiani, & Zanoli, 2004). The meta-word is a prototypical word-vector that represents a set of word-vectors that are highly discriminative for specific profiles. We find these meta-words by clustering word-vectors, and the resultant centroid of each cluster comprises a profile meta-word. To represent documents we build histograms that account for the presence of these profile meta-words.[1] For this purpose, one occurrence of a word in a document is codified as the occurrence of its most similar meta-word according to the Euclidean distance. In this methodology it is possible to obtain different number and sizes of meta-words by varying the number of clusters. The key idea is to build a set of histograms of increasing sizes of meta-words that encodes different amounts of textual evidence. The contributions of this paper are as follows:

- The introduction of the EAP task and the first reference evaluation. This is the first time that different methodologies for author profiling are evaluated under the framework of early text classification. In our study, we focus on social media data by analyzing users from Twitter. In the evaluation, the text evidence of the tweets is continuously being read –one-by-one– in order to simulate the user activity. The challenge is to model very small amounts of text in early stages (e.g., a handful of tweets), and also to exploit the additional textual evidence in late stages.
- The design of new high level features for author profiling called profile meta-words. These meta-words are low dimensional prototypes built from specialized term vectors for author profiling (López-Monroy et al., 2015). Term vectors and resulting meta-words relies in a space where each dimension is the association degree between each term and target profiles. Note that under this space, vectors of terms with little semantic relationship could be close to each other if they are relevant for specific profiles (e.g., *game, beer*, and *wife* because they are discriminative to males).
- The use, adaptation, and suitability evaluation of a multi-resolution representation in the context of early author profiling. This is a versatile representation to better encode and weight the amount of evidence in short/long texts during the different early/late stages. This representation exploits weighting schemes for short text categorization and the meta-words for author profiling. The hypothesis is that very little text evidence (early stages) can be better represented with few/coarser meta-words (e.g., predominant topics), but as more data is available, more/finer meta-words (e.g, specific topics/subtopics) are more suitable to encode larger amounts of information.

The rest of this paper is organized as follows. The next section provides a review of related work on author profiling and EAP. Section 2 describes the proposed meta-words and pyramidal representation and how they are used for EAP. Sections 3 and 4 describe the data collection and evaluation framework respectively. Sections 5 and 6 report experimental results and their discussion. Section 7 presents a more detailed analysis of the proposals of this research. Finally, Section 8 summarizes our main findings and outlines future work directions.

## 2. Related work

The Author Profiling (AP) task has been traditionally approached as a text classification problem. So far, most of the attention has been paid to exploit content and stylistic features (López-Monroy et al., 2015; Nguyen, Smith, & Rosé, 2011; Ortega-Mendoza et al., 2018; Rangel, Rosso, Potthast, & Stein, 2017). The usefulness of content features has been also demonstrated

[1] This is inspired by the Bag-of-Visual-Words (BoVW) strategy used in computer vision to represent images (Sivic & Zisserman, 2004).

in top ranked approaches at PAN forums for AP in social media (Rangel et al., 2017; Rangel, Rosso, Potthast, Stein, & Daelemans, 2015; Rangel et al., 2016). In those forums, the top performing methodologies have modeled the thematic content by means of *n*-grams (Basile et al., 2017; Meina et al., 2013) or distributional word-vectors (Alvarez-Carmona, López-Monroy, Montes-y Gómez, Villasenor-Pineda, & Escalante, 2015; op Vollenbroek et al., 2016). In general, these approaches have outperformed more complex methods including those based on topic modeling and deep learning (Rangel et al., 2014; Rangel et al., 2017; Rangel et al., 2015). Thematic features have been shown to be useful to profile gender (Ortega-Mendoza et al., 2018), age (op Vollenbroek et al., 2016), personality traits (Álvarez-Carmona, López-Monroy, Montes-y Gómez, Villaseñor-Pineda, & Meza, 2016), and native language (Basile et al., 2017).

Up to our knowledge, AP has not been addressed as an early text classification task. However there are other specific tasks on early text classification that have been gaining relevance. For example, there are specialized forums such as the eRisk-CLEF (Losada et al., 2017; 2018) that aims to study specific problems, such as early detection of depression and anorexia. According to the literature, Dulac-Arnold et al. (2011) is one of the first works to approach the early text classification task. In that work, the authors processed the documents reading sentence-by-sentence in order to predict the potential category at each sentence. The approach was a Markov Decision Process (MDP), and each sentence was represented using the Bag-of-Words (BoW). In other work, Escalante, Montes-y-Gómez, Villasenor, and Errecalde (2016) approached the problem of early sexual predator identification in chat conversations. They outperformed the approach in Dulac-Arnold et al. (2011) by proposing a method based on the Naive Bayes classifier.

In more recent works for early detection of aggressiveness and depression (Errecalde, Villegas, Funez, Ucelay, & Cagnina, 2017; Escalante et al., 2017), documents have been represented by adapting the distributional term representations, which are term vectors that exploits the distributional hypothesis to capture semantic information by observing the contextual terms. In those works, the authors successfully represented the partial information in documents by averaging word-vectors of terms that were read in each stage. In our evaluation we compare our approach with this strategy. However, instead of representing documents by averaging word-vectors, we aim to generate a representation especially suited for EAP by adapting these word-vectors into profile meta-words.

## 3. Profile based meta-words for EAP

Our methodology consists in adapting specialized word-vectors for author profiling (Section 3.1) into a multi-resolution representation for early text classification (López Monroy et al., 2018) (Section 3.2). The name comes from the goal of combining features with multiple levels of semantic granularities. The idea is to simultaneously exploit coarse and fine features (meta-words) for author profiling, with the aim of modeling the growing amount of textual evidence. Intuitively, coarse meta-words are suitable for representing content in short texts (early stages), whereas fine meta-words capture more information in longer texts (late stages). This methodology consists of two steps, the first one is to compute word vectors (Section 3.1), and the second one builds meta-words and the final document representation (Section 3.2).

### 3.1. Second order attributes for word representation

There are many alternatives to build word-vectors, two of the most popular ones are word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), and its improvement FastText (Bojanowski, Grave, Joulin, & Mikolov, 2016). The proposed methodologies for building meta-words (Section 3.2) and Pyramidal Histograms (Section 3.3) could use any word vector representation. Nonetheless, as our evaluation shows in this classification problem, we obtain better results by adapting the Second Order Attributes (SOA) from López-Monroy et al. (2015). This is not surprising since SOA is a specialized word-vector representation for AP tasks, and has been part of several state-of-the-art methodologies in different datasets (Escalante et al., 2017; Ortega-Mendoza et al., 2018; Rangel et al., 2015; Rangel et al., 2016; op Vollenbroek et al., 2016). In the rest of this section we describe an adapted version of the SOA word representation for this particular problem.

Let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_n, y_n)\}$ be a training set of labeled documents, that is, $\mathcal{D}$ is a collection of $n$ pairs of documents ($d_i$) and variables ($y_i$); where the latter indicates the profile associated with the document, with $y_i \in \mathcal{P} = \{P_1, \ldots, P_q\}$. Thus, $|\mathcal{P}| = q$ is the set of different profiles to be analyzed (e.g., *male* and *female*, or diverse language varieties). Also let $\mathcal{V} = \{v_1, \ldots, v_m\}$ denote the vocabulary of terms in the collection under analysis. SOA consists in representing the terms by their relation with each target profile (Li, Xiong, Zhang, Liu, & Li, 2011; López-Monroy et al., 2015). For this we represent each word $v_i \in \mathcal{V}$ with a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q} \rangle$, where the $t_{i,h}$ is the degree of association between word $v_i$ and profile $P_h$. Under this word vector representation, the weight $t_{i,h}$ is related to the occurrence of term $v_i$ in documents that are labeled with profile $P_h$. To compute the relationship between the $i$th word and the $h$th profile, we adapted a weighting scheme from Lavelli et al. (2004) that assigns to each term a weight proportional to its frequency in each profile. Eq. (1) formalizes the above ideas.

$$t_{i,h} = \sum_{\forall d_j : y_j == P_h} tf(v_i, d_j) \cdot \log \frac{|\mathcal{V}|}{|\mathcal{N}_j|} \tag{1}$$

where $\mathcal{N}_j \subseteq \mathcal{V}$ is the set of different terms in document $d_j$, and $tf(v_i, d_j)$ is defined in Eq. (2).

$$tf(v_i, d_j) = \begin{cases} 1 + \log(\#(v_i, d_j)) & \text{if } \#(v_i, d_j) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\#(v_i, d_j)$ indicates the frequency of term $v_i$ in $d_j$. The idea of SOA vectors is that the frequency of a term $v_i$ in documents belonging to profile $P_h$ determines the degree of association $t_{i,h}$ between them. This $t_{i,h}$ values are separately computed and aggregated for each document in Eq. (1).

Note that this adapted version of SOA weights words by using lexical diversity inside the documents in order to encode discriminative profiling information. Thus, with a larger number of different terms $|\mathcal{N}_j|$ in document $d_j$, the smaller the contribution of term $t_i$. The idea behind this is to give more relevance to terms in documents (e.g., tweets) that used few different words to convey a message. Finally, the representation of each term vector $\mathbf{t_i}$ is normalized using $||\mathbf{t_i}||_1 = 1$.

### 3.2. Meta-words and histograms

We define a meta-word as a prototypical word that summarizes a group of related words. To build these meta-words we use a clustering algorithm over a set of word vectors $T$ computed during training according to Section 3.1. Then, we use the centroids of those clusters as the meta-words. There are different alternatives for the clustering algorithm, and we explore among several effective strategies (e.g., Expectation Maximization, Hierarchical Clustering, etc.). In our evaluation we used $k$-means, which showed a good balance of performance and speed for finding clusters. More formally, we used $k$-means over $T$ to find cluster centroids, which

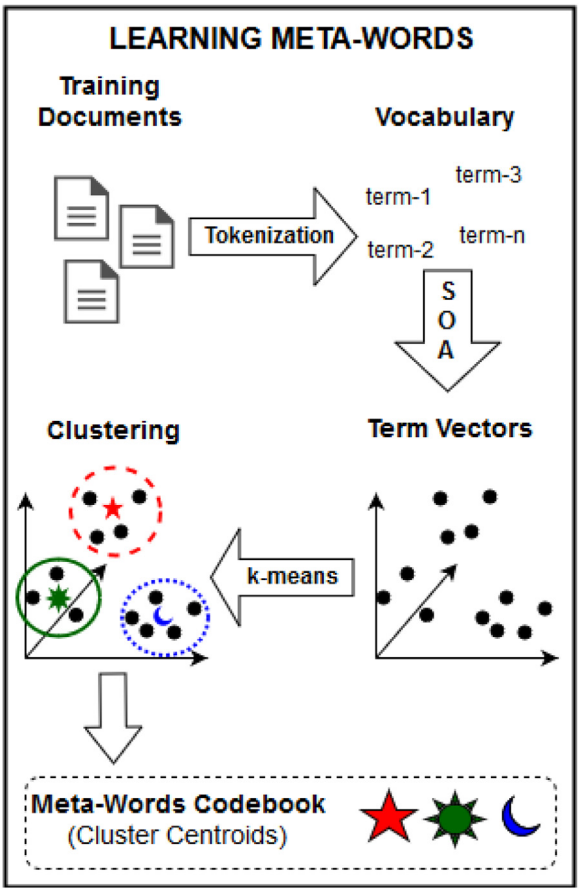



**Fig. 2.** Strategy to compute meta-words. We extract word vectors by using SOA. Then, we apply k-means in order to get $k$ meta-words (cluster centroids).

we label as meta-words $C = c_1, \ldots, c_k$ (see Fig. 2). After this procedure, we map each word $v_i$ in each document to the most similar meta-word in $C$. We do this by computing the Euclidean distance between vectors of words $t_i$ and cluster centroids in $C$. The final document representation is the histogram of meta-words occurring in each document.

In general each meta-word comprises highly discriminative information for this task. The key aspect relies in the adapted SOA word vectors from Section 3.1, where words are associated to each of the target profiles. This is crucial since the resultant clusters will be strongly driven by this property. For example, words often associated to males such as *games* and *wife* will probably be part of the same cluster (meta-word). The latter is because such word vectors under SOA will have similar representations (e.g., higher values for the *male* dimension). Intuitively, $k$-means will discover meta-words (groups of words) with different degrees of association for each target profile. Note that, under other word representations based on local contexts (e.g., word2vec), words like *games* and *wife* are likely to be in different clusters since they belong to semantically different spaces, and this will probably not help much in the task of EAP.

### 3.3. Pyramidal histograms of meta-words (PHM) and granularities

The final document representation can be seen as several histograms of meta-words with different sizes. We call this methodology Pyramidal Histograms of Meta-Words (PHM). In PHM the number of clusters defines the number and size of meta-words in each histogram. This property can be exploited to compute meta-words at different levels of granularity. In other words, fewer meta-words will result in a low dimensional representation with coarser features for the problem. For example, in a binary class setting of two meta-words, one meta/word will comprise the words more associated to class-one, whereas the other one the words related to class-two. Conversely, generating more meta-words would capture more specific aspects of each class.

The multi-resolution strategy is simple, yet very effective, and consists in computing a set of meta-vocabularies by using $k$-means, with different values of $k$, to obtain multiple levels of granularity in meta-words. The idea is to concatenate the different histograms of meta-words to build a versatile and efficient document representation with different levels of specificity. This multi-resolution representation will be more robust to represent the different amounts of text as it is being received. The motivation is that histograms with few/coarser features will better capture the little information in early stages (e.g., a handful of tweets), but as more evidence becomes available more/finer meta-words better encode the larger amounts of texts in late stages. This early prediction framework could be adapted to a wide range of contexts and applications in social media platforms. For example, as soon as a new user is detected, the system can begin to analyze the generated information in order to determine targets profiles. For example, it would be very useful to massively profile new users (e.g., gender, age, personality traits, etc.) that share harmful content or at risk of committing hazardous actions.

**Table 1**
Datasets considered for early evaluation.

| Dataset | Language | #Users | #Profiles |
|---|---|---|---|
| Gender | English | 3600 | 2 |
| | Spanish | 4200 | 2 |
| Language Variety | English | 3600 | 6 |
| | Spanish | 4200 | 7 |

## 4. Data collections

In our evaluation we used the datasets described in Table 1. We used 70% for training and 30% for test. The datasets come from the shared tasks PAN-CLEF 2017 and comprise a large number of users for the study. For each user, a set of one hundred tweets is available. We study English and Spanish languages, for the tasks of gender (male vs. female) and language variety identification. In the case of English there are variants from six countries: *Australia, Canada, Great Britain, Ireland, New Zealand,* and *United States*. For Spanish there are seven variants: *Argentina, Chile, Colombia, Mexico, Peru, Spain,* and *Venezuela*. Finally, the dataset is balanced regarding the number of documents in each target category and each user has exactly one hundred tweets.

## 5. Exploring the profile based meta-words

The aim of this section is to observe the kind of information that Meta-Words capture before we exploit them in our PHM strategy. The proposed EAP strategy relies on two components: (1) SOA, our adapted word vectors, and (2) Meta-Words, our proposed base feature for early prediction. In order to show the kind of information captured by these components, in Tables 2 and 3 we only focus on one resolution of ten meta-words. For this we ranked the meta-words by their discriminating power according to the $chi^2$ metric. Then, meta-words are named according to this rank. For example, Meta-Word-r1 is the best feature (rank=1), whereas Meta-Word-r10 is the worst one (rank=10).

In Table 2 we analyze the gender case showing four highly discriminative meta-words and the worst one (for contrast purposes). In that table we present in columns the most representative words

**Table 2**

Example of SOA word vectors for different Meta-Words in Gender Identification on the English dataset. Female Male.

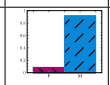
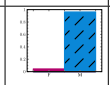
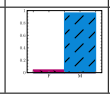
| Meta-Words for Gender Identification in English | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Meta-Word-r1 | | Meta-Word-r3 | | Meta-Word-r4 | | Meta-Word-r5 | | Meta-Word-r10 | |
| *word* | *vector* | *word* | *vector* | *word* | *vector* | *word* | *vector* | *word* | *vector* |
| *league* |  | *game* |  | *omg* |  | *family* |  | *to* |  |
| *arsenal* |  | *tax* |  | *vegan* |  | *today* |  | *i* |  |
| *tribune* |  | *beer* |  | *yoga* |  | *me* |  | *and* |  |

**Table 3**

Example of SOA word vectors and Meta-Words for Language Variety Identification in the English dataset. Australia Canada Great Britain Ireland New Zealand USA.

| Meta-Words for Language Variety Identification in English | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Meta-Word-r1 | | Meta-Word-r2 | | Meta-Word-r5 | | Meta-Word-r6 | | Meta-Word-r10 | |
| *word* | *vector* | *word* | *vector* | *word* | *vector* | *word* | *vector* | *word* | *vector* |
| *nz* |  | *edinburgh* |  | *america* |  | *beach* |  | *the* |  |
| *wellington* |  | *wales* |  | *obama* |  | *wine* |  | *me* |  |
| *earthquake* |  | *scottish* |  | *russia* |  | *mum* |  | *not* |  |
| *kiwi* |  | *hearts* |  | *fake* |  | *coffe* |  | *all* |  |
| *vlog* |  | *folk* |  | *nba* |  | *summer* |  | *dollar* |  |

in each meta-word. For this, we show word vectors of frequent words with the highest cosine similarity to the meta-word (cluster centroid). First of all, note that SOA word vectors expose associations with each profile. This is very important because the clustering that builds meta-words is driven by these associations, and not only if the words are semantically related. For example, note that Meta-Word-r1 seems to contain very specific words related to football, however the other Meta-Words contain words that are semantically distant (e.g., *tax, beer* and *game*). This confirms that meta-words based on SOA vectors are clusters of words exposing similar degrees of associations among profiles. In contrast, using other word representations such as word2vec, semantically distinct words (e.g., *family* and *today*) would be in different meta-words, even if they are highly discriminative to the same profile. Furthermore, meta-words based on word2vec could merge discriminative words for different profiles in the same cluster (e.g., *wife* and *husband*). In summary, meta-words based on SOA features are more focused in the AP problem.

In Table 3 we show a similar analysis for the language identification task. Note that we have meta-words with terms highly related to specific profiles (Meta-Words r1 and r2). However, there are also meta-words composed by terms associated to different profiles. For example, the words *beach* and *wine*, expose higher relationship to the profiles of people from *New Zealand* and *Australia*.

## 6. Evaluation

In this section we present the experimental results of the proposal and the reference evaluated methodologies. In Section 6.1 we describe the general experimental settings and the early text classification framework. In Sections 6.2 and 6.3 we present the evaluation of several methodologies in the context of early prediction.

### 6.1. Evaluation framework for early recognition

In our experimental methodology, we use a standard evaluation framework in early text classification: the chunk by chunk evaluation (Errecalde et al., 2017; Escalante et al., 2016; Losada et al., 2017; 2018). In this framework, we use the full-length training documents to build the classification model. In the test phase, we represent and classify the available information in each of the ten

chunks. Each chunk increases the number of available tweets in ten, therefore the last chunk contains the full user documents (100 tweets). In this way, we report the accuracy performance of the classification models at each chunk. Thus, the goal of early classification is to achieve good performance at early chunks. For evaluating the performance we report the accuracy, which is the most common metric used to evaluate models in author profiling tasks (Rangel et al., 2017).

In our text preprocessing, all documents were transformed to lower case and we tokenized words and punctuation marks as terms. We also discarded terms with a frequency lower than 10 in the training data (using a validation set of 30% out of the training). The base classifier for these experiments is a standard Support Vector Machine (SVM) with a linear kernel.

**Word Vector Representations:** The proposed representation can be computed using a different base word vector. Two well know word vector representations are word2vec (Mikolov et al., 2013) and its improvement, FastText, that exploits subword information (Bojanowski et al., 2016). In this work, we evaluate FastText[2] and Second Order Attributes (SOA) (López-Monroy et al., 2015). These word representations capture different aspects of the problem. However, as we will show in our experiments, the use of task-oriented vectors (SOA) provides an advantage for this particular problem. For the evaluation of the PHM, we set 5 different granularities: 10, 50, 100, 500 and 1000 meta-words. Please note that we did not fine tune these granularities and part of our future work is to automatically discover them. The idea in this evaluation is to represent each document under five distant and distinct levels of topic granularities for EAP. We will discuss the impact and number of these levels of meta-words in Section 6.5.

**Baselines:** We compare the proposed representation with Latent Semantic Analysis[3] (LSA) and BoW using the term frequency normalized by $l_1$. Furthermore, we also evaluate avg-SOA and Naive Bayes since both of them are strategies that have been used in state-of-the-art and relevant works for early prediction (Dulac-Arnold et al., 2011; Errecalde et al., 2017; Escalante et al., 2016; Escalante et al., 2017).

### 6.2. Early identification of gender results

In this section we present the evaluation for early gender identification in English and Spanish. Figs. 3 and 4 show results with several interesting findings. The first one is that our proposal (referenced as Pyramidal Histogram of Meta-Words (PHM) henceforth), independently of the word vector representation, is useful in early stages (chunks 1 to 4). This evidence shows PHM as an alternative to averaging word vectors (Avg) to represent documents in early classification scenarios. Note that in the particular case of the proposed PHM-SOA vs BoW, the improvements in chunks $1-4$ are between $\approx$9% and $\approx$4% for English and Spanish, respectively. After chunk 4, although to a lesser extent (especially for Spanish), there are also improvements in performance. In simple words PHM is more useful in early chunks, whereas in medium and late chunks, the performance is still better than traditional methods as more evidence becomes available, but the difference is smaller. Ideally early methods will be better in all chunks, but particularly beneficial in early stages.

Similar behavior can be observed when comparing PHM-SOA and the reference approach for early prediction; Avg-SOA. The improvements are $\approx$4% and $\approx$2% for English and Spanish respectively. These results confirm the usefulness of PHM to improve the use of Avg-SOA. More important, according to the Wilcoxon Signed Rank (WSR) (Demšar, 2006) test (p-value of 0.05) computed using the results at each chunk, there is a significant difference between PHM-SOA, Avg-SOA and the BoW. Regarding to PHM-FastText, results indicate improvements over Avg-FastText. Both FastText based approaches also outperformed the reference methodologies in most chunks. This might be due to the advantage in FastText to build word-vectors for words out of the vocabulary. We hypothesize that PHM-FastText did not obtain the best results because meta-words based on SOA capture thematic interests in a much broader way (we explain this in Section 6.5) and also because of the properties of SOA vectors showed in Section 5. Finally, the Naive Bayes classifier using Term Frequency (TF) and TF-IDF as weighting schemes, showed improvements over the BoW and LSA, but worse performance in early chunks than averaging word vectors and SVM (e.g., avg-SOA), which confirms findings in Dulac-Arnold et al. (2011), Escalante et al. (2016), Escalante et al. (2017).

### 6.3. Early identification of language variety results

Results for language variety identification show similar behavior to that of the gender task. In Figs. 5 and 6 we present experimental results for English and Spanish, respectively. There are several interesting outcomes in these results. First of all, we observe that PHM-SOA is the best performing approach in early stages, whereas for late chunks the performance is competitive with other methodologies (e.g., LSA and Avg-SOA). We infer that the multiple numbers and sizes of meta-words are the key of the performance at early chunks. Results also suggest that in these datasets the topic is more relevant for profiling, which results in a dramatic improvement of LSA, BoW and avg-SOA. These methods are well suited for modeling thematic information and achieved some of the best performances in late chunks.

Another interesting observation is the drop in performance of FastText. Note that, in gender identification FastText outperformed BoW and LSA in early stages. However in this problem there are much more categories with very specific thematic content (e.g., topics in Mexico are very different from topics in Spain). We infer that FastText fails in capturing these finer thematic interests associated to the profiles of users from different countries. In Section 6.5 we explain how meta-words based on SOA successfully capture this thematic information associated to target profiles. Finally, the Naive Bayes approaches, NB-TF and NB-TFIDF showed a similar behavior than in gender detection by outperforming BoW and LSA respectively. It is worth noting that Naive Bayes solutions tend to have a more stable improvements in performance than other baselines as more information is available.

### 6.4. The relevance of histogram granularities

In this section we study the impact of having histograms of meta-words with different granularities. Note that, our interest is not to fine tune the number and size of meta-vocabularies, but to study the impact of fine and coarse meta-words. Thus, we separately observe the performance across the different histogram granularities. In Table 4 we focus in the early gender identification showing the results for PHM. For this, we evaluate each of the five different histograms of meta-words ($H_1 = 10, H_2 = 50, H_3 = 100, H_4 = 500, H_5 = 1000$).

The evidence is clear; as the number of meta-words increases the performance in early stages decreases. This suggests that finer features are unnecessary in early stages, where short texts tend to produce scarce representation. Also note that the lower the number of meta-words, the better the performance in early stages. Similarly, the higher the number of meta-words, the more chunks are needed to increase the performance. In general, the experimental

[2] We used gensim to train n-grams and build word vectors from training. We tested pretrained embeddings (Godin, Vandersmissen, De Neve, & Van de Walle, 2015) from Twitter, but the performance did not show improvement.

[3] We empirically set 200 concepts in a validation set.

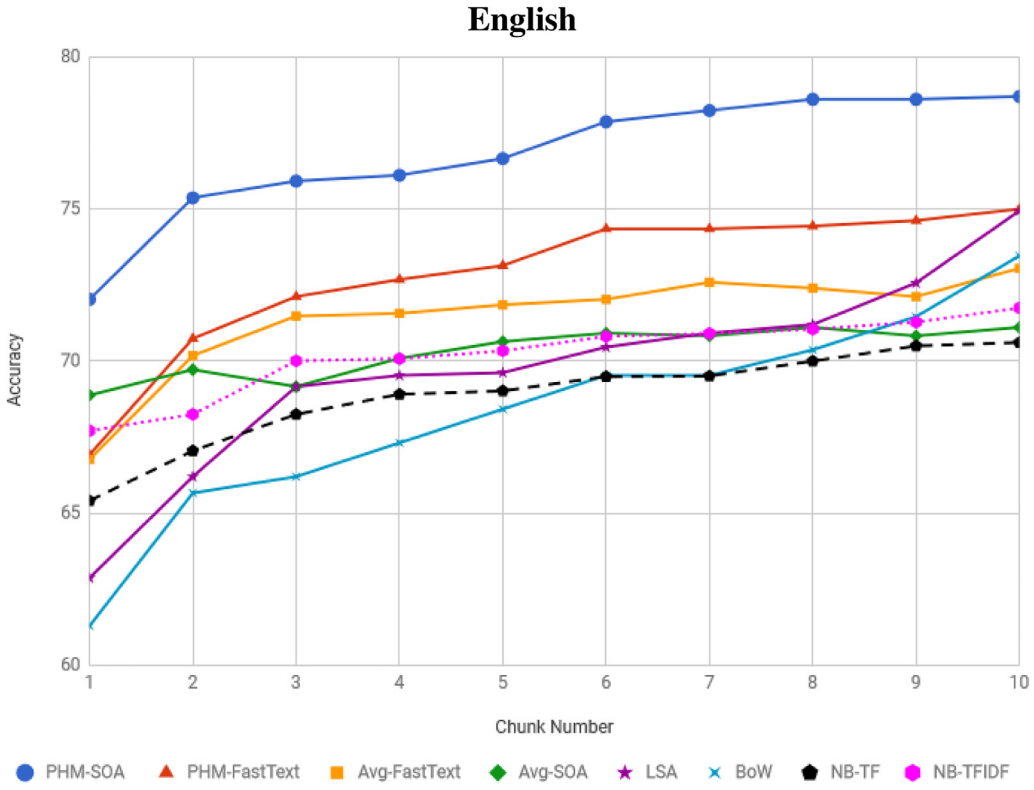



**Fig. 3.** Early Gender Identification in English. Accuracy results for the chunk by chunk evaluation. The *x* axis represents the chunk number. The *y* axis is the accuracy obtained by each methodology.

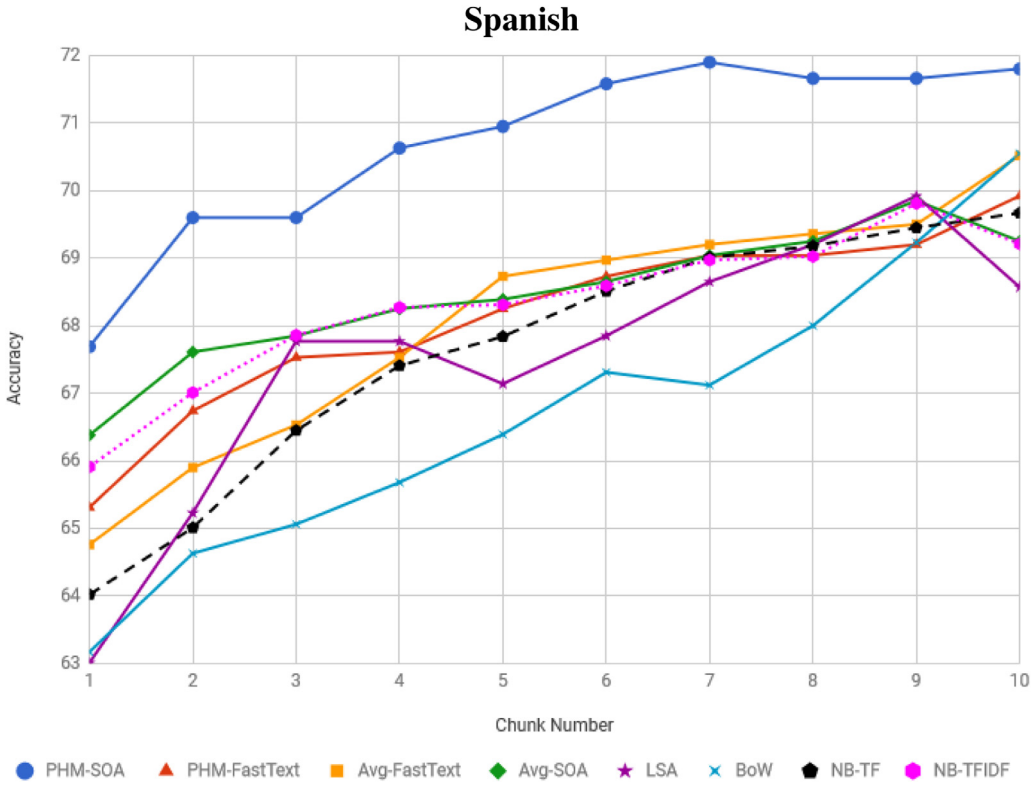



**Fig. 4.** Early Gender Identification in Spanish. Accuracy results for the chunk by chunk evaluation. The *x* axis represents the chunk number. The *y* axis is the accuracy obtained by each methodology.

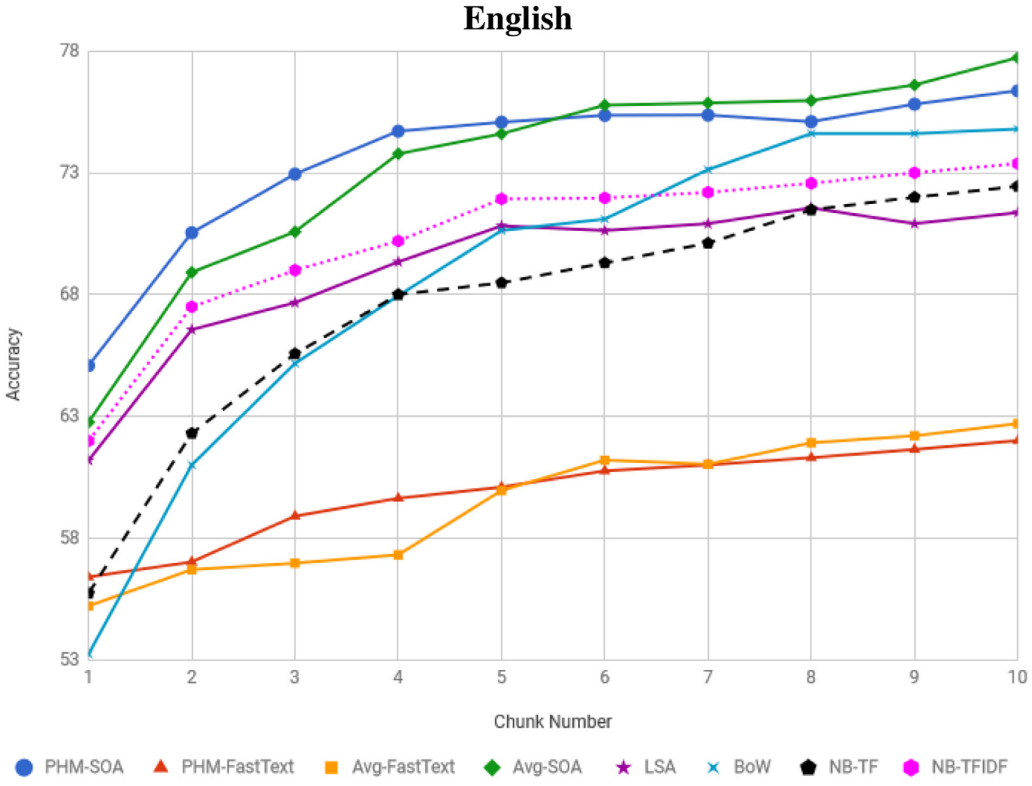



**Fig. 5.** Early Language Identification in English. Accuracy results for the chunk by chunk evaluation. The *x* axis represents the chunk number. The *y* axis is the accuracy obtained by each methodology.

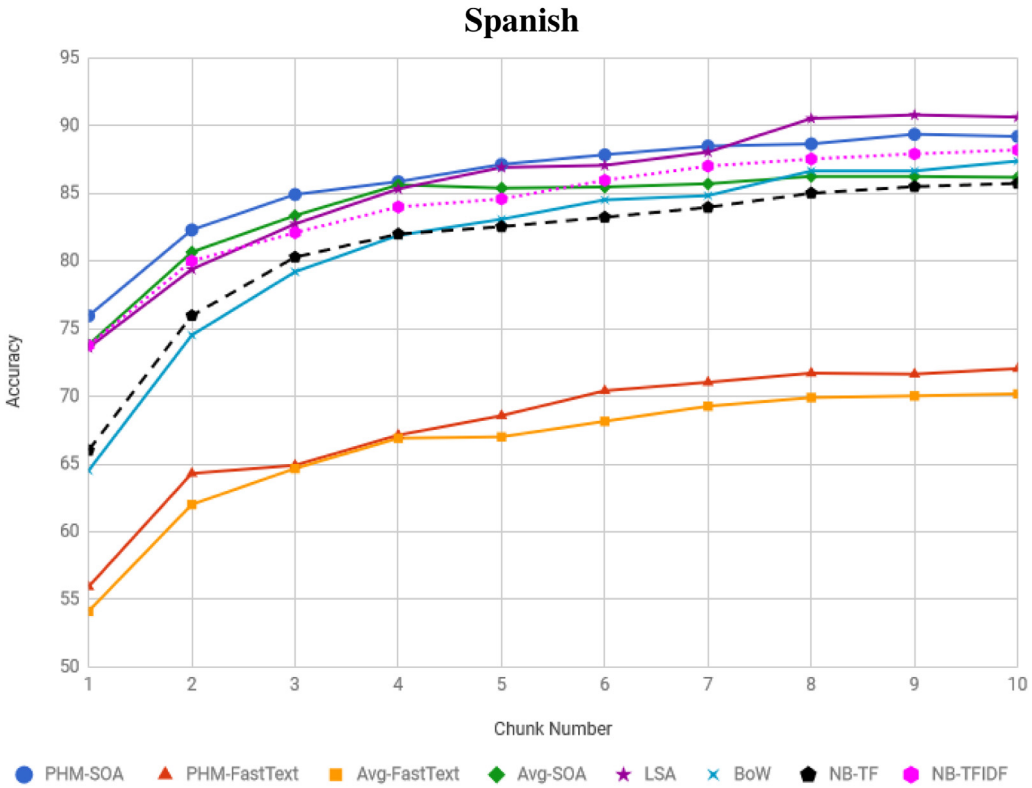



**Fig. 6.** Early Language Identification in Spanish. Accuracy results for the chunk by chunk evaluation. The *x* axis represents the chunk number. The *y* axis is the accuracy obtained by each methodology.

**Table 4**
Individual performance of histogram granularities in PHM. Accuracy results for the chunk by chunk evaluation in Early Gender Identification. The *All* experiment means the concatenation of all previous sizes.

| Histogram | $ch_1$ | $ch_2$ | $ch_3$ | $ch_4$ | $ch_5$ | $ch_6$ | $ch_7$ | $ch_8$ | $ch_9$ | $ch_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\|H_1\| = 10$ | 71.9 | 74.4 | 74.3 | 74.6 | 75.6 | 75.5 | 75.7 | 75.9 | 76.0 | 76.5 |
| $\|H_2\| = 50$ | 70.1 | 73.2 | 73.3 | 74.4 | 73.8 | 75.1 | 74.3 | 75.2 | 74.9 | 74.6 |
| $\|H_3\| = 100$ | 69.8 | 72.1 | 72.6 | 73.7 | 74.0 | 74.8 | 74.8 | 74.6 | 74.9 | 74.8 |
| $\|H_4\| = 500$ | 66.3 | 69.8 | 70.0 | 71.5 | 71.9 | 73.1 | 73.3 | 73.7 | 74.2 | 74.6 |
| $\|H_5\| = 1000$ | 65.7 | 68.0 | 69.2 | 70.4 | 70.7 | 72.4 | 72.1 | 72.5 | 72.4 | 72.5 |
| $\|All\| = 1660$ | **72.0** | **75.3** | **75.9** | **76.1** | **76.6** | **77.8** | **78.2** | **78.6** | **78.6** | **78.7** |

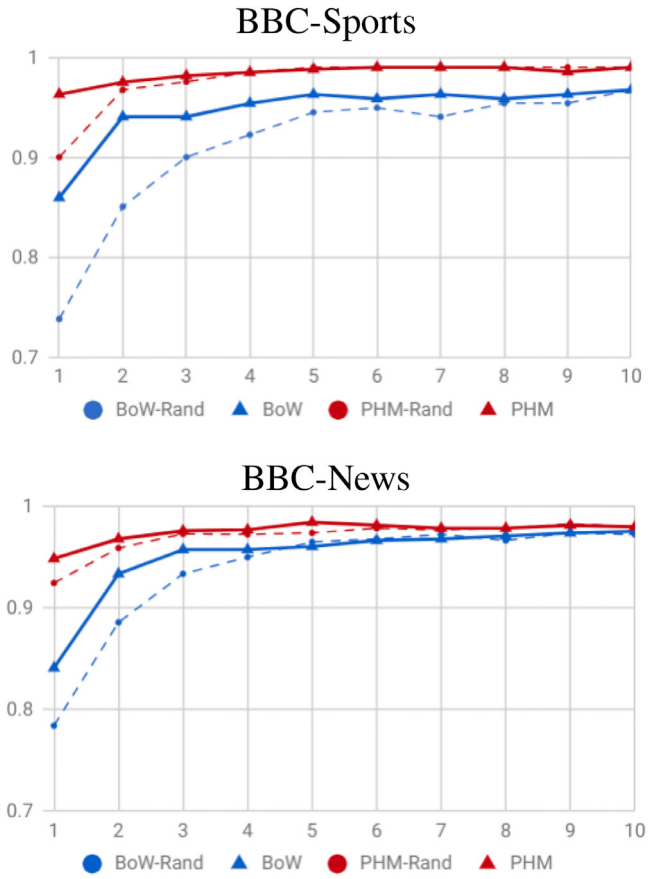



**Fig. 7.** PHM (red) vs. BoW (blue) on BBC corpora. Dotted lines show the performance in a corpus version where words in documents where shuffled. The *x* axis represents the percentage of text read at each chunk. The *y* axis is the accuracy obtained by each methodology.

evaluation shows better results if we combine the different resolutions compared to just using one of them. Finally, experimental results excluding some histogram also showed worse performance, therefore all of them are essential in the overall classification.

### 6.5. Capturing the thematic information

The aim of this section is to show that our proposal of adapting SOA vectors into multi-resolutions meta-words are also useful to capture the thematic information, which have proven to be useful for author profiling tasks. For this purpose, we design and experiment using two highly thematic datasets for *news*. In *news*, it is quite common to have the most relevant thematic information at the beginning (e.g., the title or the first sentences). We exploit this peculiarity to show that our approach successfully captures the topic in early chunks. The hypothesis is that such property makes it easier to obtain a higher performance in early chunks. Thus, we show that PHM captures thematic information even if such property is removed. For this we artificially build one extra version of each thematic dataset: The *"Rand"* version. In these datasets we shuffled the words contained in documents in order to lose the aforementioned property.

We used two thematic corpora about news from BBC (Greene & Cunningham, 2006). The first dataset is BBC-Sports, which has 737 documents with the following categories: *athletics, cricket, football, rugby* and *tennis*). The second dataset is BBC-News and contains 2225 documents from five categories: (*business, politics, football, sport* and *tech*). In Fig. 7 we show the performances of PHM (red) vs BoW (blue) on the BBC datasets. The solid line represents experiments on the original corpus, whereas the dotted line represents the dataset with shuffled words in documents. From these results we can outline the following findings:

1. PHM outperformed BoW at every chunk independently of the dataset ($\approx$10% in chunk one, and $\approx$5% in the other chunks).
2. Experiments on *Rand* datasets decreased the performance at early chunks. However, the drop was significantly less for PHM than in BoW.
3. PHM-Rand outperformed BoW and BoW-Rand at every chunk. In a few words, PHM is more suitable for capturing the thematic information. Also note that the difference is more evident with small amounts of text.
4. In BBC-News the decrease in performance was smaller. This is not surprising because the thematic among categories in BBC-New is broader and easier to classify by topic, whereas in BBC-Sports is narrower and more fine grained (news are about specific sports).

## 7. Conclusions

In this paper we presented the Pyramidal Histograms of Meta-Words (PHM) representation that captures thematic-related-profile information by means of histograms with different number of meta-words. The number and size of these meta-words can expose thematic content from Second Order Attributes (SOA) vectors with different levels of granularity. These granularity levels are useful to capture the relevant information in short and large documents across different early stages. The experimental evaluation revealed findings with impact in two early detection tasks; early prediction of gender and language variety. In this work we also showed valuable evidence about the importance of having a specialized word vector representation for the target problem. In the particular case of Author Profiling, SOA term vectors are helpful to capture valuable relationships between terms and profiles. The clustering method exploits these degrees of associations to build clusters and meta-words highly relevant for specific profiles. In this regard, we think that meta-words and the information that they capture are key aspects to understand the Early Author Profiling task. For future work we are interested in exploring strategies to automatically discover the optimal granularity level (number and

sizes of meta-words) instead of setting this manually in *k*-means. Finally, we also have an interest in studying the impact of external aspects such as text length, the number of categories and other types of social documents.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Credit authorship contribution statement

**A. Pastor López-Monroy:** Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Fabio A. González:** Conceptualization, Writing - review & editing, Supervision. **Thamar Solorio:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

## References


Alvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villasenor-Pineda, L., & Escalante, H. J. (2015). Inaoe's participation at pan'15: Author profiling task. In Working notes papers of the clef.

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 *Proceedings of the IBERO-American conference on artificial intelligence* (pp. 151–162). Springer.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017). N-gram: New Groningen author-profiling model. CoRR arXiv:1707.03764

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv:1607.04606.

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

Dulac-Arnold, G., Denoyer, L., & Gallinari, P. (2011). Text classification: A sequential reading approach. In *Proceedings of the 33rd european conference on ir research advances in information retrieval, (ECIR'11)*. In *LNCS: 6611* (pp. 411–423). Springers.

Errecalde, M. L., Villegas, M. P., Funez, D. G., Ucelay, M. J. G., & Cagnina, L. C. (2017). Temporal variation of terms as concept space for early risk prediction. In Clef (working notes).

Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., y Gómez, M. M., Morales, E. F., & Martínez-Carranza, J. (2015). Term-weighting learning via genetic programming for text classification. *Knowledge-Based Systems, 83*, 176–189. doi:10.1016/j.knosys.2015.03.025. URL: http://www.sciencedirect.com/science/article/pii/S0950705115001197.

Escalante, H. J., Montes-y-Gómez, M., Villasenor, L., & Errecalde, M. L. (2016). Early text classification: a naive solution. In *Proceedings of the NAACL-HLT 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 91–99).

Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications, 89*(Supplement C), 99–111. doi:10.1016/j.eswa.2017.07.040. URL: http://www.sciencedirect.com/science/article/pii/S0957417417305171.

Farias-Anzaldua, A. A., Montes-Y-Gómez, M., Lopez-Monroy, A. P., & Gonzalez-Gurrola, L. C. (2017). UACH-INAOE participation at eRisk2017 – Proceedings for eRisk at CLEF 2017. In *Proceedings of the CLEF, Dublin, Ireland*. CEUR-WS.org.

Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015). Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the workshop on noisy user-generated text* (pp. 146–153).

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on machine learning (ICML'06)* (pp. 377–384). ACM Press.

Lavelli, A., Sebastiani, F., & Zanoli, R. (2004). Distributional term representations: an experimental comparison. In *Proceedings of the CIKM* (pp. 615–624). ACM.

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.

López-Monroy, A. P., y Gómez, M. 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. doi:10.1016/j.knosys.2015.06.024. URL: http://www.sciencedirect.com/science/article/pii/S0950705115002427.

López Monroy, A. P., González, F. A., Montes, M., Escalante, H. J., & Solorio, T. (2018). Early text classification using multi-resolution concept representations. In *Proceedings of the conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1216–1225). Association for Computational Linguistics. doi:10.18653/v1/N18-1110. URL: http://aclweb.org/anthology/N18-1110

Losada, D. E., Crestani, F., & Parapar, J. (2017). Erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, … N. Ferro (Eds.), *Proceedings of the 8th International Conference of the CLEF Association Experimental IR Meets Multilinguality, Multimodality, and Interaction, CLEF, Dublin, Ireland, September 11–14, 2017, Proceedings* (pp. 346–360)). Cham: Springer International Publishing. doi:10.1007/978-3-319-65813-1_30.

Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of erisk: Early risk prediction on the internet. In *Proceedings of the international conference of the cross-language evaluation forum for European languages* (pp. 343–361). Springer.

Meina, M., Brodzinska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based classification for author profiling using various features. In Clef (working notes).

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

Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*. In *LaTeCH '11* (pp. 115–123). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=2107636.2107651.

Ortega-Mendoza, R. M., López-Monroy, A. P., Franco-Arcega, A., & y Gómez, M. M. (2018). Emphasizing personal information for author profiling: new approaches for term selection and weighting. *Knowledge-Based Systems, 145*, 169–181. doi:10.1016/j.knosys.2018.01.014. URL: http://www.sciencedirect.com/science/article/pii/S0950705118300224.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., … Daelemans, W. (2014). Overview of the author profiling task at PAN 2014. In *Proceedings of the Clef (online working notes/labs/workshop)* (pp. 898–927).

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. In L. Cappellato, N. Ferro, L. Goeuriot, & T. Mandl (Eds.), *Proceedings of the CEUR Workshop Proceedings*. CLEF and CEUR-WS.org. Working notes papers of the clef 2017 evaluation labs

Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. In *Proceedings of the CLEF*. sn.

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 *Proceedings of the CEUR workshop proceedings/Balog* (pp. 750–784). Working notes papers of the clef 2016 evaluation labs

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of the AAAI spring symposium on computational approaches for analyzing weblogs* (pp. 199–205).

Sivic, J., & Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the CVPR: 1* (pp. I–488). IEEE.

op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., & Nissim, M. (2016). GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016. In K. Balog, L. Cappellato, N. Ferro, & C. Macdonald (Eds.), *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org.