On the Impact of Emotions on Author Profiling

Francisco Rangel[1,2] and Paolo Rosso[1]

[1] NLE Lab, Universitat Politècnica de València, Camino de Vera, S/N, Valencia, Spain
prosso@dsic.upv.es,
http://www.dsic.upv.es/~prosso
[2] Autoritas Consulting, C/ Lorenzo Solano Tendero 7, Madrid, Spain
francisco.rangel@autoritas.es,
http://www.kicorangel.com

Abstract. In this paper, we investigate the impact of emotions on author profiling, concretely identifying age and gender. Firstly, we propose the EmoGraph method for modelling the way people use the language to express themselves on the basis of an emotion-labelled graph. We apply this representation model for identifying gender and age in the Spanish partition of the PAN-AP-13 corpus, obtaining comparable results to the bests performing systems of the PAN Lab of CLEF.

Keywords: affective processing, author profiling, emotion-labelled graphs, EmoGraphs

1 Introduction

Seth Godin[3] says "Marketing is no longer about the stuff that you make, but about the stories you tell". In this sense, world is rapidly changing, social media are growing on daily basis and customers are becoming users looking for new experiences. Thus the need of automatically processing the affective content of social media is acquiring a growing importance in order to know what the users want and need.

The potentiality offered by social media from many perspectives such as marketing, security or health is undeniable. But the information users include about themselves, if they include it, may lack of credibility. Age, gender, affiliation, likes... many users simply make them up. Getting to know the demographic and psychosocial profile of users on the basis of their writing style is an opportunity for organizations and companies, and a challenge for natural language processing technologies, due to the fact that the unique certainty we can have is what we can obtain from what the users write and share in social media.

Studies like [18] link the use of the language with demographics such as the gender of the author, but the vast majority of such investigations are limited to English. This investigation presents a method for automatically identifying emotions in social media, from a different perspective. Our hypothesis is that the way users express their emotions about topics depends on their age and gender. We aim at modelling the way they express them on the basis of a graph-based approach. The main motivation for

[3] http://heidicohen.com/seth-godin-7-truths-at-the-heart-of-marketing-how-to-use-them

using a graph-based approach is its capacity to analyse complex language structures. Pennebaker's [30] investigations are our main inspiration, where the style of writing is associated with personal attributes, such as demographics in our case. He employed a set of psycho-linguistic features obtained from texts, such as parts-of-speech, sentiment words and so forth. Our aim is to go further analysing the writing style from the perspective people combine the different parts-of-speech in a text, the kind of verbs they employ, the topics they mention, the emotions and sentiments they express (and where they do in the text), etc. Based on the differences between genders and among ages pointed out by Pennebaker (e.g. men used more prepositions than women because they tend to describe more in depth their environment), our intuition is that men will use more prepositional syntagmas than women, writing about different topics and with different emotionality, and this will confer a special importance to the sequence of *preposition* + *determinant* + *noun* + *adjective*. In this vein, we build a graph with the different parts-of-speech of user's texts and enrich it with semantic information with the topics they speak about, the type of verbs they use and the emotions they express. We model the whole text as only one single graph, considering also punctuation signs in order to capture how a writer may start and end sentences (i.e., how she connects her concepts in sentences). Once the graph is built, we obtain several properties from the graph and used them as features for a machine learning approach. Although the focus of this work is on Spanish, the proposed methodology may be applied to other languages.

The rest of the paper is structured as follows. In Section 2 we describe the related work on affective processing, author profiling and graph theory applied to text processing. In Section 3 we present our proposal for modelling the writing style to automatically identify emotions, age and gender (in Appendix A we analyse the complexity of the feature extraction). In Section 4 the evaluation framework is presented. Results are presented in Section 5 and discussed in Section 6, where we aim at identifying specific differences among age groups and gender with respect to used words, emotions and topics. In Section 7 we draw some conclusions.

2 Related Work

The main objective of this investigation is to show how the emotions expressed by users help us to profile them. We propose a novel approach based on graph theory for modelling the way people express their emotions. Therefore, the related work should be seen from three different perspectives: affective processing, author profiling and graph-based text processing.

2.1 Affective Processing

Automatic processing of affectivity has been focused mainly on sentiment analysis. However, there are a series of studies oriented to classify documents in the corresponding emotional category, usually based on the six basic emotions of Ekman (anger, disgust, joy, surprise, sadness, fear) [9]. At SemEval 2007 a task on the identification of emotions in news headline was organized. In [6] the authors used the Stanford syntactic parser for identifying what the main topic was about, estimating the polarity of each word with the help of Senti Wordnet [10] and Wordnet Affect [40]. In [19] the authors utilised three search engines for searching all the words in the headline combined with each emotion, and then calculated the Pointwise Mutual Information according to the number of returned documents. In [17] the authors used a supervised system based on unigrams and trained with another 1,000 news manually annotated by them. They used the Roget thesaurus to expand synonyms and extract the features. In [41] the results of SemEval are compared with other proposals, for instance with an approach where Latent Semantic Analysis (LSA) is employed to calculate similarity between a text and each of the six basic emotions.

In [7] the authors used the identification of imperative sentences, exclamation signs, the use of capital letters or the use of present and future, in order to identify polarity and emotional category. In a similar way, in [42] the authors used nouns, adjectives and verbs with the identification of keywords and types of sentences in Japanese in order to identify emotions. In [13] the authors used the ANEW affective dictionary. In [8] a method based on the Spanish Emotion Lexicon (SEL) is described for identifying emotions in short stories in Spanish. SEL consists of 2,036 words associated with the measure of "Probability Factor of Affective use" (PFA) related to one of the six basic emotions of Ekman [9]: joy, disgust, anger, fear, sadness, surprise. It defines four possible degrees of relationship with each emotion (null, low, medium, high) and 19 annotators indicated these values for each word. The PFA was calculated as an average of the percentages assigned to each degree. Finally, in [25] emotions and gender are investigated in three kind of emails: love letters; hate emails; and suicide notes.

2.2 Author Profiling

The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as psychology, linguistics and, more recently, natural language processing. In [29] the authors related language use with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding the gender and age of its author. In [2] the authors analysed formal written texts extracted from the British National Corpus, combining function words with part-of-speech features for gender prediction.

With the rise of the social media, the focus is on other kinds of writings, more colloquial, less structured and formal. Some researchers [26] manually labelled the collection with some risk of bias. In other cases, researchers took into account information provided by the authors themselves, with the risk of being deceived. For example: in [28] the authores retrieved a dataset from Netlog[4], where users report their gender and exact age; or in [18] the authors retrieved a dataset from Blogspot[5] and investigated the problem of automatically determining an author's gender by proposing combinations of simple lexical and syntactic features. In [37] the authors studied the effect of age and gender in the writing style in blogs; they obtained a set of stylistic features such as non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain.

[4] http://www.netlog.com

[5] http://blogspot.com

They demonstrated that language features in blogs correlates with age, as reflected in, for example, the use of prepositions and determiners.

Author profiling is in vogue in the research community and several are the tasks organised on different demographic aspects in the last two years: *a*) native language identification at BEA-8 workshop at NAACL-HT 2013 [6]; *b*) personality recognition at ICWSM 2013[7]; *c*) and age and gender identification, both in English and Spanish, at PAN 2013[8] and 2014[9].

Majority of approaches at PAN-AP 2013 [35] used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, and so on, together with POS tags and content-based features such as bag of words, TF-IDF, dictionary-based words, topic-based words, entropy-based words, or content-based features obtained with LSA. Only two participants used the occurrence of sentiment or emotional words as features. It is interesting to highlight the approach that obtained the overall best results using a representation that considered the relationship between documents and author profiles [27]. The best results in English were obtained with collocations [23].

In a similar vein, the interest in the industry in author profiling is evident in the Kaggle[10] platform, where companies and research centers can share their needs and independent researchers can join the challenge of solving them. We can find challenges as: *a*) psychopathy prediction based on Twitter usage[11]; *b*) personality prediction based on Twitter stream[12]; *c*) or gender prediction from handwriting[13].

In [34] we investigated emotions and gender. Concretely we investigated if the features used to identify the six basic emotions (see Section 3.1) could help to identify also the gender.

2.3 Graph-based Text Processing

Graphs have been widely used in information retrieval and natural language processing. For example, in [22] it is described in detail how to learn models based on graphs for tasks such as document retrieval, document classification, collaborative filtering, unified link analysis or image retrieval. In [12] a new graph-based model for the language-independent representation of documents is proposed together with a similarity measure between documents in order to compare documents written in different languages. In [31] authors represent documents with graphs considering multiple linguistic levels (lexical, morphological, syntactical and semantics) and they use the minimum path to

[6] https://sites.google.com/site/nlisharedtask2013/

[7] http://mypersonality.org/wiki/doku.php?id=wcpr13

[8] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html

[9] http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-profiling.html

[10] http://www.kaggle.com/

[11] http://www.kaggle.com/c/twitter-psychopathy-prediction

[12] http://www.kaggle.com/c/twitter-personality-prediction

[13] http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting

extract useful text patterns. For a comprehensive description about graph theory, networks and algorithms for information retrieval and several different natural language processing tasks, the interested reader may refer to [24].

With respect to affective processing, most of the graph-based approaches focused on sentiment analysis. One of the first approaches addressed the problem of automatically identifying the polarity of adjectives [15]. In [44] the relationship between word senses and subjectivity is investigated, whereas in [1] the authors model sentences as a graph of co-occurrence of words, disregarding the punctuation and stopwords. They obtain topological properties of the graph and use a machine learning algorithm to distinguish between positive and negative opinions. Finally, in [39] the authors build discourse-level opinion graphs to perform opinion analysis.

To the best of our knowledge, ours is the first time that an emotion-labelled graph is used to model age and gender. Our preliminary attempt of using only style-based features for gender identification was described in [34]. In this paper we will approach the problem of gender and age identification employing both style-based and emotion-labelled graph features.

3 Representation Models

Following we describe the style-based and the emotion-labelled graph-based features we will need for our document representation models.

3.1 Style-based Features

On the basis of what was already investigated for English by authors such as Pennebaker [29], we carried out some experiments to investigate the use of the different morphosyntactic categories in Spanish, depending on the type of media [33], the identification of emotions [34] and age and gender [36]. The complete set of style-based features is described below. Each item is a list of individual features represented by frequencies and combined into a vector space model. We obtained frequencies, punctuation marks and emoticons using regular expressions, whereas the morphosyntactic categories were obtained with the Freeling library[14]. In brackets the name of the set of features we will refer to in the experiments.

- (F)requencies: Ratio between number of unique words and total number of words; words beginning with capital letter; words all in capital letters; length of the words; number of capital letters and number of words with flooded characters (e.g. Heeeelloooo).
- (P)unctuation marks: Frequency of use of dots; commas; colon; semicolon; exclamation marks; question marks and quotes.

[14] http://nlp.lsi.upc.edu/freeling/ It is important to highlight that although the official accuracy of the Freeling POS tagger when morphological annotations are carried on is about 94-95%, when dealing with social media texts these numbers might decrease

- Morphosyntactic (C)ategories or Part-of-Speech (POS): Frequency of use of each grammatical category; number and person of verbs and pronouns; mode of verb; number of occurrences of proper nouns (Named Entity Recognition) and out of vocabulary (OOV), that is, words not found in dictionary.
- (E)moticons[15]: Ratio between the number of emoticons and the total number of words; number of the different types of emoticons representing emotions: joy, sadness, disgust, angry, surprise, derision and dumb.
- (SEL) Spanish Emotion Lexicon [38]: We obtained the probability factor of affective use value from the SEL dictionary for each lemma of each word. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion and build one feature per each emotion.
- Semantic classification of (V)erbs: We search for the semantic classification of verbs. On the basis of what was investigated in [21], we have manually annotated 158 verbs with one of the following semantic categories: *a*) *perception* (see, listen, smell...); *b*) *understanding* (know, understand, think...); *c*) *doubt* (doubt, ignore...); *d*) *language* (tell, say, declare, speak...); *e*) *emotion* (feel, want, love...); *f*) and *will* (must, forbid, allow...). We add six features with the frequencies of each verb type.

We do not use any content/context dependent features in order to obtain total independence from topics.

3.2 Emotion labelled graph-based features

Our goal is to build a graph-based representation, to apply graph analysis to extract some features, and to use such features to train a machine learning model. We obtain all the texts written by each author. For each text, we carry out a morphological analysis with Freeling, obtaining parts-of-speech and lemmas of the words. Freeling describes each part-of-speech with an Eagle label[16]. We model each part-of-speech as a node (N) of the graph (G), and each edge (E) defines the sequence of parts-of-speech in the text as directed links between the previous part-of-speech and the current one. For example, let us consider a simple text like the following:

El gato come pescado y bebe agua. (The cat eats fish and drinks water)

It generates the following sequence of Eagle labels:

DA0MS0->NCMS000->VMIP3S0->NCMS000->CC->VMIP3S0->NCMS000->Fp

[15] http://en.wikipedia.org/wiki/List_of_emoticons

[16] The Eagles group (http://www.ilc.cnr.it/EAGLES96/intro.html) proposed a series of recommendations for the morphosyntactic annotation of corpora. We used the Spanish version (http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html) where for example: in the sentence "El gato come pescado y bebe agua. (The cat eats fish and drinks water.), the word "gato" (cat) is returned as NCMS000 where NC means common noun, M means male, S means singular, and 000 is a filling until 7 chars; or the word "come" (eats) is returned as VMIP3S0 where V means verb, M means main verb (not auxiliary), I means indicative mode of the verb, P means present time, 3 means third person, S means singular, and 0 is a filling until 7 chars.

We model such sequence as the graph showed in Fig 1. Due to the fact that the link VMIP3S0 -> NCMS000 is produced twice, the weight of this edge is double than the rest.

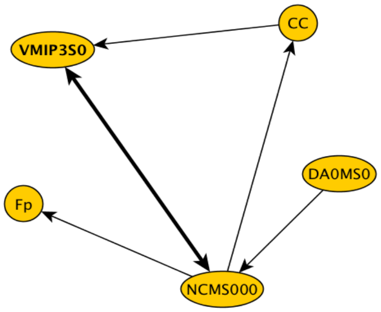


Fig. 1. POS Graph of "El gato come pescado y bebe agua." (The cat eats fish and drinks water.)

The following step is to enrich the described graph with semantic and affective information. For each word in the text, we look for the following information:

- **Wordnet Domains**[17]**:** If the word is a common noun, adjective or verb, we search for the domain of its lemma. We use Wordnet Domains linked to the Spanish version of the Euro Wordnet[18] in order to find domains of Spanish lemmas. If the word has one or more related topics, a new node is created for each topic and a new edge from the current Eagle label to the created node(s) is added. In the previous example, *gato* (cat) is related both to biology and animals, thus two nodes are created and a link is added from NCMS000 to each of them (NCMS000 -> biology & animals).
- **Semantic classification of verbs:** If the word is a verb we search for the semantic classification of its lemma as described in Section 3.1. We create a node with the semantic label and we add an edge from the current Eagle label to the new one. For example, if the verb is a perception verb, we would create a new node named "perception" and link the node VMIP3S0 to it (VMIP3S0 -> perception).
- **Polarity of words:** If the word is a common noun, adjective, adverb or verb, we look for its polarity in a sentiment lexicon[19]. For example, let us consider the following sentence:

  Ella es una amiga increíble. (She is an incredible friend.)

  It has the following sequence of Eagle labels:

[17] http://wndomains.fbk.eu/

[18] http://www.illc.uva.nl/EuroWordNet/

[19] With the help of experts on opinion analysis, we manually translated and adapted the sentiment lexicon described in [16] to Spanish

PP3FS000->VSIP3S0->DI0FS0->NCFS000->AQ0CS0(->positive & negative)->Fp

The adjective node AQ0CS0 has links both to the positive and negative tags, because *increíble* (incredible) could be both positive and negative depending on the context. Therefore, from a polarity viewpoint it is an ambiguous word which gives us two nodes (and two edges).

- **Emotional words:** If the word is a common noun, adjective, adverb or verb, we look for its relationship to one emotion in the Spanish Emotion Lexicon. We create a new node for each of them. See the following sentence as an example:

  He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público (I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public)

  The EmoGraph of the previous sentence is shown in Figure 2. The sequence may be followed by starting in VAIP1S0 node. Nodes size depends on their eigenvector and nodes colour on their modularity.

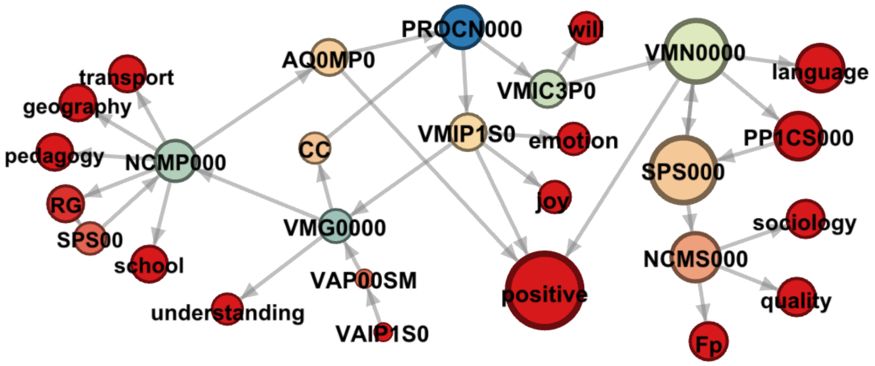


Fig. 2. EmoGraph of "He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público" (*"I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public"*).

Finally, we link the last element of the sentence (e.g. Fp) with the first element of the next one, since we are also interested in how people use sentence splitters (e.g. . ; :) and any other information prone to model how people use their language.

Once the graph is built, our objective is to use a machine learning approach to classify texts into the right gender and age categories. Therefore, we have first to extract features from the graph. We obtain such features on the basis of graph analysis in two ways: *a*) general properties of the graph describing the overall structure of the modelled texts; *b*) and specific properties of its nodes and how they are related to each other, that is, how they describe specific properties of how users use the language.

3.2.1 Graph-based Features

In this subsection we describe how to obtain general properties of the graph in order to build a set of eight features to feed our machine learning approach with.

- **Nodes-Edges ratio**. We calculate the ratio between the number of nodes N and the number of edges E of the graph G={N,E}. The maximum possible number of nodes (429) is given by: *a*) the total number of Eagle labels (247); *b*) the total number of topics in Wordnet Domains (168); *c*) the total number of verb classes (6); *d*) the total number of emotions (6); *e*) and the total number of sentiment polarities (2). The maximum possible number of edges (183,612) in a directed graph is theoretically calculated as:

  $$max(E) = N * (N - 1)$$

  where N is the total number of nodes. Thus, the ratio between nodes and edges gives us an indicator of how connected the graph is, or in our case, how complicated the structure of the discourse of the user is.
- **Average degree** of the graph, which indicates how much interconnected the graph is. The degree of a node is the number of its neighbours; in our case, this is given by the number of other grammatical categories or semantic information preceding or following each node. The average degree is calculated by averaging all the node degrees.
- **Weighted average degree** of the graph is calculated as the average degree but by dividing each node degree by the maximum number of edges a node can have (N-1). Thus, the result is transformed in the range [0,1]. The meaning is the same than the average degree but in another scale.
- **Diameter** of the graph indicates the greatest distance between any pair of nodes. It is obtained by calculating all the shortest paths between each pair of nodes in the graph and selecting the greatest length of any of these paths. That is:

  $$d = max_{n \in N} \varepsilon(N)$$

  where $\varepsilon(n)$ is the eccentricity or the greatest geodesic distance between n and any other node. In our case, it measures how far one grammatical category is from others, for example how far a topic is from an emotion.
- **Density** of the graph measures how close the graph is to be completed, or in our case, how dense is the text in the sense of how each grammatical category is used in combination to others. Given a graph G=(N,E), it measures how many edges are in set E compared to the maximum possible number of edges between the nodes of the set N. Then, the density is calculated as:

  $$D = \frac{2 * |E|}{(|N| * (|N| - 1))}$$

- **Modularity** of the graph measures the strength of division of a graph into modules, groups, clusters or communities. A high modularity indicates that nodes within modules have dense connections whereas they have sparse connections with nodes in other modules. In our case may indicate how the discourse is modelled in different structural or stylistic units. Modularity is calculated following the algorithm described in [3].
- **Clustering coefficient** of the graph indicates the transitivity of the graph, that is, if *a* is directly linked to *b* and *b* is directly linked to *c*, the probability that *a* is also linked to *c*. It indicates how nodes are embedded in their neighbourhood, or in our case, how the different grammatical categories (or semantic information such as emotions) are related to each others. For each node, the cluster coefficient (cc1) may be calculated with the Watts-Strogatzt formula [43]:

  $$cc1 = \frac{\sum_{i=1}^{n} C(i)}{n}$$

  Each $C(i)$ measures how close the neighbours of node i are to be a complete graph. It is calculated as follows:

  $$C(i) = \frac{|\{e_{jk} : n_j, n_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

  Where e_{jk} is the edge which connects node n_j with node n_k and k_i is the number of neighbours of the node i. Finally, we calculate the global clustering coefficient as the average of all node's coefficients, excluding nodes with degree 0 or 1, following the algorithm described in [20].
- **Average path length** of the graph is the average graph-distance between all the pairs of nodes and could be calculated following [5]. It gives us an indicator on how far some nodes are from others or in our case how far some grammatical categories are from others.

3.2.2 Node-based Features

For each node in the graph, we calculate two centrality measures: betweenness and eigenvector. We use each obtained value as the weight of a feature named respectively BTW-xxx and EIGEN-xxx, where xxx is the name of the node (e.g. AQ0CS0, positive, enjoyment, animal and so on).

- **Betweenness centrality** measures how important a node is by counting the number of shortest paths of which it is part of. The betweenness centrality of a node x is the ratio of all shortest paths from one node to another node in the graph that pass through x. We calculate it as follows:

  $$BC(x) = \sum_{i,j \in N - \{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$$

  where $\sigma_{i,j}$ is the total number of shortest paths from node i to node j, and $\sigma_{i,j}(n)$ is the total number of those paths that pass through n. In our case, if one node has a high betweenness centrality means that it is a common element used for link among parts-of-speech, for example, prepositions, conjunctions, or even verbs or nouns. This measure may give us an indicator of what the most common links in the linguistic structures used by authors are.

- **Eigenvector centrality** of a node measures the influence of such node in the graph [4]. Given a graph and its adjacency matrix $A = a_{n,t}$ where $a_{n,t}$ is 1 if a node n is linked to a node t, and 0 otherwise, we can calculate the eigenvector centrality score as:

  $$x_n = \frac{1}{\lambda} \sum_{t \in M(n)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{n,t} x_t$$

  where λ is a constant representing the greatest eigenvalue associated with the centrality measure, $M(n)$ is a set of the neighbours of node n and x_t represents each node different to x_n in the graph. This measure may give us an indicator of what are the grammatical categories with the most central use in the authors' discourse, for example nouns, verbs, adjectives, etc.

4 Evaluation Framework

In the following sections we describe the PAN-AP-13 corpus of the PAN Lab[20] of CLEF[21] and the methodology employed for identifying age and gender.

4.1 PAN-AP-13 Corpus

As described in [35], this corpus was collected from public repositories, retrieving posts labelled with author demographics such as gender and age[22]. Posts were grouped by author and chunking in different files those authors with more than 1,000 words in their posts. Authors with very few and short posts were also included in order to maintain a realistic evaluation framework. For age detection, we followed what was previously done in [37] and three classes where considered: 10s (13-17), 20s (23-27) and 30s (33-47). The corpus was balanced by gender and imbalanced by age group. In Table 1 statistics of the Spanish partition of the corpus are shown. In Figure 3 statistics of the number of words per author are given. In both cases the distribution follows a power distribution, with most of the authors using few words.

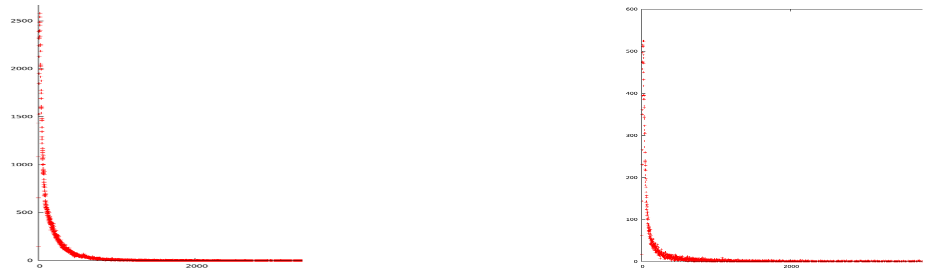
Age	**Gender**	**No. of Authors**	
		Training	Test
10s	male	1 250	144
	female	1 250	144
20s	male	21 300	2 304
	female	21 300	2 304
30s	male	15 400	1 632
	female	15 400	1 632
Σ		75 900	8 160

Table 1: PAN-AP-13 corpus statistics for its Spanish partition (training and test).

[20] Evaluation lab on uncovering plagiarism, authorship, and social software misuse: http://pan.webis.de

[21] Conference and Labs of the Evaluation Forum: http://www.clef-initiative.eu/

[22] Age and gender is given by the author without further verification and it is assumed to be true.



Training

Min.	Max.	Avg.	Std.
0	22 736	335	208

Test

Min.	Max.	Avg.	Std.
2	68 712	325	481

Fig. 3. Distribution of the number of words per author. Minimum, maximum, average and standard deviation.

4.2 Age and Gender Identification

We tested several different machine learning algorithms implemented in Weka[23]. The best results with the training set were obtained with: *a*) Support Vector Machine with a Gaussian Kernel with g=0.20 and c=1 for gender identification; *b*) and Support Vector Machine with a Gaussian Kernel with g=0.08 and c=1 for age identification. To compare our results with the ones of the participants in the PAN-AP-13 task, we evaluated our models on the test set. The accuracy measure was employed. In order to make sure of the statistical significance of our results, we applied the t-Student test assuming the null hypothesis $H_0 : p_1 = p_2$ which means that both classifiers are equal accurate, rejecting the null hypothesis otherwise. Although Part-of-Speech n-grams may already capture the syntactic structure of the discourse, our intuition is that graphs not only capture the syntactic structure but also its location in the text. We iterate n from 1 to 10 and select the best result for the POS n-grams.

5 Experimental Results

Our hypothesis is that the way people write about topics expressing emotions may help us to identify their demographics, concretely their age and gender. In Section 3 we described our proposed features for modelling the language on the basis of stylistic features and emotion-labelled graphs. The objective of the experiments is to demonstrate that the new approach, where both style-based and emotion-labelled graphs features are used, allows to identify age and gender with a very good accuracy. The results obtained in both gender and age identification show the competitiveness of our emotion-labelled graph approach.

[23] http://www.cs.waikato.ac.nz/ml/weka/

5.1 Gender Identification

We carried out the experiments with the Spanish partition of the PAN-AP-13 corpus. In Table 2 the results are shown. Rangel-S is our style-based approach [36], Rangel-nG our POS n-Grams and Rangel-EG is our new proposal that employs both style-based and emotion-labelled graph-based features. Rangel-EG outperforms significantly Rangel-S and Rangel-nG in gender identification. Furthermore, the proposed approach is competitive since it obtains the second position in the ranking.

Ranking	Team	Accuracy
1	Santosh	0.6473
2	**Rangel-EG**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
6	Flekova	0.6103
7	**Rangel-nG**	0.6016
8	Jankowska	0.5846
9	**Rangel-S**	0.5800
...	...	
19	Baseline	0.5000
...	...	
24	Gillam	0.4784

Table 2: Results in accuracy for gender identification in PAN-AP-13 corpus (Spanish)

We carried out the t-Student test in order to verify the significance of the positions. Results comparing Santosh to Rangel-EG ($z_{0.05} = 1.4389 < 1.960$) show no significant difference between methods at 95% of confidence. Santosh approached the task by combining content-based features (word n-grams), style-based features (POS n-grams and stylometry measures) and topic-based features (LDA topics). The comparison of Santosh with Pastor ($z_{0.05} = 2.3135 > 1.960$; $z_{0.01} = 2.3135 < 2.576$) shows that there is no difference at 95% of confidence but it does at 99% of confidence. Pastor approached the task with a second order representation based on the relationship between documents and profiles. Comparing Rangel-EG with Pastor ($z_{0.05} = 0.8748 < 1.960$), it is possible to appreciate there is no difference at 95% of confidence. The conclusion is that, no matter the ranking, the first three approaches obtained similar results with 95% and 99% of confidence respectively. The rest of the participants approached the task with classical features. For example, Haro and Flekova used BOW and Jankowska used character n-grams. As can be shown, methods such as EmoGraph using more elaborated features outperform those methods that use classical ones.

5.2 Age Identification

Another demographics of interest is age identification. We carried out some experiments with the Spanish partition of the PAN-AP-13 corpus. Results are shown in Table 3.

Ranking	Team	Accuracy
1	**Rangel-EG**	0.6624
2	Pastor	0.6558
3	Santosh	0.6430
4	**Rangel-S**	0.6259
5	Haro	0.6219
6	**Rangel-nG**	0.6162
7	Flekova	0.5966
...	...	
21	Baseline	0.3333
...	...	
23	Mechti	0.0512

Table 3: Results in accuracy for age identification in PAN-AP-13 corpus (Spanish)

As shown, the proposed approach outperforms the style-based approach and the POS n-grams one, and also is competitive since it obtains the first position in the ranking. We carried out the t-Student test showing no significant difference between Rangel-EG and Pastor ($z_{0.05} = 0.8894 < 1.960$) at 95% of confidence and Pastor and Santosh ($z_{0.05} = 1.7139 < 1.960$) but a significant difference between Rangel-EG and Santosh ($z_{0.05} = 2.6027 > 1.960$). In this case there is a slightly higher difference among the first three approaches although statistically there is not a significant difference between the first (our approach) and the second one. Similar to gender identification, those methods employing more elaborated features outperform approaches based on classical ones.

6 Discussion

In the previous section, we showed the effectiveness of the proposed features for age and gender identification. In this section we aim at investigating the differences in terms of usage of words as well as topics, verbs and emotions. We finalize this section investigating further the impact of emotions on author profiling and concretely of the emotion-labelled graph. At the end, the most discriminating features for gender and age identification are illustrated.

6.1 Differences on the Basis of Gender and Age

It is common sense that we write differently depending on our gender and age, but in most cases these are cliches without a scientific background. Here we show how people differ in the use of their language depending on their gender and age[24].

We are interested in identifying the words that females and males use more frequently. Figures 4 and 5 show top words for females and males respectively. As can be

[24] Conjectures are drawn on the basis of the studied corpora, and they should not be generalised

seen, there are no big differences between genders. No matter the gender, people seem to worry about life (*vida*) and what they love (*amor*), want (*quiero*) and hope (*espero*).

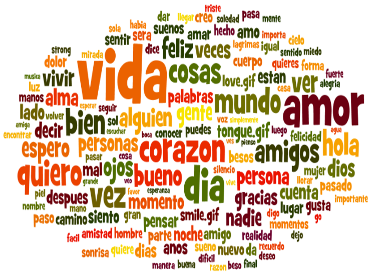


Fig. 4. Top words for females in PAN-AP-13 corpus

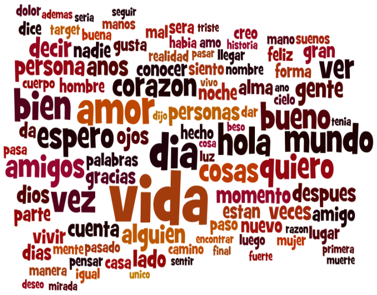


Fig. 5. Top words for males in PAN-AP-13 corpus

We obtained the topics people write about with the help of Wordnet Domains. We removed the most frequent topics[25] because not so informative being at the top of the domain hierarchy. The corresponding word clouds are shown in Figures 6, 7 and 8 for females in each age group (10s, 20s and 30s), and in Figures 9, 10 and 11 for males in the same age groups.

Younger people tend to write more about many different disciplines such as physics, linguistics, literature, metrology, law, medicine, chemistry and so on, maybe due to the fact that this is the stage of life when people mostly speak about their homework. Females seem to write more about chemistry or gastronomy, and males about physics or law. Both write about music and play. On the contrary of what one could might think, 10s females write about sexuality whereas males do not, and the contrary for commerce (shopping). As they grow up, both females and males show more interest in buildings (maybe due to the fact that they look for independence), animals, gastronomy, medicine, and about religion, although in a highest rate among males.

[25] e.g. biology, quality, features, psychological, economy, anatomy, period, person, transport, time and psychology

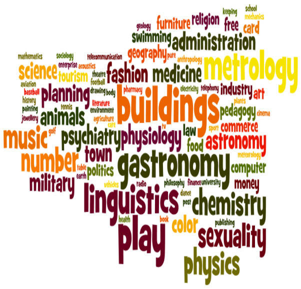


Fig. 6. Top domains for 10s females in PAN-AP-13 corpus

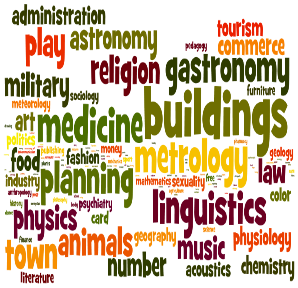


Fig. 7. Top domains for 20s females in PAN-AP-13 corpus

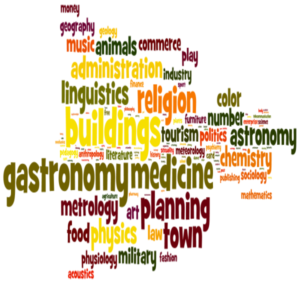


Fig. 8. Top domains for 30s females in PAN-AP-13 corpus

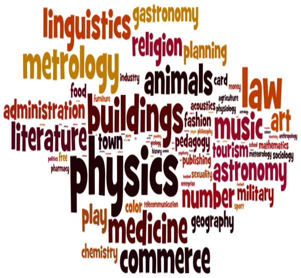


Fig. 9. Top domains for 10s males in PAN-AP-13 corpus

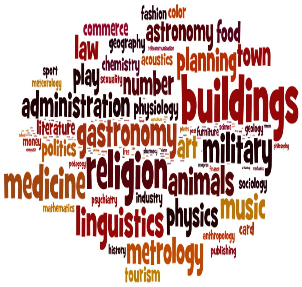


Fig. 10. Top domains for 20s males in PAN-AP-13 corpus

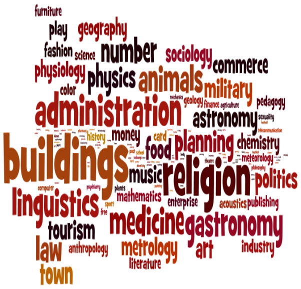


Fig. 11. Top domains for 30s males in PAN-AP-13 corpus

In Figure 12 we show the proportion in the use of emotional words per gender. Females and males do not seem to convey emotions in a very different way, although we can observe that females seem to express more disgust than males, who express more sadness.

With respect to the use of verb types, we are interested in investigating what kind of actions (verbs) females and males mostly refer to and how this changes over time. Figure 13 illustrates that females use more *emotional* verbs (e.g. feel, want, love...) than males, who use more *language* verbs (e.g. tell, say, speak...). This is an interesting conclusion because, although Figure 12 shows that females and males use emotional words in a similar proportion, Figure 13 depicts that females convey more verbal emotions than males.

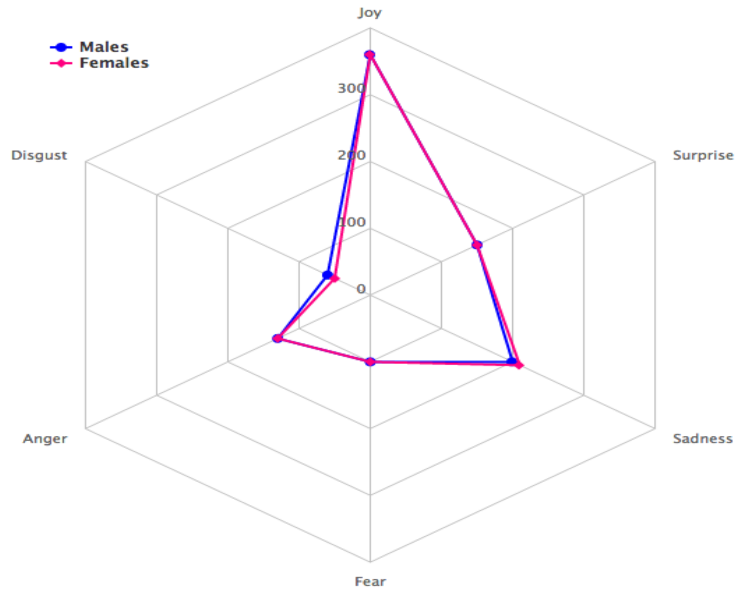


Fig. 12. Emotional words per gender in PAN-AP-13 corpus

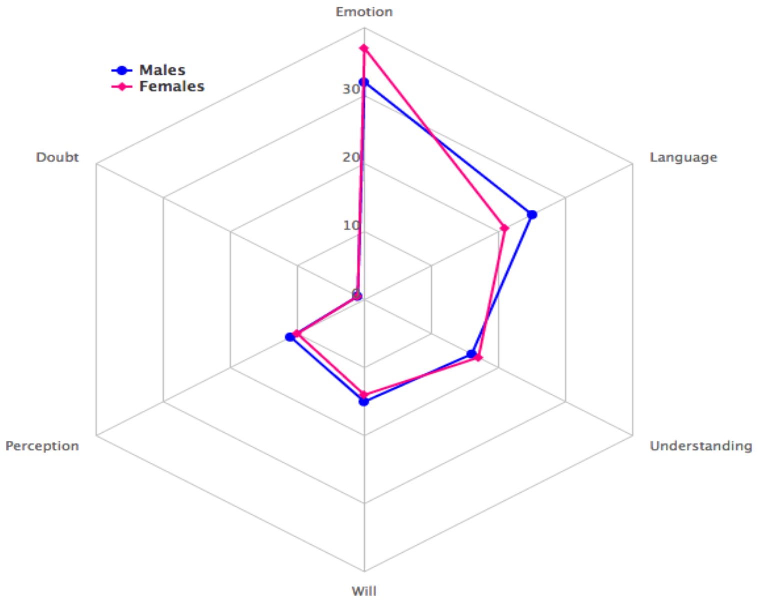


Fig. 13. Use of verb types per gender in PAN-AP-13 corpus

Since the use of the type of verbs seems to give us some differences between genders, we analyse the evolution in their use over the ages. Figures 14 and 15 show the evolution through 10s, 20s and 30s. The use of *emotional* verbs decreases over years, whereas verbs of *understanding* (e.g. know, understand, think...) seems to increase for males and remains stable for females, but it has to be said that females started using more verbs of understanding already in the early age at a similar ratio than males do later. Similarly, verbs of *will*[26] (e.g. must, forbid, allow...) increase for both genders, but at a higher rate for males. We can assert that females use more *emotional* verbs than males in any stage of life, and the contrary happens with verbs of *language*.

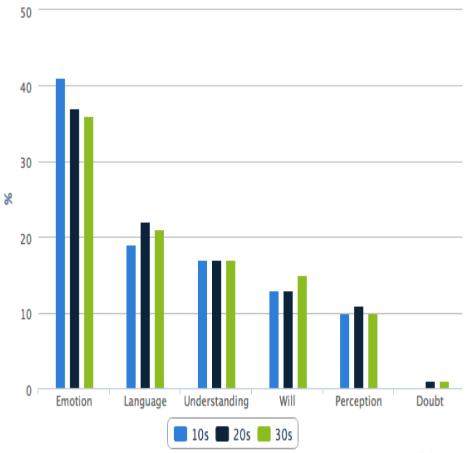


Fig. 14. Evolution in the Use of Verb Types for Females in PAN-AP-13 corpus

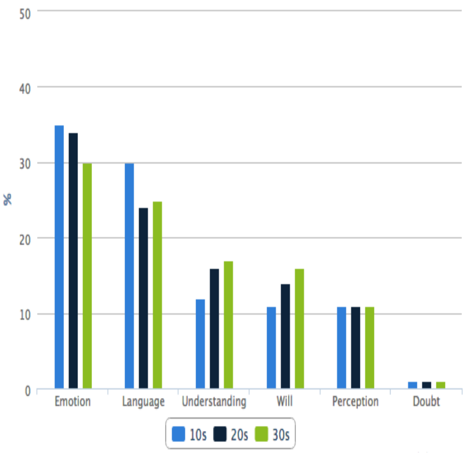


Fig. 15. Evolution in the Use of Verb Types for Males in PAN-AP-13 corpus

6.2 Most Discriminating Features

We analyse the most discriminating features for the identification of gender and age on the basis of information gain [45]. Table 4 shows the top 20 features over 1100. Betweenness (*BTW-xxx*) and eigenvector (*EIGEN-xxx*) features are among the top features[27]. We can identify a higher number of *eigen* features (mainly for verbs, nouns and adjectives) in gender identification in comparison to the higher number of *betweenness* features (mainly prepositions or punctuation marks) in age identification. This means that features describing the important nodes in the discourse provide more information to gender identification, whereas features describing the most common links in the

[26] "Verbs of will": verbs that suggest interest or intention of doing things (such as *must, forbid, allow*). Verbs of will do not have any relationship with *will* as the auxiliary verb for the future in English

[27] Refer to Section 3.2 where the features names are explained and to the Eagles website to know the meaning of the morphosyntactic annotations: http://www.cs.upc.edu/~nlp/tools/parole-sp.html

discourse provide more information to the age identification. In other words, the selection of the position in the discourse for words such as nouns, verbs or adjectives, which mainly give the meaning of the sentence, is the best discriminating features for gender identification, whereas the selection of connectors such as prepositions, punctuation marks or interjections are the best discriminating features for age identification. It is important to notice the amount of features related to emotions (SEL-sadness, SEL-disgust, SEL-anger) for gender identification and the presence of certain grammatical categories (Pron, Intj, Verb) for age identification.

Ranking	Gender	Age	Ranking	Gender	Age
1	punctuation-semicolon	words-length	11	BTW-NC00000	EIGEN-SPS00
2	EIGEN-VMP00SM	Pron	12	BTW-Z	BTW-NC00000
3	EIGEN-Z	BTW-SPS00	13	EIGEN-DA0MS0	punctuation-exclamation
4	EIGEN-NCCP000	BTW-NCMS000	14	BTW-Fz	emoticon-happy
5	Pron	Intj	15	BTW-NCCP000	BTW-Fh
6	words-length	EIGEN-Fh	16	EIGEN-AQ0MS0	punctuation-colon
7	EIGEN-NC00000	BTW-PP1CS000	17	SEL-disgust	punctuation
8	EIGEN-administration	EIGEN-Fpt	18	EIGEN-DP3CP0	BTW-Fpt
9	Intj	EIGEN-NC00000	19	EIGEN-DP3CS0	EIGEN-DA0FS0
10	SEL-sadness	EIGEN-NCMS000	20	SEL-anger	Verb

Table 4: Most discriminating features for gender and age identification

6.3 Error Analysis

In this subsection we aim at shedding some light on to what extent the error rate depends on the number of words per author. In order to analyse it more in depth we calculate the error rate taking into account the bins of 50 words (in Appendix B the detailed error rates depending on the number of words per author is presented). The results are shown in Figure 16. It is clearly shown that most of errors are committed when we have less than 50 words.

Looking at cases where the prediction fails and the author has less than 50 words we can see texts such as:

- *Bibliografia libros* (books bibliography)
- *El amor es una mentira* (love is a lie)
- *Hola a todos los cocolinos de corazon* (hello to all "cocolinos" of heart)

As it is expected, in these cases the prediction is random chance. Therefore, the complexity of the task is very dependent on the number of words per author. As the figure shows, age identification is more dependent than gender identification when a low number of words is available. In general, due to the fact that there are three categories in age identification instead of two in gender identification and the number of errors is very similar, we can conclude that gender identification is even a more difficult task than age identification. With respect to joint identification, results are more dependent on the number of words. For example, when the number of words increases, the joint identification error rate decreases with respect to the individual tasks, unlike when there are few words where the joint identification is almost double than the individual ones.

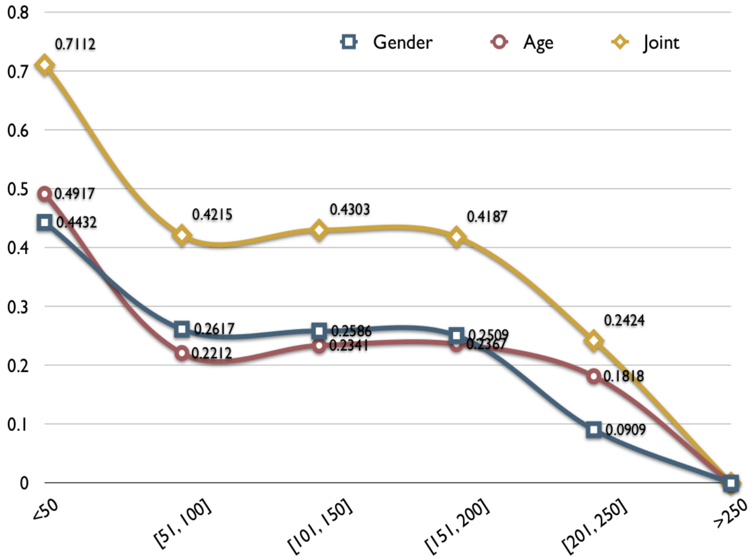


Fig. 16. Error rate per bin of 50 words

6.4 The Impact of the Emotion-labelled Graph

In Section 5 we already showed the improvement in the ranking when emotion-labelled graph features were used together with style-based ones. In order to understand further the impact of the emotion-labelled graph features vs. the style-based ones we carried out another experiment with another corpus, the EmIroGeFB [32] corpus of Facebook comments used in [34]. We also investigated the impact of the emotion-labelled graph (EmoGraph) vs. other variations of the graph, namely:

- **Simple Graph**: A graph built only with the grammatical category of the Eagle labels (the first character of the Eagle label), that is, verb, noun, adjective and so on;
- **Complete Graph**: A graph built only with the Eagle labels, without topics, verbs classification and emotions;
- **Semantic Graph**: A graph built with all the features described above (Eagle labels, topics and verbs classification) but without emotions.

We combine all the graph-based features with the (S)tyle-based ones (see Section 3.1): *a*) (F)requency; *b*) (P)unctuation marks; *c*) Grammatical (C)ategories; *d*) (E)moticons; *e*) (SEL): Spanish Emotion Lexicon. Results are shown in Table 5.

Features	Accuracy
F + P	0.5292
C	0.5542
F + P + C	0.5625
F + P + C + E + SEL	**0.5909**
Simple Graph + S	0.5083
Complete Graph + S	0.5192
Semantic Graph + S	0.5501
EmoGraph + S	**0.6596**

Table 5: Results for gender identification in accuracy; **c**orpus: EmIroGeFB (Facebook in Spanish)

The EmoGraph improves significantly ($z_{0.05} = 3.4764 > 1.960$) the task of identifying gender and strengthen even more our graph-based approach (see subsection 3.2).

7 Conclusions and Future Work

In this paper we investigated the impact of emotions on gender and age identification. We proposed an emotion-labelled graph to model the way people use the language and the emotions when writing. We compared the proposed model with state-of-the-art approaches obtaining respectively the first and the second best results on the Spanish partition of the PAN-AP-13 corpus. We presented the most discriminating features showing the importance of graph-based features, mainly the eigen-features for identifying gender and betweenness-features for identifying age.

We investigated further the obtained results in order to see whether it was possible to identify a different use of the language and the emotions. Females seem to make a higher use of emotional verbs (e.g. feel, want, love...) than males who instead use more language verbs (e.g. tell, say, speak...). With respect to the variation of the language and the addressed topics in the different age groups, it is interesting to highlight for instance how females in their 10s seem to worry more about sexuality than males.

Although the presented approach was tested on Spanish data, it can be applied to other languages if tools and resources such as POS tagger, emotion dictionary, etc. are available. This is an interested research line to be investigated because no matter the ability of graph to model and analyse language and emotions, the way people express their emotions could be not only language-dependent but also culture-dependent (e.g. the expression of emotions in oriental cultures [14]).

As future work we aim at applying the proposed method to address other problems related to author profiling such as language variety identification [46]. We also plan to investigate further the graph-based representation. For example, building and analysing heterogeneous and multimodal graphs with the aim at mining morphosyntactics and semantics information separately.

ACKNOWLEDGEMENTS

The work of the first author was partially funded by Autoritas Consulting SA and by Spanish Ministry of Economics under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the second author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS: Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. A special mention to María Dolores Rangel Pardo for her linguistic contribution to this investigation.

References

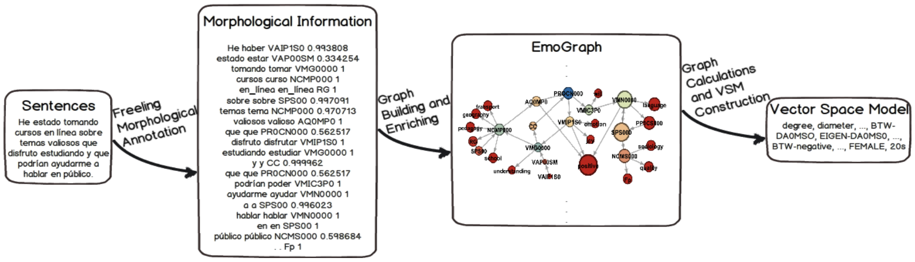
1. Amancio, D., Fabbri, R., Oliveira, O., Nunes, M., Costa, L. Distinguishing between Positive and Negative Opinions with Complex Network Features. TextGraphs-5: Graph-based Methods for Natural Language Processing, ACL Workshop. Uppsala, Sweden, pp. 83-87 (2013)
2. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, Genre, and Writing Style in Formal Written Texts. TEXT, vol. 23, pp. 321-346 (2003)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment, vol. 2008 (10), pp. 10008 (2008)
4. Bonacich, P. Factoring and Weighting Approaches to Clique Identification. In: Journal of Mathematical Sociology 2 (1), pp. 113-120 (1972)
5. Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25(2), pp. 163-177 (2001)
6. Chaumartin, F. Upar7: A knowledge-based system for headline sentiment tagging. In Proceedings of SemEval2007, Prague, Czech Republic, pp. 422-425 (2007)
7. Dhaliwal, K., Gillies, M., O'Connor, J., Oldroyd, A., Robertson, D., Zhang, L.: Facilitating online role-play using emotionally expressive characters. Artificial and Ambient Intelligence, Proceedings of the AISB Annual Convention, pp. 179-186 (2007)
8. Díaz Rangel, I.: Detección de afectividad en texto en español basada en el contexto lingüístico para síntesis de voz. Tesis Doctoral. Instituto Politécnico Nacional. México (2013) (in Spanish)
9. Ekman, P.: Universals and cultural differences in facial expressions of emotion. Symposium on Motivation, Nebraska, pp. 207-283 (1972)
10. Esuli, A., Sebastiani, F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06), pp. 417-422 (2006)
11. Forner, P., Navigli, R., Tufis, D. editors. CLEF 2013 Evaluation Labs and Workshop. Working Notes Papers, September, Valencia, Spain. CEUR-WS.org, vol. 1179 pp. 23-26 (2013)
12. Franco-Salvador, M., Rosso, P., Navigli, R. A Knowledge-based Representation for Cross-Language Document Retrieval and Categorization. In: Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL, Gothenburg, Sweden, pp. 26-30 (2014)
13. García, D., Alías, F.: Emotion identification from text using semantic disambiguation. Procesamiento del Lenguaje Natural. (50), pp. 75-82 (2008)
14. Harris, M. Emotions and their expression in Chinese culture. Journal of Nonverbal Behavior, Kluwer Academic Publishers-Human Sciences Press, vol. 17(4) pp.245-262 (1993)
15. Hatzivassiloglou, V., McKeown, K. Predicting the semantic orientation of adjectives. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, pp. 174-181 (1997)
16. Hu, M., Liu, B. Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Washington, USA, pp. 168-177 (2004)

17. Katz, P., Singleton, M., Wicentowski, R.: Swat-mp:the semeval-2007 systems for task 5 and task 14. In Proceedings of SemEval-2007, Prague, Czech Republic, pp. 308-313 (2007)
18. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literay and Linguistic Computing 17 (4), pp. 401-412 (2003)
19. Kozareva, Z., Navarro, B., Vazquez, S., Montoyo., A.: Ua-zbsa: A headline emotion classification through web information. In Proceedings of SemEval-2007, Prague, Czech Republic, pp. 334-337 (2007)
20. Latapy, M. Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) Graphs. In: Theoretical Computer Science (TCS) 407 (1-3), pp. 458-473 (2008)
21. Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993)
22. Liu Yi. Graph-based Learning Models for Information Retrieval: A Survey. Available at: http://www.cse.msu.edu/ rongjin/semisupervised/graph.pdf (2006)
23. Meina, M., Brodzinska, K., Celmer, B., Czokow, M., Patera, M., Pezacki, J., Wilk, M. Ensemble-based Classification for Author Profiling Using Various Features Notebook for PAN at CLEF 2013. In Forner et al. [11]
24. Mihalcea, R., Radev, D. Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press (2011)
25. Mohammad, S.M., Yang, T.: Tracking sentiment in mail: how gender differ on emotional axes. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Portland, Oregon, pp. 70-79 (2011)
26. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T. "How Old Do You Think I Am?"; A Study of Language and Age in Twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. (2013)
27. Pastor Lopez-Monroy, A., Montes-Gomez, M., Jair Escalante, H., Villasenor-Pineda, L., Villatoro-Tello, E. INAOEs Participation at PAN13: Author Profiling task. Notebook for PAN at CLEF 2013. In Forner et al. [11]
28. Peersman, C., Daelemans, W., Vaerenbergh, L.V. Predicting Age and Gender in Online Social Networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, ACM. Glasgow, Scotland, UK, pp. 37-44 (2011)
29. Pennebaker, J. W., Mehl, M. R., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, (54), pp. 547-577 (2003)
30. Pennebaker, J.W. The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press. (2011)
31. Pinto, D., Gómez-Adorno, H., Vilariño, D., Kumar Singh, V. A graph-based multi-level linguistic representation for document understanding In: Pattern Recognition Letters, pp. 1-9 (2013)
32. Rangel F., Hernández I., Rosso P., Reyes A. Emotions and Irony per Gender in Facebook. In: Proc. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD), LREC-2014, Reykjavík, Iceland, May 26-31. pp. 68-73 (2014)
33. Rangel, F., Rosso, P. El uso del lenguaje en los diferentes canales de Internet. In: Proceedings Comunica 2.0. Gandia, Spain, February 21-22. (2013) (in Spanish)
34. Rangel, F., Rosso. P. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. In: Proc. ESSEM Workshop on Emotion and Sentiment in Social and Expressive Media, AIxIA, CEUR-WS.org, vol. 1096, pp. 34-46 (2013)
35. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179 (2013)
36. Rangel, F., Rosso, F.: Use of Language and Author Profiling: Identification of Gender and Age. In: 10th International Workshop on Natural Language Processing and Cognitive Sciences NLPCS 2013 CIRM, Marseille, France, October 13-17. pp. 177-186 (2013)
37. Schler, J., Koppel, M., Argamon, S, Pennebaker, J.W. Effects of Age and Gender on Blogging. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, AAAI, pp. 199-205 (2006)

38. Sidorov, G., Miranda, S., Viveros, F., Gelbukh, A., Castro, N., Velásquez, F., Díaz, I., Suárez, S., Treviño, A., Gordon, J.: Empirical Study of Opinion Mining in Spanish Tweets. 11th Mexican International Conference on Artificial Intelligence, MICAI, pp. 1-14 (2012)
39. Somasundaran, S., Namata, G., Getoor, L., Wiebe, J. Opinion Graphs for Polarity and Discourse Classification. TextGraphs-4: Graph-based Methods for Natural Language Processing, ACL Workshop. Singapore, India. pp. 66-74 (2009)
40. Strapparava, C., Valitutti, A.: Wordnetaffect: an affective extension of wordnet. Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisboa. pp. 1083-1086 (2004)
41. Strapparava, C. Mihalcea, R.: SemEval- 2007 Task 14: Affective Text. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). pp. 70-74 (2008)
42. Sugimoto, F., Yoneyama, M.: A method for classifying emotion of text based on emotional dictionaries for emotional reading. Proceedings of the 24th IASTE D International Multi-Conference Artificial Intelligence and Applications. Innsbruck. pp. 91-96 (2006)
43. Watts, D.J., Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature 393 (6684): pp. 409-410 (1998)
44. Wiebe, J., Mihalcea, R. Word sense and subjectivity. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Sydney, Australia. pp. 1065-1072 (2006)
45. Yang, Y., Pedersen, J.O. A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML. pp. 412-420 (1997)
46. Zampieri, M., Gebrekidan-Gebre, B. Automatic Identification of Language Varieties: The Case of Portuguese. In Proceedings of the Conference on Natural Language Processing. Vienna, September 19-21. pp. 233-237 (2012)

Appendix A Complexity Analysis

The feature extraction is carried out in three steps: *(i)* the morphological annotation of the text with Freeling; *(ii)* the construction of the graph with the morphological labels and its enrichment with emotions, topics, sentiments and verb types; *(iii)* the calculation of different graph-based measures. The three-step feature extraction is shown in Figure B1. Following, each step is described and its complexity calculated.



He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

```
He haber VAIP1S0 0.993808
estado estar VAP00SM 0.334254
tomando tomar VMG0000 1
cursos curso NCMP000 1
en_línea en_línea RG 1
sobre sobre SPS00 0.997091
temas tema NCMP000 0.970713
valiosos valioso AQ0MP0 1
que que PR0CN000 0.562517
disfruto disfrutar VMIP1S0 1
estudiando estudiar VMG0000 1
y y CC 0.999962
que que PR0CN000 0.562517
podrían poder VMIC3P0 1
ayudarme ayudar VMN0000 1
a a SPS00 0.996023
hablar hablar VMN0000 1
en en SPS00 1
público público NCMS000 0.598684
. . Fp 1
```

degree, diameter, ..., BTW-DA0MSO, EIGEN-DA0MS0, ..., BTW-negative, ..., FEMALE, 20s

Fig. B1. Three-step feature extraction

Step (i). The morphological annotation with Freeling

Freeling morphological annotator is a standard Viterbi algorithm on a Hidden Markov Model, so the complexity is $O(W \cdot S^2)$, where W is the number of words of the sentence and S is the number of POS tags (in this case, average POS tags per word)[28]

Step (ii). The construction and enrichment of the graph

The construction of the graph is as follows. Each morphological annotation is a node and each couple of nodes has a relationship of precedence. In case the word is a common noun, adjective, adverb or verb, then we search for polarity and emotion information. If the word is just a common noun, adjective or verb, we search also for the domain. Moreover, if the word is a verb, then we search for the semantic category. The search is done in a hashtable therefore the complexity in the worst case is $O(W)$, where W is the number of words annotated.

Step (iii). The calculations in the graph

Once the graph is built, most of the calculations may be done in parallel. Calculations such as the nodes-edges ratio with complexity $O(1)$, average degree and weighted average degree with $O(N)$, or clustering coefficient with $O(N)$, where N is the number of nodes. Following, we calculate the complexity of the other two graph-based features (diameter and modularity) and the two node-based ones related with the centrality concept (betweenness and eigenvector).

[28] http://nlp.lsi.upc.edu/freeling/index.php?Itemid=65&func=view&id=3998

- Diameter: The shortest paths must be calculated and the diameter is the greatest length of any of these paths. To calculate the shortest paths the Bellman-Ford[29] algorithm is used because it handles directed graphs. Its complexity is $O(N \cdot E)$, where N is the number of nodes, and E the number of edges.
- Modularity: The method described in [3] is used. In short, the algorithm initialises each node as a different community and works in two steps: *a*) for each node and its neighbours the gain on modularity is calculated according to a given formula; *b*) once the step one is done, each community is treated as a node and repeated the step (a) several times. As the number of communities decreases in each step, the complexity may be estimated as $O(N \cdot logN)$, even though their authors estimated the average complexity as linear.
- Betweenness centrality: According to the formula indicated in Section 3.2.2, for each node x we need to compute the ratio between all the shortest paths from one node to another and those which pass through node x. The computation of the shortest paths with the Bellman-Ford algorithm has a complexity of $O(N \cdot E)$. The complexity of the betweenness formula, once the shortest paths were calculated, is $O(N^2)$. Here we take advantage of the previously calculated shortest paths in the diameter calculation. Although we need to store all the shortest paths, due to the maximum possible number of nodes in our graph (429) and due to the fact that we are modelling text so the possible number of edges is very limited, the spatial cost is negligible.
- Eigenvector centrality: Given a square matrix A (the adjacency matrix of the graph), an eigenvector is a non-zero vector v that when multiplied with A yields an eigenvalue, a scalar multiple of itself named λ. The relationship is $A \cdot v = \lambda \cdot v$. The power iteration algorithm[30] is used to calculate the λ with the greatest absolute value. Its cost is $O(N^2)$, where N is the number of nodes.

The two most complex calculations are the ones for the morphological annotation and the centrality measures: $O(W \cdot S^2)$ + [$O(N \cdot E)$ + $O(N^2)$], where W is the number of words of the sentence, S is the number of POS tags, N the number of nodes and E the number of edges.

Although the above overall complexity for feature extraction is higher than the one we would have extracting n-grams, our approach allows for a dimensionality reduction to *always* 1100 features. This is an important aspect when dealing with large datasets.

[29] http://en.wikipedia.org/wiki/Bellman-Ford_algorithm
[30] http://en.m.wikipedia.org/wiki/Power_iteration

Appendix B Error Rates per Number of Words per Author

In Figures B2, B3 and B4 we can appreciate that the error rate seems to be : *a*) higher and more concentrated when there are less than 50 words per author; *b*) tend to be close to 0 when there are more than 150 or 200 words; *c*) more sparse otherwise.

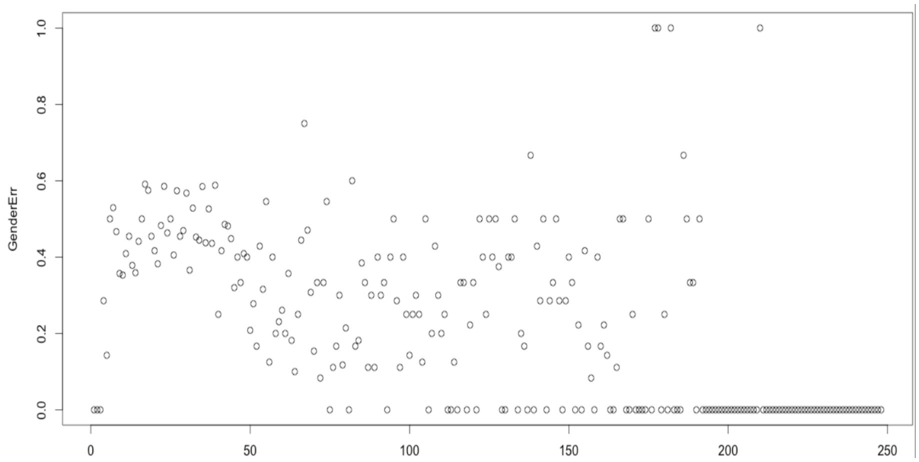


Fig. B2. Error rate (Y-axis) depending on the number of words per author for gender identification

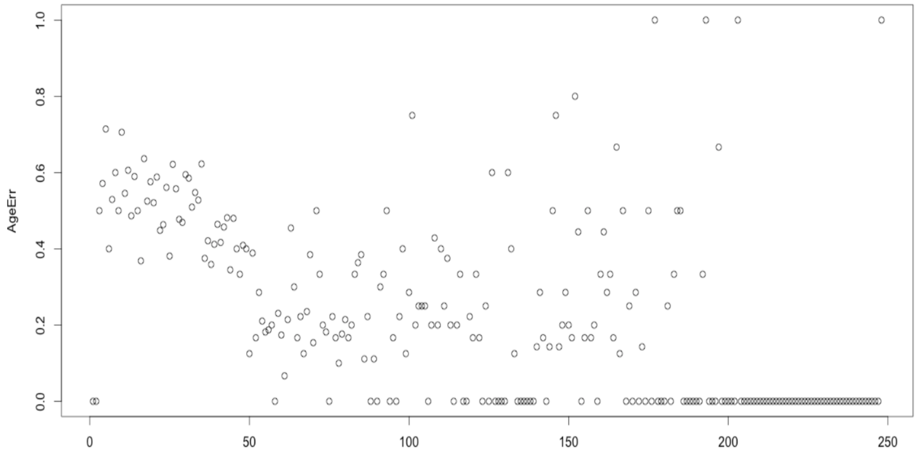


Fig. B3. Error rate (Y-axis) depending on the number of words per author for age identification

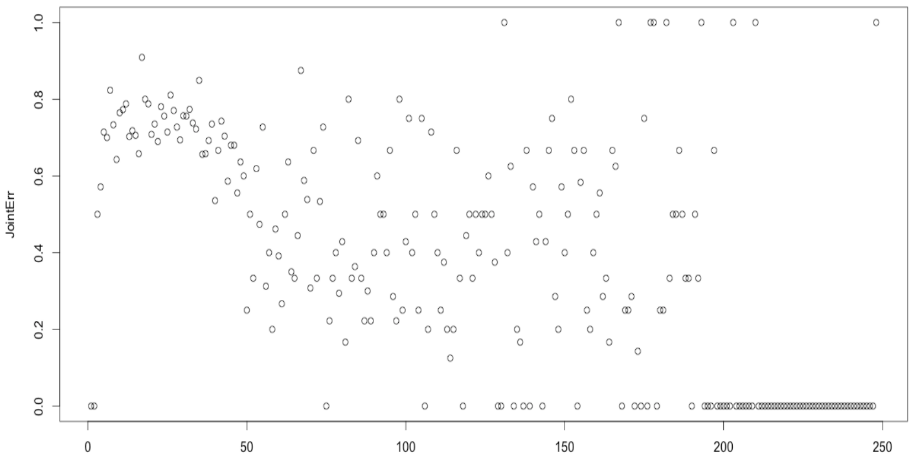


Fig. B4. Error rate (Y-axis) depending on the number of words per author for the joint identification