Rosso, P.; Rangel-Pardo, FM. (2017). Author Profiling in Social Media: The Impact of Emotions on Discourse Analysis. Lecture Notes in Computer Science. 10583:3-18. doi:10.1007/978-3-319-68456-7_1

Author Profiling in Social Media: The Impact of Emotions on Discourse Analysis

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

[1] PRHLT Research Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es
http://www.dsic.upv.es/~prosso
[2] Autoritas Consulting, Spain
francisco.rangel@autoritas.es
http://www.kicorangel.com

Abstract. In this paper we summarise the content of the keynote that will be given at the 5th International Conference on Statistical Language and Speech Processing (SLSP) in Le Mans, France in October 23-25, 2017. In the keynote we will address the importance of inferring demographic information for marketing and security reasons. The aim is to model how language is shared in gender and age groups taking into account its statistical usage. We will see how a shallow discourse analysis can be done on the basis of a graph-based representation in order to extract information such as how complicated the discourse is (i.e., how connected the graph is), how much interconnected grammatical categories are, how far a grammatical category is from others, how different grammatical categories are related to each other, how the discourse is modelled in different structural or stylistic units, what are the grammatical categories with the most central use in the discourse of a demographic group, what are the most common connectors in the linguistic structures used, etc. Moreover, we will see also the importance to consider emotions in the shallow discourse analysis and the impact that this has. We carried out some experiments for identifying gender and age, both in Spanish and in English, using PAN-AP-13 and PAN-PC-14 corpora, obtaining comparable results to the best performing systems of the PAN Lab at CLEF.

Keywords: author profiling, graph-based representation, shallow discourse analysis, EmoGraph

1 Author Profiling in Social Media

Often social media users do not explicitly provide demographic information about themselves. Therefore, due to the importance that is for marketing, but also for security or forensics, this information needs to be inferred somehow, for instance on the basis of how language is generally used among group of people that may share a more commong writing style (e.g. adolescents vs. adults).

Studies like [8] linked the use of language with demographic traits. The authors approached the problem of gender and age identification combining function words with part-of-speech (POS) features. In [15] the authors related the language use with

personality traits. They employed a set of psycho-linguistic features obtained from texts, such as POS, sentiment words and so forth. In [22] the authors studied the effect of gender and age in the writing style in blogs. They obtained a set of stylistic features such as non-dictionary words, POS, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. They showed that language features in blogs correlates with age, as reflected in, for example, the use of prepositions and determiners.

More recently, at PAN 2013[3] and 2014[4] gender and age identification have been addressed in the framework of a shared task on author profiling in social media. Majority of approaches at PAN-AP 2013 [18] and PAN-AP 2014 [19] 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, dictionary-based words, topic-based words, entropy-based words, etc. 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 [14]. The best results in English were obtained employing collocations [12].

Following, in Section 2 we describe how discourse features can be extracted from a graph-based representation of texts, and in Section 3 we show the impact that considering emotions in the framework of discourse analysis may have. Finally, in Section 4 we draw some conclusions.

2 Discourse Analysis in Author Profiling

Very recently, discourse features started to be used in author profiling [23], [24]. Rhetorical Structure Theory (RST)[5] has been applied for the characterization of the writing style of authors. Features have been extracted from the discourse trees, such as the frequencies of each discourse relation per elementary discourse unit, obtaining interesting results when used in combination with other features. Unfortunately, no comparison has been made with any state-of-the-art method, for instance on the PAN-AP-13 and PAN-AP-14 corpora, and it is difficult to fully understand the impact that the use of discourse features may have on author profiling, but the preliminary results that have been obtained are quite promising.

Our aim is instead to extract discourse features after modelling the use of language of a group of authors with a graph-based representation. These features will indicate the discourse complexity, the different structural and stylistic units the discourse is modelled in, etc. Concretely, our aim is to analyse the writing style from the perspective people combine the different POS in a text, the kind of verbs they employ, the topics they mention, the emotions and sentiments they express, etc. As Pennebaker pointed

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

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

[5] RST is a descriptive linguistic approach to the organization of discourse based on the linguistic theory formulated by Mann and Thompson in 1988 [11]

out [16], men generally use more prepositions than women and, for instance, they may use more prepositional syntagmas than women (e.g. preposition + determinant + noun + adjective). In the proposed approach, we build a graph with the different POS of authors' 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 text of authors of a given gender or age group as a single graph, considering also punctuation signs in order to capture how a gender or age group of authors connects concepts in sentences. Once the graph is built, we extract from the graph structure and discourse features we feed a machine learning approach with. Moreover, we will see that the way authors express their emotions depends on their age and gender. The main motivation for using a graph-based approach is its capacity to analyse complex language structures and discourses.

2.1 EmoGraph graph-based representation

For each text of a group of authors, we carry out a morphological analysis with Freeling[6][4, 13], obtaining POS and lemmas of the words. Freeling describes each POS with an Eagle label[7]. We model each POS as a node (N) of the graph (G), and each edge (E) defines the sequence of POS 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)

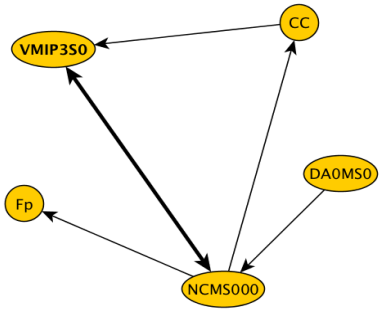

Fig. 1. POS Graph of "El gato come pescado y bebe agua." (The cat eats fish and drinks water)

[6] http://nlp.lsi.upc.edu/freeling/

[7] The Eagles group (http://www.ilc.cnr.it/EAGLES96/intro.html) proposed a series of recommendations for the morphosyntactic annotation of corpora. For Spanish, we used the Spanish version (http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html). 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.

It generates the following sequence of Eagle labels:

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

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.

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**[8] **:** 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[9] 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:** Semantic classification of (V)erbs: We search for the semantic classification of verbs. On the basis of what was investigated in [10], 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.

  If the word is a verb we search for the semantic classification of its lemma. 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. For example, let us consider the following sentence:

  She is an incredible friend.

  It has the following sequence of Eagle labels:

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

  The adjective node AQ0CS0 has links both to the positive and negative tags, because *incredible* could be both positive and negative depending on the context.

[8] http://wndomains.fbk.eu/
[9] http://www.illc.uva.nl/EuroWordNet/

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, for texts in English we look for its relationship to one emotion in Wordnet Affect[10] [25] and for texts in Spanish in the Spanish Emotion Lexicon [5]. 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 representation of the previous sentence with our graph-based approach, that will call EmoGraph, 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.

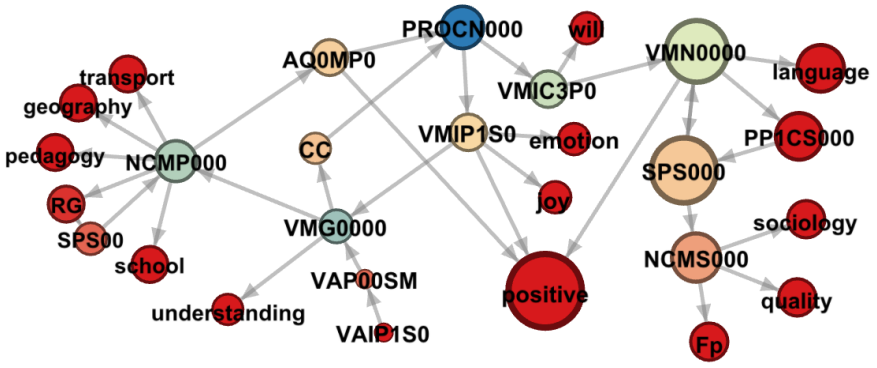

Fig. 2. EmoGraph of "He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público" (*"I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public"*).

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 model texts of gender and age groups in order to be able to classify a given text later into the right class. 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

[10] http://wndomains.fbk.eu/wnaffect.html

describing the overall structure of the modelled texts; *b*) and specific properties of its nodes and how they are related to each other, that describe how authors use language.

Following, we describe how to extract the structure-based features from the graph and what they describe from a discourse-based perspective:

- **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 [1].

- **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*. The clustering coefficient 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 [26]:

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

  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 [9].

- **Average path length** of the graph is the average graph-distance between all the pairs of nodes and could be calculated following [3]. 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.

Moreover, 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 [2]. 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.

2.2 Experiments with PAN-AP-13 and PAN-AP-14 corpora

Below we present the results that have been obtained for gender and age identification with a Support Vector machine with a Gaussian Kernel on the PAN-AP-13 and PAN-AP-14 corpora.

We carried out the experiments with the Spanish partition of the PAN-AP-13 social media corpus. In Table 1 the results for gender identification are shown. The proposed graph-based approach obtained competitive results with respect to the two best performing systems (with no statistically significant difference). In Table 2 EmoGraph shows a better performance than the system that was ranked first at the shared task was obtained for age identification (10s, 20s and 30s), although statistically with no significant difference (t-Student test).

Ranking	Team	Accuracy
1	Santosh	0.6473
2	**EmoGraph**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
6	Flekova	0.6103
7	Jankowska	0.5846
...	...	
17	Baseline	0.5000
...	...	
22	Gillam	0.4784

Table 1: Results in accuracy for gender identification in PAN-AP-13 corpus (Spanish)

Ranking	Team	Accuracy
1	**EmoGraph**	0.6624
2	Pastor	0.6558
3	Santosh	0.6430
4	Haro	0.6219
5	Flekova	0.5966
...	...	
19	Baseline	0.3333
...	...	
21	Mechti	0.0512

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

We studied what topics the different group of authors wrote about in the corpus (we removed the most frequent topics[11] because not so informative being at the top of the domain hierarchy). We obtained the topics with the help of Wordnet Domains. The corresponding word clouds are shown in Figures 3, 4 and 5 for females in each age group (10s, 20s and 30s), and in Figures 6, 7 and 8 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 flats to rent), animals, gastronomy, medicine, and about religion, although in a highest rate among males.

Fig. 3. Top domains for 10s females in PAN-AP-13 corpus

Fig. 4. Top domains for 20s females in PAN-AP-13 corpus

Fig. 5. Top domains for 30s females in PAN-AP-13 corpus

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

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

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

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

With respect to the use of verb types, we were interested in investigating what kind of actions (verbs) females and males mostly refer to and how this changes over time. Figure 9 illustrates that males use more *language* verbs (e.g. tell, say, speak...), whereas females use more *emotional* verbs (e.g. feel, want, love...) conveying more verbal emotions than males.

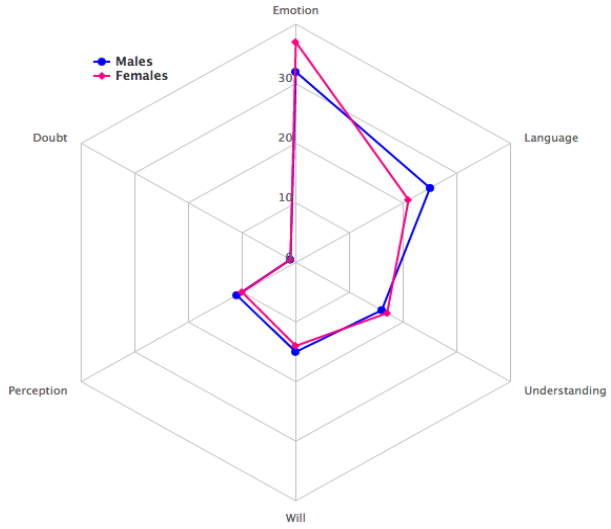

Fig. 9. Use of verb types per gender in PAN-AP-13 corpus

Moreover, we analysed the evolution of the use of verbs over the age. Figures 10 and 11 show the evolution through 10s, 20s and 30s. The use of *emotional* verbs decreases over years, although we can assert that females use more *emotional* verbs than males in any stage of life. The contrary happens with verbs of *language*. Verbs of *understand-*

ing (e.g. know, understand, think...) seem to increase for males and remain 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*[12] (e.g. must, forbid, allow...) increase for both genders, but at a higher rate for males.

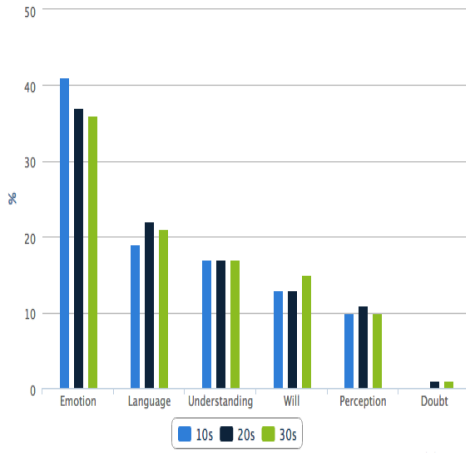

Fig. 10. Evolution in the use of verb types for females in PAN-AP-13 corpus

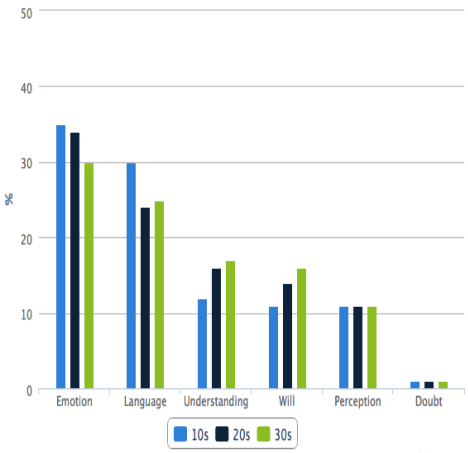

Fig. 11. Evolution in the use of verb types for males in PAN-AP-13 corpus

Finally, we analyse the most discriminative features for the identification of gender and age on the basis of information gain [27]. Table 3 shows the top 20 features over 1100. Betweenness (*BTW-xxx*) and eigenvector (*EIGEN-xxx*) features are among the top features. 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 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 discriminative features for gender identification, whereas the selection of connectors such as prepositions, punctuation marks or interjections are the best discriminative 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.

[12] "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

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 3: Most discriminating features for gender and age identification

Following, we tested further the robustness of the EmoGraph approach on the the PAN-AP-14 corpus, both in Spanish and in English. This corpus is composed of four different genres: *i*) social media (such as in the PAN-AP-13 corpus); *ii*) blogs; *iii*) Twitter; *iv*) and hotel reviews. All corpora were in English and in Spanish, with the exception of the hotel reviews (in English only). In 2014 the age information was labelled in a continuous way (without gaps of 5 years), and the following classes were considered: *i*) 18-24; *ii*) 25-34; *iii*) 35-49; *iv*) 50-64; *v*) and 65+ .

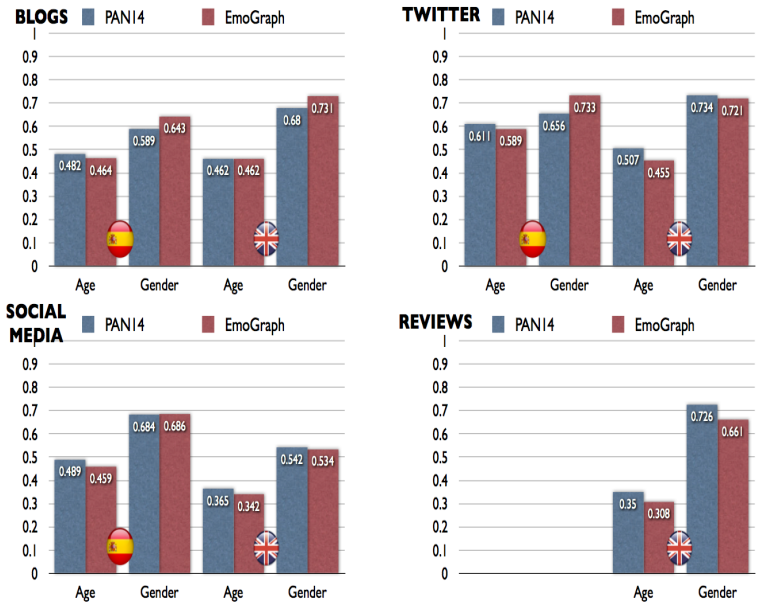

Fig. 12. Results in accuracy in PAN-AP-14 corpus: EmoGraph vs. the best team

Results are shown in Figure 12. Results for Spanish are in general better than for English. This may be due to the higher variety of the morphological information ob-

tained with Freeling for Spanish. In fact, Freeling obtains 247 different annotations for Spanish whereas it obtains 53 for English. For example, in the Spanish version the word "cursos" (courses) for the given example in Figure 2 is returned as NCMP000 where NC means common noun, M means male, P means plural, and 000 is a filling until 7 chars; in the English version, the word "courses" is annotated as NNS.

With respect to the results obtained in the PAN-AP-13 corpus for Spanish, the results for age are lower due to the higher number of classes (3 classes in 2013 vs. 5 continuous ones in 2014). Results for Twitter and blogs are better than for social media and reviews. This is due to the quality of the annotation, in fact both blogs and Twitter corpora were manually annotated, ensuring that the reported gender and age of each author was true. On the contrary, in social media and reviews what the authors reported was assumed to be true. Furthermore, in blogs and also in Twitter there were enough texts per author in order to obtain a better profile. In fact, although in Twitter each tweet is short (as much 140 characters), we had access to hundreds of tweets per author. The worst results were obtained for the reviews. Besides the possibility of deceptive information regarding age and gender in the reviews corpus, it is important to know that reviews were bounded to the hotel domain and just to two kinds of emotions: complain or praise.

3 The Impact of Emotions

In order to understand further the impact of emotions in our graph-based representation of texts, we carried out a further experiment with another corpus, the EmIroGeFB [17] corpus of Facebook comments in Spanish, that we previously annotated with the Ekmans's six basic emotions [6]. We compared the proposed approach where emotions are taken into account with other variations of the graph-based representation that take into acount some of the structure and discourse features:

- **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 complete Eagle labels, but 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.

Features	Accuracy
EmoGraph	0.6596
Semantic Graph	0.5501
Complete Graph	0.5192
Simple Graph	0.5083

Table 4: Results for gender identification in accuracy on the EmIroGeFB corpus (in Spanish)

Results for gender identification are shown in Table 4. The best results were obtained when also emotions that were used in the discourse were considered in the graph-based approach.

4 Conclusions

In this paper we tried to summarise the main concepts that will be addressed in the keynote at the 5th International Conference on Statistical Language and Speech Processing (SLSP) that will be held in Le Mans, France in October 23-25, 2017. Our aim was to show that with a graph-based representation of texts is possible to extract discourse features that describe how complicated the discourse is, how the discourse is modelled in different structural or stylistic units, what are the grammatical categories with the most central use in the discourse of a demographic group, where in the discourse emotion-bearing words have been used, etc. *Eigen* features describing the important nodes in the discourse (e.g. the position in the discourse of words such as nouns, verbs or adjectives, which mainly give the meaning of the sentence) showed to help in gender identification, whereas *betweenness* features describing the most common links in the discourse (e.g. connectors such as prepositions, punctuation marks or interjections) helped more in age identification.

A more complete description of the EmoGraph graph-based approach and the experiments carried out on the PAN-AP-13 and PAN-AP-14 can be found in [21] and in [20].

ACKNOWLEDGEMENTS

We thank the SLSP Conference for the invitation for giving the keynote on Author Profiling in Social Media. The research work described in this paper was partially carried out in the framework of the SomEMBED project (TIN2015-71147-C2-1-P), funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO).

References

1. 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)
2. Bonacich, P. Factoring and Weighting Approaches to Clique Identification. In: Journal of Mathematical Sociology 2 (1), pp. 113-120 (1972)
3. Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical Sociology 25(2), pp. 163-177 (2001)
4. Carreras, X., Chao, I., Padró, L., Padró, M. FreeLing: An Open-Source Suite of Language Analyzers In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), 2004
5. Díaz Rangel, I. Sidorov, G., Suárez-Guerra, S.: Creación y evaluación de un diccionario marcado con emociones y ponderado para el español.Onomazein, 29, p. 23 (2014) (in Spanish)
6. Ekman, P.: Universals and cultural differences in facial expressions of emotion. Symposium on Motivation, Nebraska, pp. 207-283 (1972)

7. 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)
8. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literay and Linguistic Computing 17 (4), pp. 401-412 (2003)
9. 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)
10. Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993)
11. Mann, W. C. and Thompson, S. A. Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243–281 (1988).
12. 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. [7]
13. Padró, L., Stanilovsky, E. FreeLing 3.0: Towards Wider Multilinguality In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), 2012
14. 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. [7]
15. 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)
16. Pennebaker, J.W. The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press. (2011)
17. 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, Reykjavik, Iceland, May 26-31. pp. 68-73 (2014)
18. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In Forner et al. [7]
19. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Cappellato L., Ferro N., Halvey M., Kraaij, W.(Eds.), Notebook Papers of CLEF 2014 LABs and Workshops. CEUR-WS.org, vol. 1180, pp. 951-957 (2014)
20. Rangel, F., Rosso, P.: On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media. In: 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction, CLEF 2015, Springer-Verlag, LNCS(9283), pp. 274-280 (2015)
21. Rangel, F., Rosso, P. On the Impact of Emotions on Author Profiling In: Information Processing & Management, 52(1): 73-92 (2016)
22. 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)
23. Soler-Company, J. Wanner, L. Use of Discourse and Syntactic Features for Gender Identification. In The Eighth Starting Artificial Intelligence Research Symposium. Collocated with the 22nd European Conference on Artificial Intelligence, pp. 215–220 (2016)
24. Soler-Company, J. Wanner, L. On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In 15th Conference of the European Chapter of the Association for ComputationalLinguistics (EACL), pp. 681–687. valencia, Spain (2017)
25. 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)

26. Watts, D.J., Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature 393 (6684): pp. 409-410 (1998)
27. 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)