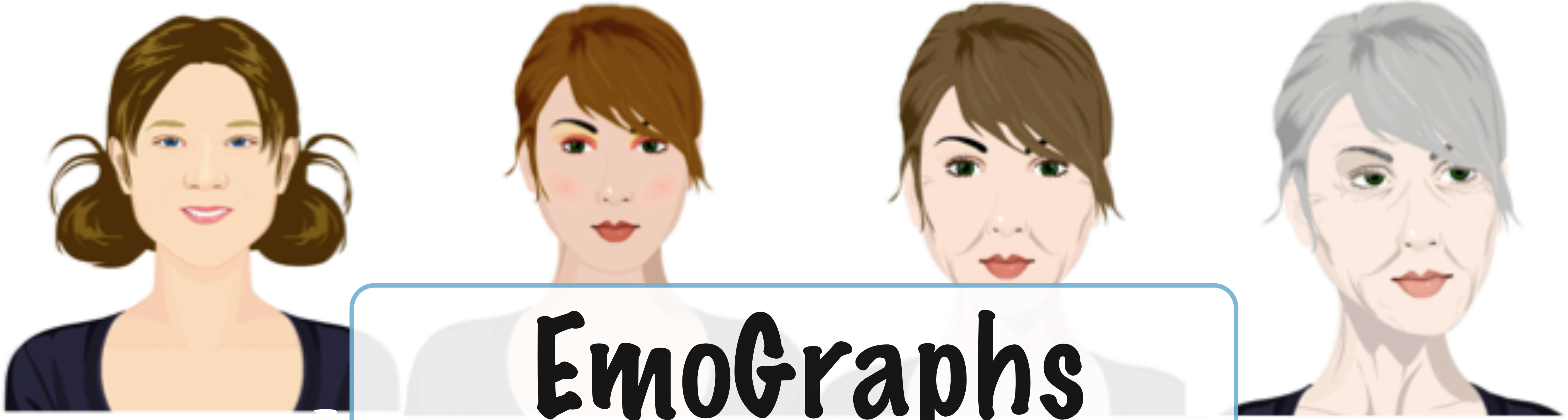


EmoGraphs

for Age and Gender Identification

Francisco Rangel

autoritas
nuevas ideas, nuevas soluciones

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

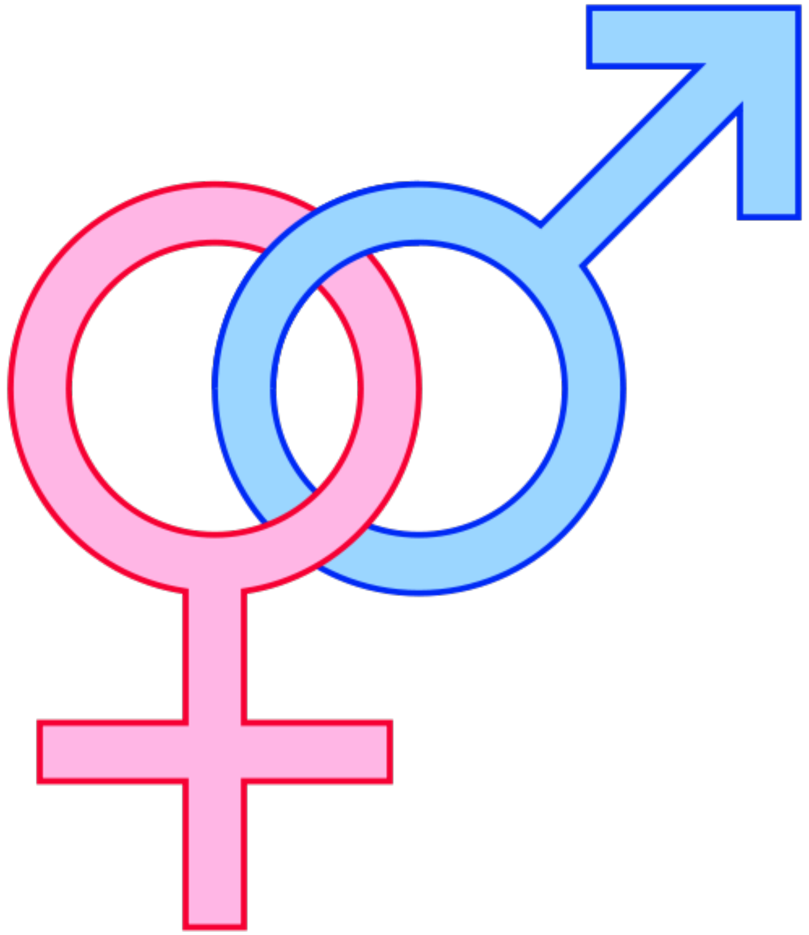
Introduction to Author Profiling

- Author profiling use sociolect aspects to distinguish among classes of authors [1]. E.g.
  - Age, gender, native language, emotional profile, personality type...
- Author profiling is important in:
  - Forensics
  - Security
  - Marketing

[1] Pennebaker, J.W.: The secret life of pronouns: What our words say about us. Bloomsbury Press (2011)

Research Aim

- Our aim is at investigating how people use the language, and especially how they convey verbal emotions, to determine their age and gender



Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

Related Work

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Holmes & Meyerhoff, 2003	Formal texts	-	Age and gender	
Burger & Henderson, 2006	Blogs	Posts length, capital letters, punctuations. HTML features.	They only reported: "Low percentage errors"	Two age classes: [0,18[,[18,-]
Koppel et al., 2003	Blogs	Simple lexical and syntactic functions	Gender: 80% accuracy	Self-labeling
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 80% accuracy Age: 75% accuracy	
Goswami et al., 2009	Blogs	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/ sentence length, POS, word factor analysis	Gender: 72,10 accuracy	
Nguyen et al., 2011 y 2013	Blogs & Twitter	Unigrams, POS, LIWC	Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable
Peersman et al., 2011	Netlog	Unigrams, bigrams, trigrams and tetagrams	Gender+Age: 88.8 accuracy	Self-labeling, min 16 plus 16,18,25

PAN task at CLEF (http://pan.webis.de)

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
PAN 2013 [1]	Social Media	Style-based features (frequencies, readability, POS...) Content-based features (LDA, topics, BOW...) n-grams, language models Collocations IR Features Second Order Representations	Gender: ~64% accuracy Age: ~64% accuracy	English & Spanish Age, Gender
PAN 2014 [2]	Social Media, Blogs, Twitter, Reviews		Gender: ~72% accuracy Age: ~61% accuracy	English & Spanish Age, Gender
PAN 2015	Twitter	-	-	English, Spanish, Italian & Dutch Age, Gender, Personality Traits

[1] Rangel,F.,Rosso,P.,Koppel,M.,Stamatatos,E.,Inches,G.:Overviewoftheauthorprofiling task at pan 2013. In: In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179 (2013)

[2] 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: In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180 (2014)

Outline

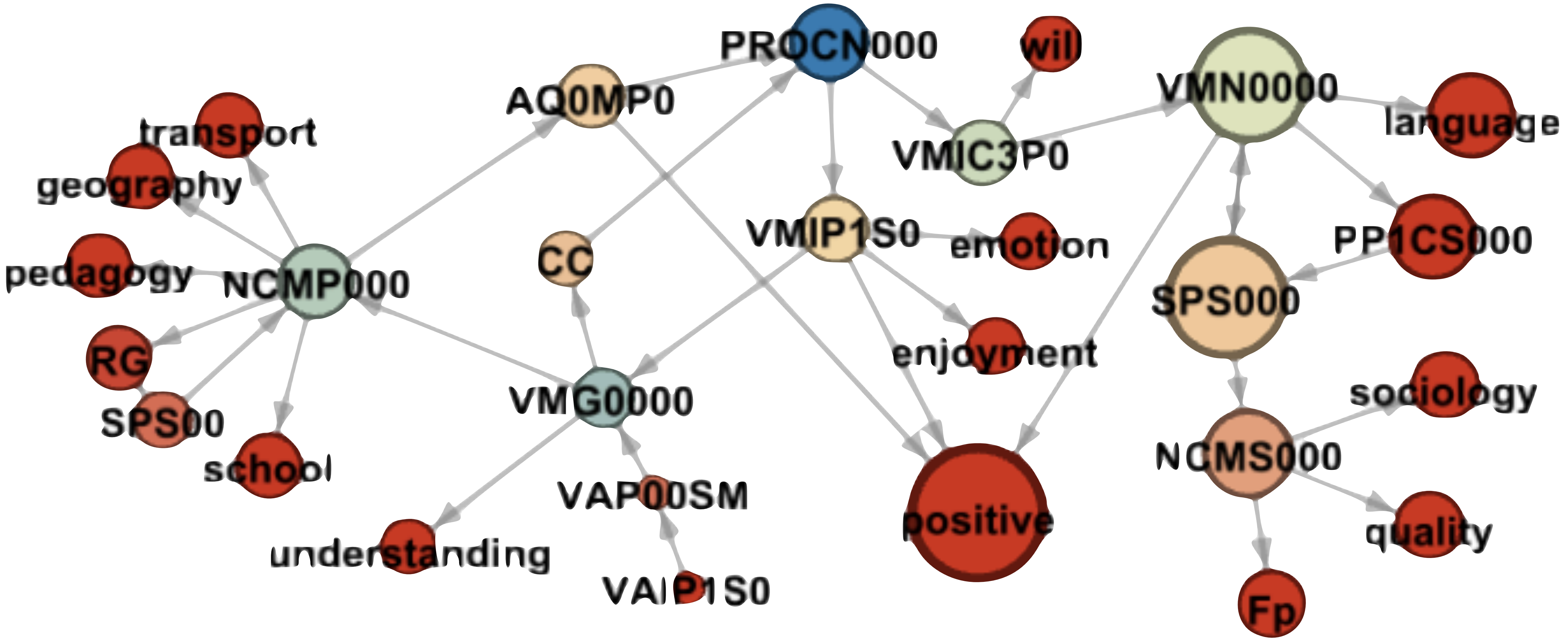
- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

Style-based Features

PART-OF-SPEECH (GRAMMATICAL CATEGORIES)	Frequency of use of each grammatical category, number and person of verbs and pronouns, mode of verb, proper nouns (NER) and non-dictionary words (words not found in dictionary);
FREQUENCIES	Ratio between number of unique words and total number of words, words starting with capital letter, words completely in capital letters, length of the words, number of capital letters and number of words with flooded characters (e.g. Heeeelloooo);
PUNCTUATION MARKS	Frequency of use of dots, commas, colon, semicolon, exclamations, question marks and quotes;
EMOTICONS	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, surprised, derision and dumb;
SPANISH EMOTION LEXICON (SEL)	We obtained the lemma for each word and then its *Probability Factor of Affective Use* value from the SEL dictionary. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion, building one feature per emotion.

IMPORTANT NOTE: NONE OF THE FEATURES IS TOPIC DEPENDENT

EmoGraph



“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”

Resources

Freeling	http://nlp.lsi.upc.edu/freeling/
WordNet Domains (+EuroWordnet)	http://wndomains.fbk.eu/ http://www.illc.uva.nl/EuroWordNet/
Semantic Classification of Verbs	Levin, B. English Verb Classes and Alternations. University of Chicago Press, Chicago. (1993) 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...)
Polarity Lexicon	Hu, M., Liu, B. Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Seattle, Wash- ington, USA, pp. 168-177 (2004)
Spanish Emotion Lexicon	Sidorov,G.,Miranda,S.,Viveros,F.,Gelbukh,A.,Castro,N.,Velásquez,F.,Díaz,I.,Su árez, S., Treviño, A., Gordon, J.: Empirical Study of Opinion Mining in Spanish Tweets. 11th Mex- ican International Conference on Artificial Intelligence, MICAI, pp. 1-14 (2012)

Graph-based Features

Given a graph G={N,E} where:

- N is the set of nodes
- E is the set of edges

we obtain a set of:

- graph-based features from global measures of the graph
- node-based features from node specific measures

Graph-based Features

Nodes-edges ratio	It gives an indicator of how connected the graph is. In our case, how complicated the discourse is.	Theoretical maximum: $max(E) = N * (N - 1)$
Average degree Weighted average degree	It gives an indicator on how much interconnected the graph is. In our case, how much interconnected the grammatical categories are.	Averaging all nodes degrees. Scaling it to [0,1]
Diameter	It indicates the greatest distance between any pair of nodes. In our case, how far a grammatical category is from others, or how far a topic is from an emotion.	$d = max_{n \in N} \varepsilon(N)$ where E(N) is the eccentricity
Density	It indicates how close the graph is to be completed. In our case, how dense is the text in the sense of how each grammatical category is used in combination to others.	$D = \frac{2*\|E\|}{(\|N\|*(\|N\|-1))}$
Modularity	It indicates different divisions of the graph into modules. One node has dense connections within the module and sparse with nodes in other modules. In our case, it may indicate how the discourse is modelled in different structural or stylistic units.	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)
Clustering coefficient	It indicates the transitivity of the graph. If a is directlyy linked to b and b is directly linked to c, what's the probability that a is directly linked to c. In our case, how different grammatical categories or semantic information is related to each others	Watts-Strogatzt: $cc1 = \frac{\sum_{i=1}^{n} C(i)}{n}$
Average path length	It indicates how far some nodes are from others. In our case, how far some grammatical categories are from others, or for example how far some topics are from some emotions	Brandes, U. A Faster Algorithm for Betweenness Centrality. In: Journal of Mathematical So- ciology 25(2), pp. 163-177 (2001)

Node-based Features

EigenVector	It gives a measure of the influence of each node. In our case, it may give what are the grammatical categories with the most central use in the author's discourse, for example, which nouns, verbs or adjectives	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: $x_n = \frac{1}{\lambda} \sum_{t \in M(n)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{n,t} x_t$ where λ is a constant representing the greatest eigenvalue associated with the centrality measure.
Betweenness	It gives a measure ofthe importance of a each node depending on the number of shortest paths of which it is part of. 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 and nouns. Hence, this measure may give us an indicator of what the most common connectors in the linguistic structures used by authors	It is the ratio of all shortest paths from one node to another node in the graph that pass through x: $BC(x) = \sum_{i,j \in N - \{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$ Where $\sigma_{i,j}$ is the total number of shortest paths from node i to j, and $\sigma_{i,j}(n)$ is the total number of those paths that pass through n.

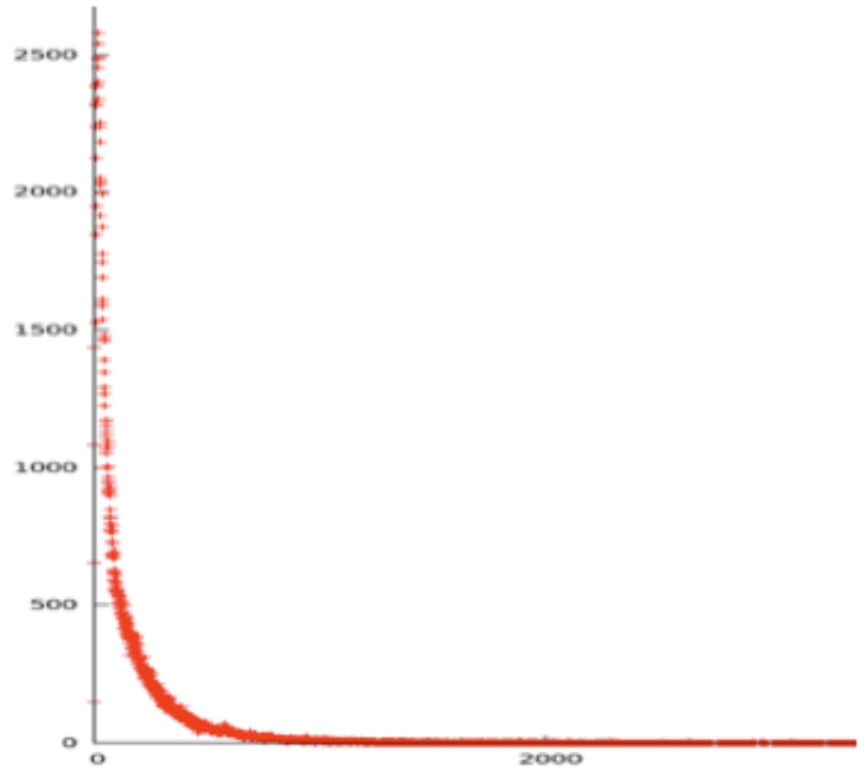
Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

PAN-AP-2013 Corpus (Spanish)

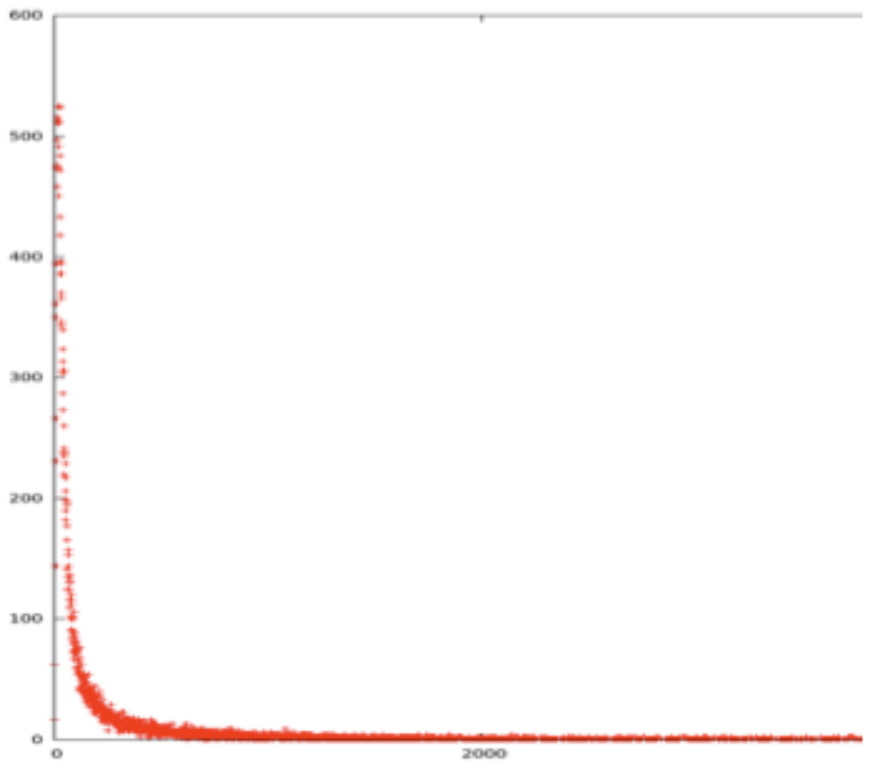
Age	Gender	No. of Authors: Training	No. of Authors: Test
10s	male	1 250	144
	female	1 250	144
20s	male	21 300	2 304
	female	21 300	2 304
30s	male	15 400	1 632
	female	15 400	1 632
Σ		75 900	8 160

Training



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

Test



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

Machine Learning Approach

Weka toolkit

Gender Identification	Support Vector Machine Gaussian Kernel g=0.20 c=1
Age Identification	Support Vector Machine Gaussian Kernel g=0.08 c=1

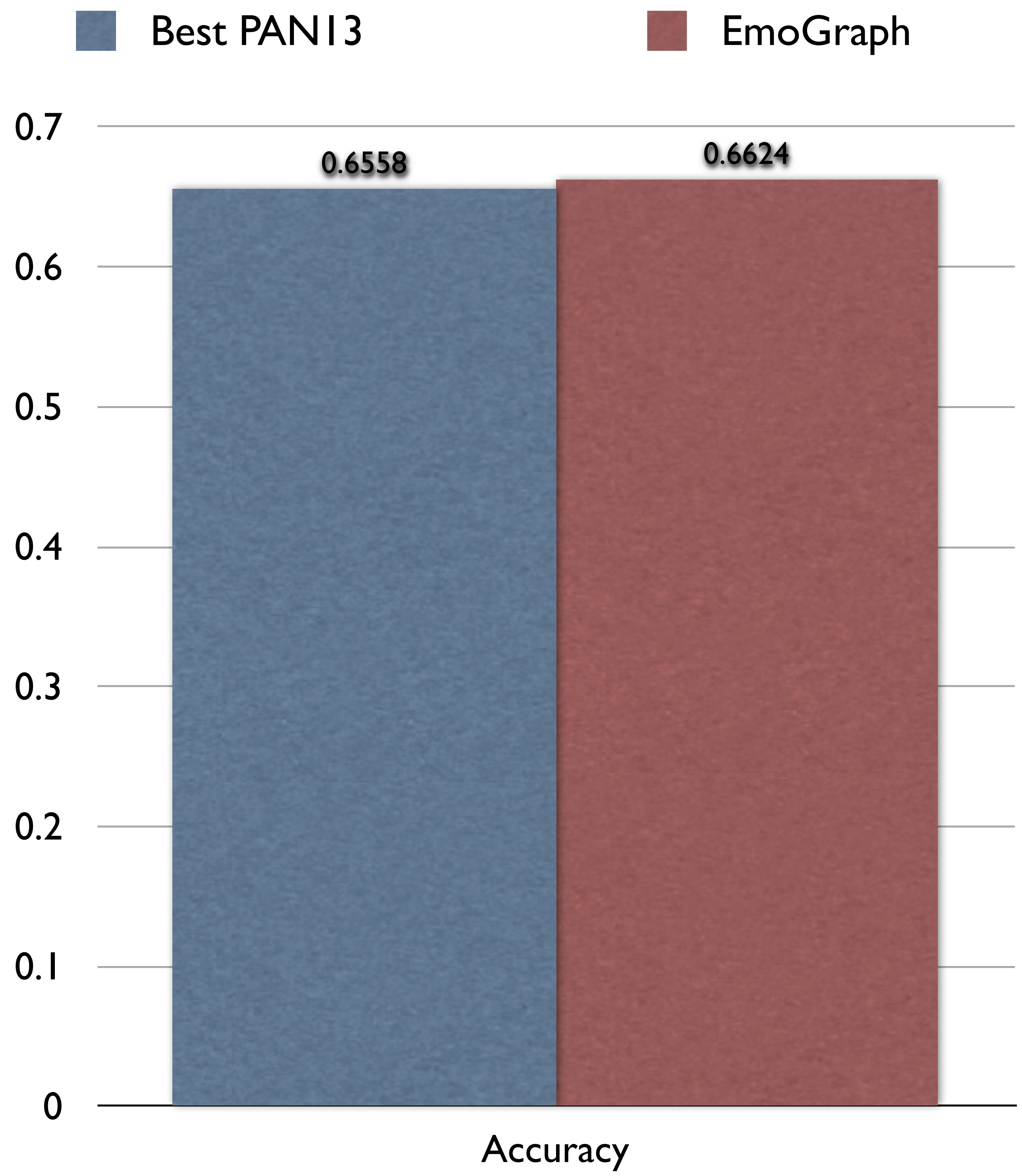
Evaluation measure:
Accuracy

t-Student
$H_0: p_1=p_2$

Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

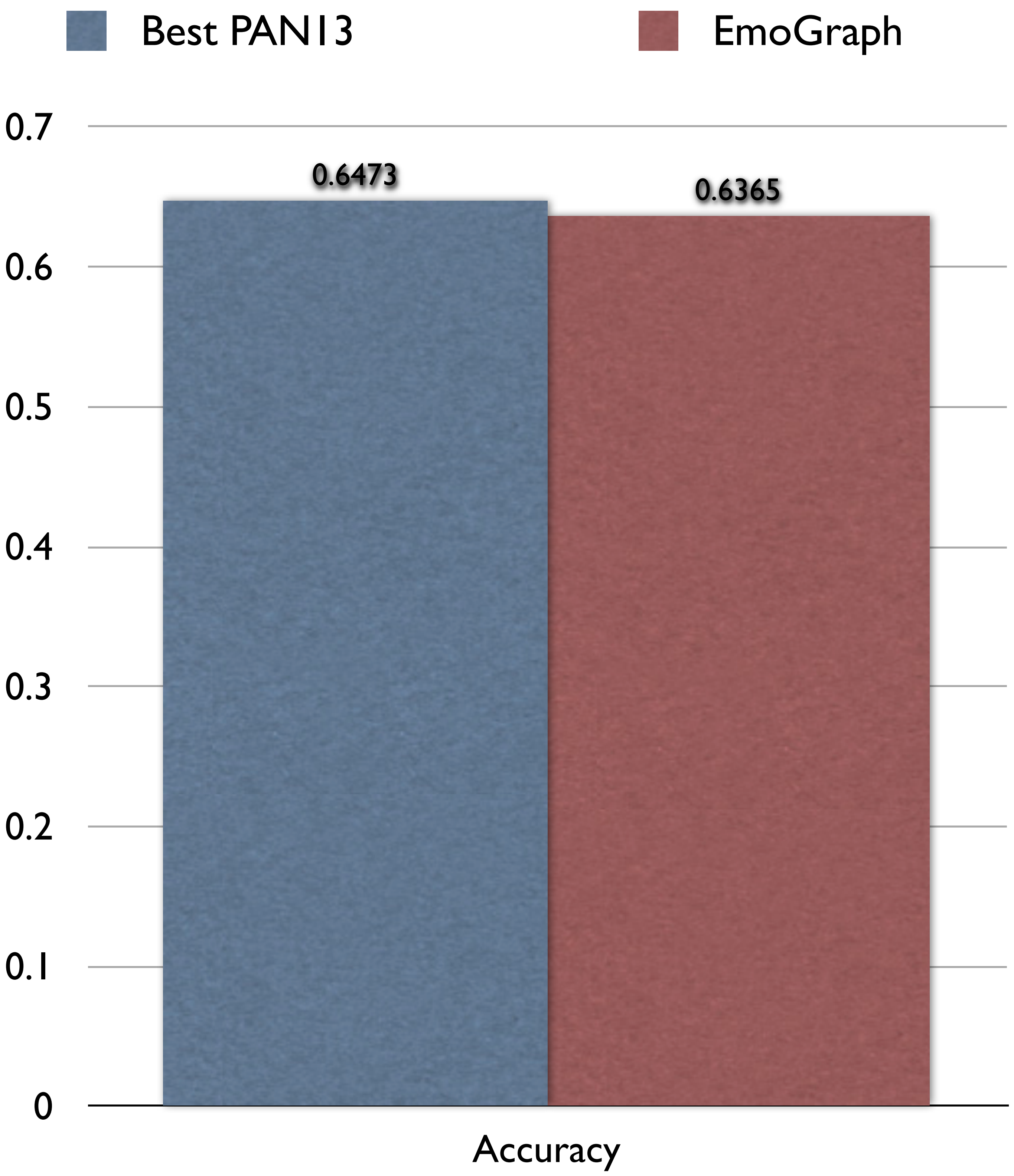
Age Identification



Ranking	Team	Accuracy
1	**Rangel-EG**	0.6624
2	Pastor	0.6558
3	Santosh	0.6430
4	**Rangel-S**	0.6350
5	Haro	0.6219
6	**Rangel-nG**	0.6162
7	Flekova	0.5966
...	...	
21	Baseline	0.3333
...	...	
23	Mechti	0.0512

$(z_{0.05} = 0.8894 < 1.960)$

Gender Identification



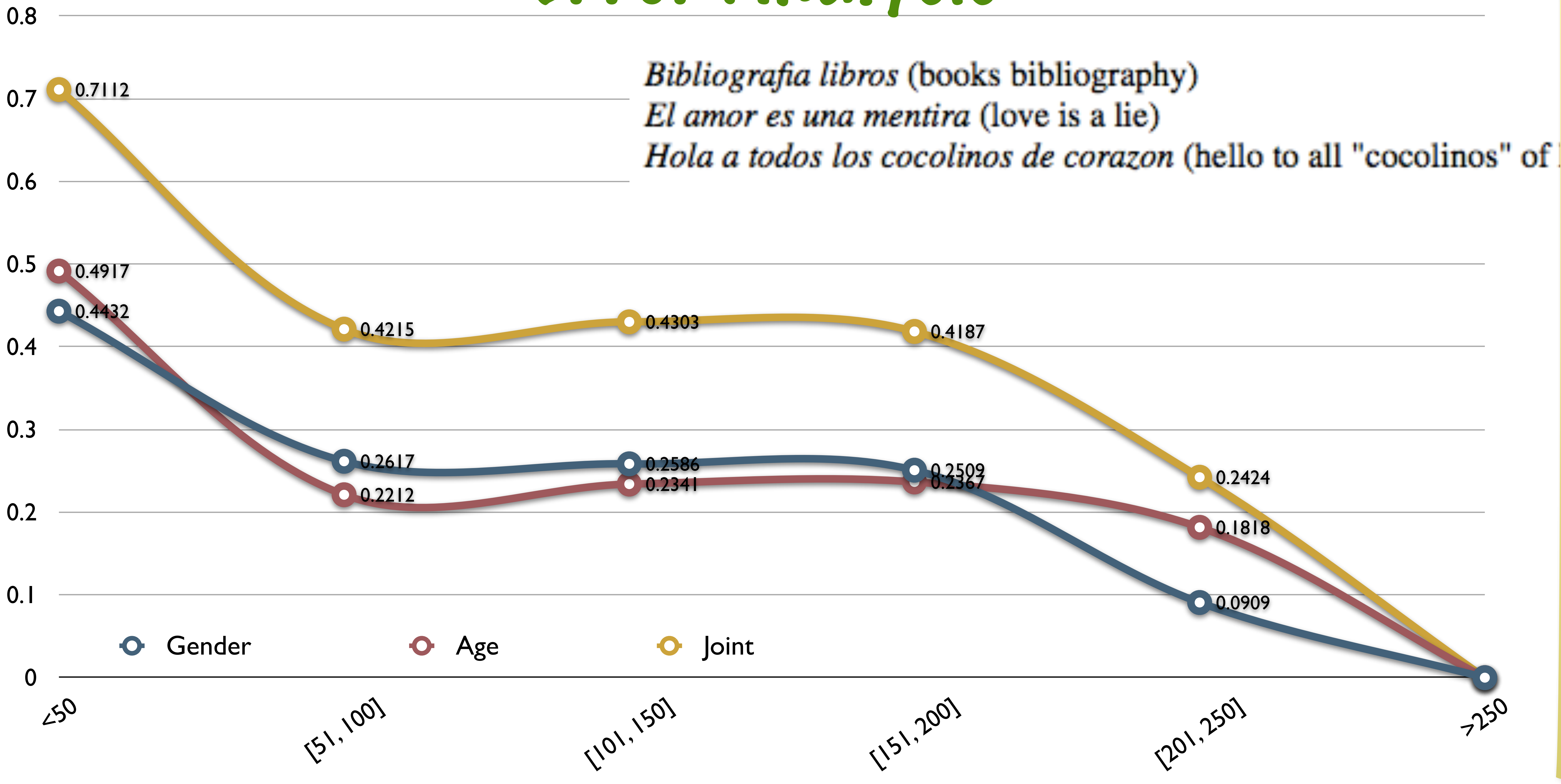
Ranking	Team	Accuracy
1	Santosh	0.6473
2	**Rangel-EG**	0.6365
3	Pastor	0.6299
4	Haro	0.6165
5	Ladra	0.6138
6	Flekova	0.6103
7	**Rangel-nG**	0.6016
8	Jankowska	0.5846
9	**Rangel-S**	0.5713
...	...	
19	Baseline	0.5000
...	...	
24	Gillam	0.4784

$(z_{0.05} = 1.4389 < 1.960)$

Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

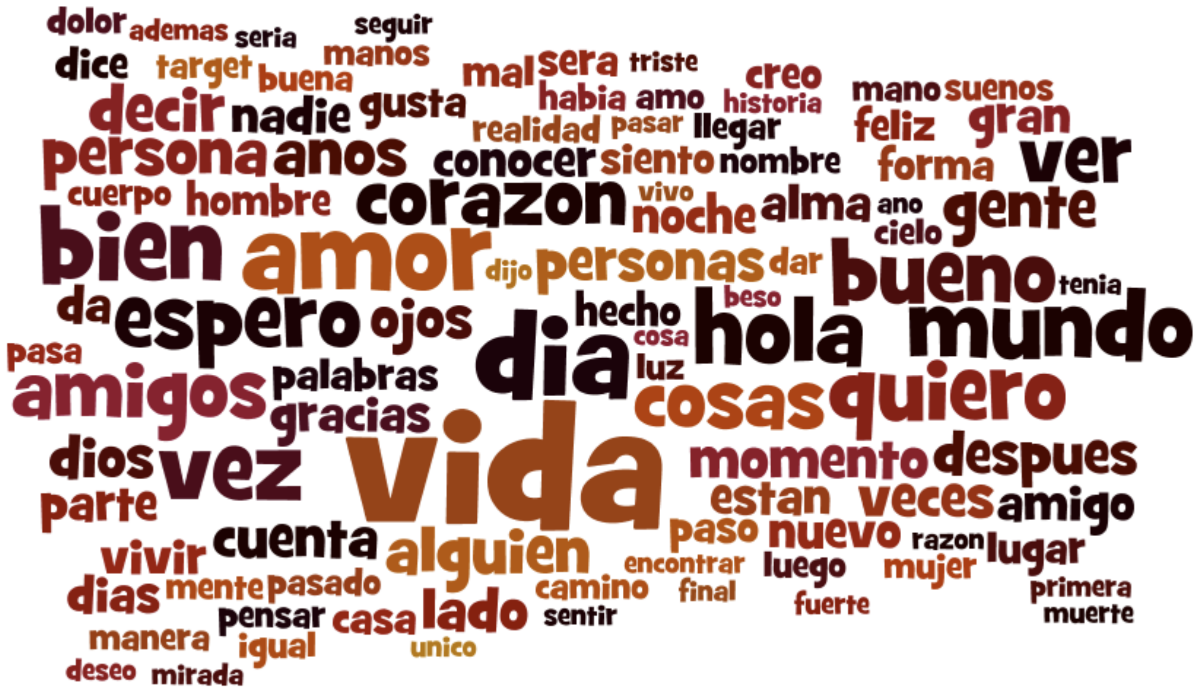
Error Analysis



Topics per Gender

Females

Males



- No significative differences between genders
- No matter the gender, people seem to worry about life (vida), love (amor), want (quiero) and hope (espero)

Evolution of Topics per Age

Females 10s

Females 20s

Females 30s

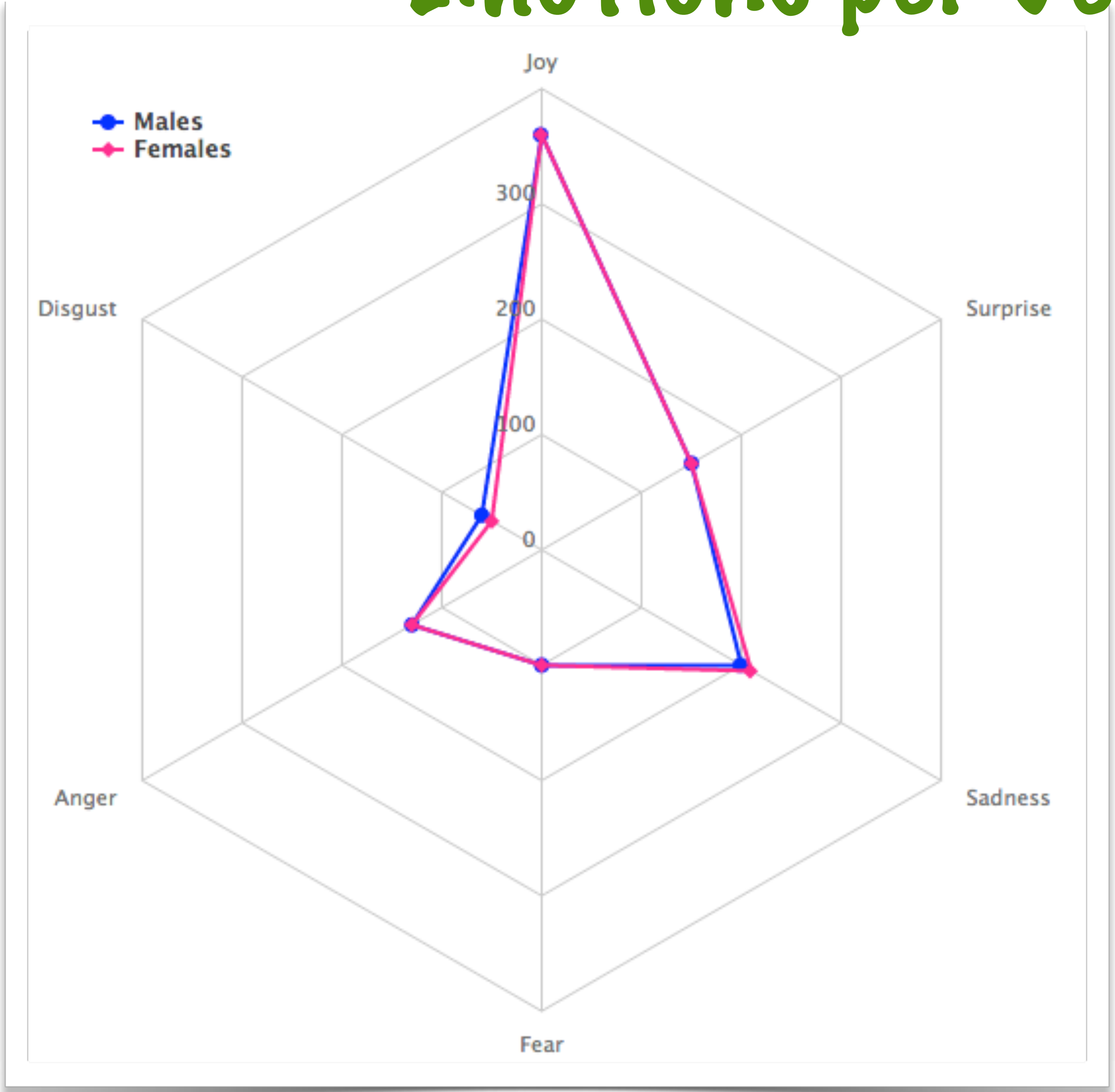
Males 10s

Males 20s

Males 30s

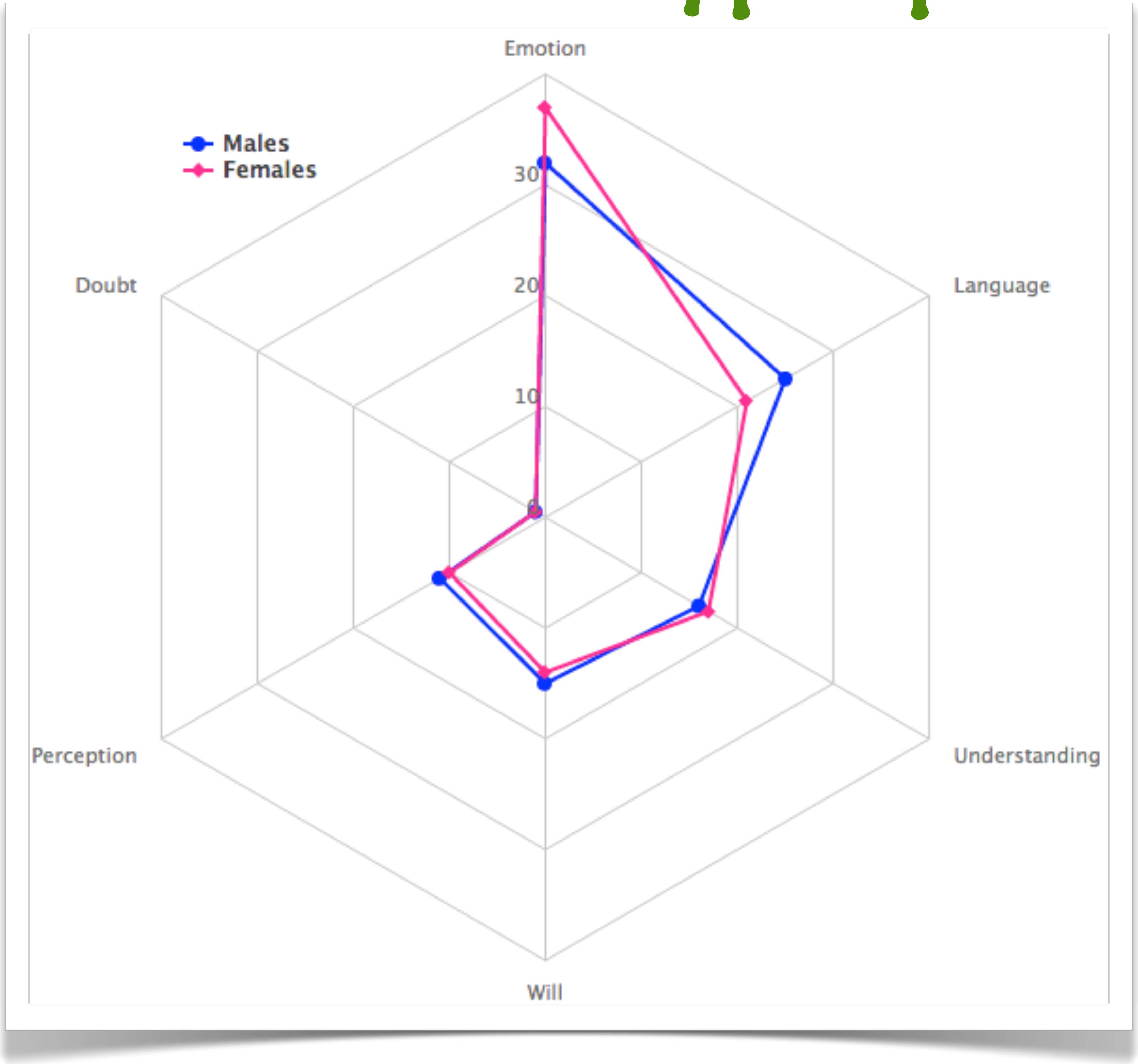
- Younger people tend to write more about disciplines such as: (males) physics, law... (females) chemistry, linguistics...
- 10s females talk more about sexuality whereas 10s males talk about shopping
- As they grow both males and females are interested in buildings, animals, gastronomy, medicine or religion

Emotions per Gender



- No significative differences between genders
- Females seem to express more disgust than males
- Males seem to express more sadness

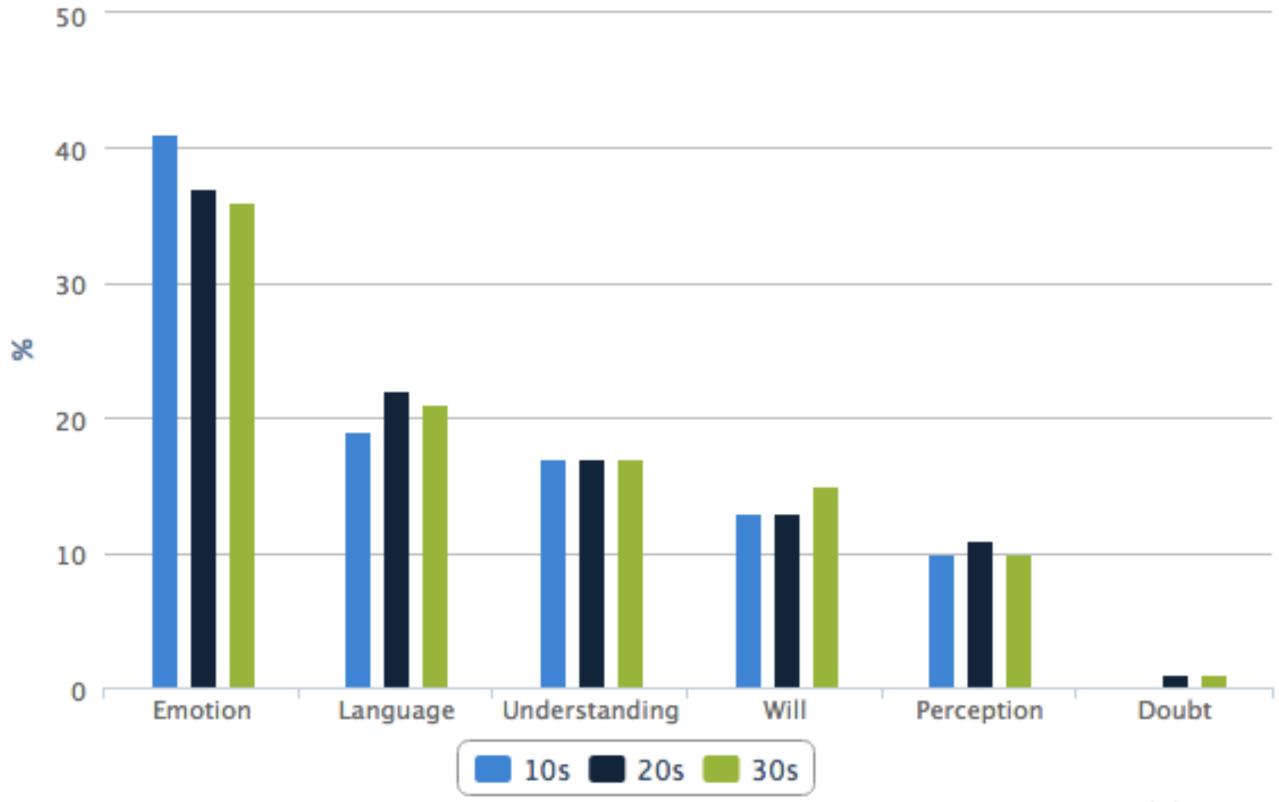
Verb Types per Gender



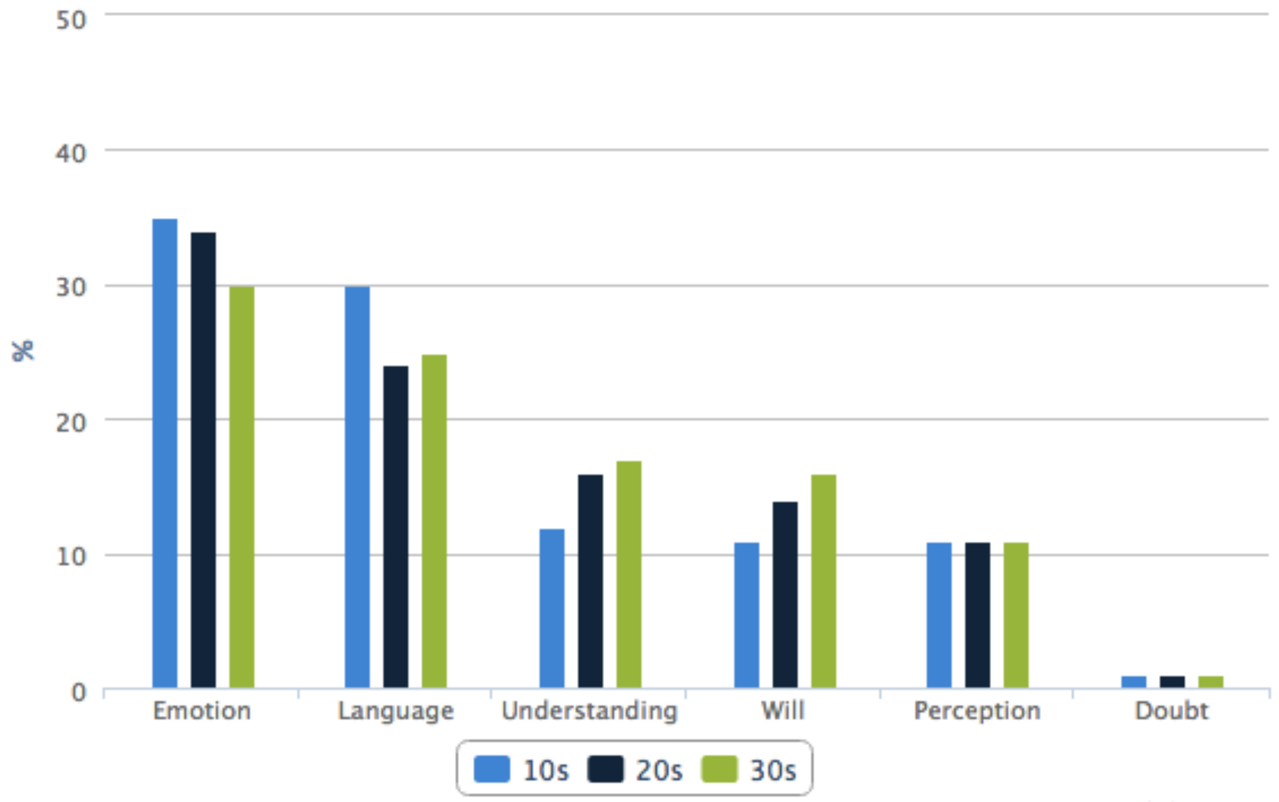
- Females use more emotional verbs (feel, want, love...)
- Males use more language verbs (tell, say, speak...)

Evolution of Verb Types per Gender

Females



Males



- The use of emotional verbs decreases over years
- Females start using verbs of understanding at higher rate than males
- Verbs of understanding seems to increase for males and remains stable for females
- Verbs of will increases for both genders, but more for males
- Females use emotional verbs more than males in any stage of life vs. males use more verbs of language

Most Discriminating Features

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

- Eigen features in gender vs. betweenness in age
- Verbs, nouns and adjectives in gender vs. prepositions and punctuation marks in age
- Higher presence of emotion-based features in gender identification

Outline

- Related work
- Representation models
- Experimental setup
- Experimental results
- Analysis
- Conclusions and future work

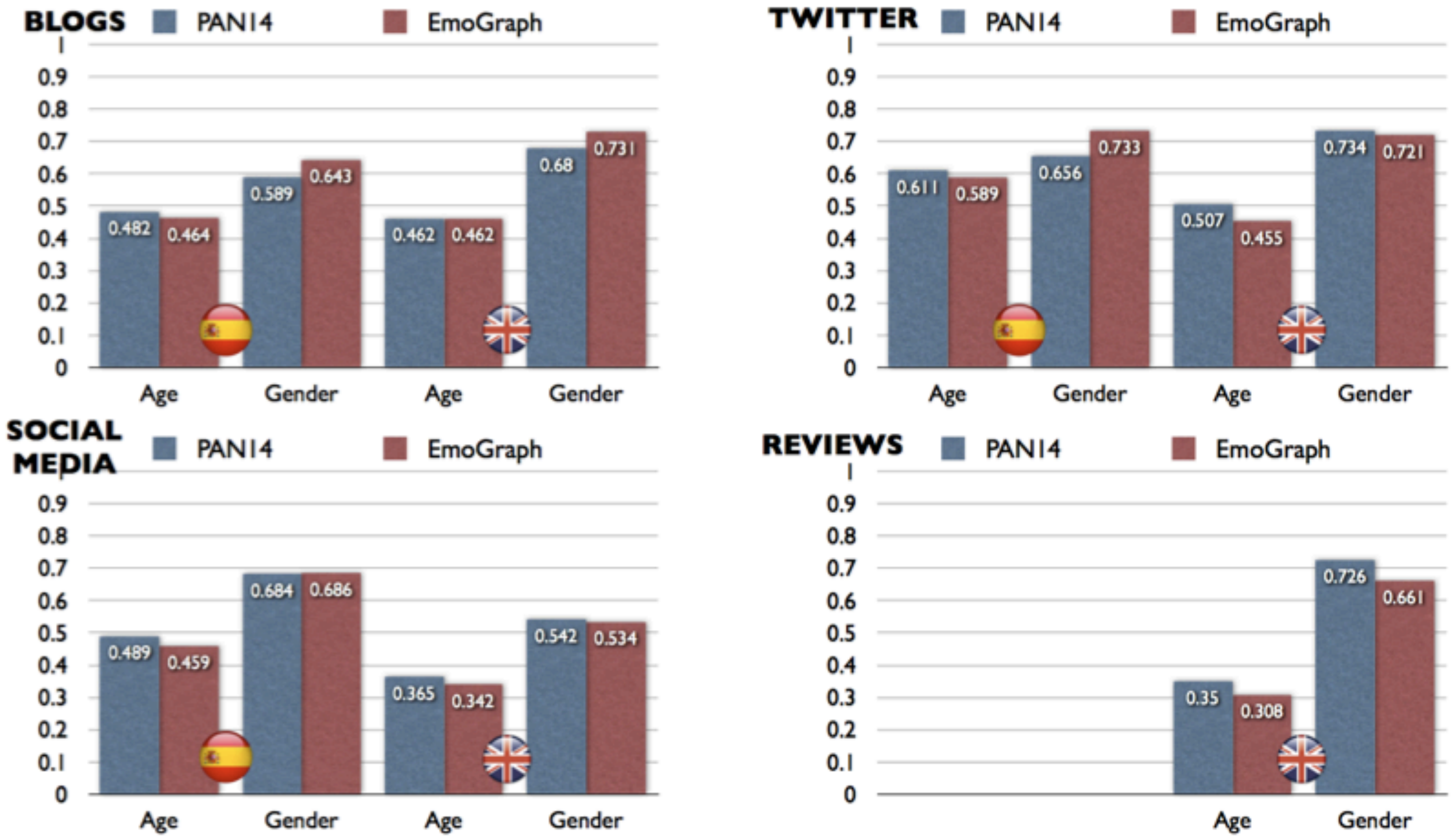
Conclusions

- We investigated the impact of emotions on gender and age identification
- An emotion-labeled graph (EmoGraph) has been proposed
- Results are competitive with the state-of-the-art
- The most discriminating features show the importance of emotions and graph-based model
- Some conclusions were drawn with respect to the use of the language depending age and gender

Future Work

- To test the robustness of the method in:
  - Another language: English
  - Another corpora: PAN-AP14
- To apply the method to other author profiling tasks:
  - Language Variety Identification
- To combine the method with probabilistic features

Future Present...



Accuracies of the best PAN14 team vs. EmoGraph on different languages and genres.