UNIVERSITAT ROVIRA i VIRGILI

Aix*Marseille université

A Fuzzy Sociolinguistic Model for Gender Prediction in Spanish Social Network Texts

PhD

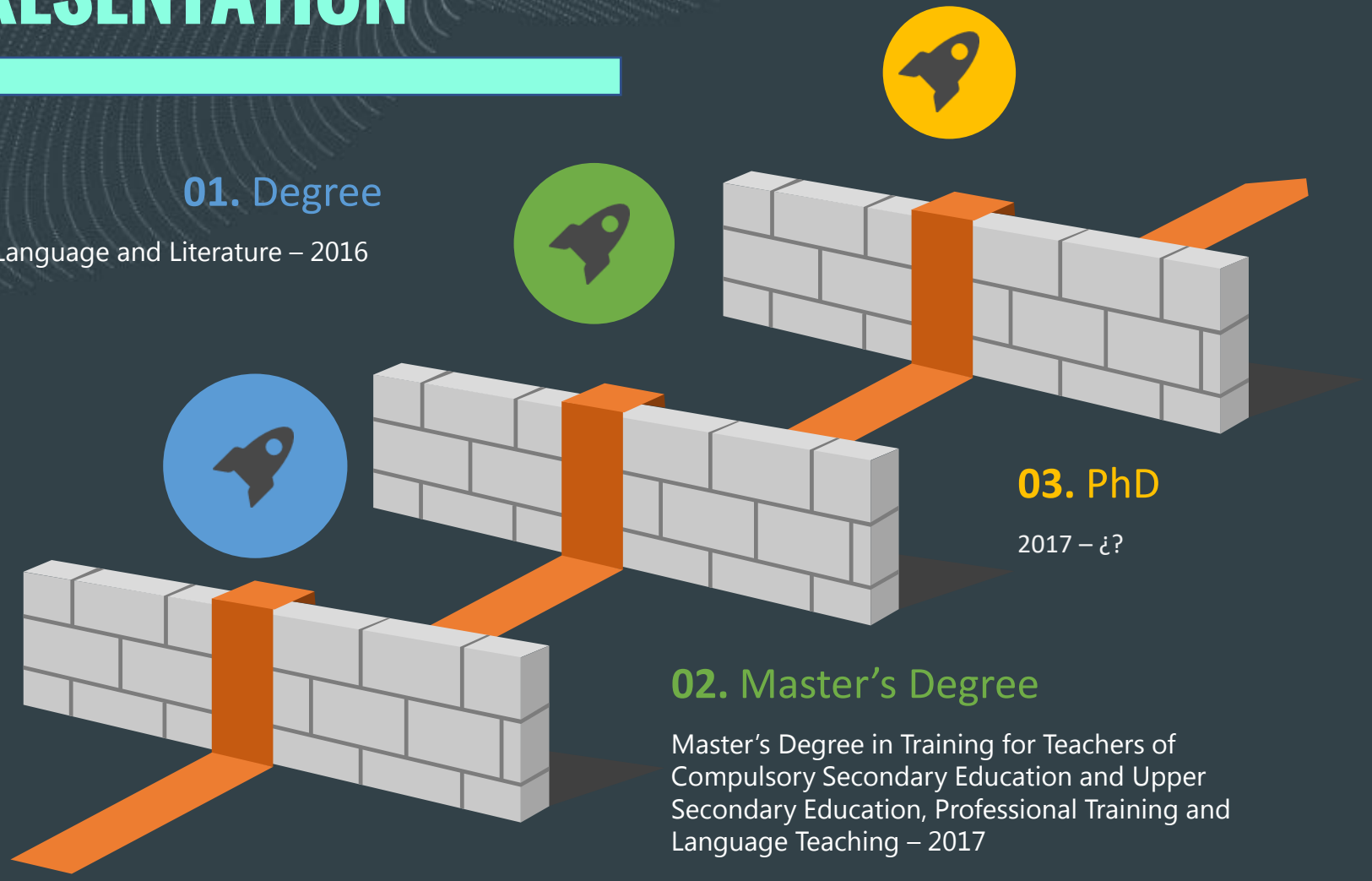
Author: Morales Sánchez, Damián
Supervisor: Jiménez López, María Dolores
Co-supervisor: Moreno Ribas, Antonio

July 2021

SELF-PRESENTATION

01. Degree

Degree in Spanish Language and Literature – 2016



02. Master's Degree

Master's Degree in Training for Teachers of Compulsory Secondary Education and Upper Secondary Education, Professional Training and Language Teaching – 2017

03. PhD

2017 – ¿?

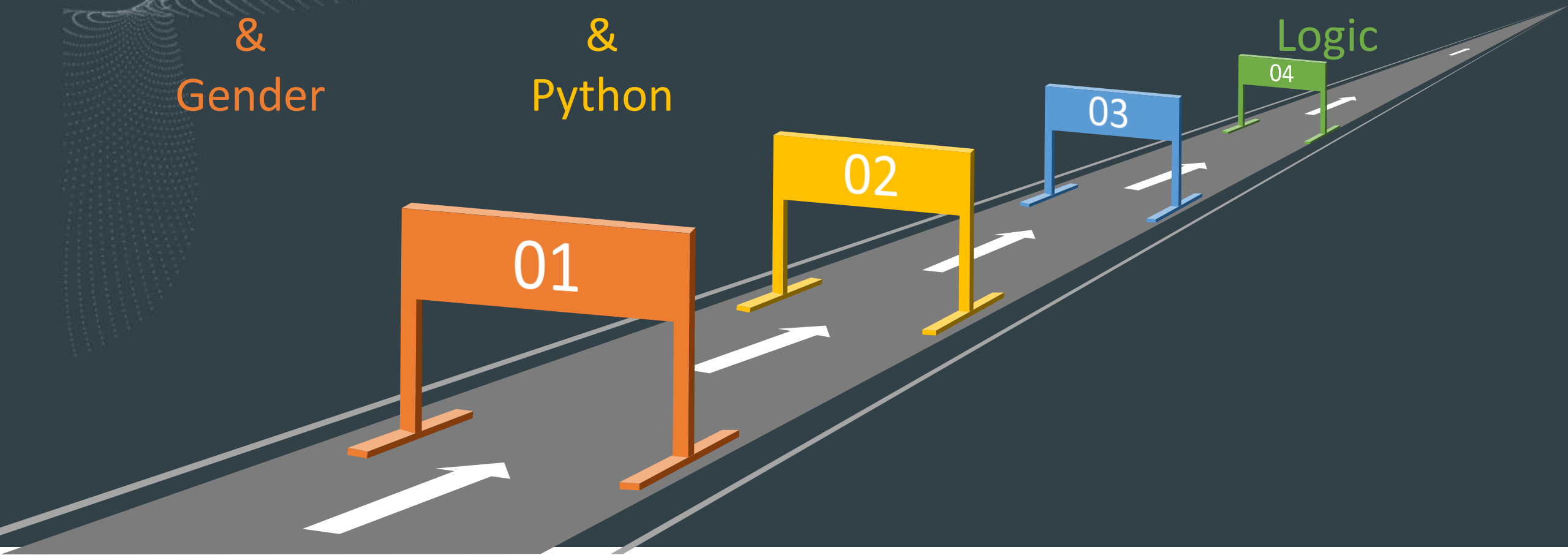
SELF-PRESENTATION

2017-18
Sociolinguistics
&
Gender

2018-19
NLP
&
Python

2020-21
ML

2021-22
DL & Fuzzy
Logic



SELF-PRESENTATION

01

02

03

04

05

TABLE OF CONTENTS

INTRODUCTION

TABLE OF CONTENTS

THEORETICAL FRAMEWORK

TABLE OF CONTENTS

COMPUTATIONAL ANALYSIS

01 Orthographic

02 Morphological

03 Lexical

04 Syntactic

- Hypotheses
- Features
- Pre-processing Techniques
- Exploratory Data Analysis
- Supervised Classification Model
  - Decision Tree
  - Random Forest
- Gender Formalization: Fuzzy Model
- Hypotheses Evaluation and Sociolinguistic Discussion

TABLE OF CONTENTS

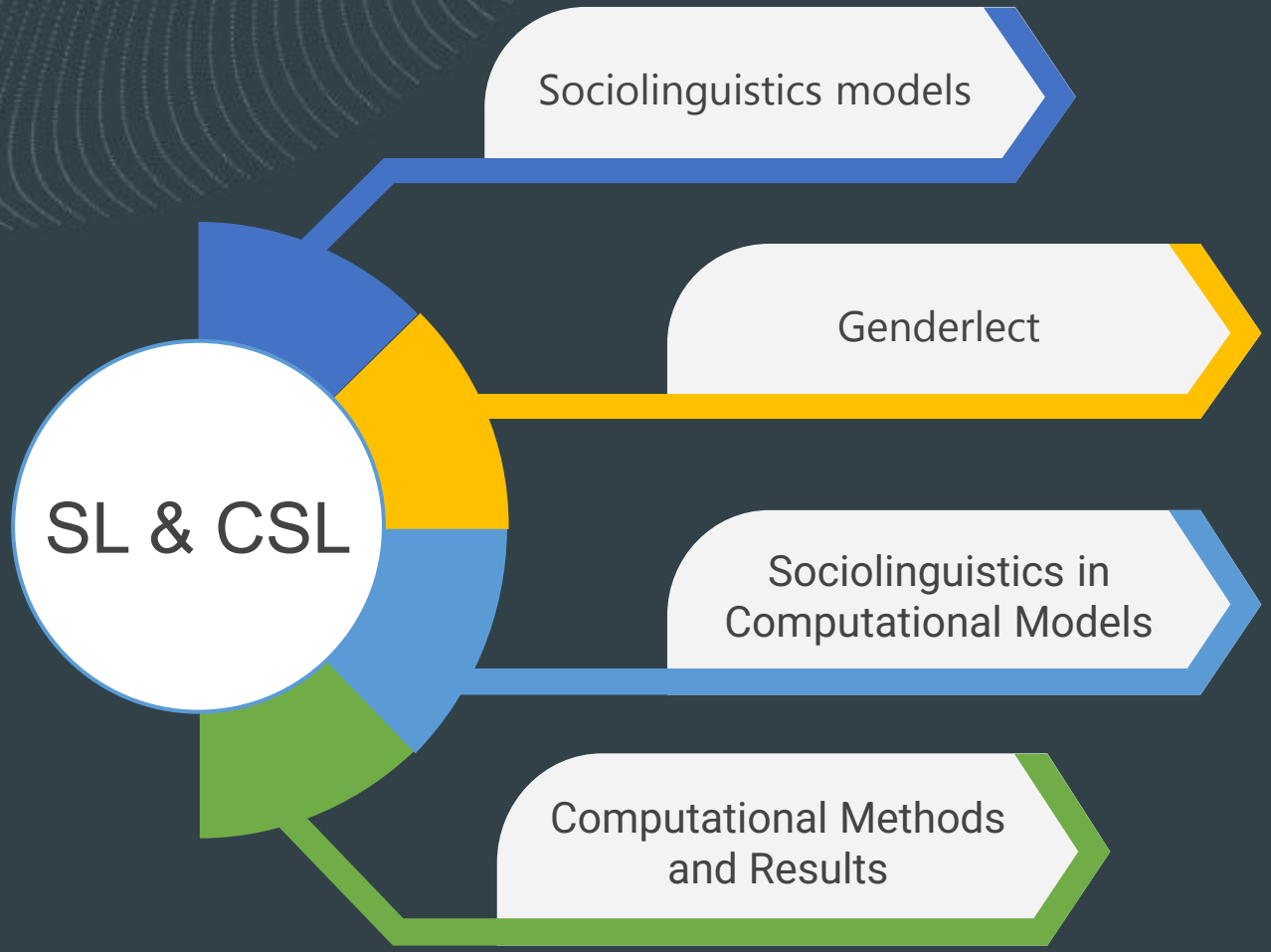
COMPUTATIONAL ANALYSIS

INTRODUCTION

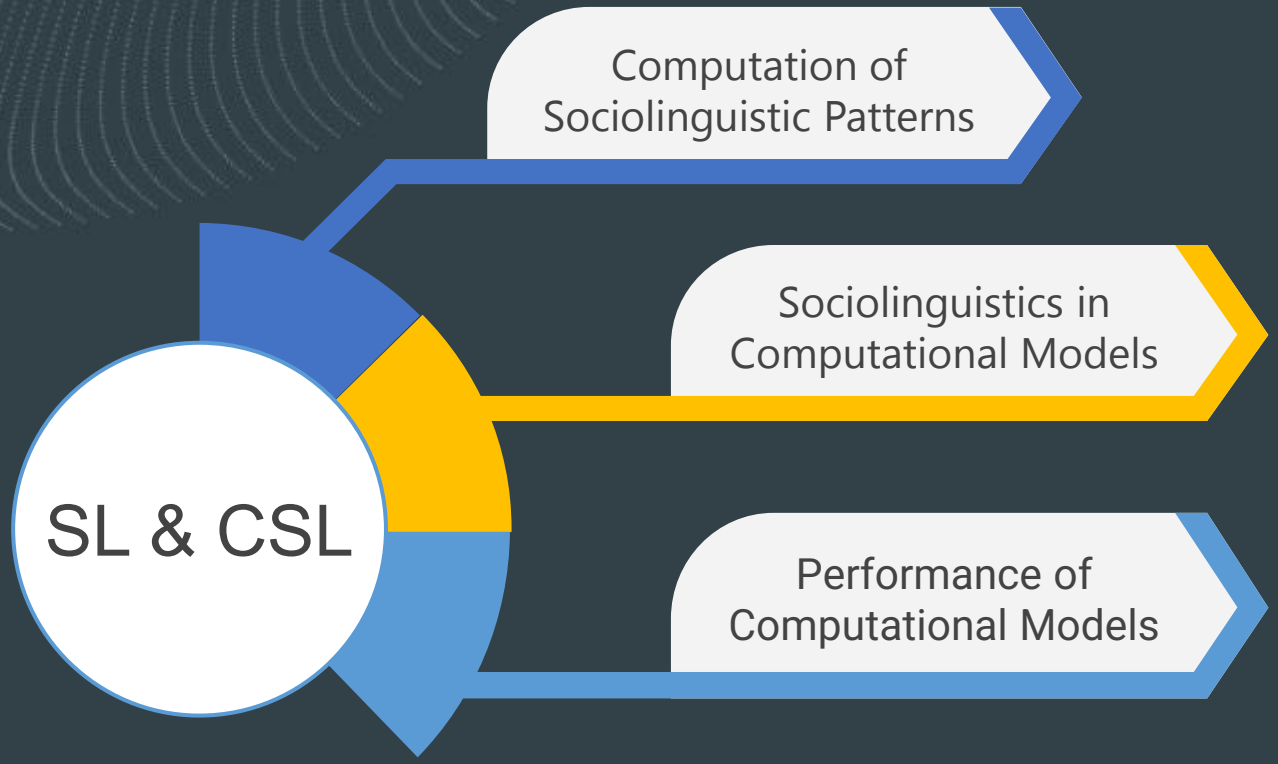
MOTIVATION


01
Evaluation of Sociolinguistics Theories
Sociolinguistics
Gender
Fuzzy Gender Conception
02
03
Spanish NLP
NLP / ML
Computational Sociolinguistics
Better Computational Results
04

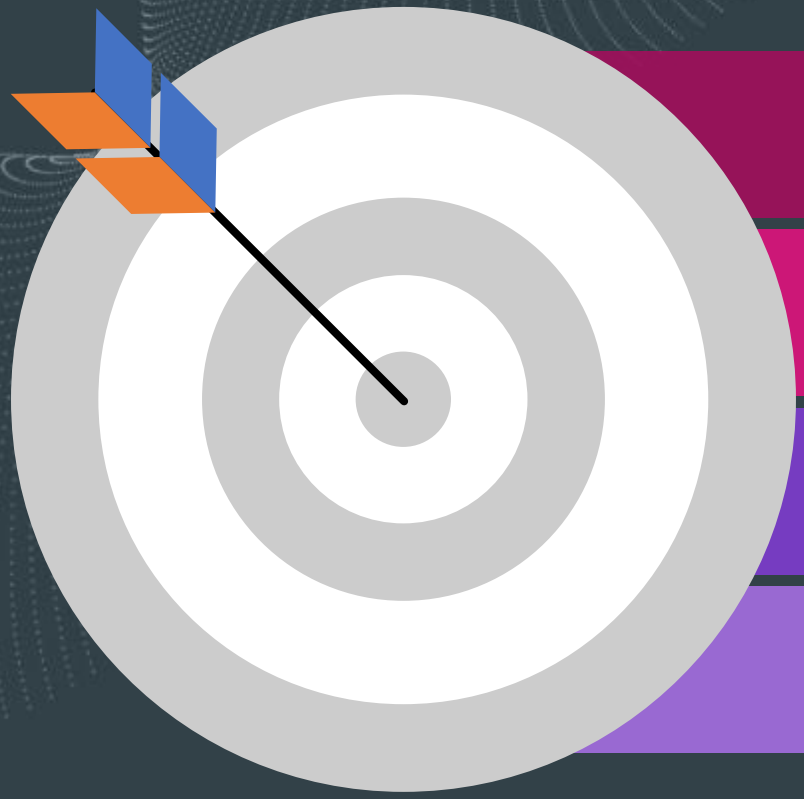
RESEARCH QUESTIONS



HYPOTHESES


Computation of Sociolinguistic Patterns
Sociolinguistics in Computational Models
Performance of Computational Models
SL & CSL

OBJECTIVES

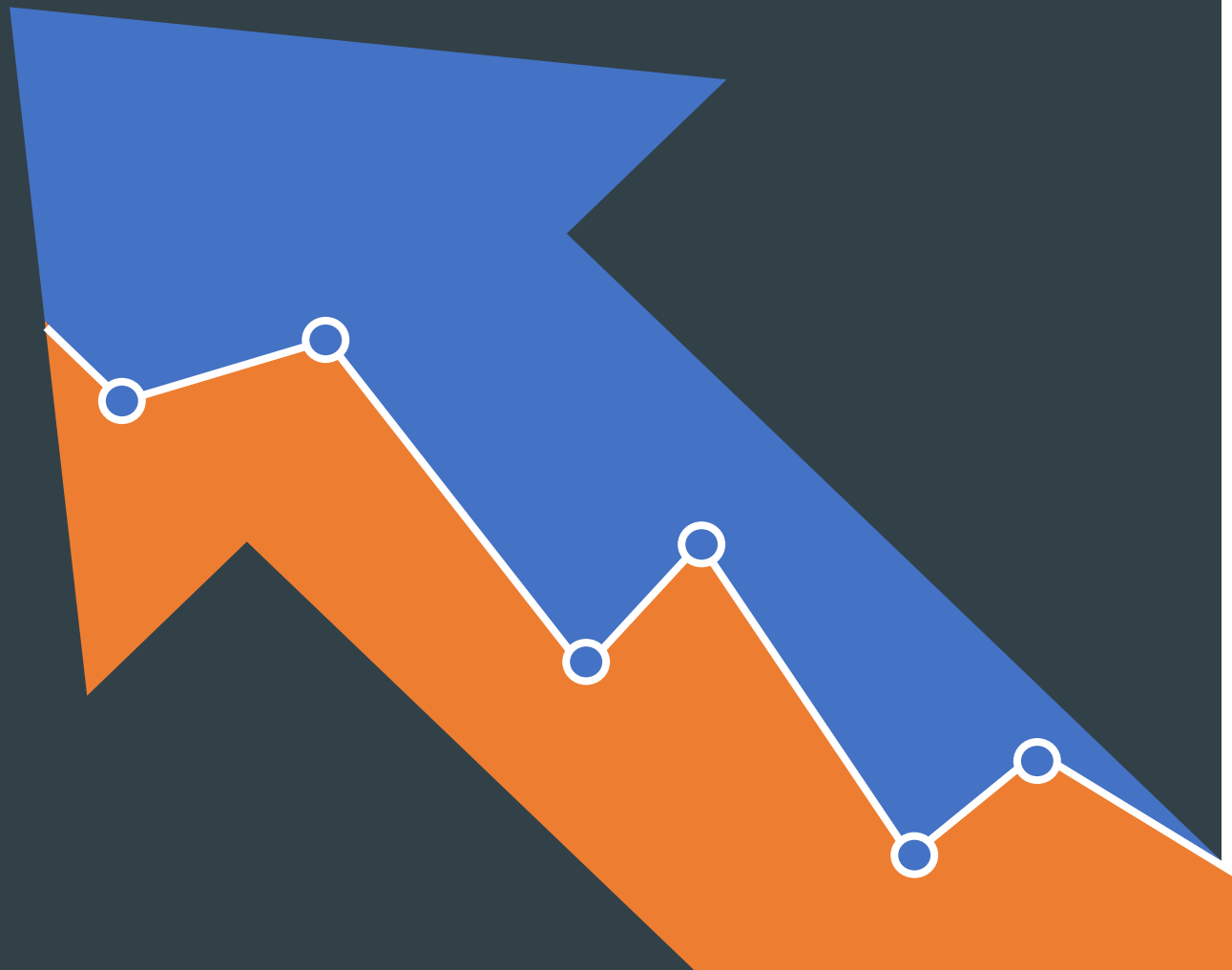


- Fuzzy Sociolinguistic Model for Gender Prediction
- Sociolinguistics Gender Theories
- Computational Sociolinguistics Methodologies and Results
- Gender Formalization

RESEARCH APPLICATION

Legal investigation

Marketing



THEORETICAL FRAMEWORK

GENDERLECT

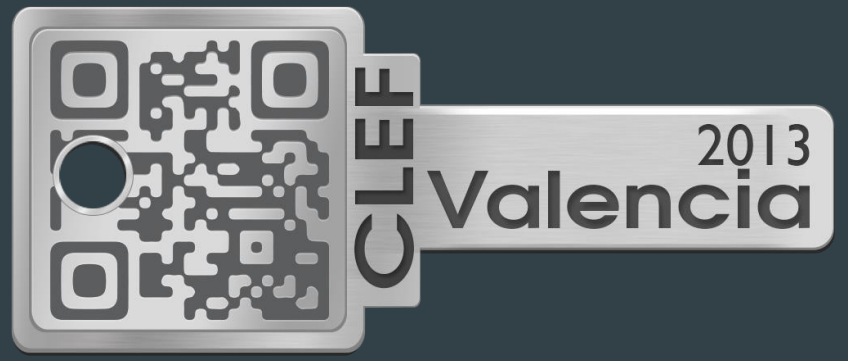
Generolectos

GENERCOLECTO FEMENINO	GENEROLECTO MASCULINO
PRAGMÁTICO-DISCURSIVAS	
Estrategias conversacionales cooperativas; creación de la intimidad a través de la conversación, con infrecuente aparición de la controversia, el desafío o la jocosidad (Coates 2016; Muhkerjee & Liu 2010; Blas-Arroyo 2005; Cameron 2003; Coates 2003; Eckert & McConnell-Ginet 2003; García Mouton 2003; Holmes & Stubbe 2003; Corney *et al*. 2002; García Mouton 1999; Wodak & Benke 1998; Corson 1997; Coates 1996; Hall 1995; Holmes 1995; Scheibman 1995; Thorne 1993; McCauslan & Kleiner 1992; Pilkington 1992; Cox *et al*. 1990; Tannen 1990; Franken 1983; Walker 1981; Spender 1980; Aries 1976; Lakoff 1975).	**Estrategias conversacionales competitivas; creación de la jerarquía y la independencia a través de la conversación, con frecuente aparición de la controversia, el desafío y la jocosidad; frecuentes implicaturas de carácter despectivo relacionadas con la homosexualidad y la feminidad** (Coates 2003; Cameron 1997; Pujolar i Cos 1997). (Coates 2016; Muhkerjee & Liu 2010; Blas-Arroyo 2005; Lozano 2005; Coates 2003; Eckert & McConnell-Ginet 2003; García Mouton 2003; Herring 2003; Holmes & Stubbe 2003; Corney *et al*. 2002; Páramo 2002; Holmes 2000; García Mouton 1999; Wodak & Benke 1998; Corson 1997; Johnson & Finlay 1997; Kiesling 1997; Herring *et al*. 1995; Tannen 1994; Ainsworth-Vaughn 1992; Case 1991; Tannen 1990; Spender 1980; Mitchell-Kernan 1973, Dundes, 1972; Labov 1972). Publicaciones críticas: Talbot 2020; Coates 2003; Freed 2003; Kiesling 2003; Tannen 2003; Páramo 2002; Cameron 1997; Hewitt 1997; Freed & Greenwood 1996; Greenwood & Freed 1992.
Menor frecuencia de la interrupción como estrategia de dominio conversacional (Cameron 2003; Páramo 2002; Fitzpatrick *et al*. 1995; Tannen 1990; Esposito 1979; Natale *et al*. 1979; Zimmerman & West 1975). Publicaciones críticas: Coates 1988, 1987; Roger & Nesshoever 1987; James & Clarke 1983; Beattie 1981; McCarrick *et al*. 1981.	**Mayor frecuencia de la interrupción como estrategia de dominio conversacional** (Lozano 2005; Cameron 2003; García Mouton 2003; Holmes & Stubbe 2003; Páramo 2002; Holmes 1995; Case 1988; Wood 1988; West 1984; West & Zimmerman 1983; Esposito 1979; Natale *et al*. 1979; Zimmerman & West 1975). Publicaciones críticas: Greenwood 2013; Freed 2003; Coates 1988, 1987; Roger & Nesshoever 1987; James & Clarke 1983; Beattie 1981; McCarrick *et al*. 1981.

COMPUTATIONAL ANALYSIS

DATASET

Language	Age group	Gender	Number of Authors		
			Training	Early Bird	Test
Spanish	10s	male	1,250	120	144
		female	1,250	120	144
	20s	male	21,300	1,920	2,304
		female	21,300	1,920	2,304
	30s	male	15,400	1,360	1,632
		female	15,400	1,360	1,632
TOTAL			75,900	6,800	8,160



DATASET PRE-PROCESSING

HTML Cleaning to obtain plain text	5 teams: [gopal-patra][moreau][meina] [weren][pavan]
Deletion of documents with at least 0.1% of spam words	1 team: [flekova]
Principal Component Analysis to reduce dimensionality	1 team: [yong-lim]
Subset selection during training to reduce dimensionality	5 teams: [caurcel-diaz][flekova][moreau] [hernandez-farias][sapkota]
Discrimination between human-like posts and spam-like posts (chatbots)	1 team: [meina]

DATASET FEATURES

Stylistic features: frequencies of punctuation marks, capital letters, quotations...	9 teams: [yong-lim][cruz][pavan][gopal-patra][de-arteaga][meina][flekova][aleman][santosh]
+ POS tags	5 teams: [yong lim][meina][aleman][cruz][santosh]
HTML-based features like image urls or links	3 teams: [santosh][sapkota][meina]
Readability	7 teams: [gopal-patra][yong-lim][meina][flekova][aleman][weren][gillam]
Emoticons	2 teams: [aleman][hernandez-farias] *[sapkota] explicitly discarded them

DATASET FEATURES

Content features: LSA, BoW, TF-IDF, dictionary-based words, topic-based words, entropy-based words...	11 teams: [sapkota][gopal-patra][yong-lim][seifeddine][caurcel-diaz][flekova][meina][cruz][santosh][pavan][hernandez-farias]
Named entities	1 team: [flekova]
Sentiment words	1 team: [gopal-patra]
Emotions words	1 team: [meina]
Slang, contractions and words with character flooding	4 teams: [flekova][caurcel-diaz][aleman][hernandez-farias]

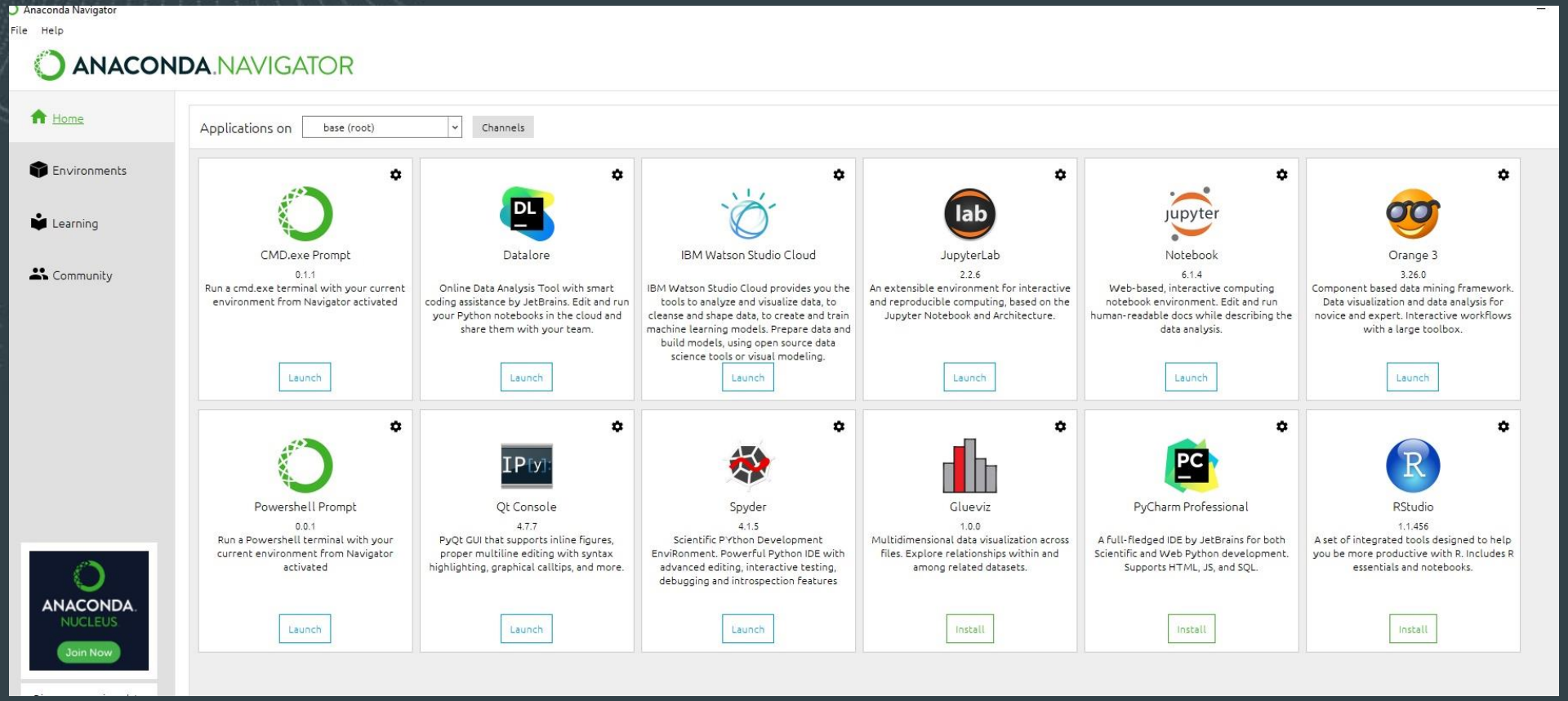
DATASET CLASSIFICATION METHODS

Decision Trees	5 teams: [santosh][gopal-patra] [seifeddine][gillam][weren]
Support Vector Machines	3 teams: [yong-lim][cruz][sapkota]
Logistic Regression	2 teams: [de-arteaga][flekova]
Naïve Bayes	1 team: [meina]
Maximum Entropy	1 team: [pavan]
Stochastic Gradient Descent	1 team: [caurcel-diaz]
Random Forest	1 team: [aleman]
Information Retrieval	1 team: [weren]

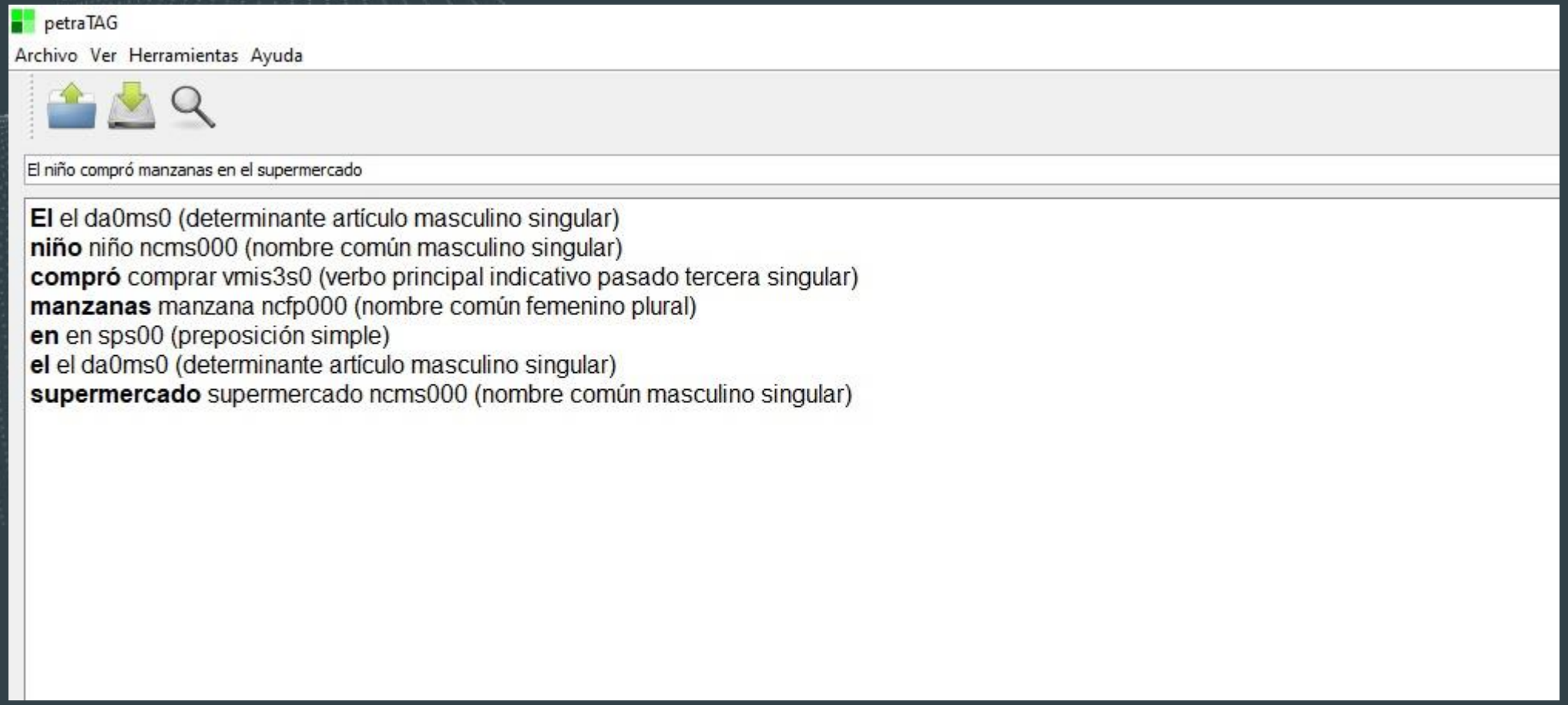
DATASET 2013 RESULTS

Team	Gender		Team	Gender
Santosh	0.6473		Yong Lim	0.5468
Pastor L.	0.6299		Seifeddine	0.5455
Cruz	0.6165		Weren	0.5362
Ladra	0.6138		Meina	0.5287
Flekova	0.6103		Sapkota	0.5116
Jankowska	0.5846		Pavan	0.5000
Kern	0.5706		Caurcel Diaz	0.5000
De-Arteaga	0.5627		H. Farias	0.4982
Aleman	0.5526		Moreau	0.4967
Cagnina	0.5516		Guillam	0.4784

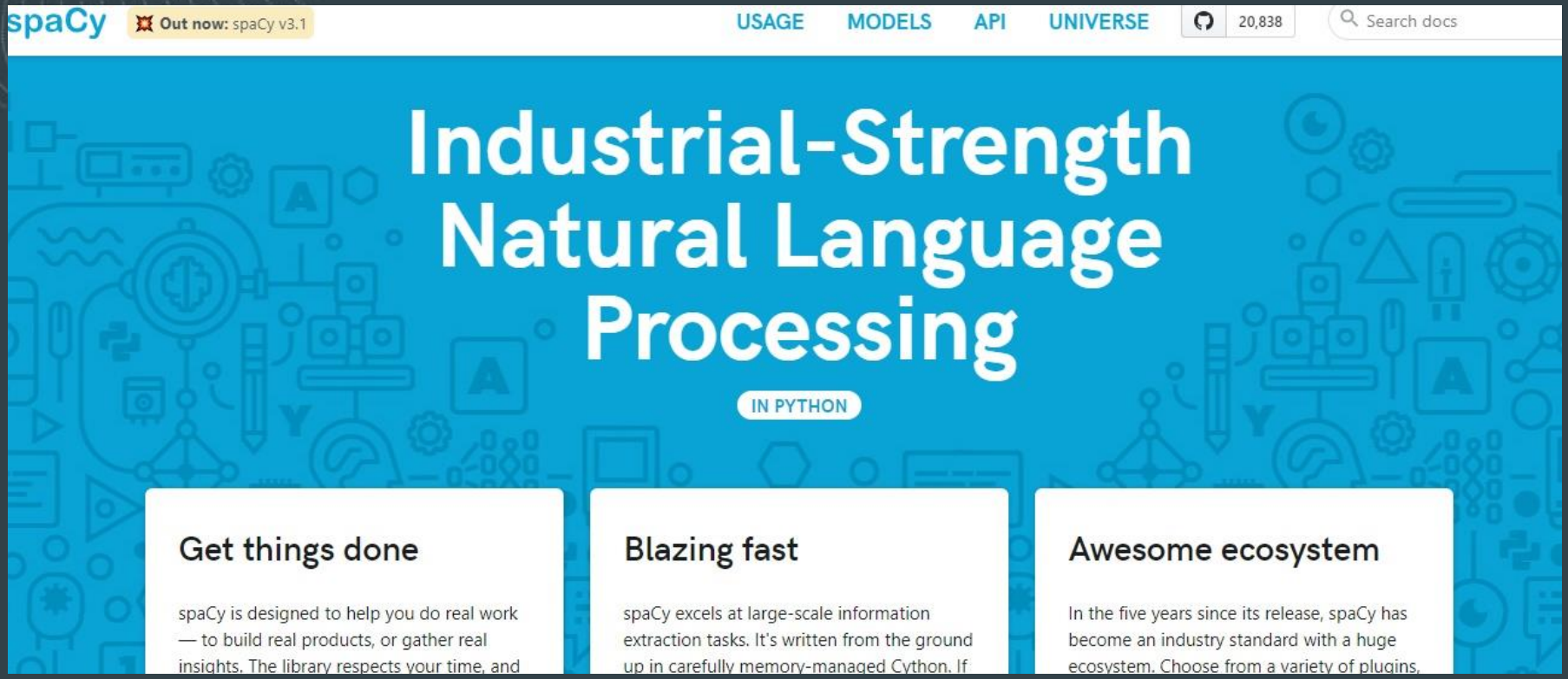
FEATURES EXTRACTION

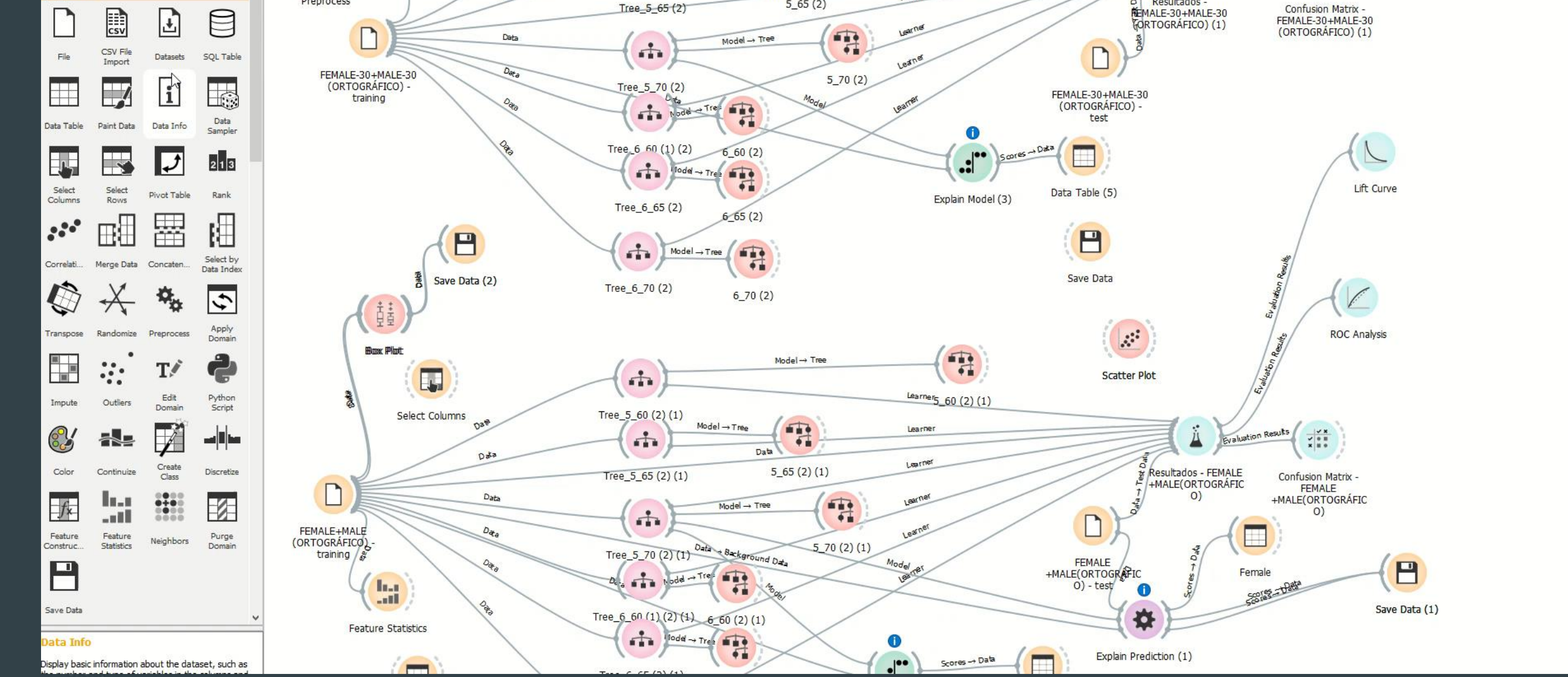


POS TAGGER


petraTAG
Archivo Ver Herramientas Ayuda
El niño compró manzanas en el supermercado
El el da0ms0 (determinante artículo masculino singular)
niño niño ncms000 (nombre común masculino singular)
compró comprar vmis3s0 (verbo principal indicativo pasado tercera singular)
manzanas manzana ncfp000 (nombre común femenino plural)
en en sps00 (preposición simple)
el el da0ms0 (determinante artículo masculino singular)
supermercado supermercado ncms000 (nombre común masculino singular)

SPACY


spaCy
Out now: spaCy v3.1
USAGE
MODELS
API
UNIVERSE
20,838
Search docs
Industrial-Strength Natural Language Processing
IN PYTHON
Get things done
spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and
Blazing fast
spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If
Awesome ecosystem
In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins,


File
CSV File Import
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Pivot Table
Rank
Correlati...
Merge Data
Concaten...
Select by Data Index
Transpose
Randomize
Preprocess
Apply Domain
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Construc...
Feature Statistics
Neighbors
Purge Domain
Save Data
Data Info
Display basic information about the dataset, such as
FEMALE-30+MALE-30 (ORTOGRÁFICO) - training
Tree_5_65 (2)
5_65 (2)
Tree_5_70 (2)
5_70 (2)
Tree_6_60 (1) (2)
6_60 (2)
Tree_6_65 (2)
6_65 (2)
Tree_6_70 (2)
6_70 (2)
Model → Tree
Learner
Data
Model
Explain Model (3)
Scores → Data
Data Table (5)
FEMALE-30+MALE-30 (ORTOGRÁFICO) - test
Resultados - FEMALE-30+MALE-30 (ORTOGRÁFICO) (1)
Confusion Matrix - FEMALE-30+MALE-30 (ORTOGRÁFICO) (1)
Save Data (2)
Box Plot
Select Columns
Save Data
Lift Curve
ROC Analysis
Scatter Plot
Evaluation Results
Tree_5_60 (2) (1)
5_60 (2) (1)
Tree_5_65 (2) (1)
5_65 (2) (1)
Tree_5_70 (2) (1)
5_70 (2) (1)
Tree_6_60 (1) (2) (1)
6_60 (2) (1)
FEMALE+MALE (ORTOGRÁFICO) - training
Feature Statistics
Resultados - FEMALE +MALE(ORTOGRÁFICO)
Confusion Matrix - FEMALE +MALE(ORTOGRÁFICO)
Data → Test Data
FEMALE +MALE(ORTOGRÁFICO) - test
Female
Scores → Data
Explain Prediction (1)
Save Data (1)
Data → Background Data

ORTHOGRAPHIC LEVEL

01

Pre-processing

HTML tags

02

Features extraction

29

03

EDA

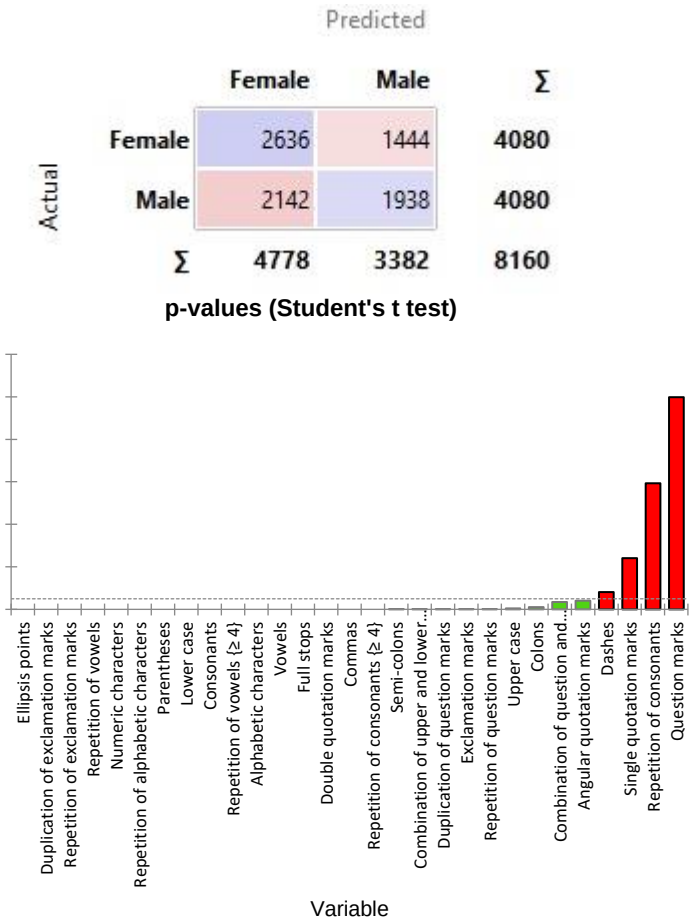
25 p-value < 0.05

04

Classification Model

56.05%

ORTHOGRAPHIC TRAINING	Female	Male	Student's t	P-value	Statistical significance	Gender
Single quotation marks	0.327	0.418	1.172	0.241	No	Male
Double quotation marks	1.073	1.336	5.929	0.000	Yes	Male
Angular quotation marks	0.096	0.139	2.038	0.042	Yes	Male
Commas	14.115	15.430	5.548	0.000	Yes	Male
Full-stops	6.884	7.771	6.289	0.000	Yes	Male
Colons	1.389	1.553	2.549	0.011	Yes	Male
Semi-colons	0.665	0.802	4.982	0.000	Yes	Male
Question marks	1.140	1.141	0.003	0.998	No	Male
Exclamation marks	0.591	0.515	3.729	0.000	Yes	Female
Parentheses	1.264	1.828	7.511	0.000	Yes	Male
Dashes	3.760	4.660	1.743	0.081	No	Male
Ellipsis points	3.764	2.841	8.567	0.000	Yes	Female
Duplication of question marks	0.060	0.029	3.796	0.000	Yes	Female
Duplication of exclamation marks	0.138	0.080	9.596	0.000	Yes	Female
Repetition of question marks	0.115	0.057	3.546	0.000	Yes	Female
Repetition of exclamation marks	0.318	0.195	12.968	0.000	Yes	Female
Combination of question and exclamation marks	0.019	0.015	2.108	0.035	Yes	Female
Vowels	498.823	546.184	6.508	0.000	Yes	Male
Consonants	584.344	645.232	7.086	0.000	Yes	Male
Alphabetic characters	1,083.167	1,191.416	6.823	0.000	Yes	Male
Repetition of vowels	0.066	0.036	8.809	0.000	Yes	Female
Repetition of vowels {≥ 4}	0.122	0.083	7.011	0.000	Yes	Female
Repetition of consonants	0.067	0.069	0.534	0.593	No	Male
Repetition of consonants {≥ 4}	0.064	0.047	5.267	0.000	Yes	Female
Repetition of alphabetic characters	0.319	0.235	8.001	0.000	Yes	Female
Upper case	56.448	52.007	2.836	0.005	Yes	Female
Lower case	1,026.719	1,139.409	7.407	0.000	Yes	Male
Combination of upper and lower case	0.536	0.338	4.471	0.000	Yes	Female
Numeric characters	4.285	6.418	9.992	0.000	Yes	Male



Feature – Female & Male	SHAP value
Ellipsis points	0.0678634
Numeric characters	0.0611684
Repetition of exclamation marks	0.0155106
Commas	0.00605814
Consonants	0.00519023
Upper case	0.00375069
Dashes	0.00370815
Lower case	0.00326851
Single quotation marks	0.00296966
Alphabetic characters	0.00260533
Double quotation marks	0.00255959
Duplication of exclamation marks	0.00234925
Exclamation marks	0.0022933
Vowels	0.00228896
Semi-colons	0.0022705
Angular quotation marks	0.00225496
Colons	0.00223365
Repetition of alphabetic characters	0.00221037
Full stops	0.00219836
Combination of upper and lower case	0.00217356
Question marks	0.00214819
Parentheses	0.0021346
Repetition of vowels	0.000640054
Repetition of vowels {≥ 4}	0.000106018
Repetition of question marks	8.65809e-05
Repetition of consonants {≥ 4}	8.00013e-05
Repetition of consonants	6.72595e-05
Combination of question and exclamation marks	4.28965e-05
Duplication of question marks	1.6337e-05

MORPHOLOGICAL LEVEL

01 *Pre-processing*
POS Tagger

02 *Features extraction*
434 (29)

03 *EDA*
p-value < 0.05

04 *Classification Model*
54.5% (*54.8%*)

MORPHOLOGICAL TRAINING	Female	Male	Student's t	P-value	Statistical significance	Gender
Proper nouns	2.966	3.605	9.327	0.000	Yes	Male
Masculine singular common nouns	10.255	10.197	0.541	0.588	No	Female
Feminine singular common nouns	9.637	9.501	1.343	0.179	No	Female
Gender invariable singular common nouns	0.091	0.106	4.283	0.000	Yes	Male
Masculine plural common nouns	3.476	3.573	2.256	0.024	No	Male
Feminine plural common nouns	3.074	3.102	0.769	0.442	No	Male
Gender invariable plural common nouns	0.051	0.066	5.684	0.000	Yes	Male
Masculine common nouns	13.731	13.770	0.273	0.785	No	Male
Feminine common nouns	12.712	12.603	0.821	0.441	No	Female
Gender invariable common nouns	0.142	0.173	6.246	0.000	Yes	Male
Singular common nouns	19.984	19.804	0.877	0.380	No	Female
Plural common nouns	6.601	6.742	1.848	0.065	No	Male
Common nouns	26.585	26.546	0.143	0.886	No	Female
Nouns	29.551	30.151	1.928	0.054	No	Male
Masculine singular qualificative adjectives	2.402	2.538	4.916	0.000	Yes	Male
Feminine singular qualificative adjectives	1.407	1.328	4.354	0.000	Yes	Female
Gender invariable singular qualificative adjectives	2.400	2.403	0.115	0.908	No	Male
Masculine plural qualificative adjectives	0.607	0.642	3.307	0.001	Yes	Male
Feminine plural qualificative adjectives	0.430	0.448	2.269	0.023	No	Male
Gender invariable plural qualificative adjectives	0.609	0.689	6.871	0.000	Yes	Male
Masculine qualificative adjectives	3.009	3.180	5.034	0.000	Yes	Male
Feminine qualificative adjectives	1.837	1.776	2.688	0.007	No	Female
Gender invariable qualificative adjectives	3.009	3.093	2.254	0.024	No	Male

LEXICAL LEVEL

01 *Pre-processing*

02 *Features extraction*
38

03 *EDA*
24 p-value < 0.05

04 *Classification Model*
(57.1%)

LEXICAL TRAINING	Female	Male	Student's t	P-value	Statistical significance	Gender
Slang words	**0,165**	**0,190**	2.858	0.004	Yes	Male
Slang emoticons	**0,077**	**0,068**	1.066	0.287	No	Female
Social media slang expressions	**0,242**	**0,259**	1.290	0.197	No	Male
Joy words	**5,585**	**4,923**	9.310	0.000	Yes	Female
Anger words	**0,735**	**0,773**	2.529	0.011	No	Male
Fear words	**0,893**	**0,878**	0.748	0.454	No	Female
Revulsion words	**0,400**	**0,428**	2.622	0.009	No	Male
Surprise words	**0,625**	**0,627**	0.124	0.901	No	Male
Sadness words	**2,680**	**2,566**	2.850	0.004	Yes	Female
Emotive lexicon	**10,918**	**10,194**	5.038	0.000	Yes	Female
Modal verbs	**1,187**	**1,148**	2.012	0.044	No	Female
Semi-auxiliary verbs	**0,041**	**0,060**	7.573	0.000	Yes	Male
Epistemic verbs	**0,798**	**0,727**	4.979	0.000	Yes	Female
First person singular elided epistemic verbs	**0,008**	**0,008**	0.788	0.431	No	-
First person plural elided epistemic verbs	**0,016**	**0,016**	0.000	1.000	No	-
First person singular explicit epistemic verbs	**0,009**	**0,008**	1.692	0.091	No	Female
First person plural explicit epistemic verbs	**0,000**	**0,000**	-	-	-	-
Probability adjectives	**0,053**	**0,065**	4.920	0.000	Yes	Male
Probability adverbs	**0,022**	**0,028**	4.019	0.000	Yes	Male
Approximators	**0,062**	**0,092**	9.681	0.000	Yes	Male
Conditional tense	**0,213**	**0,203**	1.516	0.130	No	Female
Subjunctive mood verbs	**2,882**	**2,545**	10.636	0.000	Yes	Female
Emitter implication expressions	**0,093**	**0,095**	0.826	0.409	No	Male
Non-personal verbs	**9,434**	**8,304**	12.225	0.000	Yes	Female
Mitigating lexicon	**5,385**	**4,994**	6.408	0.000	Yes	Female
Offensive lexicon	**0,073**	**0,079**	1.745	0.081	No	Male
Appellatives	**0,262**	**0,191**	11.273	0.000	Yes	Female
Augmentative prefixed appreciative lexicon	**3,138**	**3,641**	8.761	0.000	Yes	Male
Diminutive prefixed appreciative lexicon	**0,164**	**0,249**	9.732	0.000	Yes	Male
Prefixed appreciative lexicon	**3,302**	**3,891**	9.511	0.000	Yes	Male
Suffixed appreciative lexicon	**0,913**	**1,247**	16.847	0.000	Yes	Male
Derived appreciative lexicon	**4,215**	**5,138**	13.027	0.000	Yes	Male

Words from 1 to 3 characters	**110,063**	**116,264**	4.096	0.000	Yes	Male
Words from 4 to 6 characters	**69,523**	**72,433**	3.192	0.001	Yes	Male
Words over 6 characters	**43,957**	**50,317**	9.103	0.000	Yes	Male
Ratio letters/words	**4,317**	**4,367**	11.353	0.000	Yes	Male
Lexical diversity	**0,742**	**0,758**	11.369	0.000	Yes	Male
TTR Lemma	**0,151**	**0,165**	12.381	0.000	Yes	Male

SYNTACTIC LEVEL

01 *Pre-processing*
Spacy

02 *Features extraction*
29

03 *EDA*
17 p-value < 0.05

04 *Classification Model*
(*55.4%*)

SYNTACTIC TRAINING	Female	Male	Student's t	P-value	Statistical significance	Gender
Sentences	3.854	3.408	7.903	0.000	Yes	Female
Sentences length	71.554	74.049	3.133	0.002	Yes	Male
Repetition of words	0.239	0.227	1.180	0.238	No	Female
Word repetition in coordination	0.040	0.036	1.820	0.069	No	Female
Nominal subject	16.592	17.062	2.100	0.036	No	Male
Direct object	28.369	28.672	0.820	0.412	No	Male
Indirect object	1.211	1.133	3.959	0.000	Yes	Female
Oblique nominal	11.012	12.026	6.005	0.000	Yes	Male
Nominal modifier	9.750	12.441	13.763	0.000	Yes	Male
Adjectival modifier	11.800	14.169	11.456	0.000	Yes	Male
Adverbial modifier	15.101	14.673	2.099	0.036	No	Female
Numeric modifier	1.388	1.743	7.976	0.000	Yes	Male
Determiner	33.410	36.910	7.005	0.000	Yes	Male
Case marking	24.224	28.740	10.936	0.000	Yes	Male
Appositional modifier	3.413	4.001	7.733	0.000	Yes	Male
Clausal subject	1.944	1.841	3.855	0.000	Yes	Female
Clausal complement	4.310	4.208	1.724	0.085	No	Female
Open clausal complement	5.361	5.036	4.318	0.000	Yes	Female
Adverbial clause modifier	12.046	12.119	0.442	0.659	No	Male
Adjectival clause	5.767	6.118	4.082	0.000	Yes	Male
Coordination	11.652	12.420	4.574	0.000	Yes	Male
Juxtaposition	0.050	0.053	1.451	0.147	No	Male
Subordination	29.427	29.321	0.273	0.785	No	Female
Copulative relationships	5.626	5.701	0.978	0.328	No	Male
Fixed multiword expression	3.109	3.566	8.170	0.000	Yes	Male
Flat multiword expression	6.507	8.442	8.342	0.000	Yes	Male
Compound	0.773	0.906	2.770	0.006	No	Male
Direct object initial position	0.211	0.177	7.502	0.000	Yes	Female
Indirect object initial position	0.024	0.021	2.308	0.021	No	Female

DIGITAL LEVEL

01 *Pre-processing*

02 *Features extraction*

16

03 *EDA*

16 p-value < 0.05

04 *Classification Model*

DIGITAL TRAINING	Female	Male	Student's t	P-value	Statistical significance	Gender
JPG	0.097	0.065	6.353	<0,0001	Yes	Female
URL	0.185	0.253	5.790	<0,0001	Yes	Male
GIF	1.301	0.795	20.556	<0,0001	Yes	Female
Ratio GIF/words	0.023	0.015	20.339	<0,0001	Yes	Female
angel GIF	0.077	0.040	10.734	<0,0001	Yes	Female
biggrin GIF	0.108	0.085	5.486	<0,0001	Yes	Female
cool GIF	0.034	0.063	8.064	<0,0001	Yes	Male
hug GIF	0.037	0.013	12.844	<0,0001	Yes	Female
inlove GIF	0.043	0.017	12.334	<0,0001	Yes	Female
love GIF	0.213	0.065	18.815	<0,0001	Yes	Female
sad GIF	0.066	0.036	9.838	<0,0001	Yes	Female
smile GIF	0.103	0.060	12.494	<0,0001	Yes	Female
tongue GIF	0.117	0.071	11.476	<0,0001	Yes	Female
unsure GIF	0.034	0.023	6.468	<0,0001	Yes	Female
w00t GIF	0.045	0.051	2.553	0,011	Yes	Male
wink GIF	0.042	0.027	8.393	<0,0001	Yes	Female

Table 7 Exploratory Data Analysis F&M

PRAGMATIC-DISCURSIVE LEVEL

01 *Pre-processing*

02 *Features extraction*

65

03 *EDA*

p-value < 0.05

04 *Classification Model*

(*57.0%*)

PRAGMATIC-DISCURSIVE TRAINING	**Female**	**Male**	**Student's t**	**P-value**	**Statistical significance**	**Gender**
Commentators information structuring discursive markers	**0,005**	**0,007**				**Male**
Opening sequencers information structuring discursive markers	**0,008**	**0,011**				**Male**
Continuity sequencers information structuring discursive markers	**0,012**	**0,019**				**Male**
Closing sequencers information structuring discursive markers	**0,083**	**0,087**				**Male**
Sequencers information structuring discursive markers	**0,104**	**0,117**				**Male**
Digressers information structuring discursive markers	**0,007**	**0,008**				**Male**
Information structuring discursive markers	**0,116**	**0,132**				**Male**
Additive connectors same argumentative scale discursive markers	**0,049**	**0,072**				**Male**
Additive connectors different argumentative scale discursive markers	**0,062**	**0,085**				**Male**
Additive connectors discursive markers	**0,112**	**0,157**				**Male**
Consecutive connectors discursive markers	**0,049**	**0,067**				**Male**
Counterargumentative connectors discursive markers	**0,078**	**0,098**				**Male**
Connectors discursive markers	**0,238**	**0,321**				**Male**

SUPERVISED CLASSIFICATION MODELS

DECISION TREE

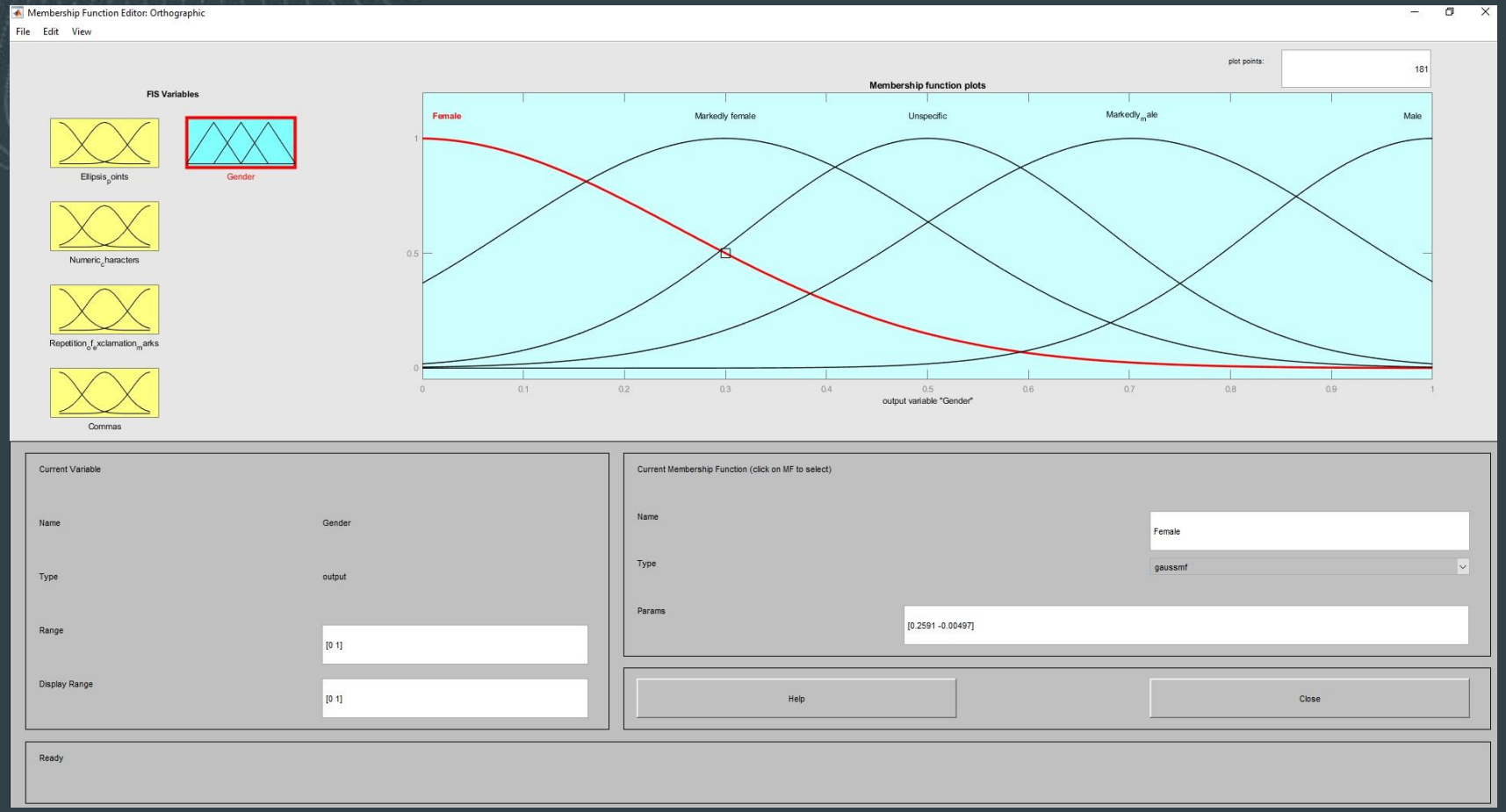
01	DIGITAL	57.7%
02	DIGITAL+LEXICAL	59.6%
03	DIGITAL+LEXICAL+ ORTHOGRAPHIC	59.7%
04	DIGITAL+LEXICAL+ ORTHOGRAPHIC+PRAGMATIC	60.3%

SUPERVISED CLASSIFICATION MODELS

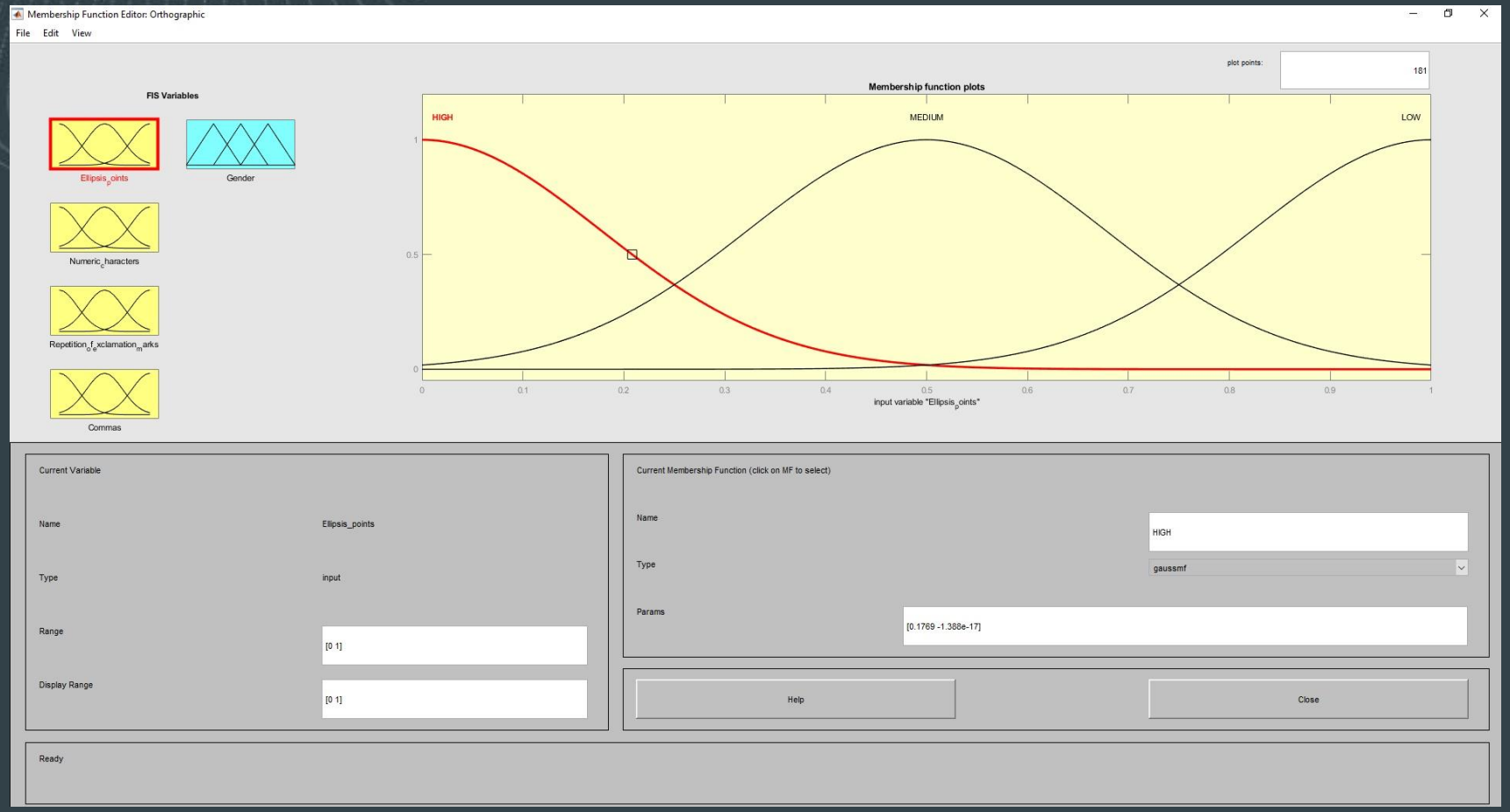
RANDOM FOREST

01	ORTHOGRAPHIC	57.5%
02	DIGITAL+LEXICAL	60.3%
03	DIGITAL+LEXICAL+ ORTHOGRAPHIC	60.7%
04	DIGITAL+LEXICAL+ ORTHOGRAPHIC+SYNTACTIC	61.3%

FUZZY MODEL: GENDER FORMALIZATION



FUZZY MODEL: GENDER FORMALIZATION



FUZZY MODEL: GENDER FORMALIZATION