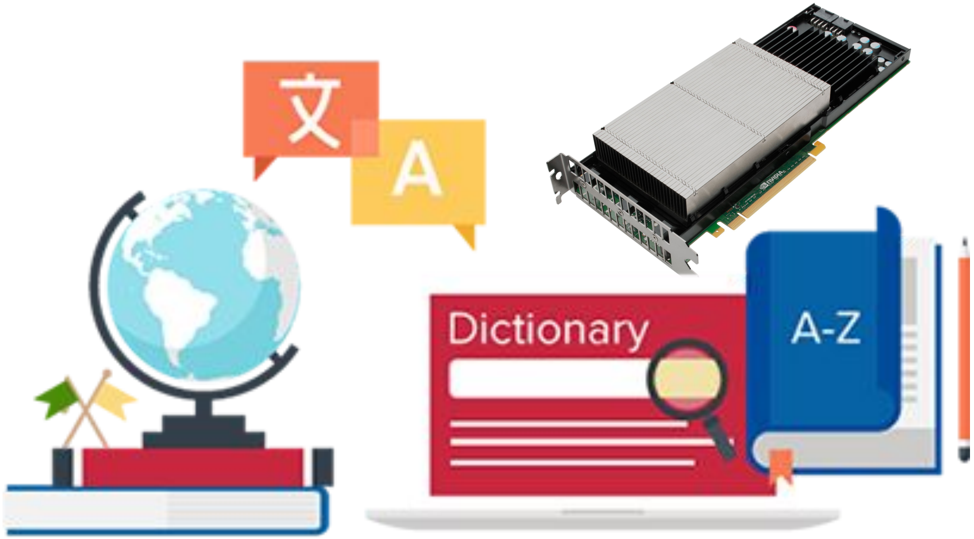
What an NLPer wishes (and does) when permeating a Translation Department.



Alberto Barrón-Cedeño
Alma Mater Studiorum - Università di Bologna
a.barron@unibo.it
@_albarron_

UABC, Mexico (remotely)
3 December, 2020

Alma Mater Studiorum-Università di Bologna

- Oldest university in the western world (est. 1088)
- The Alma Mater of all universities (Magna Charta Universitatum Europaeum, 1988)
- Home to Nicolaus Copernicus, Laura Bassi, Luigi Galvani, Giosué Carducci, Umberto Eco, and many others
- 5 campus across Emilia-Romagna + Argentina

Emilia:	Romagna:
Bologna	Cesena
	Forlì
	Ravenna
	Rimini

Department of Interpreting and Translation

- Born in 2012 (merging the SITLeC Dept. and the *Scuola Superiore di Lingue Moderne per Interpreti e Traduttori*)
- Emphasis in applied research, theoretical, practical, and didactic aspects of translation and interpreting

Degrees

- Bachelor in Intercultural and Linguistic Mediation
- Masters in Interpreting
- Masters in Specialized Translation
- PhD in Translation, Interpreting, Interculturality

About myself

Computing scientist working on

Natural Language Processing

Information Retrieval

Machine Learning

2004	**B. Eng (Computing)** U. Nacional Autónoma de México	🇲🇽
2007	**MSc in Computing Science** U. Nacional Autónoma de México	🇲🇽
2008	**MSc in Computing Science** Universitat Politècnica de València	🇪🇸
2012	**PhD in Computing Science** Universitat Politècnica de València	🇪🇸
2012-2014	**Alain Bensoussan Fellow** Universitat Politècnica de Catalunya	🇪🇸
2014-2019	**Scientist** Qatar Computing Research Institute	🇶🇦
2019-	**Senior Assistant Professor** Università di Bologna	🇮🇹

Disclaimers

1. This is my very own perception of 1.5 years of research and teaching at UniBO; it does not necessarily reflect that of the rest of the department

2. I am used to speak about these topics in English (or Italian). My apologies if I start sounding *pocho* or I miss some proper terms in Spanish

Overview

1. How CS is being *plugged* into DIT

2. Teaching initiatives

3. Three student projects

4. Closing remarks

How CS is being *plugged* into DIT

Computing scientists hiring

Spring 2019

Senior assistant professor with NLP background

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

Alberto Barrón-Cedeño

Winter 2020

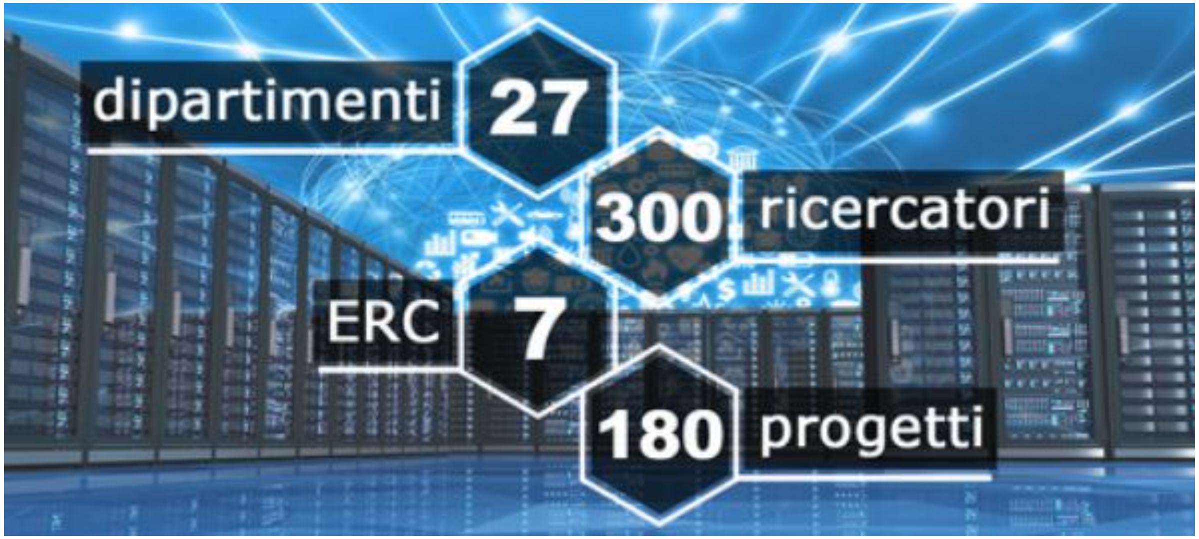
Research assistant with MT background

Microsoft® Research

Federico Garcea

Into Alma AI

DIT adhered to UniBO's Alma Human Artificial Intelligence Centre



Foundations of AI

AI and hard sciences

Humanistic AI

AI for industry

AI for health and well-being

AI for law and governance

AI and education

AI and high performance computing

Initiatives with heavy CS load

- Neural machine translation
- MT of academic websites
- Interaction with local companies in need for MT and multilingual NLP
- Webinars and workshops on MT and related technologies
- Automatic identification of propaganda
- Translation for creative and artistic documents (e.g. opera lyrics)
- Discussions on the curriculum in 5 years time

- Conveying that MT is not an enemy

Teaching initiatives

DIT Python course 2020

Official objective 1. Giving a gentle introduction to programming in python to get students in the right position to go further on their own

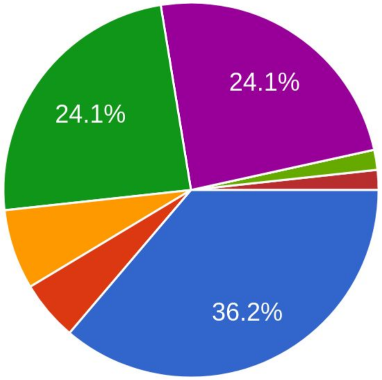
Official objective 2. Serving as propaedeutic to the computational linguistics course

Unofficial objective. Never again listen in the aisle "that would be so awesome! But stop… it needs programming"

DIT Python course 2020

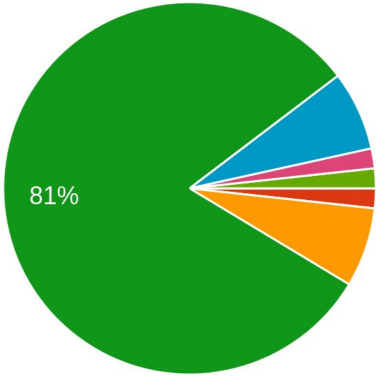
Pre-entry survey

A translator should know how to code



- ● Absolutely
- ● Yes, if (s)he wants to do research (not for industry)
- ● Yes, if (s)he wants to go for industry (n…
- ● Yes, if (s)he is targetting the software/…
- ● Maybe
- ● Nope. There is out-of-the-box softwar…
- ● Nope. Translation does not involve sof…
- ● knowing the theory when localizing so…
- ● Yes, for both industry and research pu…

Familiarity with Python



- ● I use it on a regular basis
- ● I can code a few routines, with a lot of effort
- ● I have passive knowledge (I can read it, but I cannot produce it)
- ● I've heard about it, but I don't know it…
- ● I hate it
- ● Isn't it a snake?
- ● last year I attended an online course o…
- ● So far I've only coded a few scripts for…

DIT Python course 2020

Course structure

Three 2-hour sessions

1. Presentation of concepts with the support of slides
2. Live on-screen coding of task-specific routines
3. Take-home simple coding exercises

Coding *platform*

Jupyter notebooks on Google's colab

http://colab.research.google.com

DIT Python course 2020

Session 1. The basics

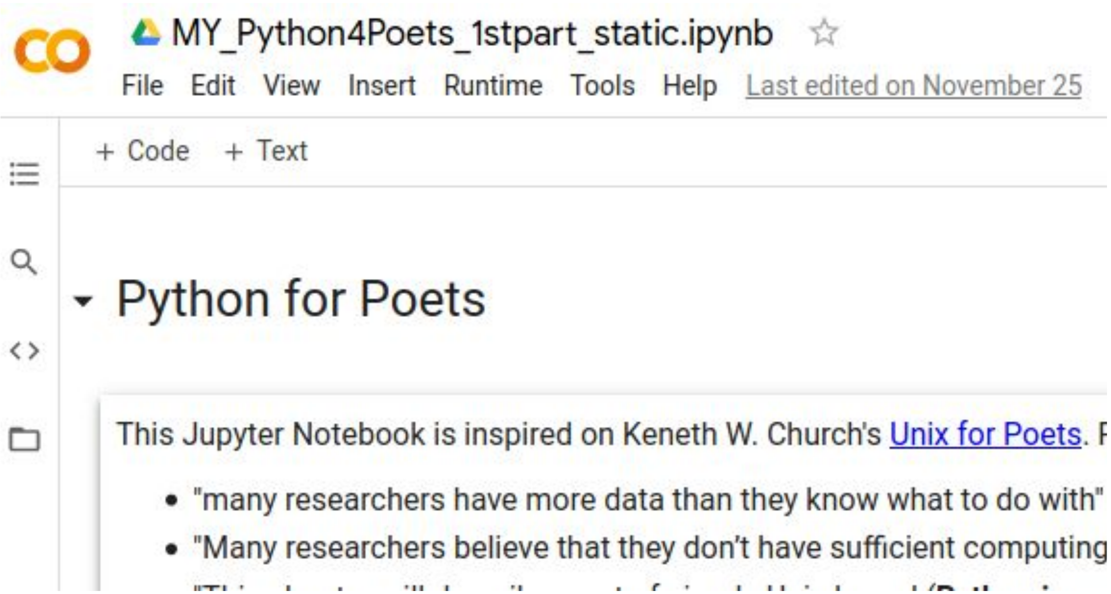
- What is a programming language?
- What is an algorithm
- ”Translating” from an algorithm into a program
- The characteristics of the python programming language
- Basic functions, variables, conditionals, loops

```
# my code
x = 0
while x < 50:
 for i in range(x):
   print('x', end="")
 print()
 x += 1
```

DIT Python course 2020

Session 2. Python 4 Poets (1/2)
(derived from K. Church's Unix for poets)

- Opening text files
- Splitting into words
- Obtaining vocabularies
- Extracting *n*-grams


MY_Python4Poets_1stpart_static.ipynb
File Edit View Insert Runtime Tools Help Last edited on November 25
+ Code + Text

Python for Poets

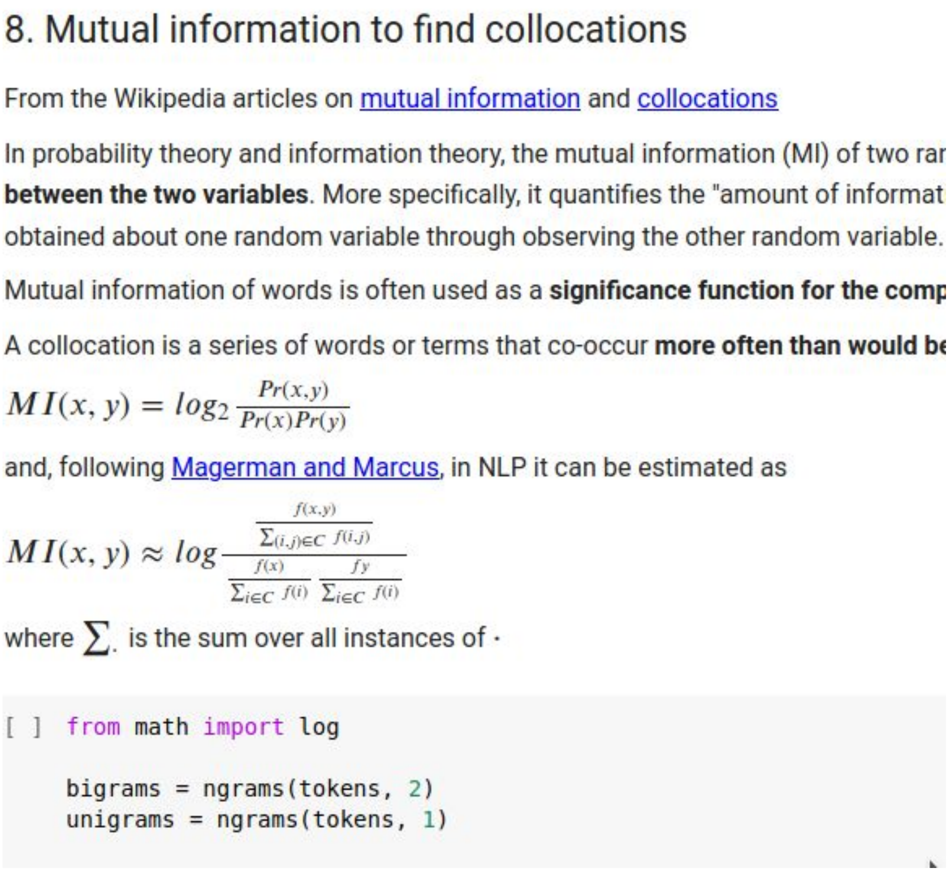
This Jupyter Notebook is inspired on Keneth W. Church's Unix for Poets. F

- "many researchers have more data than they know what to do with"
- "Many researchers believe that they don't have sufficient computing
- "This chapter will describe a set of simple Unix-based (Python i

DIT Python course 2020

Session 3. Python 4 Poets (2/2)
(derived from K. Church's Unix for poets)

- Finding specific tokens/strings
- Finding palindromes
- String substitutions
- Functions
- Collocations



8. Mutual information to find collocations

From the Wikipedia articles on mutual information and collocations

In probability theory and information theory, the mutual information (MI) of two rar **between the two variables**. More specifically, it quantifies the "amount of informat obtained about one random variable through observing the other random variable.

Mutual information of words is often used as a **significance function for the comp**

A collocation is a series of words or terms that co-occur **more often than would be**

$$MI(x,y) = log_2 \frac{Pr(x,y)}{Pr(x)Pr(y)}$$

and, following Magerman and Marcus, in NLP it can be estimated as

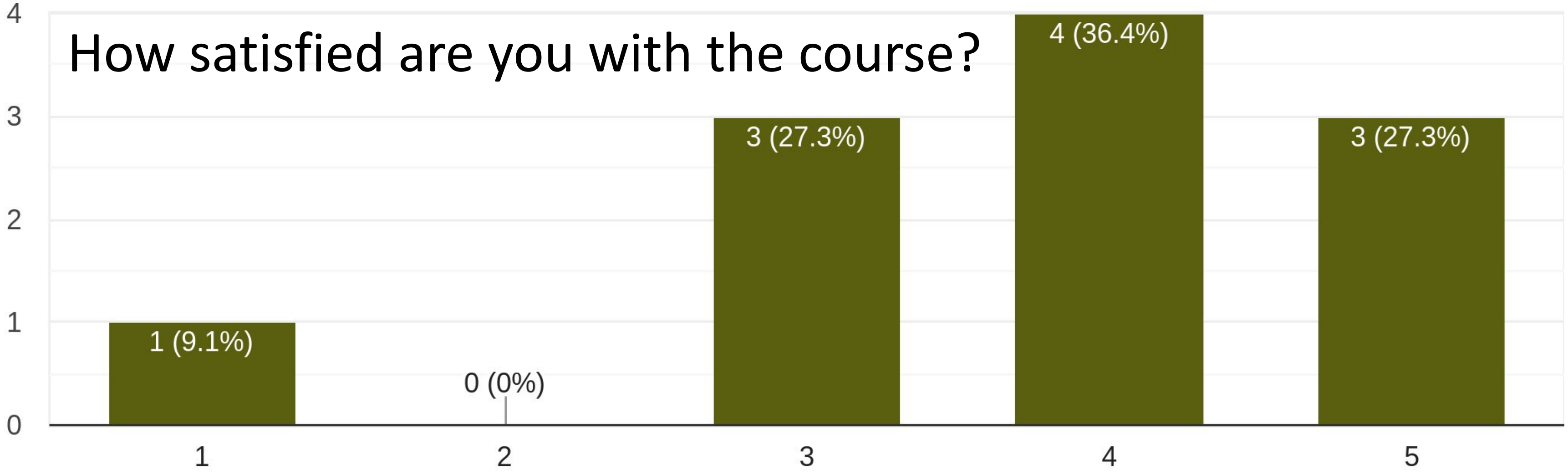
$$MI(x,y) \approx log \frac{\frac{f(x,y)}{\sum_{(i,j) \in C} f(i,j)}}{\frac{f(x)}{\sum_{i \in C} f(i)} \frac{fy}{\sum_{i \in C} f(i)}}$$

where $\sum_{\cdot}$ is the sum over all instances of ·

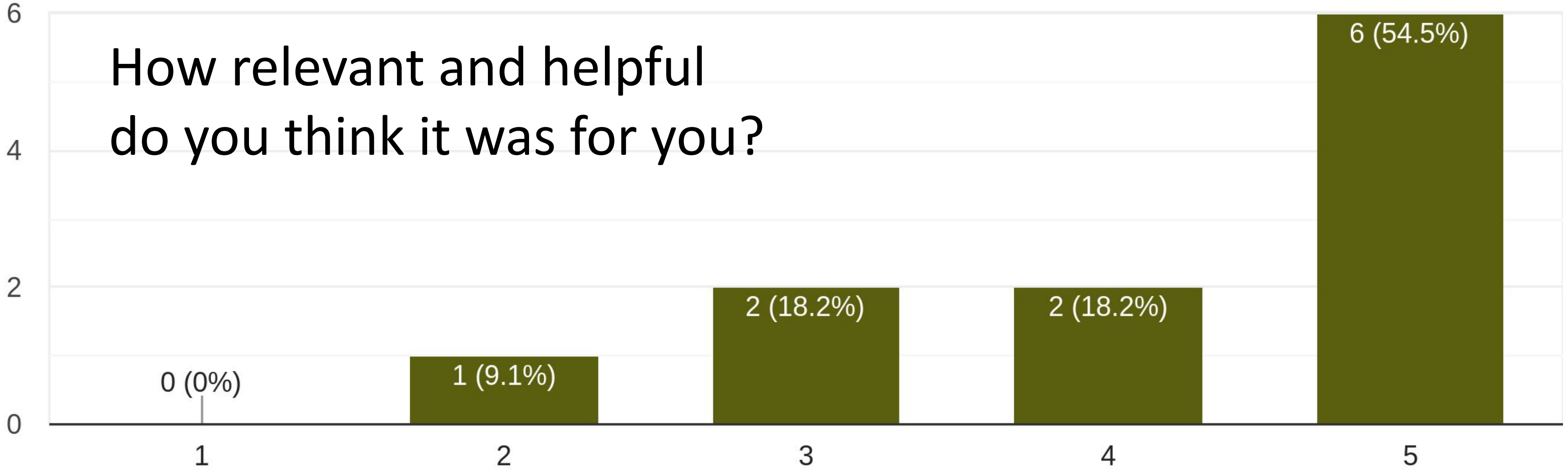
```
[ ]  from math import log

     bigrams = ngrams(tokens, 2)
     unigrams = ngrams(tokens, 1)
```

DIT Python course 2020

Closing perception



How satisfied are you with the course?

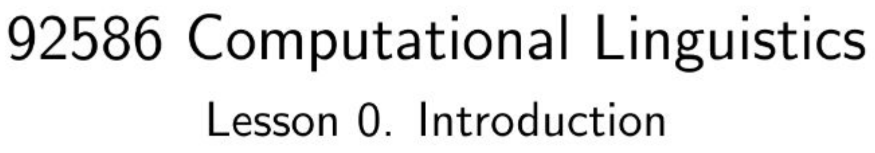


How relevant and helpful do you think it was for you?

Computational Linguistics course

Learning outcomes. [...] basic theoretical aspects of computational linguistics [...] acquire practical skills [all the way to] supervised models

Full semester (optative) course for Masters students



92586 Computational Linguistics

Lesson 0. Introduction

Computational Linguistics course

Course structure

1. Presentation of concepts with the support of slides
2. Live on-screen coding with simple running routines + voluntary homework
3. Evaluation based on one final project + poster presentation (with potential to become a publication)

Coding *platforms*

jupyter notebooks + PyCharm + BASH THE BOURNE-AGAIN SHELL

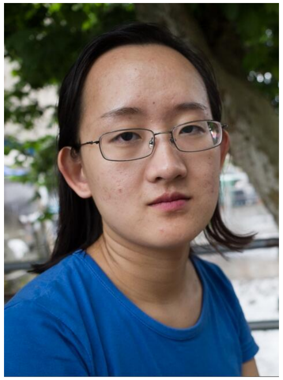
Computational Linguistics course

Rough contents: coding, statistics, and machine learning applied to text

1. Introduction to computational linguistics / python scripting
2. Tokens and the vector space model
3. The Naïve Bayes classifier
4. The training and evaluation process in machine learning
5. Word vectors
6. Latent semantic analysis
7. Neural networks
8. Word Embeddings
9. Convolutional neural networks
10. Sequential neural networks

Three student projects

AriEmozione



Shibingfeng Zhang

Francesco Fernicola

Specialized Translation Masters

Objective. Identifying the emotion transmitted in 17th/18th-century Italian opera arias at the verse level

Developed in the context of UniBO's Centro per l'Interazione con le Industrie Culturali e Creative (https://site.unibo.it/cricc/it)

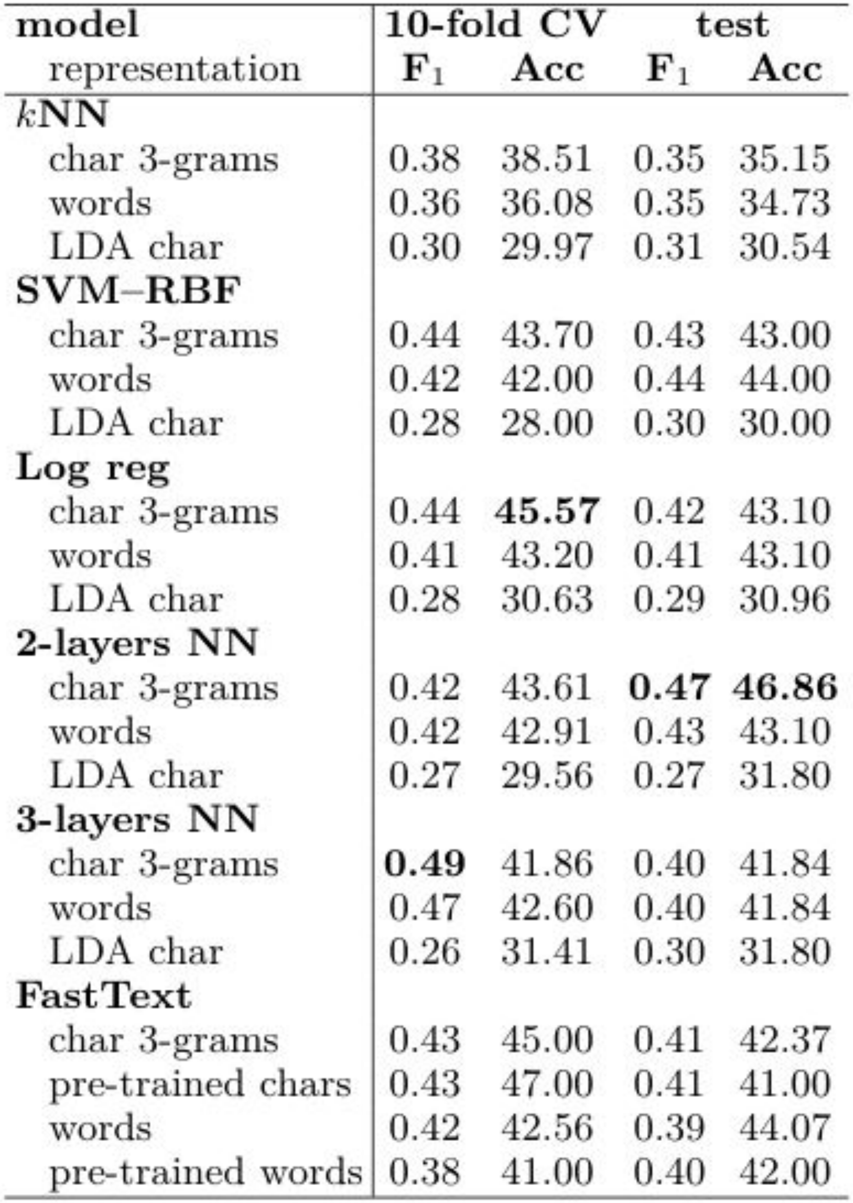
AriEmozione

AriEmozione 1.0 corpus

- 678 operas composed between 1655 and 1765
- All texts are written in Italian of the period and articulated in verses
- 2,473 verses manually annotated in six classes

    - Amore (Love)
    - Gioia (Joy)
    - Ammirazione (Admiration)
    - Rabbia (Anger)
    - Tristezza (Sadness)
    - Paura (Fear)

AriEmozione



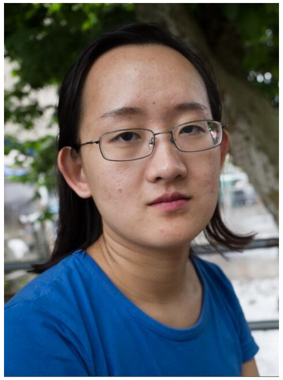
model representation	10-fold CV F$_1$	 Acc	test F$_1$	 Acc
*k*NN				
char 3-grams	0.38	38.51	0.35	35.15
words	0.36	36.08	0.35	34.73
LDA char	0.30	29.97	0.31	30.54
SVM–RBF				
char 3-grams	0.44	43.70	0.43	43.00
words	0.42	42.00	0.44	44.00
LDA char	0.28	28.00	0.30	30.00
Log reg				
char 3-grams	0.44	**45.57**	0.42	43.10
words	0.41	43.20	0.41	43.10
LDA char	0.28	30.63	0.29	30.96
2-layers NN				
char 3-grams	0.42	43.61	**0.47**	**46.86**
words	0.42	42.91	0.43	43.10
LDA char	0.27	29.56	0.27	31.80
3-layers NN				
char 3-grams	**0.49**	41.86	0.40	41.84
words	0.47	42.60	0.40	41.84
LDA char	0.26	31.41	0.30	31.80
FastText				
char 3-grams	0.43	45.00	0.41	42.37
pre-trained chars	0.43	47.00	0.41	41.00
words	0.42	42.56	0.39	44.07
pre-trained words	0.38	41.00	0.40	42.00

Results: far from striking

Main drawbacks: short amount of data in archaic language

Ongoing efforts: Increasing the size of the annotated corpus to afford to apply deep learning effectively

AriEmozione



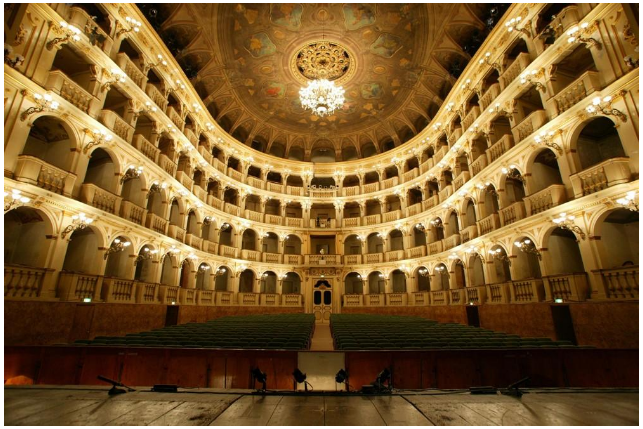
Shibingfeng Zhang

Francesco Fernicola

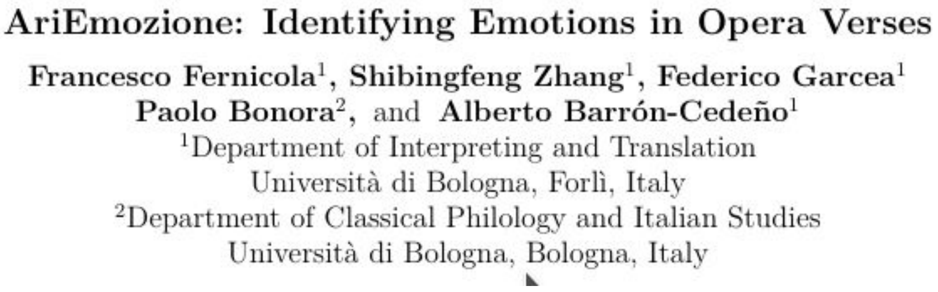
Specialized

Translation

Masters



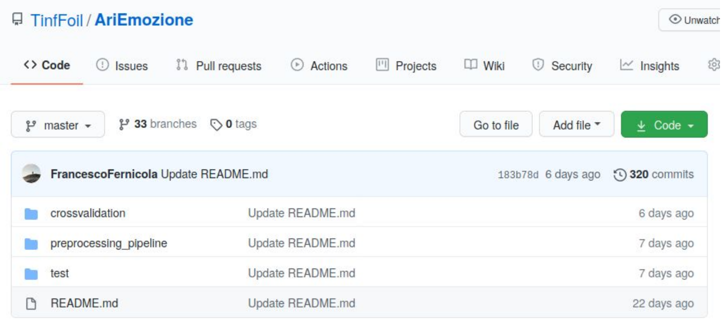
Paper to appear in CLIC-it 2020



AriEmozione: Identifying Emotions in Opera Verses

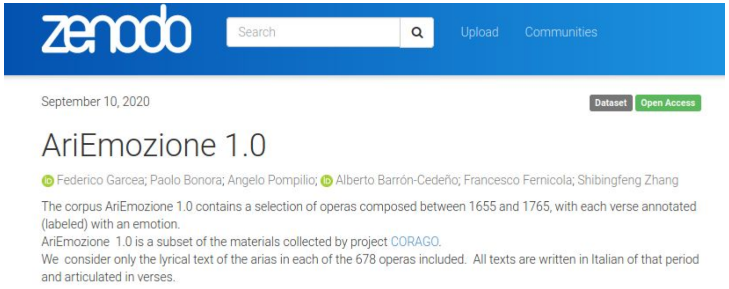
Francesco Fernicola[1], Shibingfeng Zhang[1], Federico Garcea[1]
Paolo Bonora[2], and Alberto Barrón-Cedeño[1]
[1]Department of Interpreting and Translation
Università di Bologna, Forlì, Italy
[2]Department of Classical Philology and Italian Studies
Università di Bologna, Bologna, Italy

Code available on github



FrancescoFernicola Update README.md	183b78d 6 days ago	320 commits
crossvalidation	Update README.md	6 days ago
preprocessing_pipeline	Update README.md	7 days ago
test	Update README.md	7 days ago
README.md	Update README.md	22 days ago

Corpus available on zenodo



September 10, 2020

AriEmozione 1.0

Federico Garcea; Paolo Bonora; Angelo Pompilio; Alberto Barrón-Cedeño; Francesco Fernicola; Shibingfeng Zhang

The corpus AriEmozione 1.0 contains a selection of operas composed between 1655 and 1765, with each verse annotated (labeled) with an emotion.
AriEmozione 1.0 is a subset of the materials collected by project CORAGO.
We consider only the lyrical text of the arias in each of the 678 operas included. All texts are written in Italian of that period and articulated in verses.

UniBO @ AMI 2020

Arianna Muti

Language, Society, and Communication Masters

Objective. Recognize if a tweet is misogynous and, in case of misogyny, if it expresses an aggressive attitude (Task A)

Automatic Misogyny Identification (AMI)

Shared Task at EVALITA 2020

UniBO @ AMI 2020

Two classification tasks

1. Is this tweet misogynous?

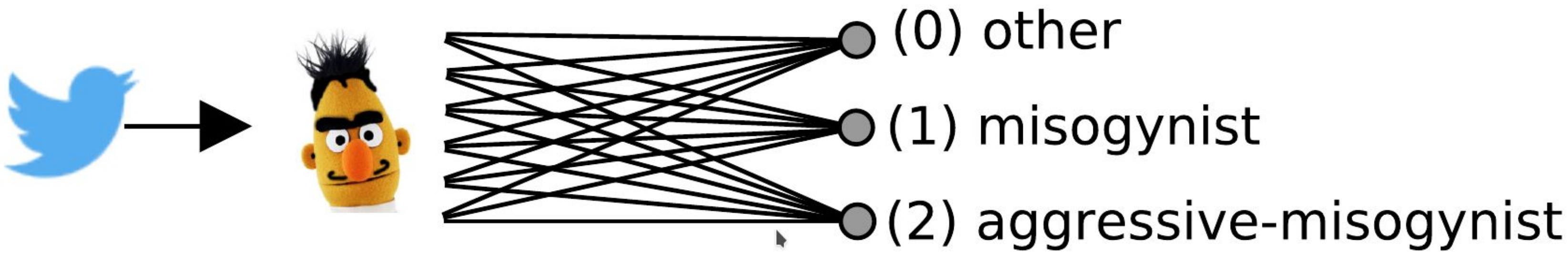
   YES/NO

2. Is this misogynous tweet aggressive?

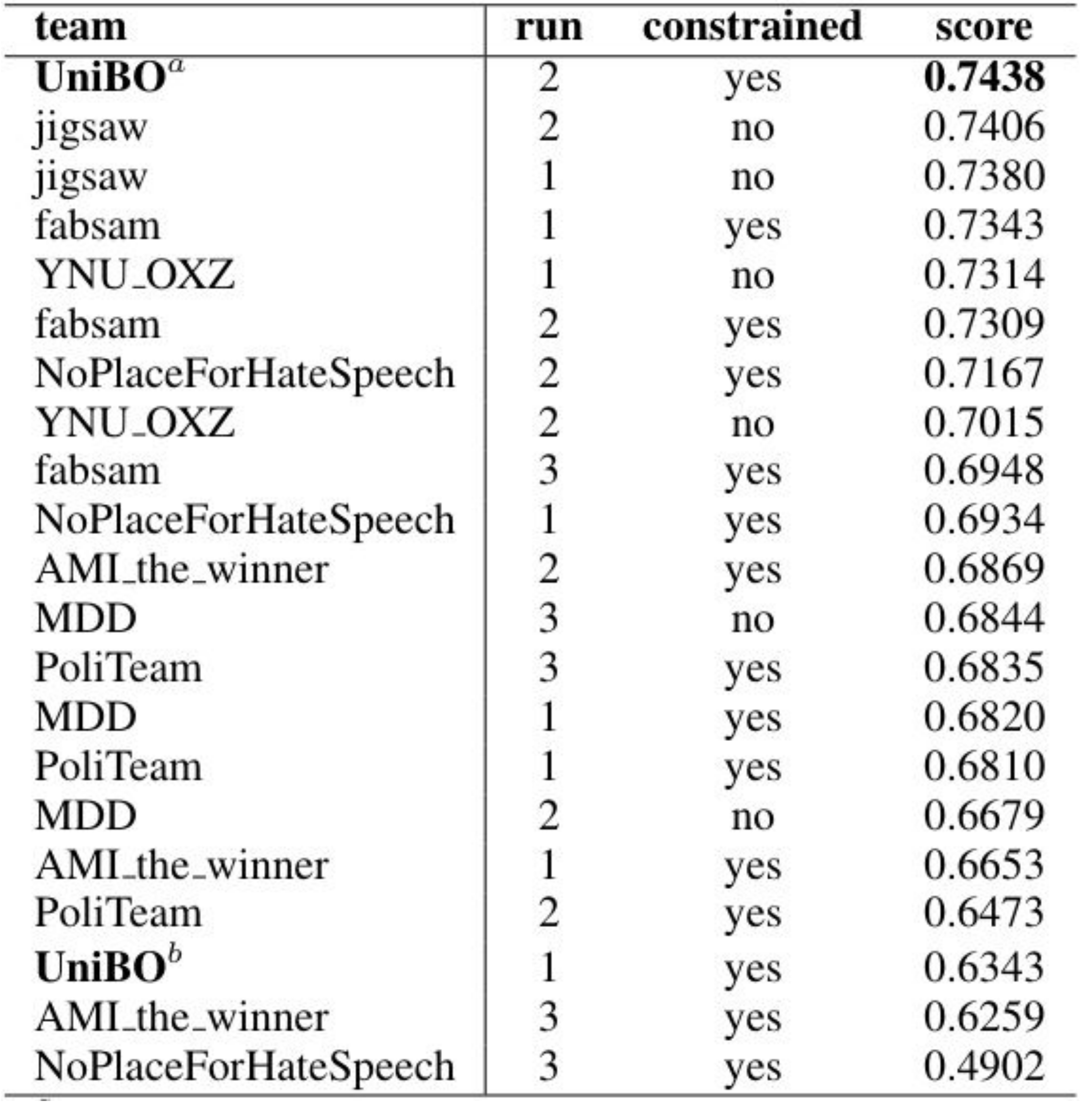
   YES/NO

Our solution: one multi-class network built on top of AlBERTo

(an Italian version of BERT)



UniBO @ AMI 2020



team	run	constrained	score
UniBO[a]	2	yes	**0.7438**
jigsaw	2	no	0.7406
jigsaw	1	no	0.7380
fabsam	1	yes	0.7343
YNU_OXZ	1	no	0.7314
fabsam	2	yes	0.7309
NoPlaceForHateSpeech	2	yes	0.7167
YNU_OXZ	2	no	0.7015
fabsam	3	yes	0.6948
NoPlaceForHateSpeech	1	yes	0.6934
AMI_the_winner	2	yes	0.6869
MDD	3	no	0.6844
PoliTeam	3	yes	0.6835
MDD	1	yes	0.6820
PoliTeam	1	yes	0.6810
MDD	2	no	0.6679
AMI_the_winner	1	yes	0.6653
PoliTeam	2	yes	0.6473
UniBO[b]	1	yes	0.6343
AMI_the_winner	3	yes	0.6259
NoPlaceForHateSpeech	3	yes	0.4902

Misogyny:	0.8102
Aggressiveness:	0.6774

Results: top-performing model

Main drawbacks: still weak against aggressiveness

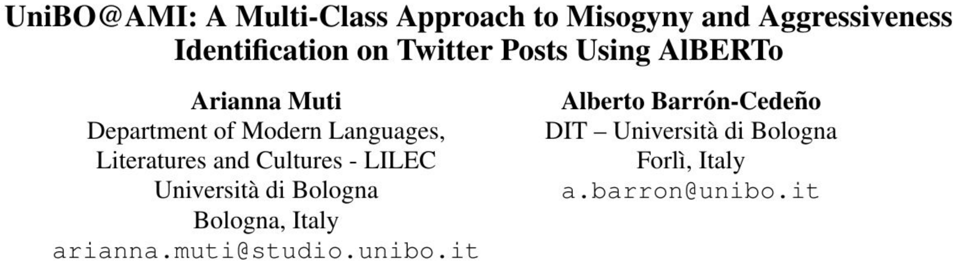
Ongoing efforts: Engineering smarter ways to combine the two decisions

UniBO @ AMI 2020 Task A

Arianna Muti

Language, Society, and Communication Masters

Paper to appear in Evalita 2020

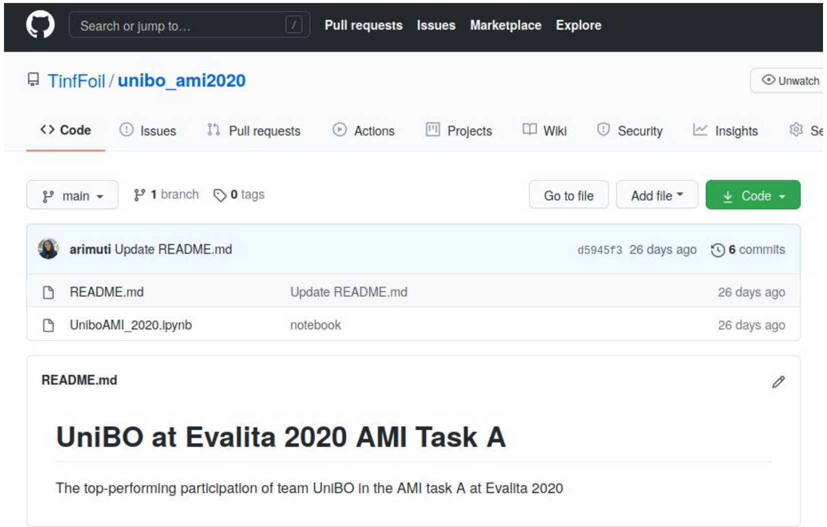


UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo

Arianna Muti
Department of Modern Languages,
Literatures and Cultures - LILEC
Università di Bologna
Bologna, Italy
arianna.muti@studio.unibo.it

Alberto Barrón-Cedeño
DIT – Università di Bologna
Forlì, Italy
a.barron@unibo.it

Code available on github



arimuti Update README.md d5945f3 26 days ago 6 commits

README.md	Update README.md	26 days ago
UniboAMI_2020.ipynb	notebook	26 days ago

README.md

UniBO at Evalita 2020 AMI Task A

The top-performing participation of team UniBO in the AMI task A at Evalita 2020

Are fictional voices *different*?

Ettore Galletti

Specialized Translation Masters

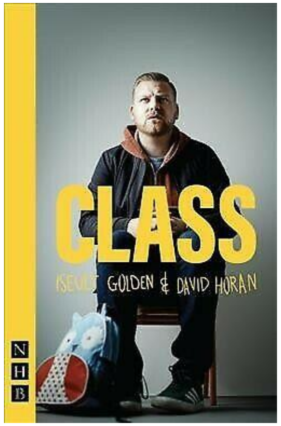
Objective. Finding out if the authors of a specific play managed to create recognisable fixional voices

Are fictional voices *different*?

Identifying characters vs **Identifying groups of characters**

Identifying characters

- McCafferty (teacher)
- Brian (father)
- Donna (mother)
- Jayden (son)
- Kaylie (Jayden's schoolmate)
- None (stage directions)



Identifying groups of characters

- Male
- Female
- None

- Adults
- Kids
- None

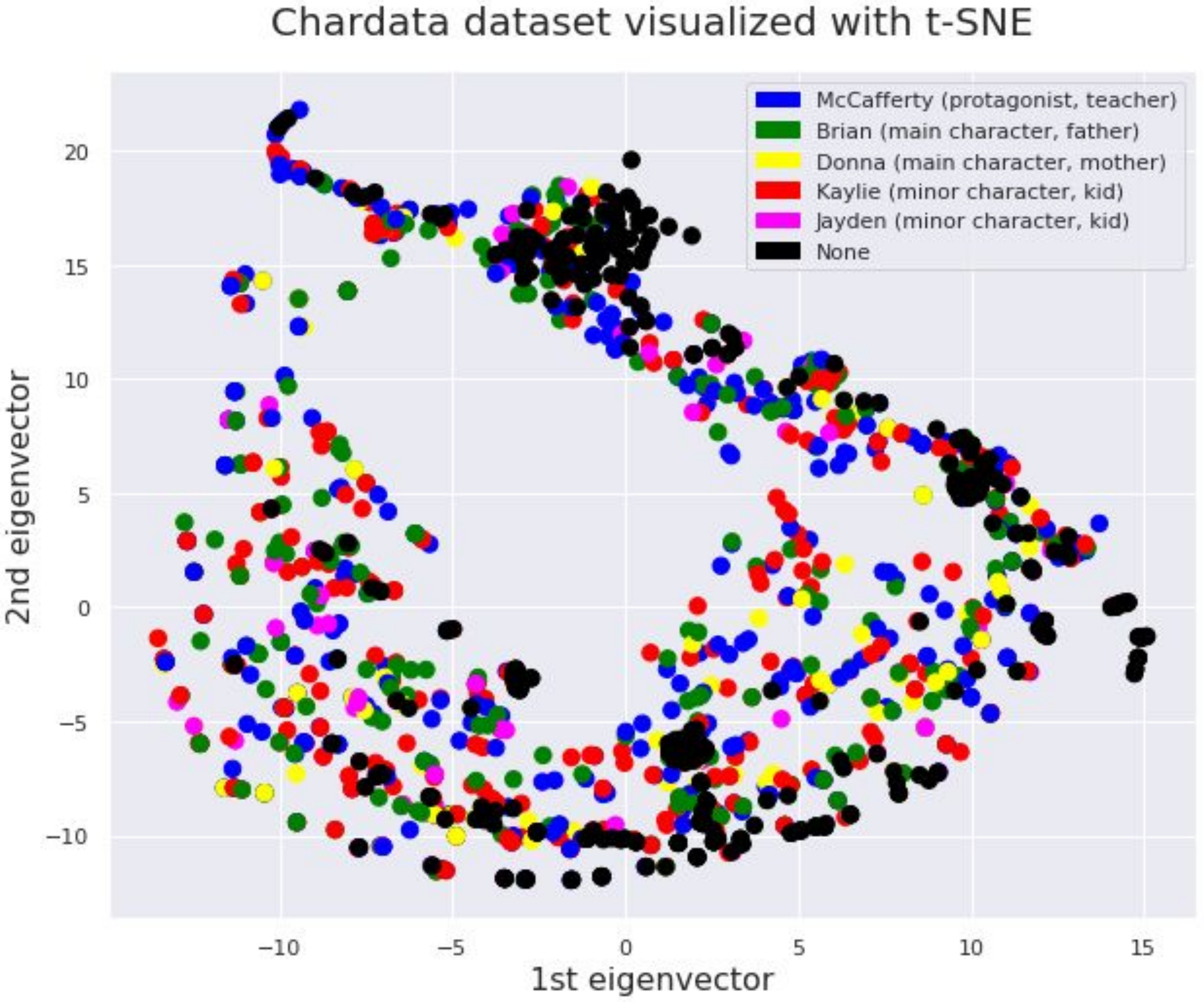
Are fictional voices *different*?

Core idea

1. Build a topic-independent representation of every character intervention
2. Observe if the representations of all interventions make the characters (clearly) differentiable

Are fictional voices *different*?

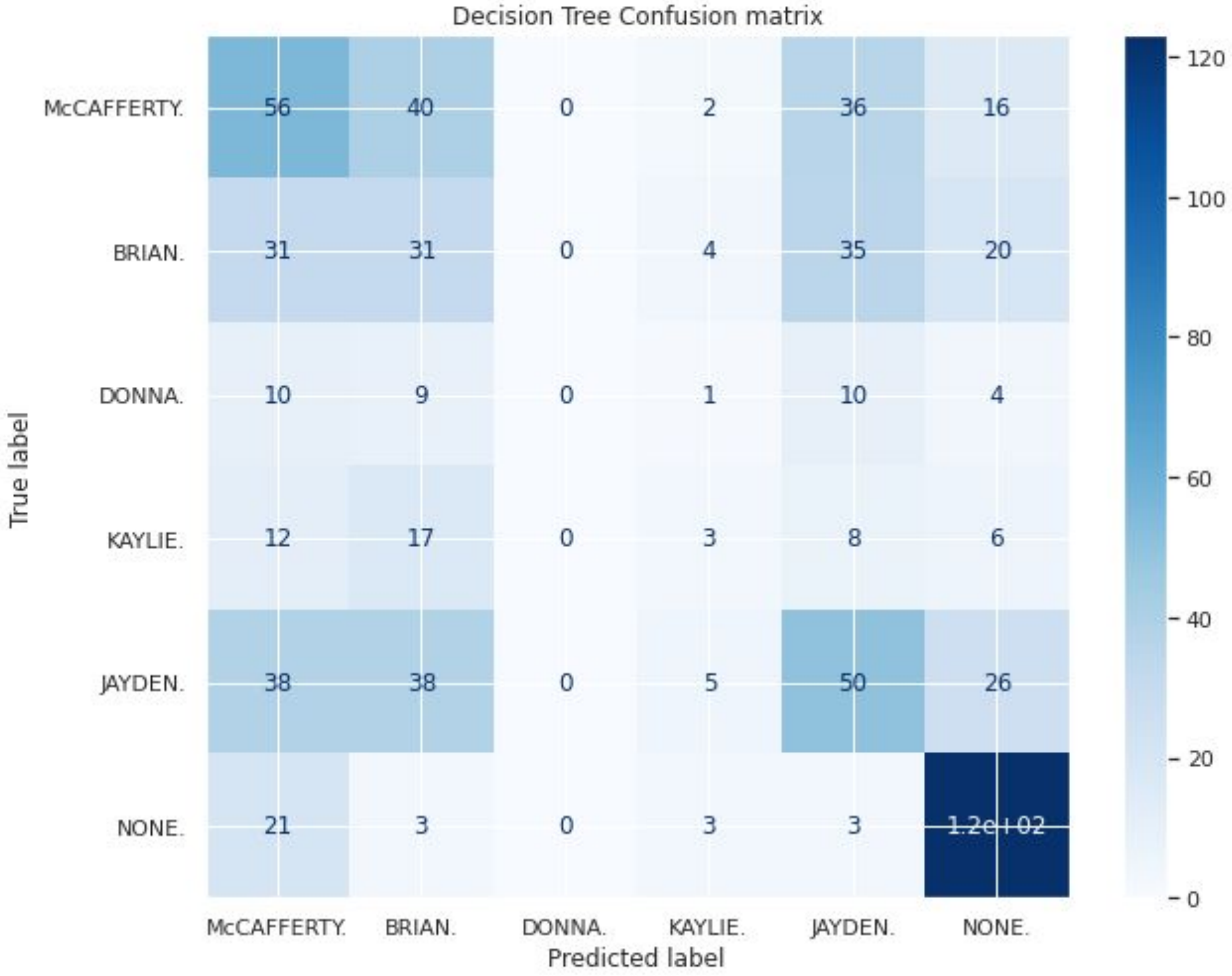
Unsupervised approach. Cluster all the instances and analyse at what extent the clusters correlate with the characters



Chardata dataset visualized with t-SNE

Are fictional voices *different*?

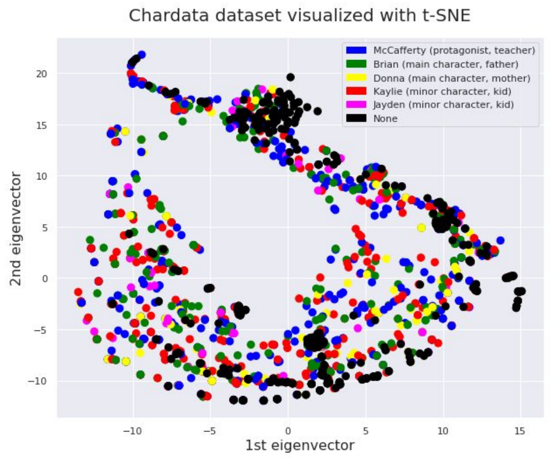
Supervised approach. Build a multi-class character classifier and study whether it manages to label the interventions accurately

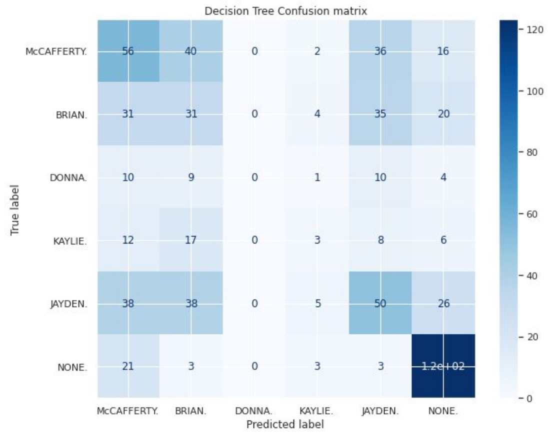


Decision Tree Confusion matrix

True label \ Predicted label	McCAFFERTY.	BRIAN.	DONNA.	KAYLIE.	JAYDEN.	NONE.
McCAFFERTY.	56	40	0	2	36	16
BRIAN.	31	31	0	4	35	20
DONNA.	10	9	0	1	10	4
KAYLIE.	12	17	0	3	8	6
JAYDEN.	38	38	0	5	50	26
NONE.	21	3	0	3	3	1.2e+02

Are fictional voices *different*?





	McCAFFERTY.	BRIAN.	DONNA.	KAYLIE.	JAYDEN.	NONE.
McCAFFERTY.	56	40	0	2	36	16
BRIAN.	31	31	0	4	35	20
DONNA.	10	9	0	1	10	4
KAYLIE.	12	17	0	3	8	6
JAYDEN.	38	38	0	5	50	26
NONE.	21	3	0	3	3	1.2e+02

Results: non-conclusive yet

Main drawbacks: we observe some hints, but we need to study the problem further

Ongoing efforts: Looking if we manage to reproduce/improve the experiments in the (professional) Italian translation of the play

Further perspectives

Projects in earlier stages

Identification of Chinese-oriented hate-speech in COVID-19 tweets

Xin Xin Yu (CL final project) 🇨🇳

Estimating the level of comprehension of texts in French by monolingual native speakers of Italian

Vera Norova Lukina (CL final project) 🇷🇺

Verifying the extent at which people can detect if a text has been machine- or human- translated

Natasha Tatta (Masters thesis at Université de Montréal) 🇨🇦

Projects in earlier stages

Implicit crowdsourcing techniques to produce linguistic resources through language learning

Lavinia Aparaschivei 🇷🇴

PhD thesis co-supervised with Eurac Research

Analysis of bias injected in the translation of news coverage

Natalia Rodriguez Blanco 🇧🇴

PhD thesis (as “computing” advisor)

Efforts without heavy DIT involvement

PROPPY

Automatic identification of propaganda in text

CheckThat!

Automatic prioritisation and verification of claims

Open issues

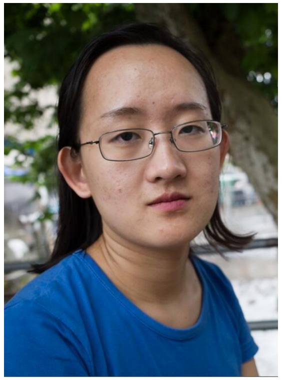
- Bigger load of translation-related topics on top of mono- and cross-language ones
- Creation of online technological demos
- Involvement of students in propaganda and verification efforts
- Further attraction of financing sources (e.g., national and European, private)
- Foster the interdisciplinary research
- Building more links with other academic institutions

Acks

Federico Garcea

Alberto Barrón-Cedeño

Computing Scientists



Shibingfeng Zhang

Francesco Fernicola

Ettore Galletti

Specialized Translation Masters

Arianna Muti

Language, Society, and Communication Masters

Interested, questions?

GET
IN
TOUCH

a.barron@unibo.it

@_albarron_

Thanks!