Une méthode en deux étapes pour la prédiction du niveau d'anglais.

Alexandre Garcia
CAP2018 shared task

English level prediction - Problem

- Goal: predict a « score » (A1,A2,B1,…,C2)
- Inputs: text + grammatical / structural descriptors
- Loss minimization:

Predicted / Real	A1	A2	B1	B2	C1	C2
A1	0	1	2	3	4	6
A2	1	0	1	4	5	8
B1	3	2	0	3	5	8
B2	10	7	5	0	2	7
C1	20	16	12	4	0	8
C2	44	38	32	19	13	0

$$\forall i,j,j' \in \{1,\ldots,6\}; |i-j'| < |i-j| \Rightarrow C_{i,j'} \leq C_{i,j}$$

English level prediction - Solution

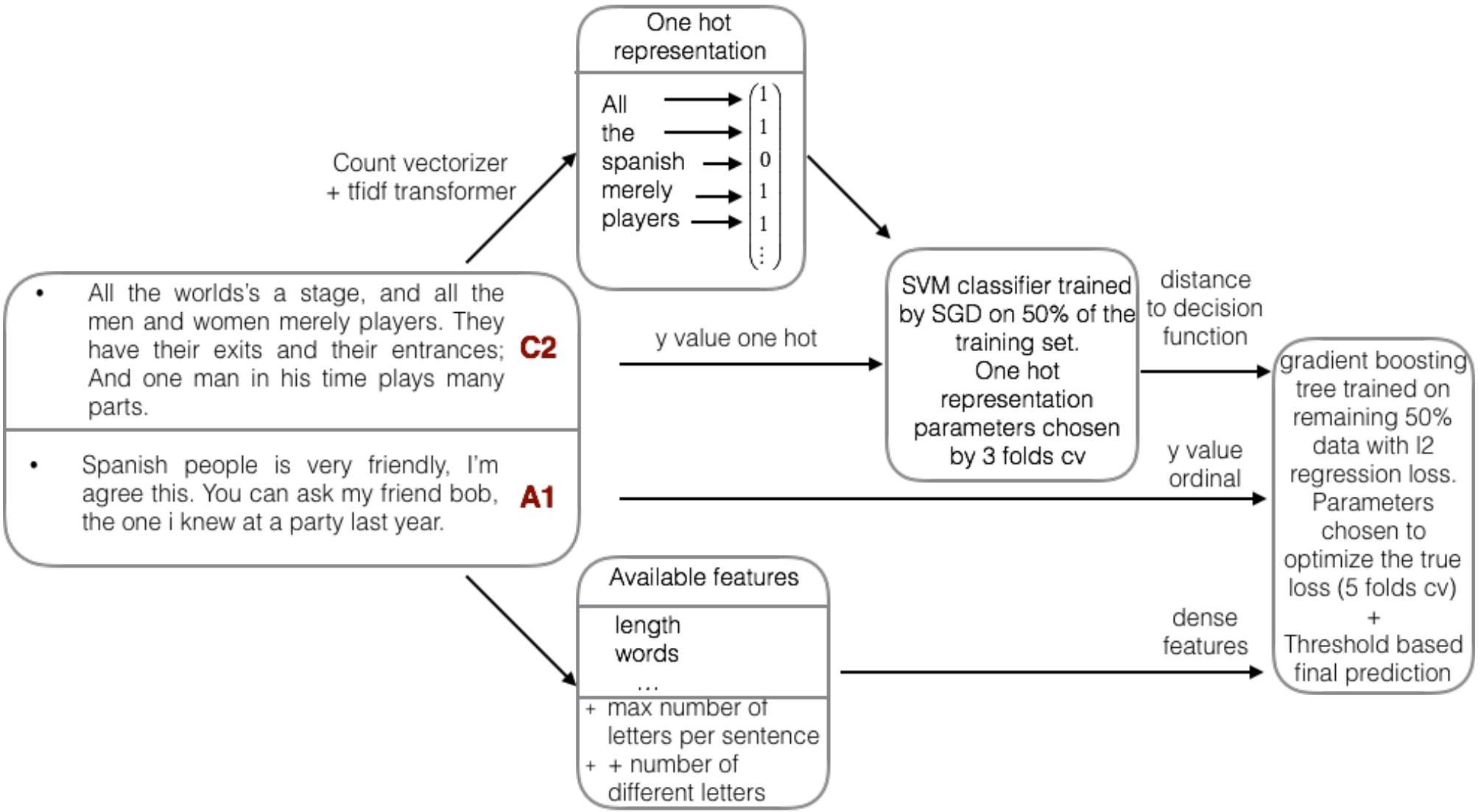
- Loss property: $\forall i, j, j' \in \{1,\ldots,6\}; |i - j'| < |i - j| \Rightarrow C_{i,j'} \leq C_{i,j}$

- Minimize an objective of the form : $\min_f \sum_{k=1}^{n} |y_k - f(x_k)|^2 + \Omega(f)$

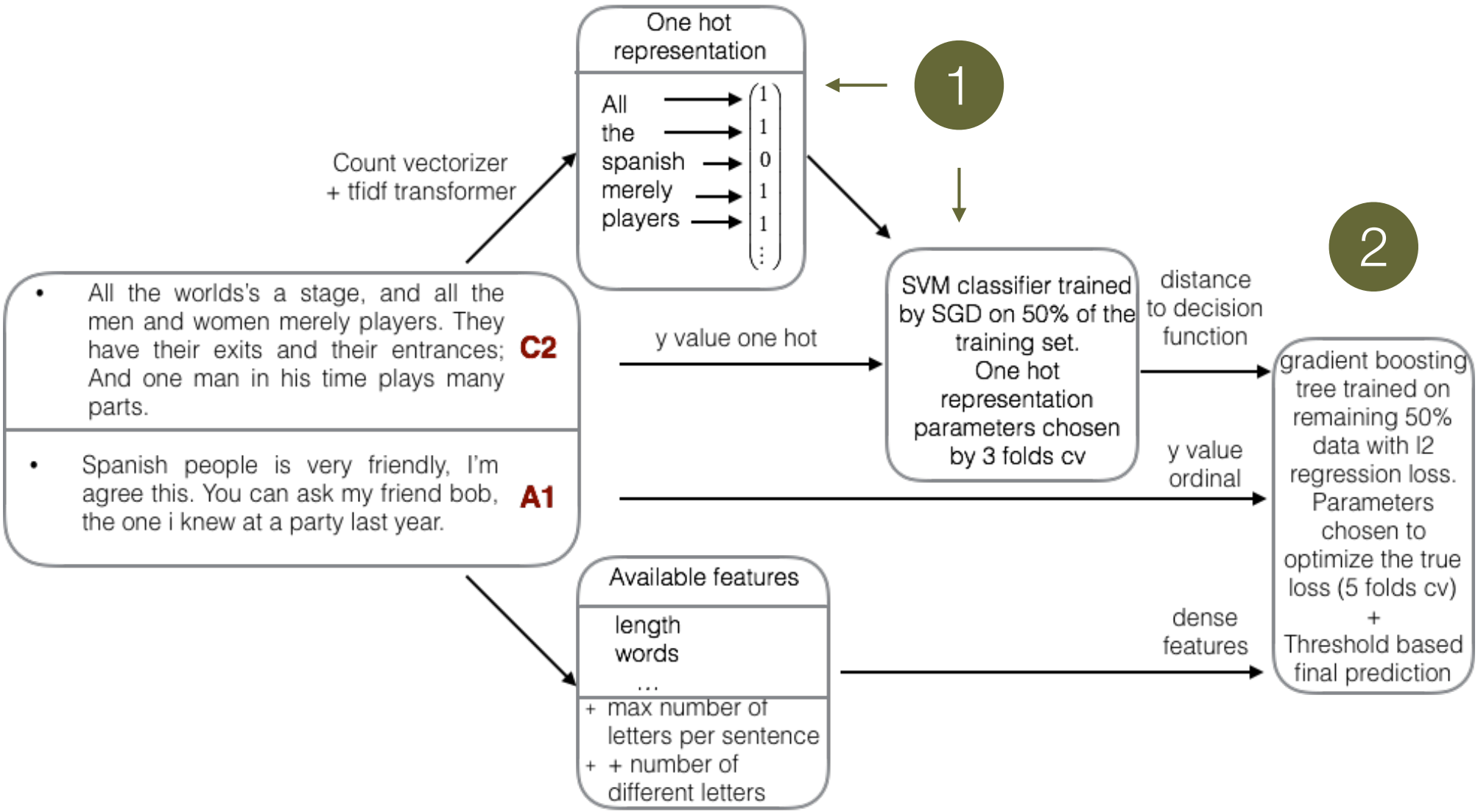
- (Ordinal regression seen as a least square regression problem)

- -> Many predictors available (Random forest, gradient boosting trees, …)

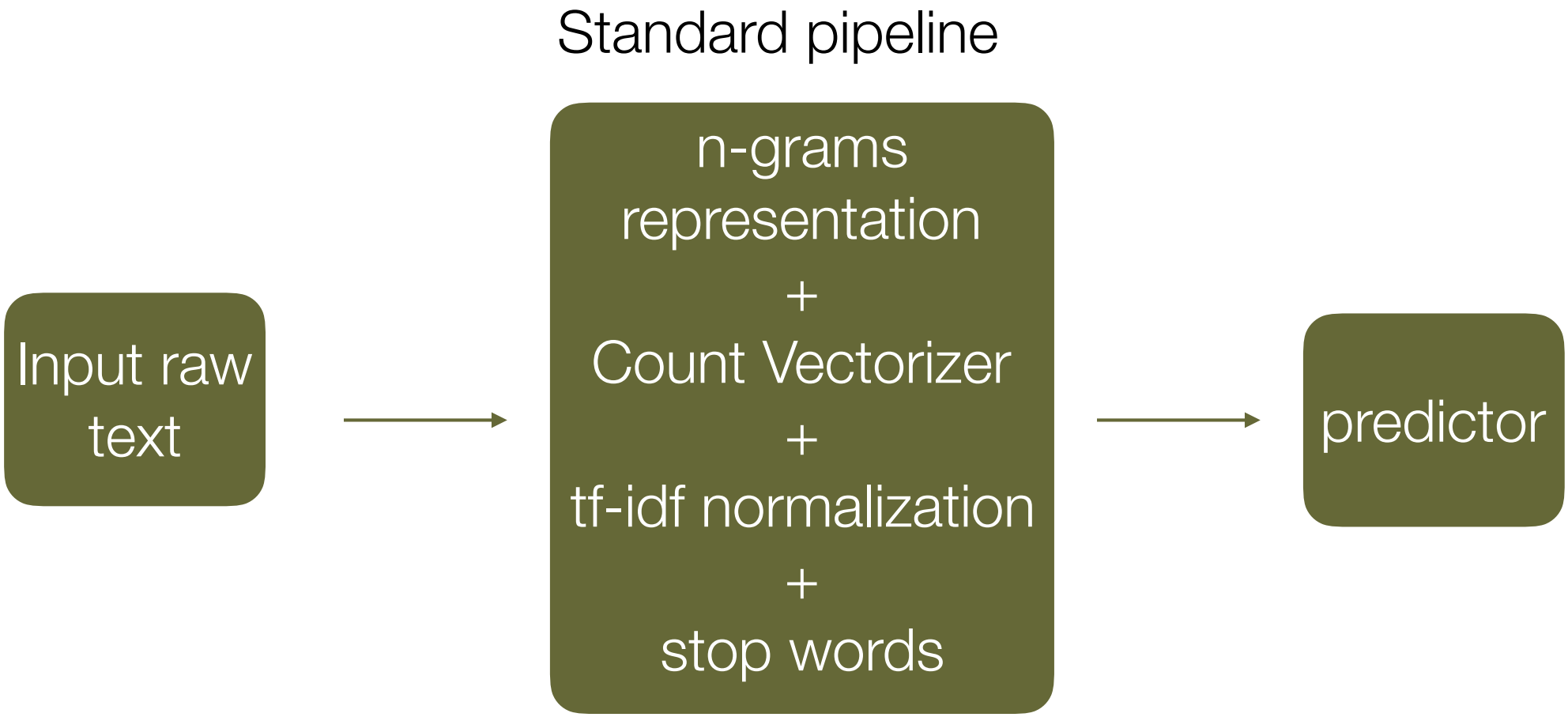
Model structure



Model structure

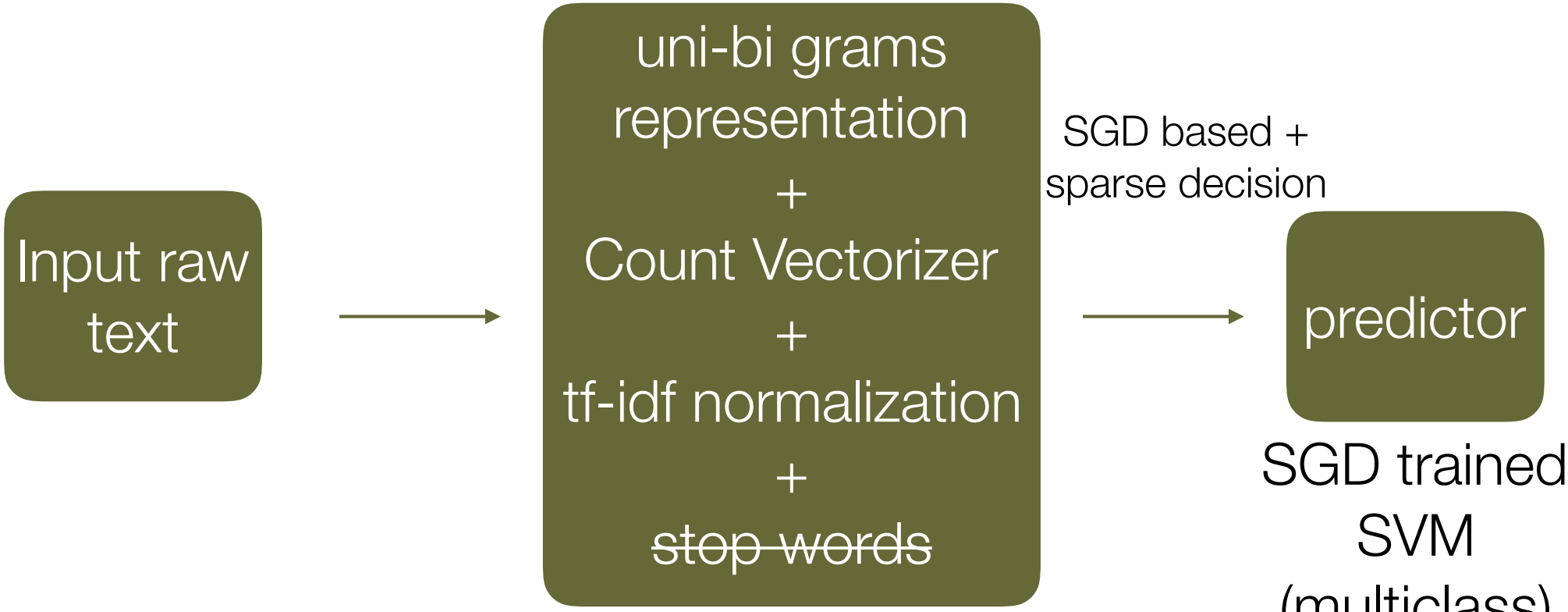


1) Sparse input representation



1) Sparse input representation

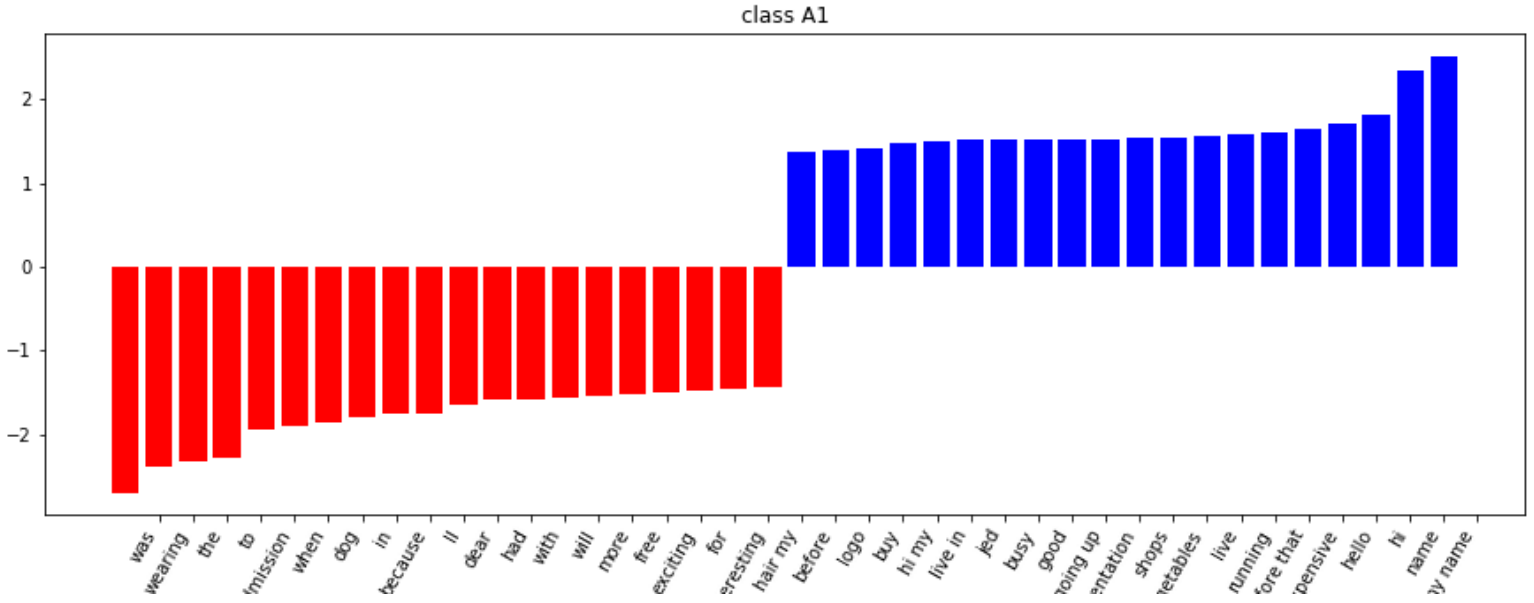
Standard pipeline



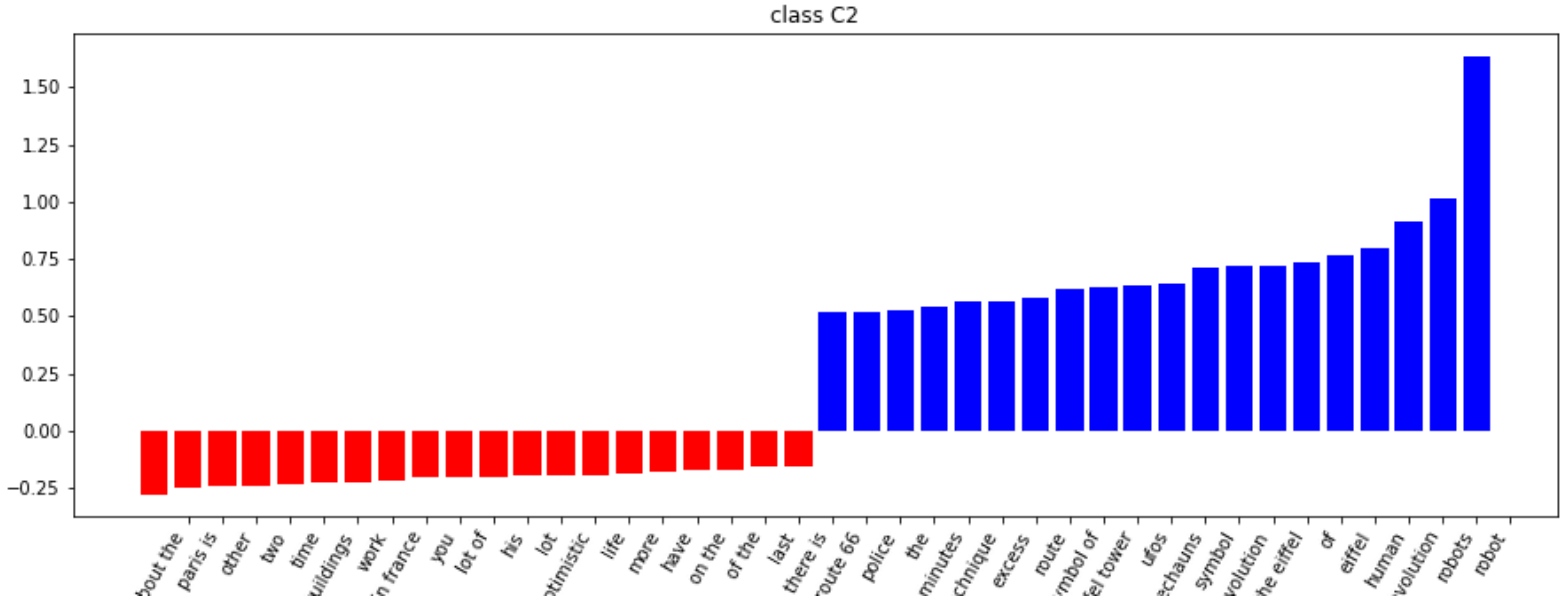
stop words give information on the english level

Top features sparse

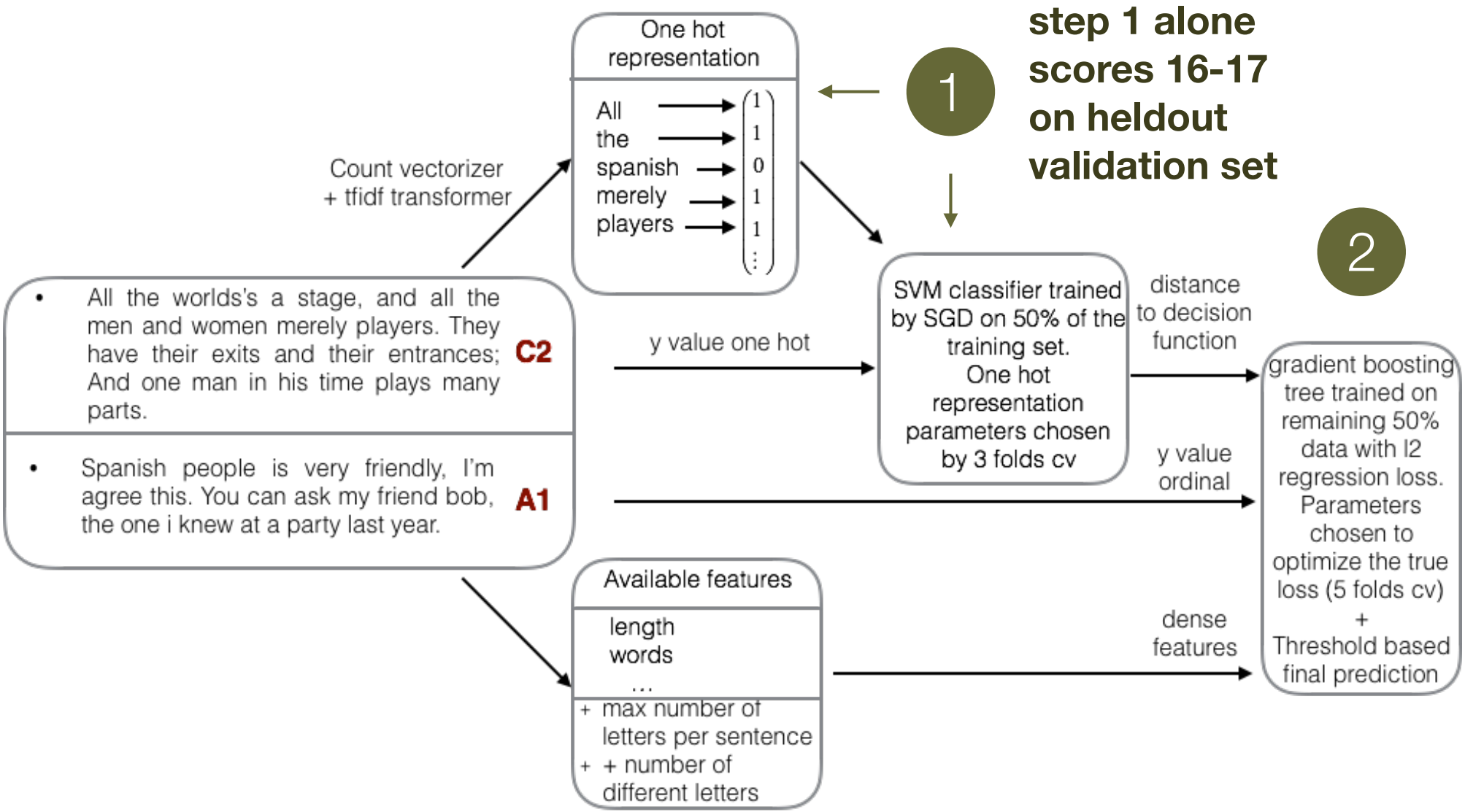
A1 -> name, my name, hi my, live in



C2 -> <u>thematic imbalance ?</u> (robot, humans, symbol, ufos) + (eiffel tower, french revolution)

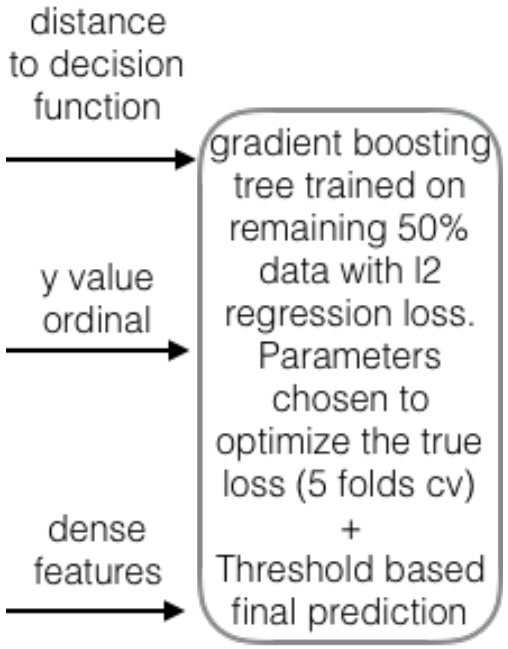


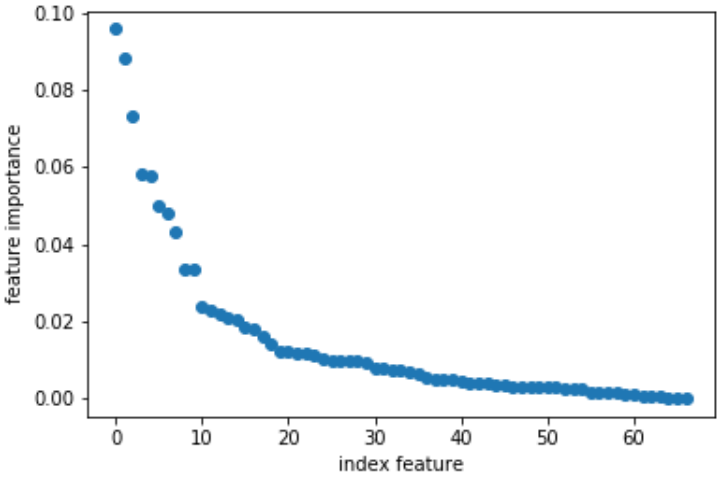
Structure du modèle



1) Dense input fusion

- xgboost implementation (small parameter grid based on depth essentially)
- Features importance (Top 10):
  - Distance to decision function
  - text,sentences,words,letters.all





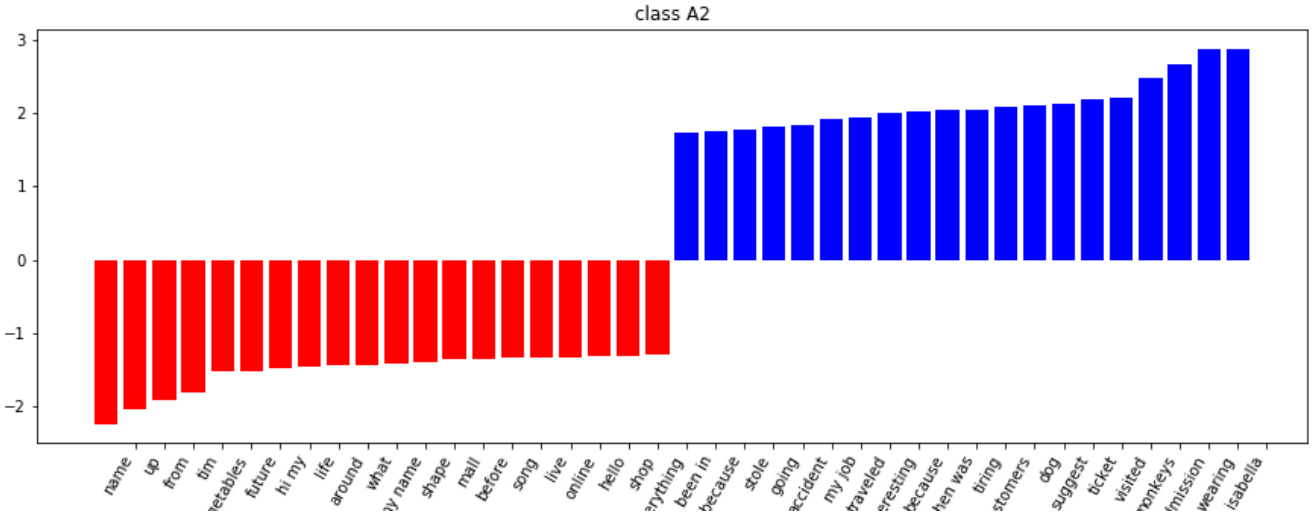
Scores with different feature sets

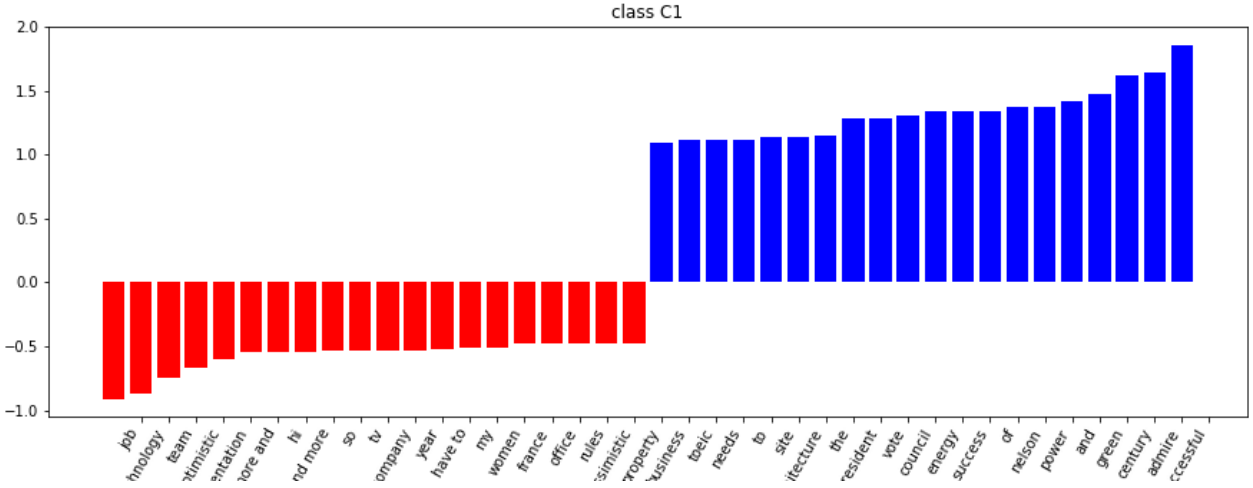
- c.v. scores
  - all features -> 7.8
  - top 10 features -> 7.5
  - sparse only -> 16.1
  - dense only -> 47
- private set scores (averaged 10 times predictions)
  - all features -> 7.28
  - dense only -> 47.23

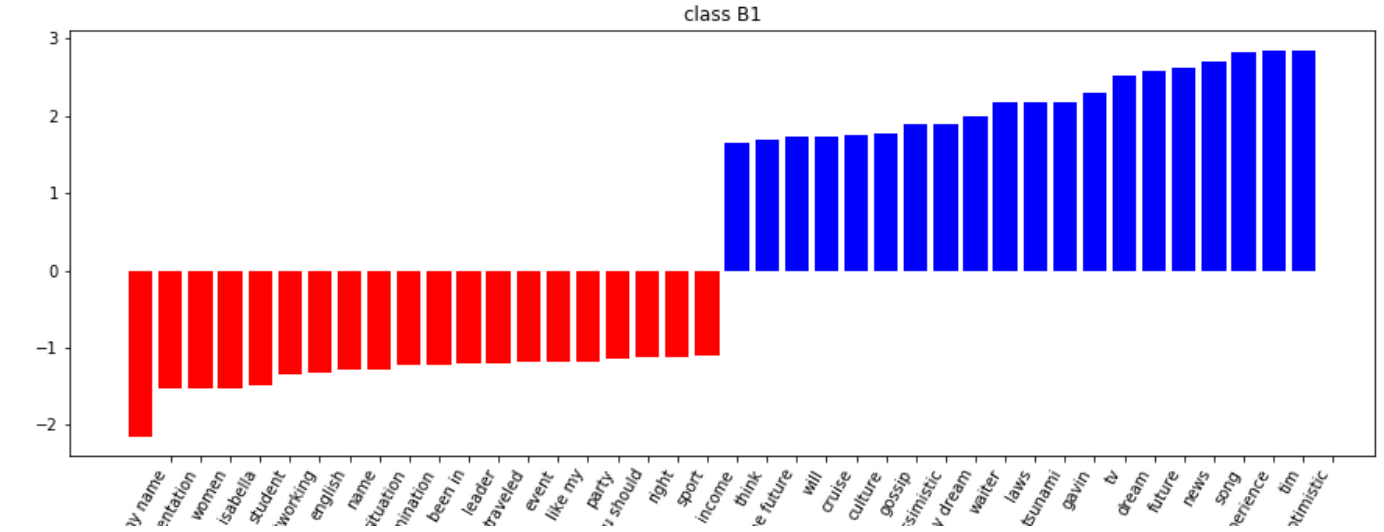
Possible improvements

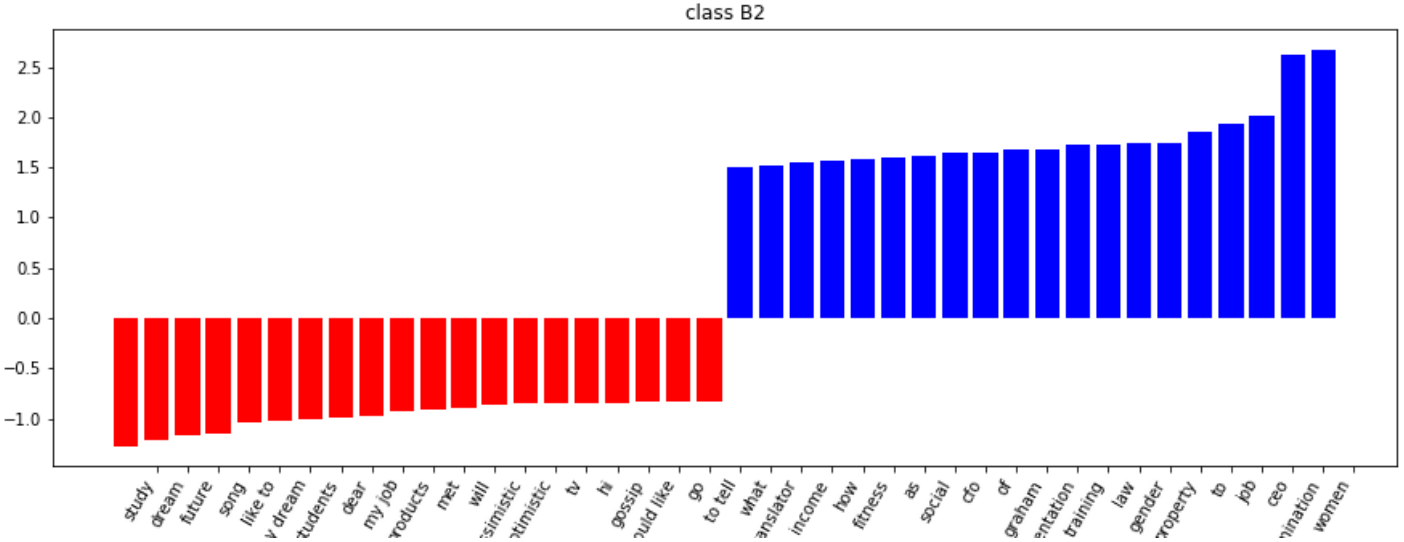
- Orthographic faults dedicated treatment
- Use external ressources (pre-trained word embeddings, pos tagger, …)
- Use word tf-idf based on word occurrences in big corpora with varying english level (Reddit - WIkipedia, 20News)
- Take into account domain knowledge (pre-classification of known low-level english faults.
- Etc …

Questions


class A2
3
2
1
0
−1
−2
name
up
from
tim
future
hi my
life
around
what
shape
mall
before
song
live
online
hello
shop
been in
because
stole
going
my job
because
then was
tiring
dog
suggest
ticket
visited
wearing
isabella


class C1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
have to
my
women
france
office
rules
needs
to
site
the
vote
council
energy
success
of
nelson
power
and
green
century
admire


class B1
3
2
1
0
−1
−2
been in
leader
traveled
event
like my
party
right
sport
income
think
will
cruise
culture
gossip
waiter
laws
tsunami
gavin
tv
dream
future
news
song
tim


class B2
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
study
dream
future
song
like to
dear
met
will
tv
hi
gossip
go
to tell
what
income
how
fitness
as
social
cto
of
training
law
gender
to
job
ceo
women