

Detecció de depressió a les xarxes socials

Martí Caixal i Joaniquet





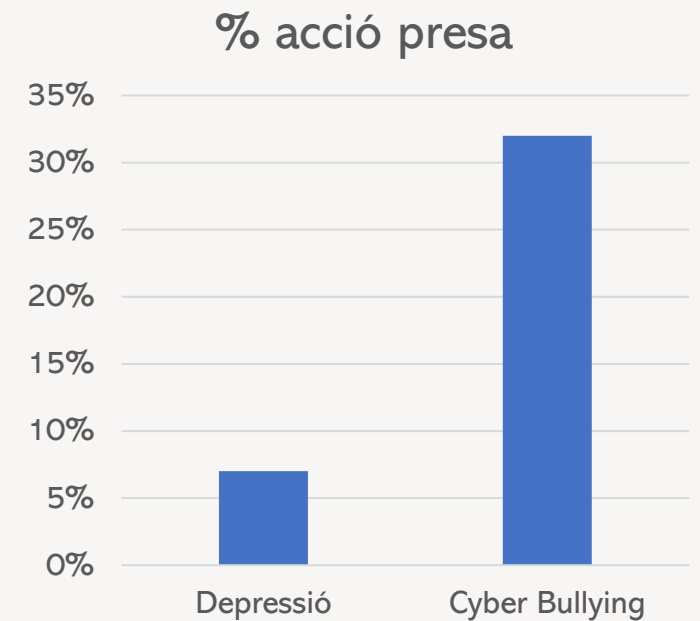
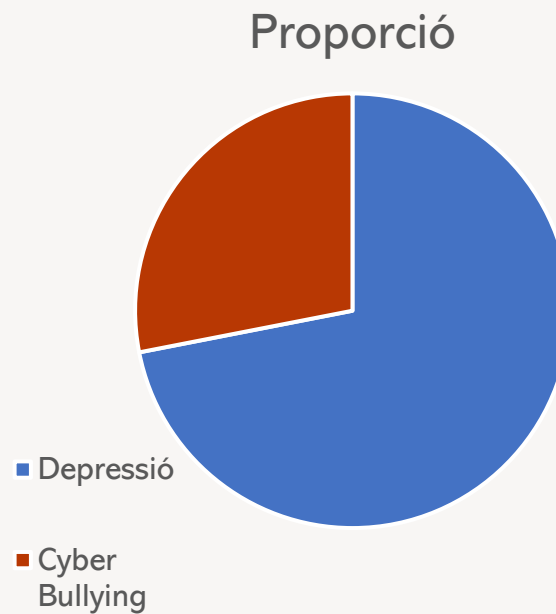
Agenda





Quin problema tenim?

- Més missatges de depressió que d'altres problemes
- Tot i això, reben menys atenció





Problema NLP

Trobar quin mètode dona els millors resultats
i les diferències de comportament entre ells

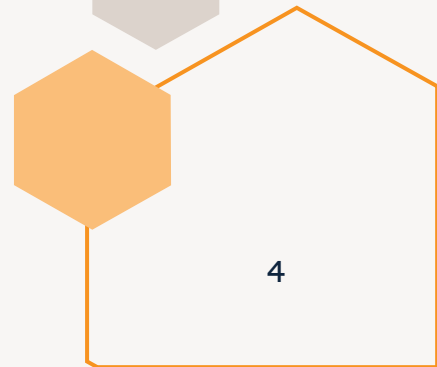
Shallow Learning:

- Naïve Bayes
- Decision Tree
- Random Forest
- SVM
- KNN

- Hyper Parameter Search

Deep Learning

- RNN
 - RNN LSTM
 - RNN GRU
- Transformers (BERT)



Planificació

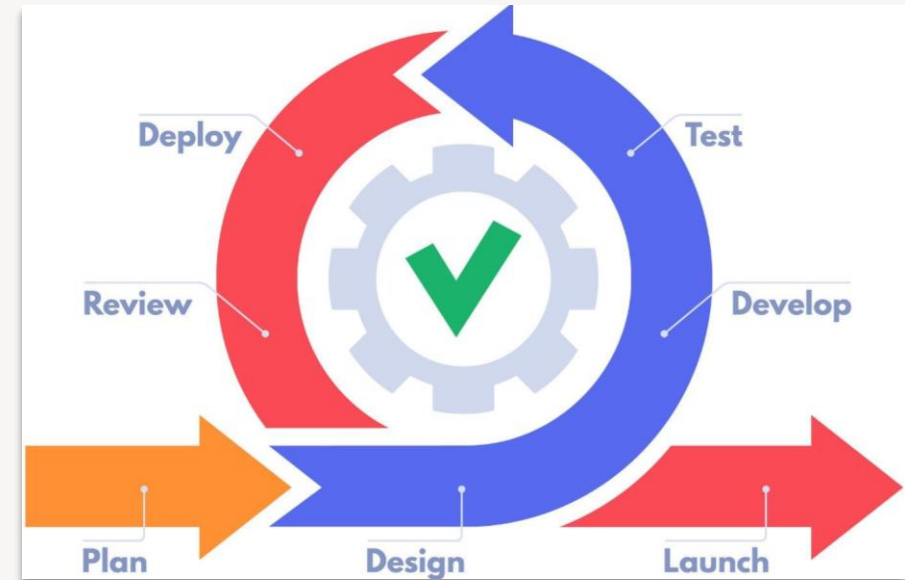
Iteracions curtes

Bon control del ritme

Amigable a canvis

Subobjectius independents

Fàcil detecció d'errors



Metodologia àgil



Dades utilitzades

Mental Health Twitter (Twitter 3)

- Només missatge i classificació

Depression Twitter (Twitter Scale)

- Regressió

Depression Reddit (Reddit)

- Netejat

- 10000 missatges
- 2 classes
- No balancejat (80/20)

*“ @cosmicgirlie Thinking of you.
Everything crossed Turn baby turn! “*





Dades utilitzades

Mental Health Twitter (Twitter 3)

- Només missatge i classificació

Depression Twitter (Twitter Scale)

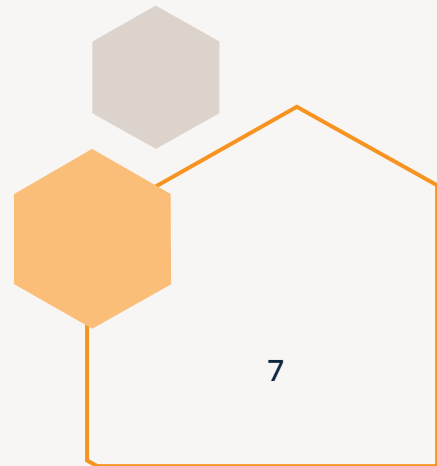
- Regressió

Depression Reddit (Reddit)

- Netejat

- 45000 missatges
- 4 classes (Escala 0 al 3)
- No balancejat (40/20/30/10)

*“ humm dodgers scored a hr stupid
dodgers i hate them”*





Dades utilitzades

Mental Health Twitter (Twitter 3)

- Només missatge i classificació

Depression Twitter (Twitter Scale)

- Regressió

Depression Reddit (Reddit)

- Netejat

- 40000 missatges
- 2 classes
- No balancejat (60/40)
- Already cleaned

“ i used to be highly functional before but it now i can barely function at all i take everything just...”





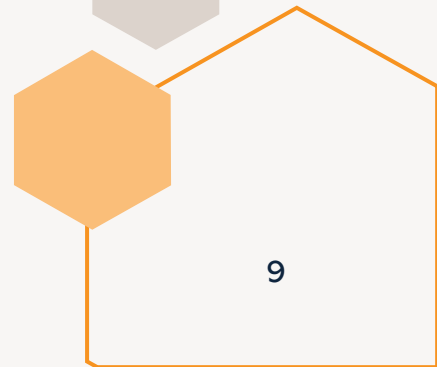
Dades utilitzades

No balançat,
classe objectiu és minoritària:

- × Undersampling
- × Oversampling



- Recall en comptes de accuracy
- Macro average





Preprocessament inicial

Eliminar noms d'usuari

Eliminar Stop Words

Eliminar números

Lemmanizació

Eliminar puntuació



Enfocaments específics

Shallow Learning

- Bag of Words
- TF-IDF

Deep Learning

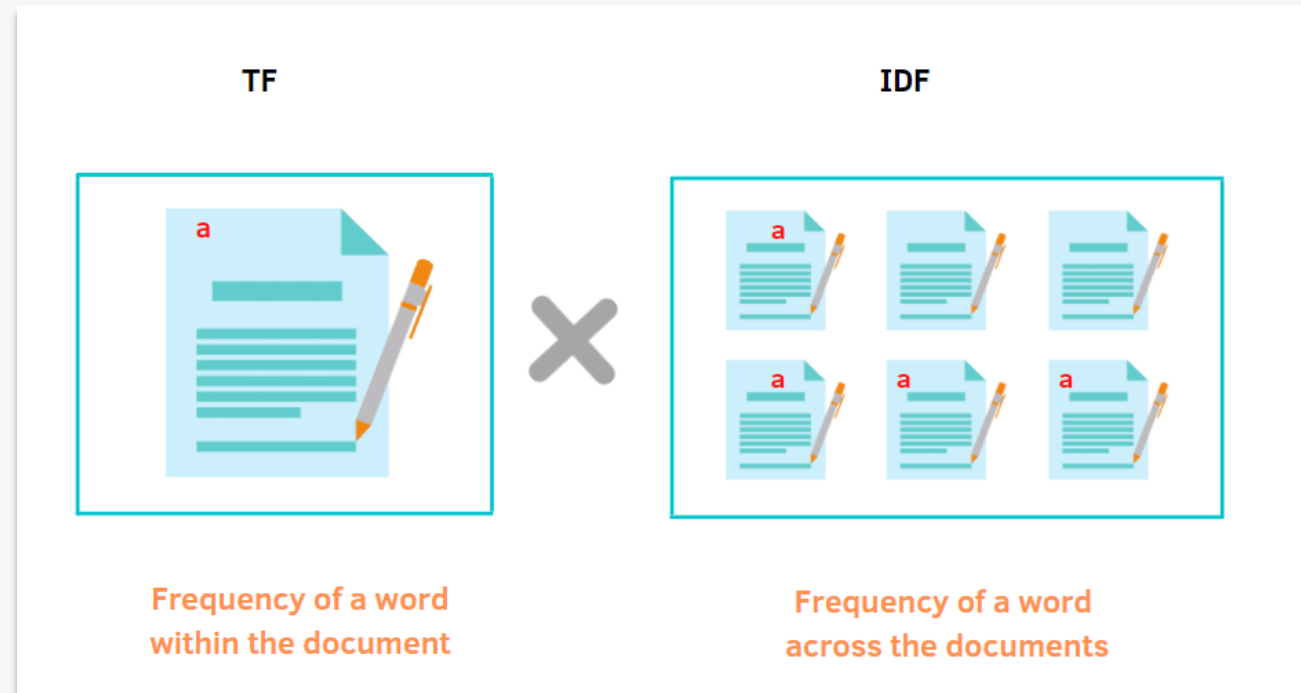
- Word Embedding (GloVe)



Bag of Words

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

TF-IDF



Resultats shallow learning

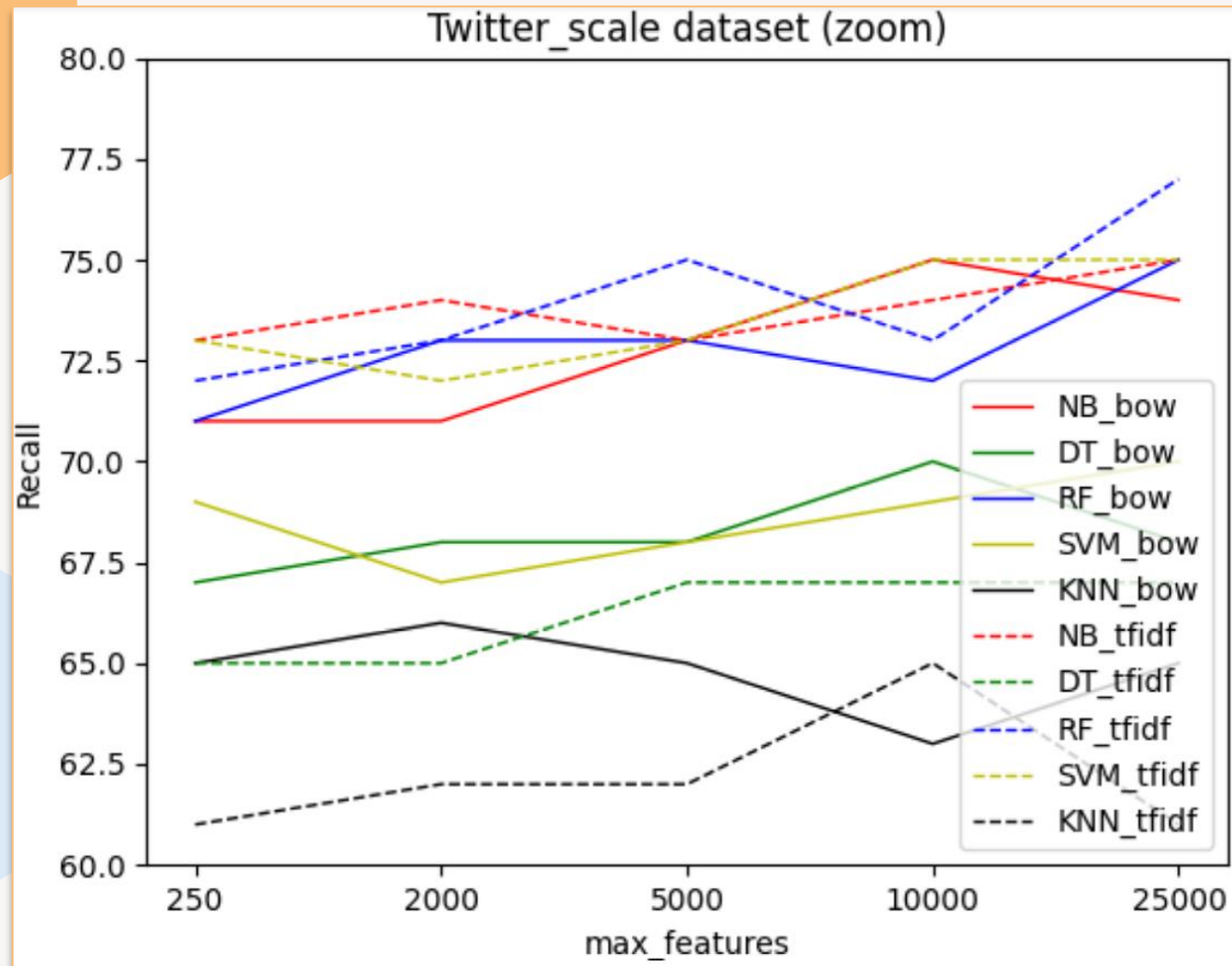
- Scikit Learn library
- Paràmetres inicials

- Naïve Bayes
- Decision Tree
- Random Forest
- SVM
- KNN

- Hyperparameter Search



TF-IDF vs BoW & feature size



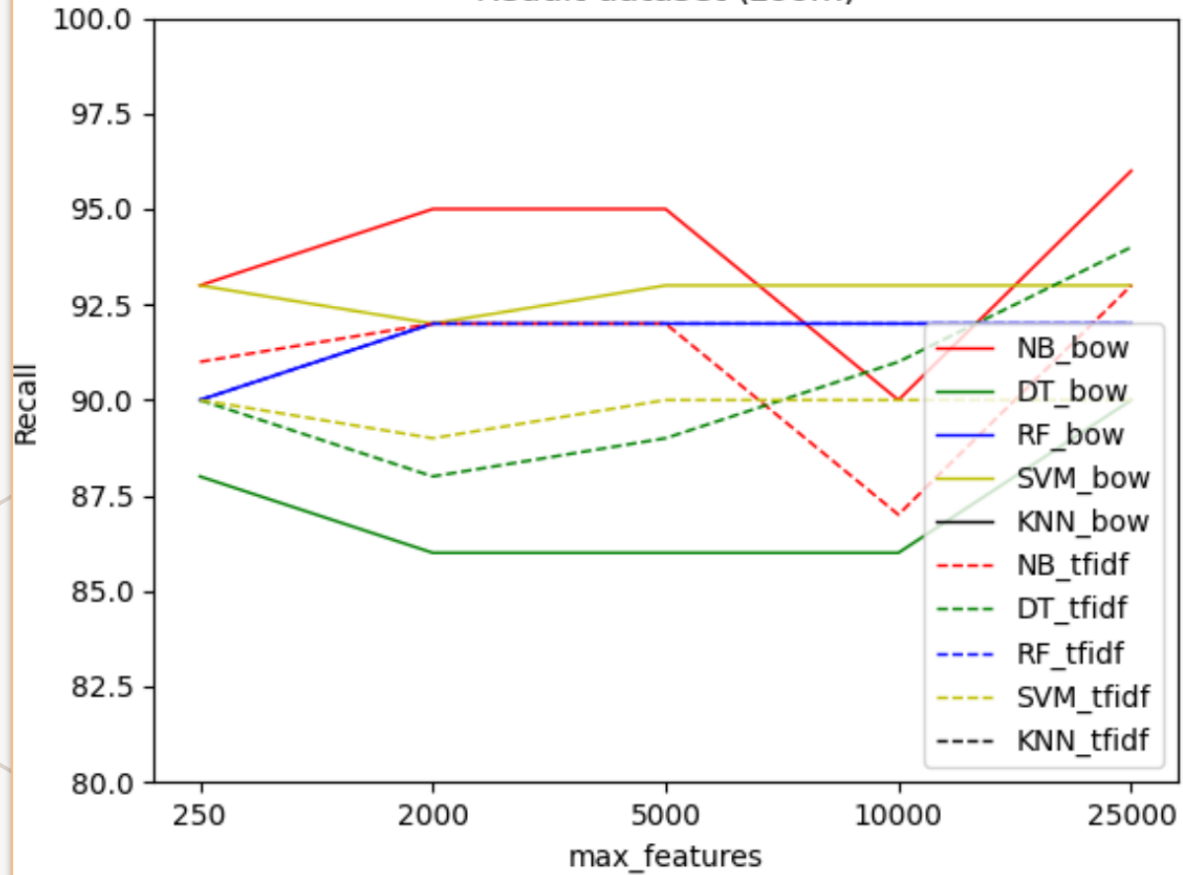
TF-IDF - - - -

BoW ————

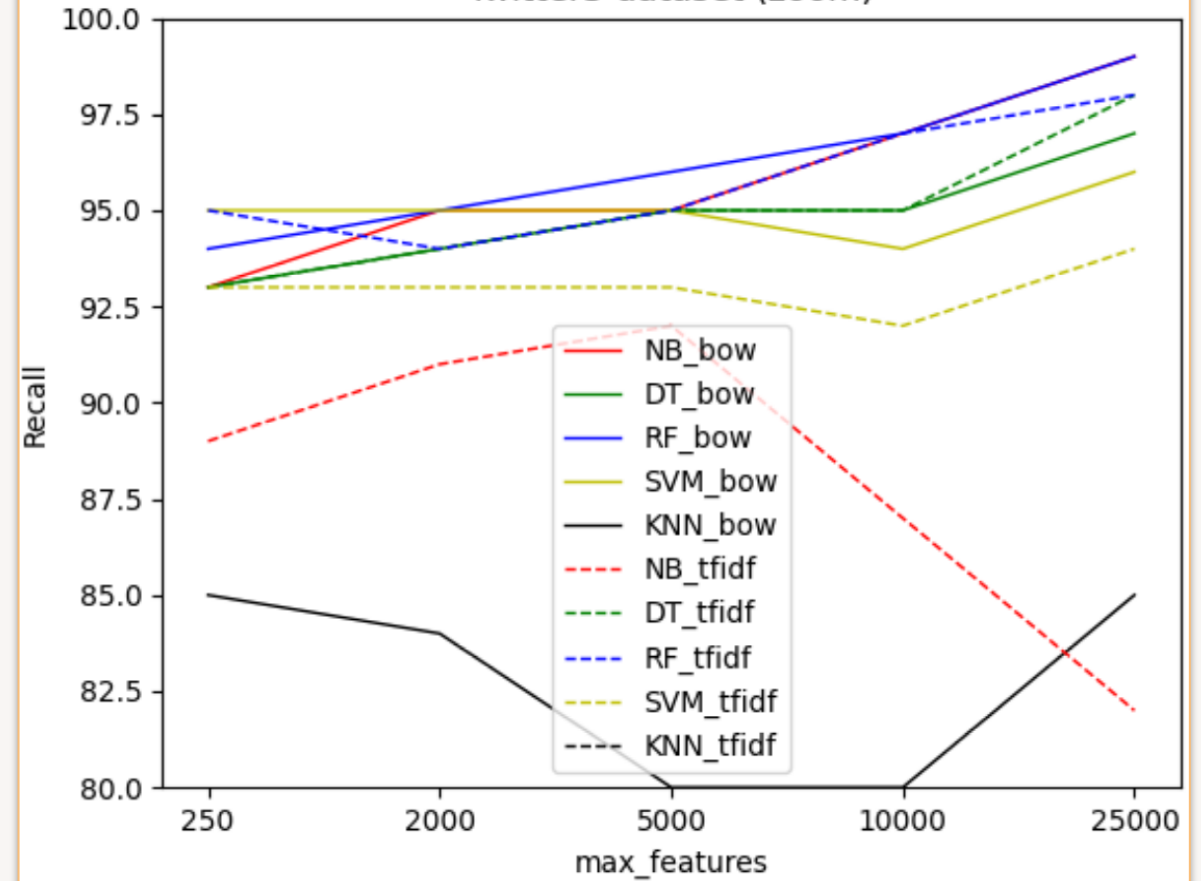
- ✓ TF-IDF lleugerament millor
- ✓ Augmentar el número de “features” millora lleugerament el resultat

Resultats

Reddit dataset (zoom)

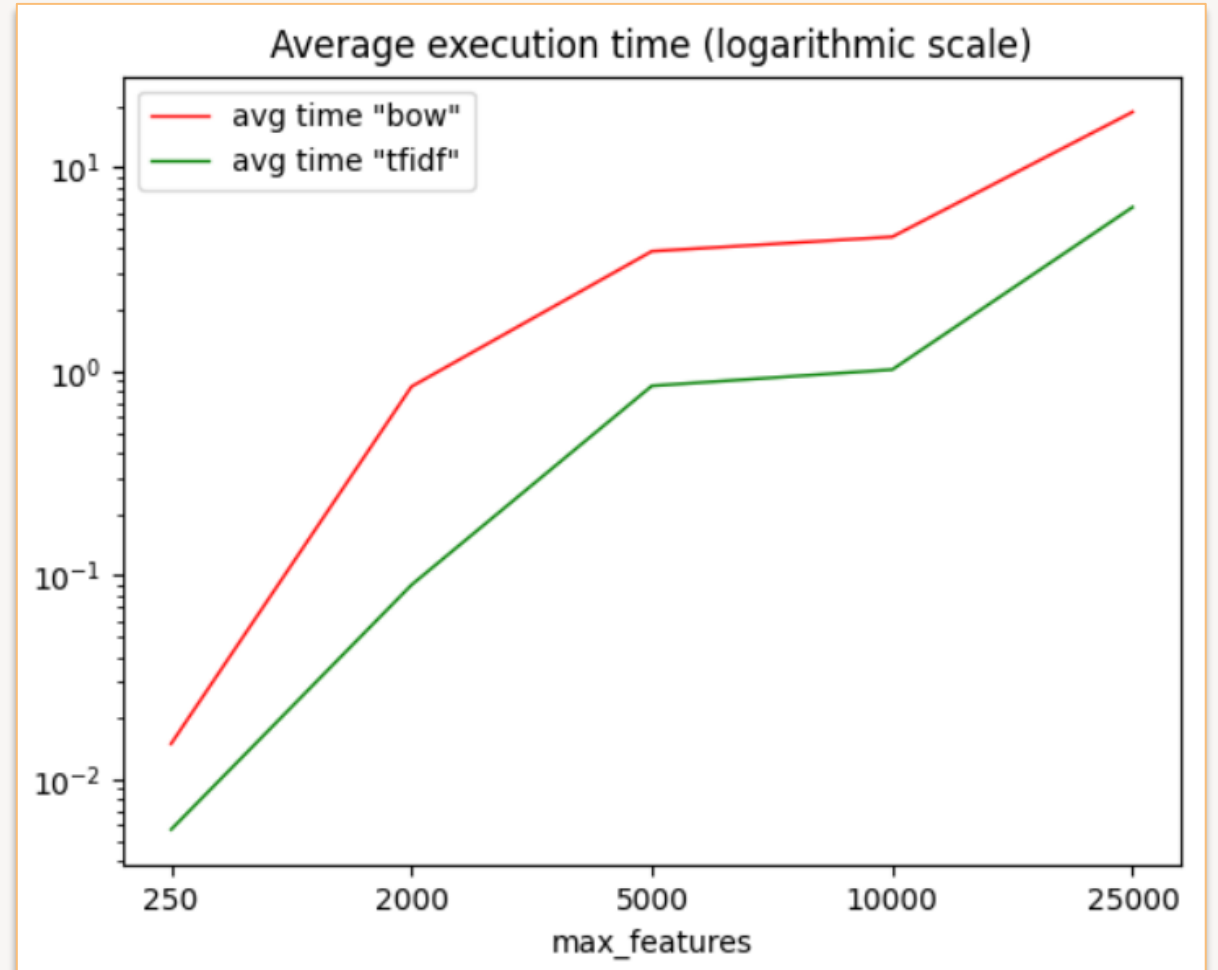


Twitter3 dataset (zoom)

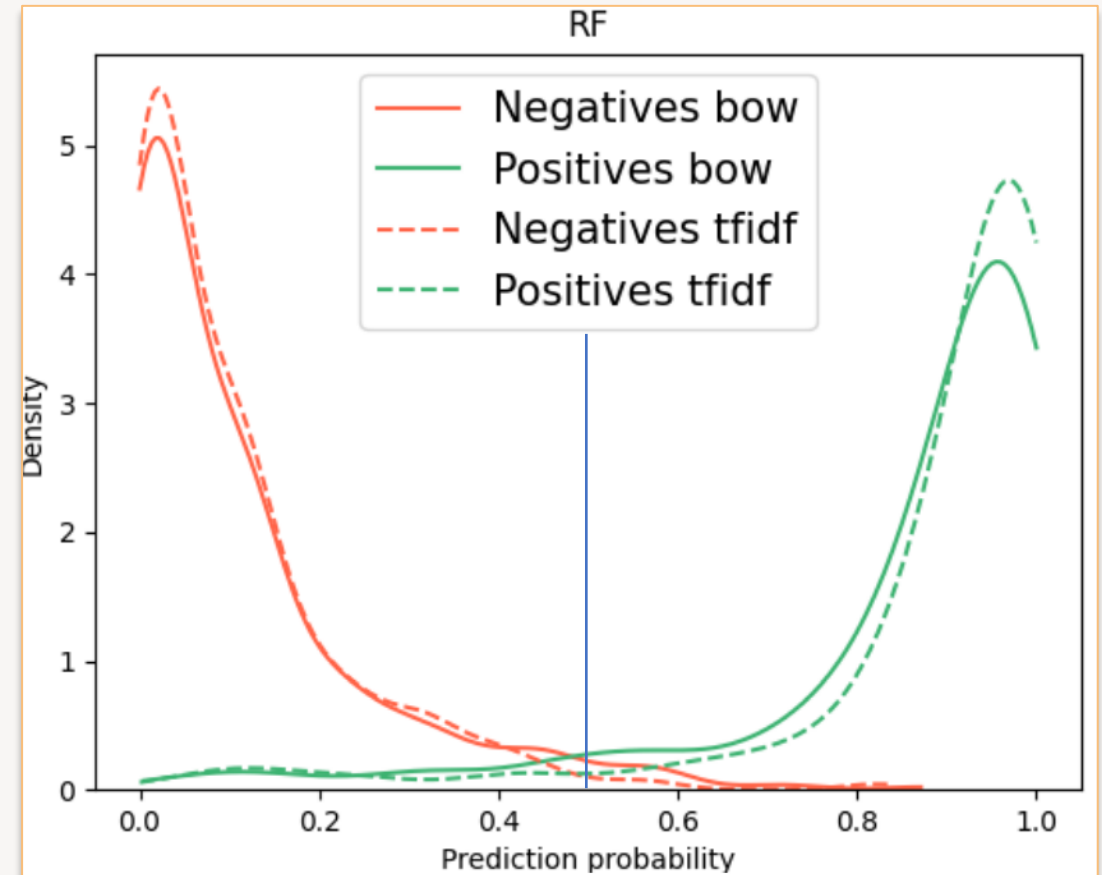
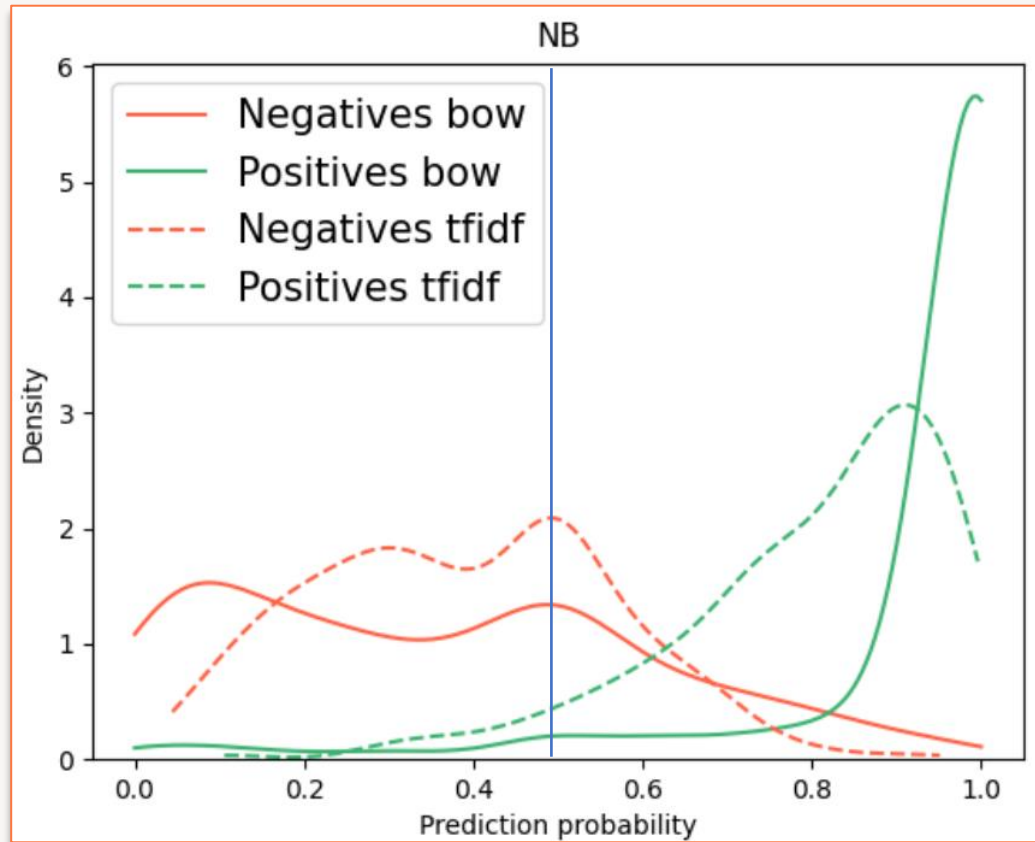


Temps d'execució

- ✓ TF-IDF lleugerament millor
- ✓ Augmentar el número de "features" millora lleugerament el resultat
- ✓ Millors temps



Confiança en les prediccions





Hyperparameter search

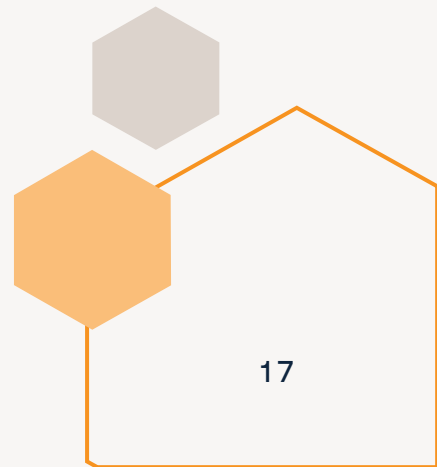
Fet amb Optuna:

- Python library
- Cerca optimitzada
- Parallelization

Model	Nº paràmetres	execucions
SVM	3	10^3
KNN	3	10^3
DT	4	8^4
RF	4	8^4
NB	1	100^1

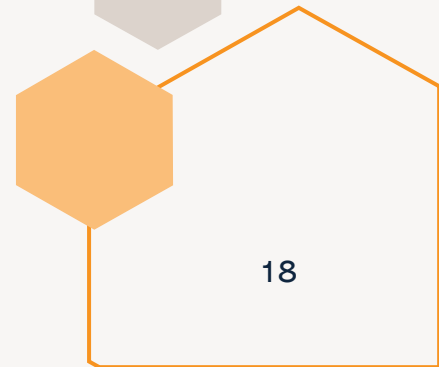
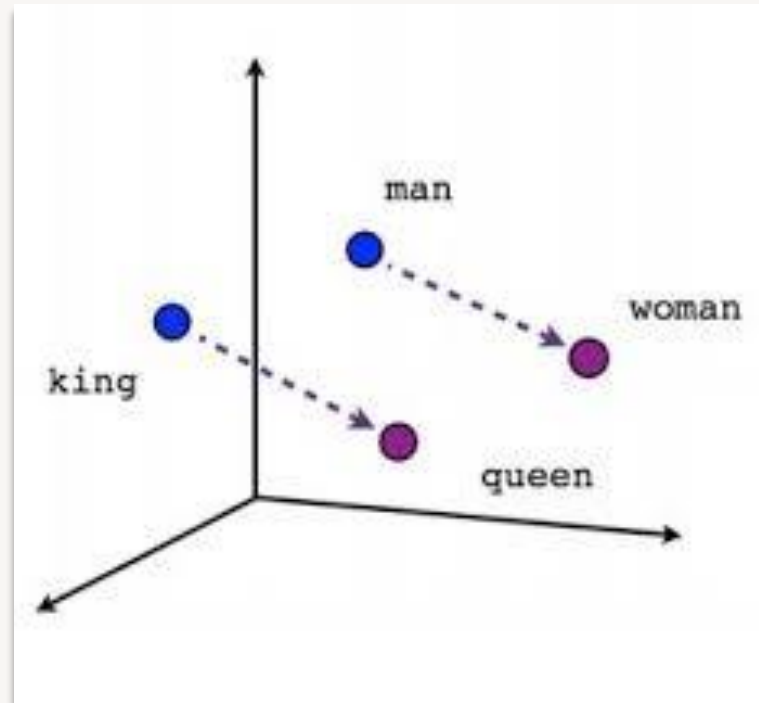


× Els resultas
no milloren





Word Vectoring





Resultats deep learnig

- Keras (Python)
- RTX 3070 Ti

- RNN
- RNN GRU
- RNN LSTM
- BERT



RNN

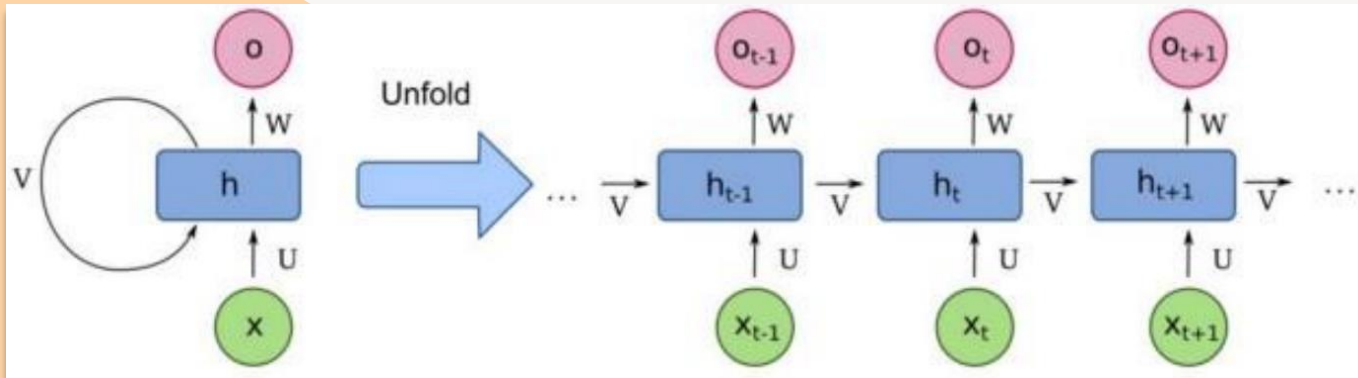
- Seqüència de capes
- Input, funció d'activació, output
- Sense memòria

RNN LSTM

- 3 portes
- Memòria

RNN GRU

- 2 portes
- Memòria
- LSTM simplificada



RNN

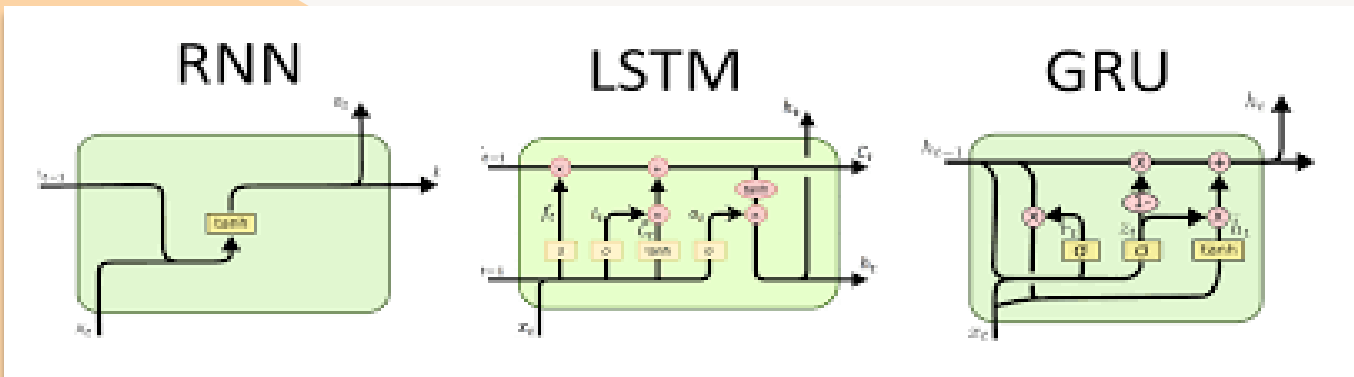
- Seqüència de capes
- Input, funció d'activació, output
- Sense memòria

RNN LSTM

- 3 portes
- Memòria

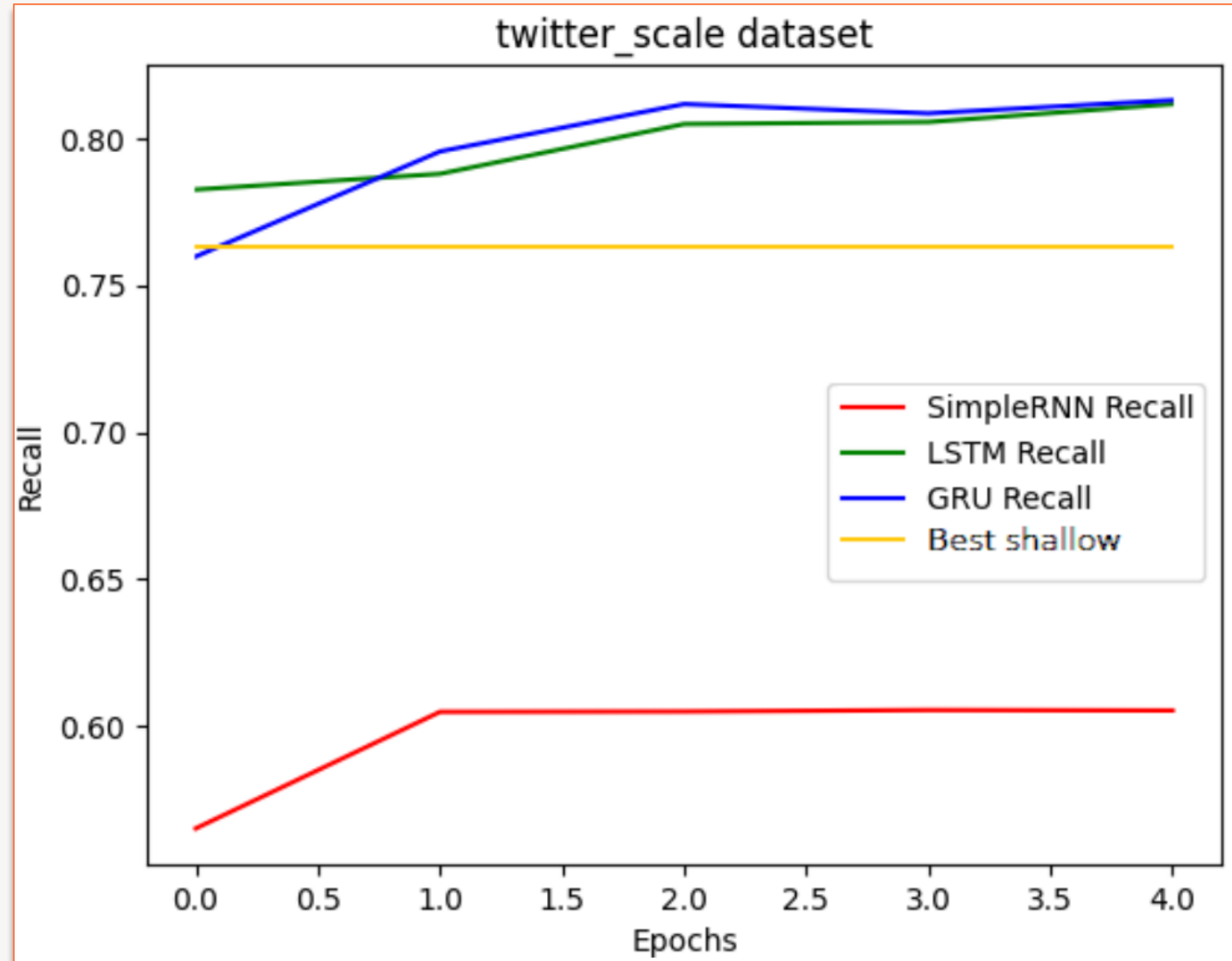
RNN GRU

- 2 portes
- Memòria
- LSTM simplificada





Resultats

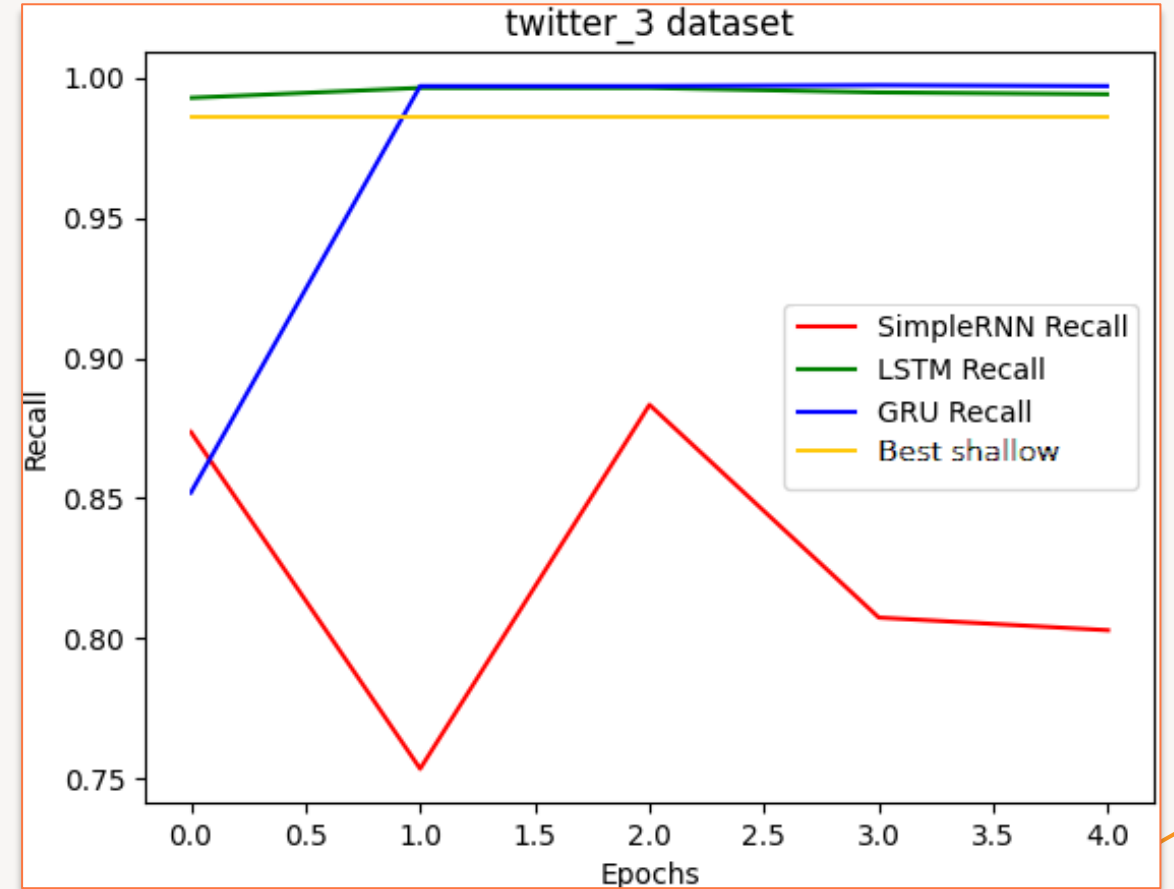
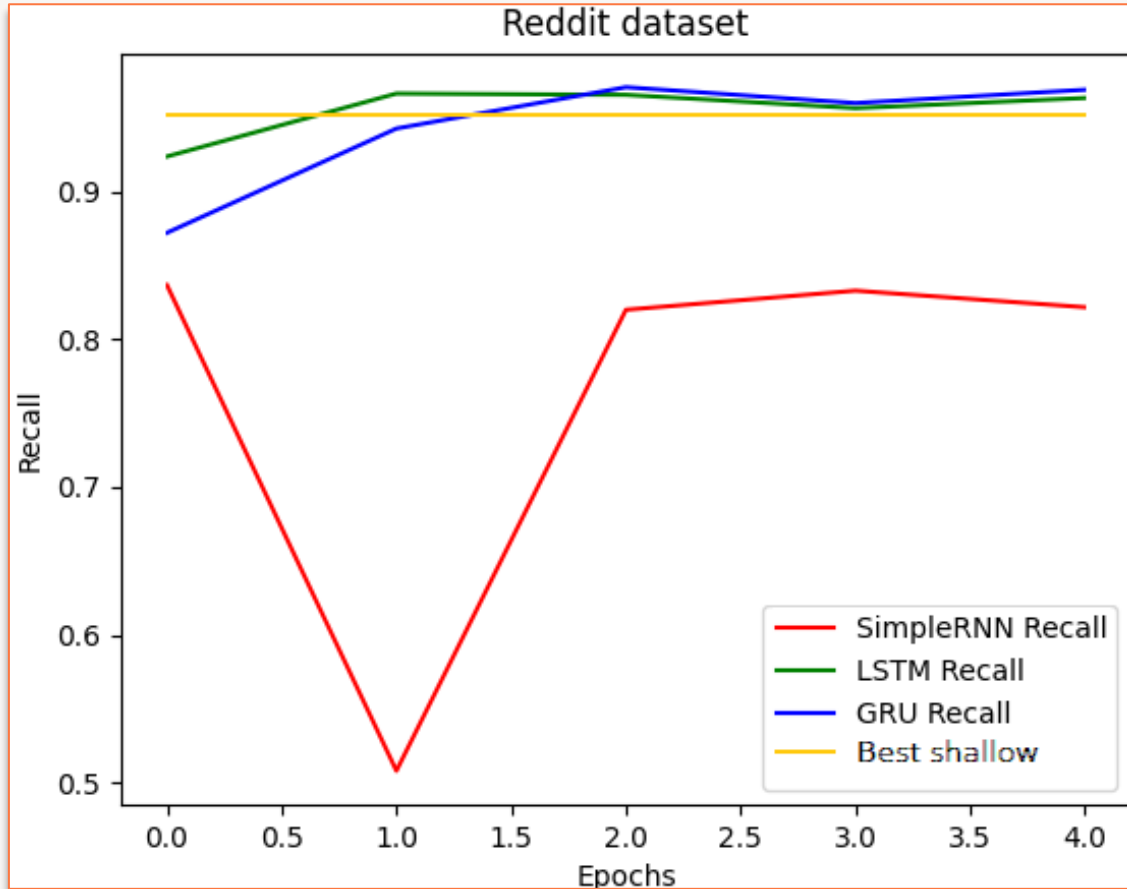


RNN

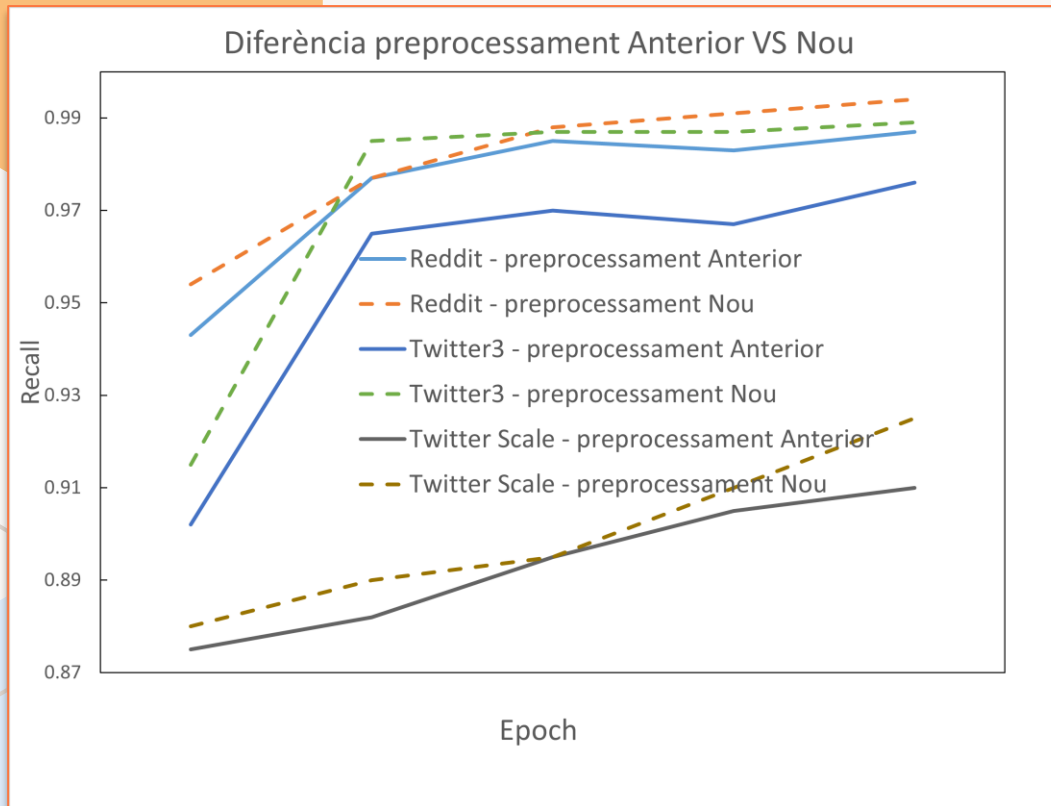




Resultats



Nou preprocessament



Eliminar noms d'usuari

~~Eliminar stopwords~~

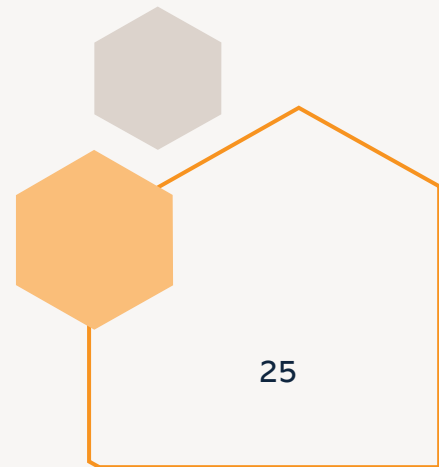
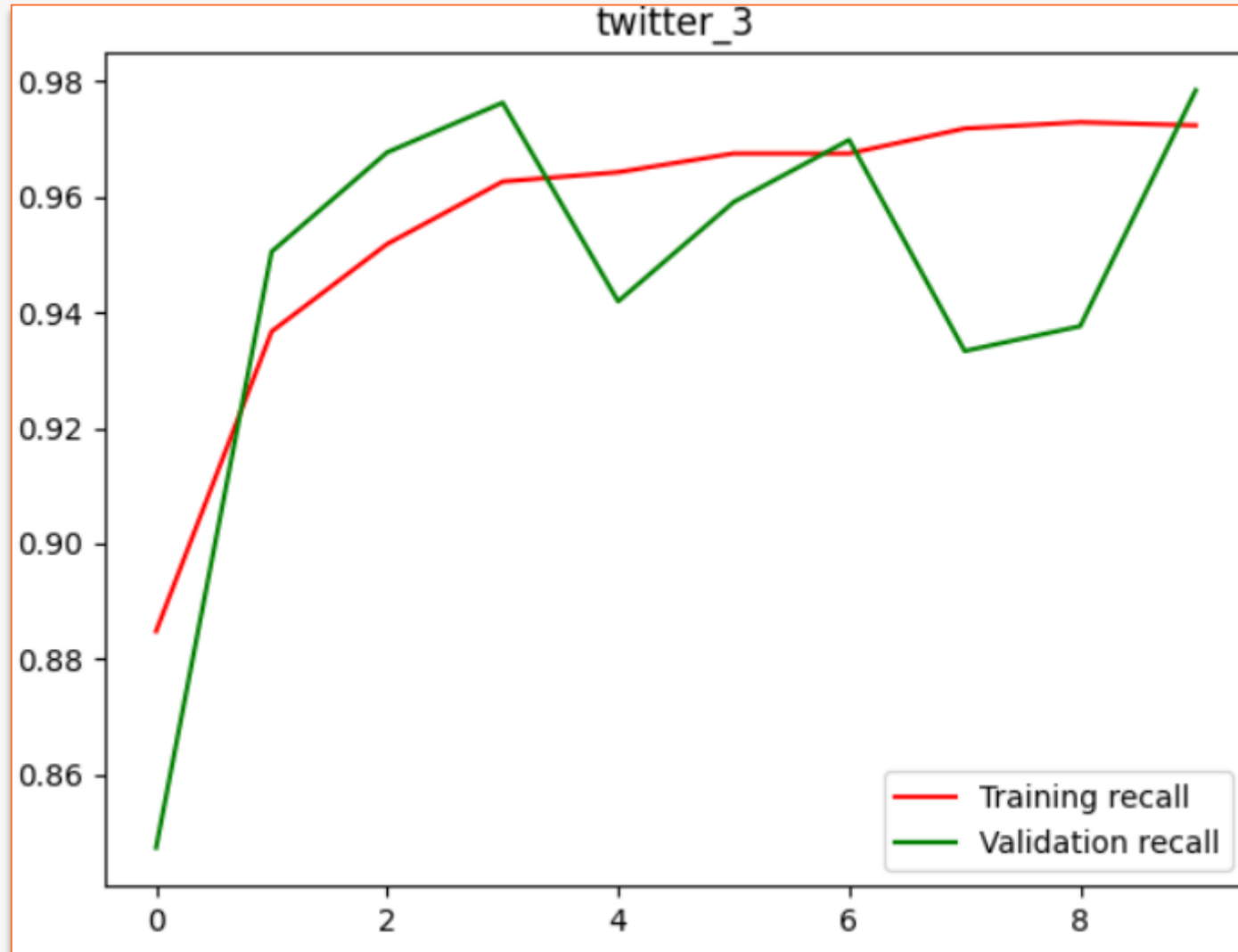
Eliminar números

~~Lemmanization~~

Eliminar puntuació

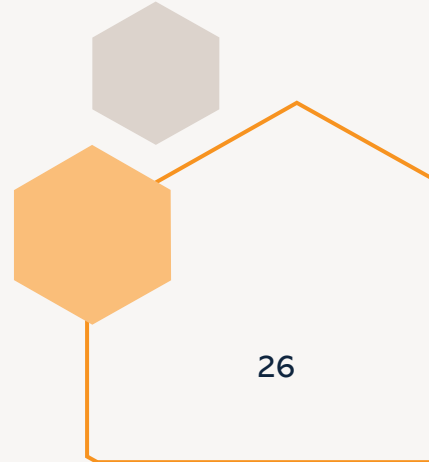


BERT (transformers)



Diferències en les prediccions

	Deep Learning	Shallow Learning
“study finds no casual relationship between cannabis and depression”	✓	✗
“dailytonic exposure to the bacteria in soil can be good for mental hearlth and could treat depression and prevent ptsd”	✓	✗
“don’t be sad, armys are here for you we will always suport you btstwt be strong”	✓	✗



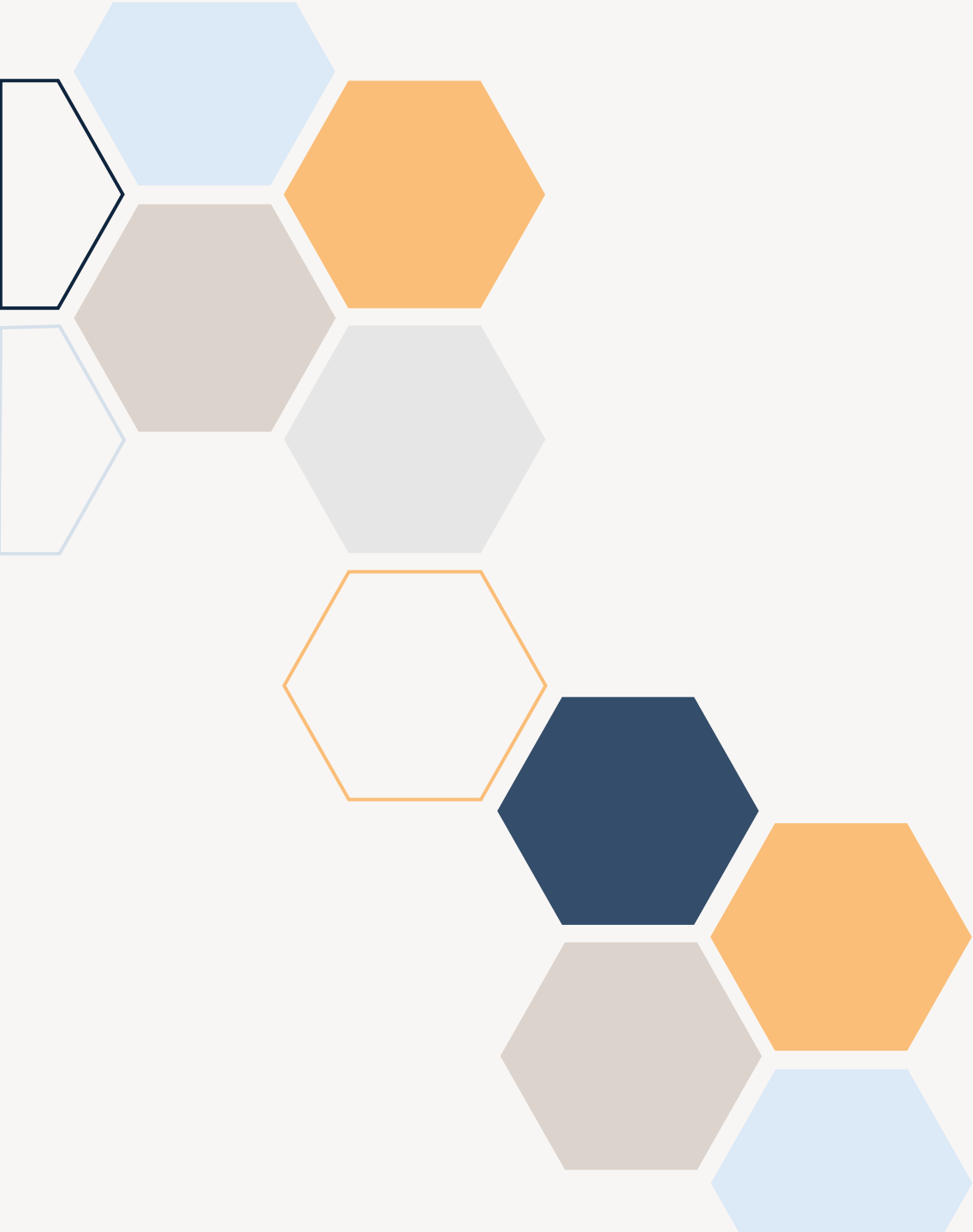


Shallow learning

- Millor: SVM i RF (relatiu a la confiança)
- Preprocessament té molta importància a les prediccions
- Extracció de característiques afecta molt al temps, però no als resultats
- Els paràmetres no són decisius

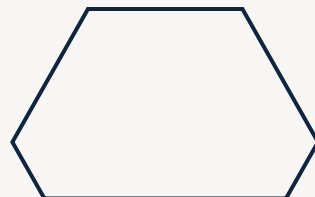
Deep learning

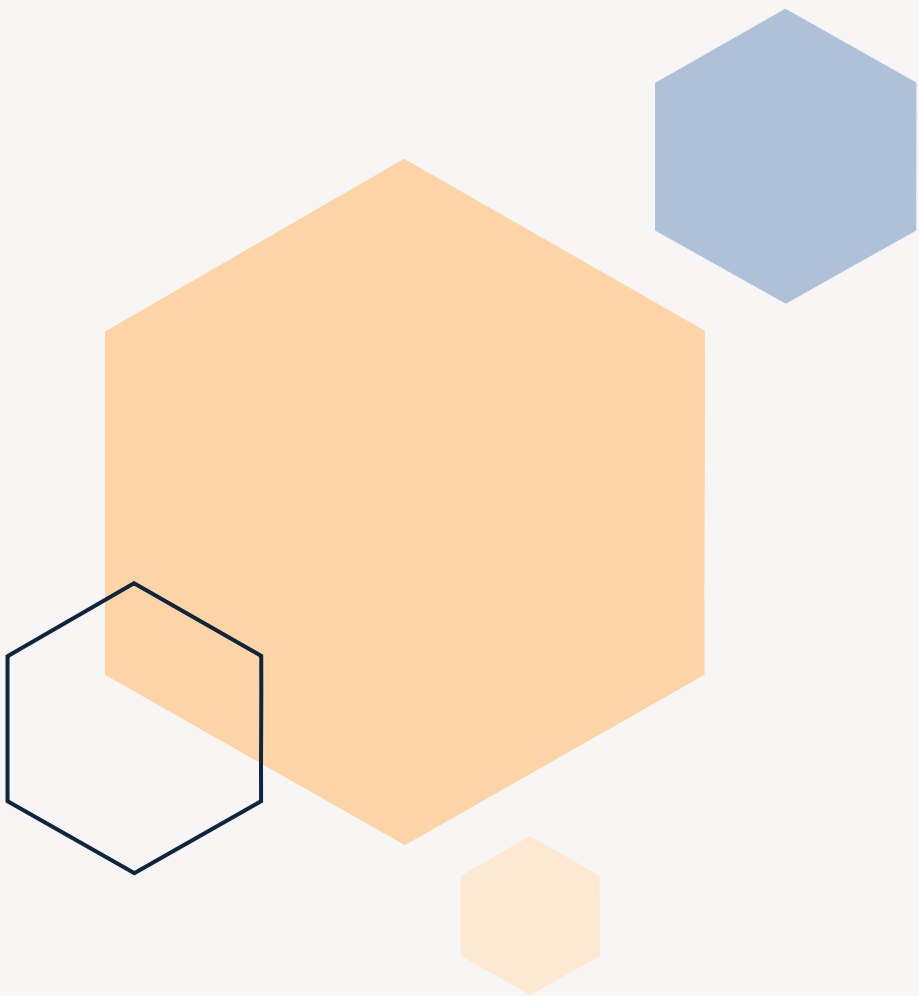
- Resultats sobre un 10% millors
- RNN simples no són bones, necessites GRU o LSTM
- LSTM millor que GRU amb missatges molt llargs
- Enten la semàntica en comptes de les relacions
- BERT necessita moltes dades i potència computacional

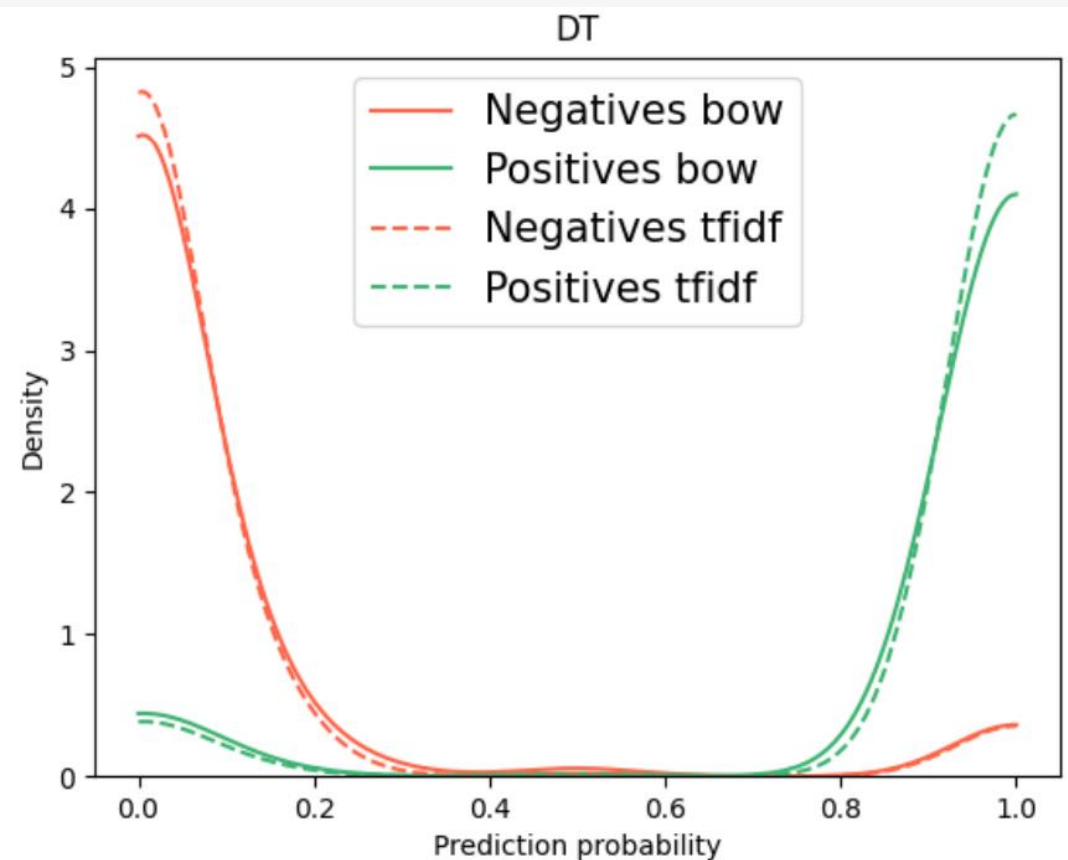
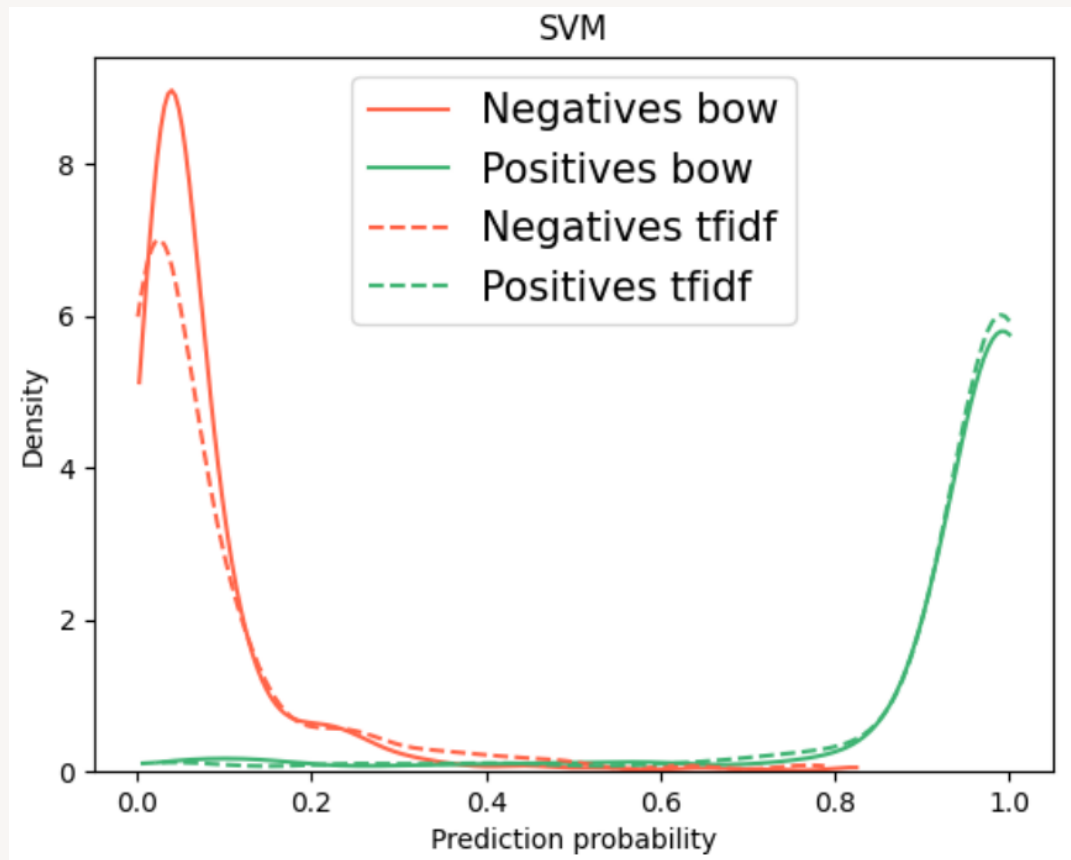


Gràcies

Martí Caixal i Joaniquet







Execution time

