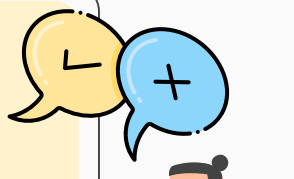
TRANSFORMERS APPLIED TO STANCE DETECTION FOR SPANISH AND CATALAN LANGUAGES

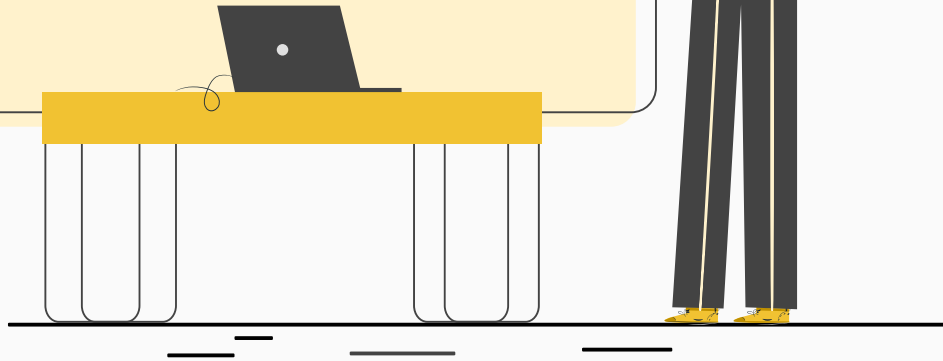
#StanceDetection #Transformers #BERT #BETO #BiLSTM #FasText #SVM #Tweets #Twitter #WordEmbeddings #SpanishLanguage #CatalanLanguage

Pavel Razgovorov and Isabel Segura Bedmar | Big Data Analytics MSc | September 2020

STANCE DETECTION

Classify the inclination of a given text regarding a particular subject or topic to be “in favor”, “against” or “neutral”





“Asumo el mandato del pueblo que Catalunya se convierta en un estado independiente en forma de república”

—CARLES PUIGDEMONT, 130TH PRESIDENT OF CATALONIA

J.Coscu
@jcoscu

Prou que m'agradaria aquesta possibilitat. Però avui està encara més difícil que fa un any. El grau de conflictivitat i radicalització del conflicte fa més difícil que el conjunt de la societat catalana accepti aquesta hipòtesi. Negar la realitat no és el camí per transformar-la

Montserrat @montsebf4
Replying to @jcoscu
Un any despres i tot és tan fàcil com dipositar un vot en una urna els del SI independència i els del NO feu una proposta que sedueixi i sortirem del "bucle".
El que no pot ser és que trieu unidad avans que democràcia.
El poble s'ha de pronunciar el país ha d'evolucionar.

7:42 AM · Sep 6, 2018
153 166 people are Tweeting about this

mercè clara NI OBLIT NI PERDÓ
@merce_clara
Millo acusa els CDR de violents. Els criminalitza i posa el relat per a que se'ns persegueixi penalment
10:24 AM · Mar 5, 2019
1 See mercè clara NI OBLIT NI PERDÓ 's other Tweets

Jesús
@Yosutumaka
Está claro que solo vais a ganarnos utilizando la violencia, con la palabra no dais ni para P3
#judicialademocracia
1:49 PM · Feb 19, 2019
See Jesús 's other Tweets

veudecatalans
@vozdecatalanes
Los indepes deben explicar si quieren una República donde no se cumpla la Ley, vivir sin democracia, sin legalidad.

Cuando defienden a los políticos presos están avalando esa tesis.
2:56 AM · Mar 5, 2019
6 See veudecatalans's other Tweets

01.
REVIEW

Previous literature

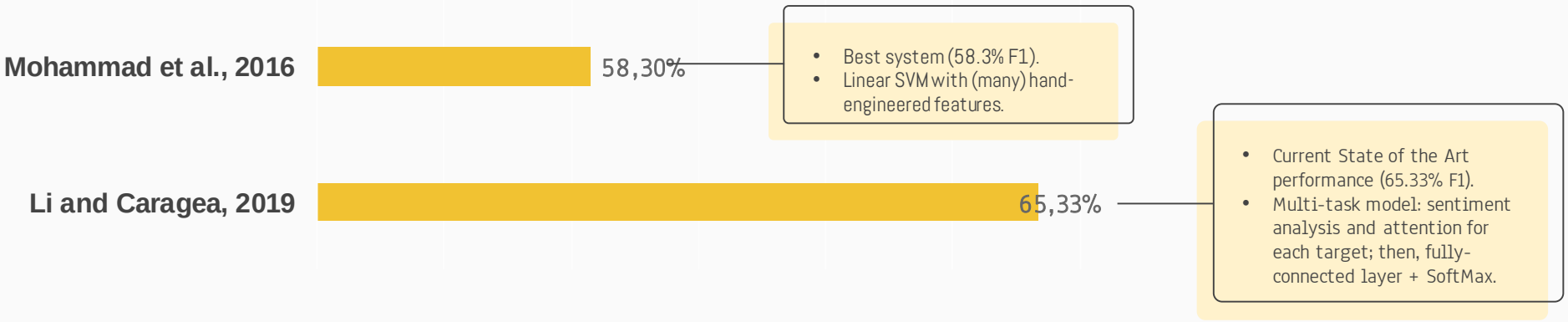
SEMEVAL 2016 CHALLENGE

FIRST SUBTASK

- Detect stance in tweets within 5 targets.
- 2,914 labelled training data instances and 1,249 for evaluation

SECOND SUBTASK

- Predict 707 Donald Trump's tweets.
- 78,000 unannotated tweets were provided.



- Best system (58.3% F1).
- Linear SVM with (many) hand-engineered features.

- Current State of the Art performance (65.33% F1).
- Multi-task model: sentiment analysis and attention for each target; then, fully-connected layer + SoftMax.

IBEREVAL CHALLENGES

STANCECAT DATASET (2017)

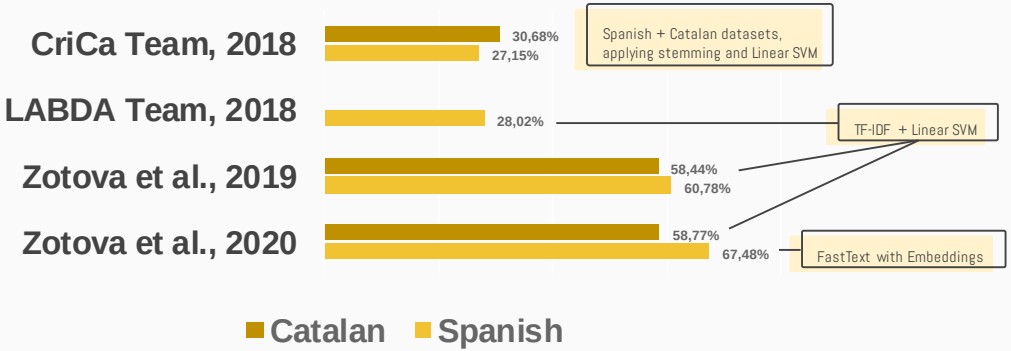
- Tweets about Catalonia independence.
- Annotated with language, stance and gender

MULTISTANCECAT DATASET (2018)

- MultiModal dataset
- Tweets about Catalonia 1Oct Referendum

IberEval 2017

IberEval 2018

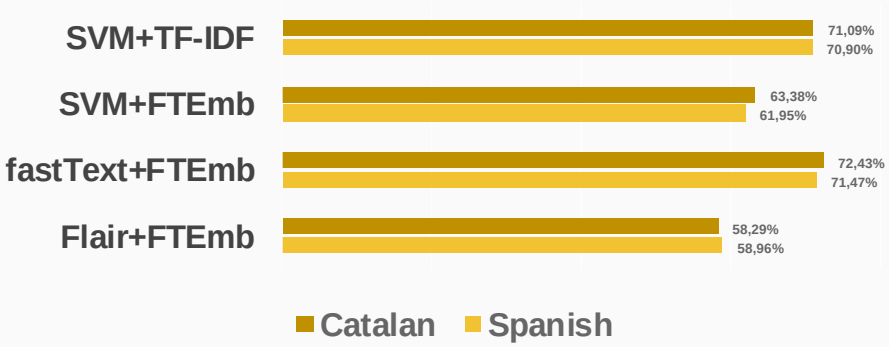


OTHER DL APPROACHES

- For IberEval 2017, many approaches: LSTM, MLP, CNN, FastText, BiLSTM.
- Casacufans Team, 2018 used CNNs to identify flags (no details provided)
- ELiRF Team, 2018 used CNNs with word embeddings and polarity/emotion lexicons

CATALONIA INDEPENDENCE CORPUS (CIC) - 2020

Scores by Zotova et al., 2020

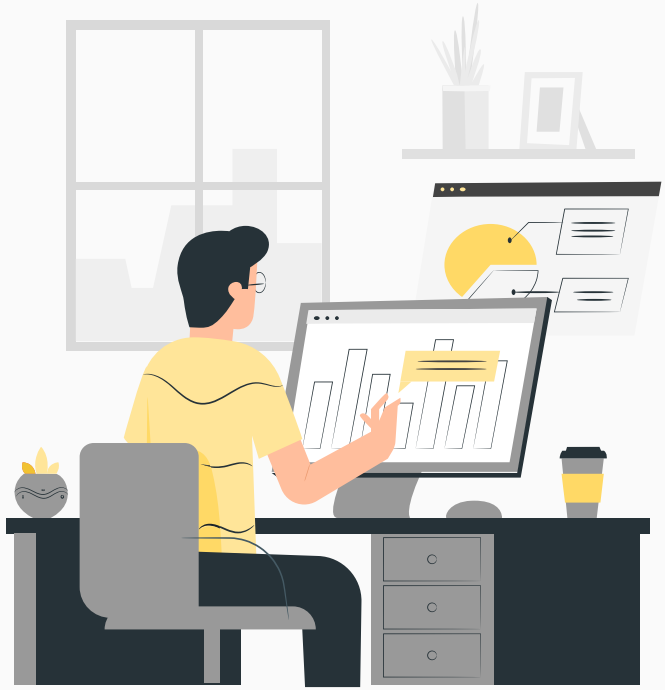


ALTERNATIVE TO MULTISTANCECAT

- Increases dataset size
- More balanced class distribution

02.
DEVELOPMENT

Proposed approaches' implementations



PROPOSED APPROACHES

TF-IDF + SVM

Most popular and performant approach

BILSTM

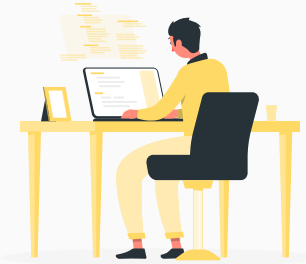
Better than LSTM for text classification

FASTTEXT

Efficient alternative to DL approaches

BETO

Novel Spanish BERT (Transformers)



TF-IDF + SVM

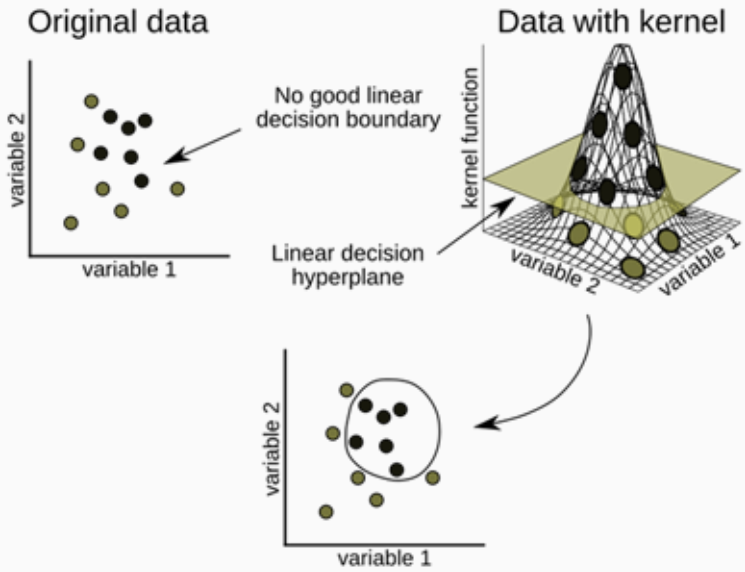
TF-IDF EXAMPLE

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Sentence 1: The car is driven on the road.
Sentence 2: The truck is driven on the highway.

SVM'S BEST PARAMETERS

Language	C	Gamma	Kernel
Spanish & Catalan	10	1	RBF
Combined	1	1	Linear



fastText

FASTTEXT VS DL APPROACHES

	Yahoo		Amazon full		Amazon polarity	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
char-CNN	71.2	1 day	59.5	5 days	94.5	5 days
VDCNN	73.4	2h	63	7h	95.7	7h
fastText	72.3	5s	60.2	9s	94.6	10s

Releasing fastText

August 18, 2016

Edouard Grave 

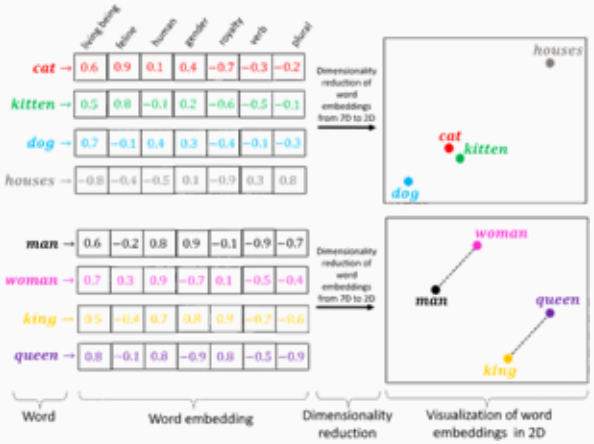
Faster, better text classification!

INPUT EXAMPLE

```
__label__NEUTRAL así es inés arrimadas su carrera en vídeos
```

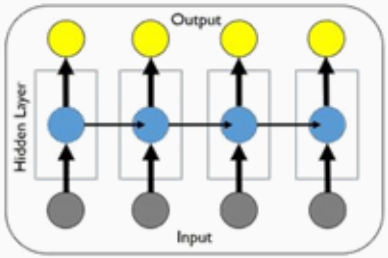
BILSTM

WORD EMBEDDINGS



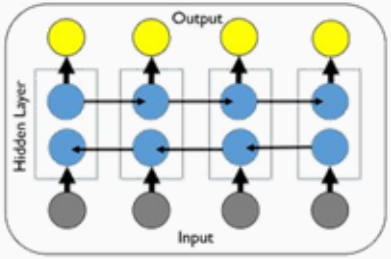
Spanish: Spanish Unannotated Corpora (FastText)
Catalan: Catalan CoNLL17 corpus (Word2Vec)

LSTM VS BILSTM



LSTM Architecture

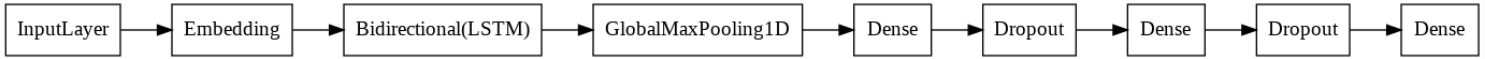
Hochreiter & Schmidhuber, 1997



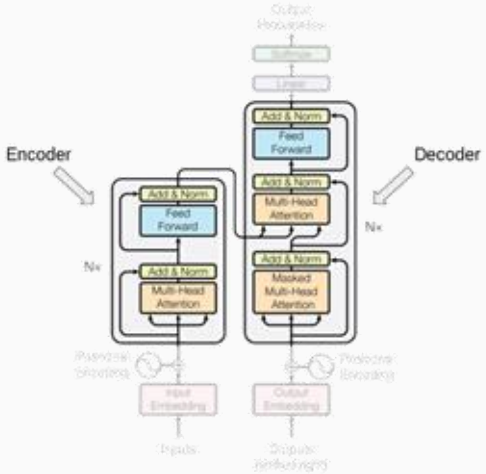
BiLSTM Architecture

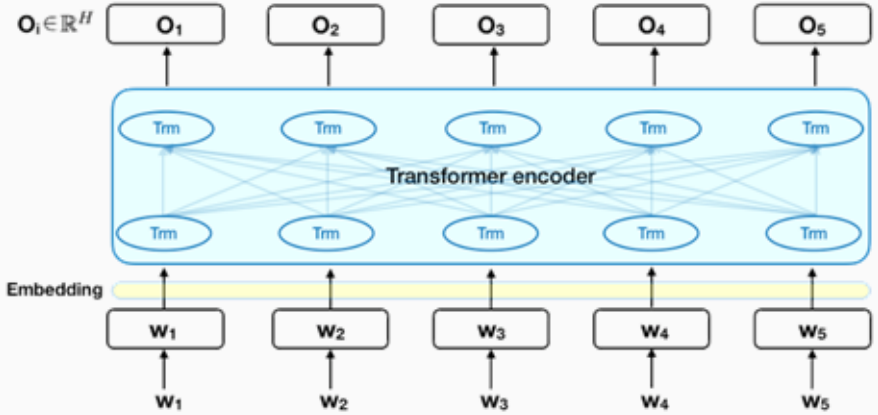
Graves & Schmidhuber, 2005

NETWORK ARCHITECTURE



BETO





Transformers

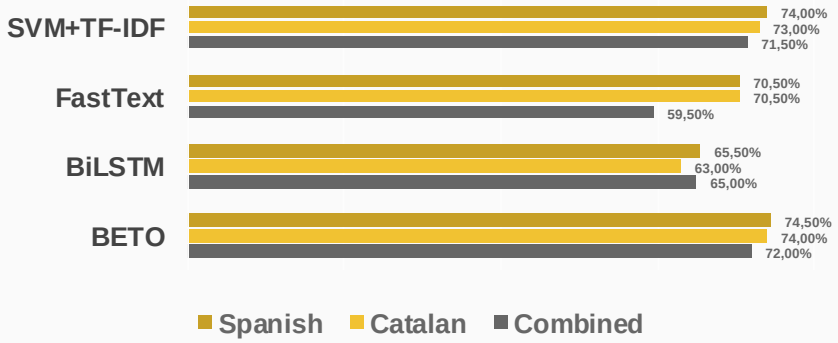
State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0

03.
RESULTS ANALYSYS

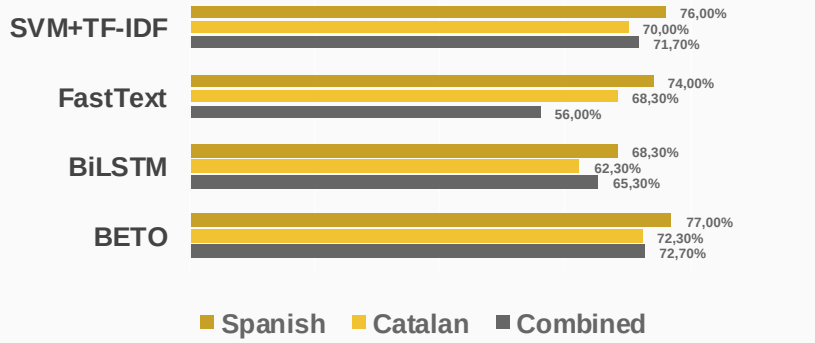
Who performed better?
Why are we having errors?

CATALONIA INDEPENDENCE CORPUS (CIC) - 2020

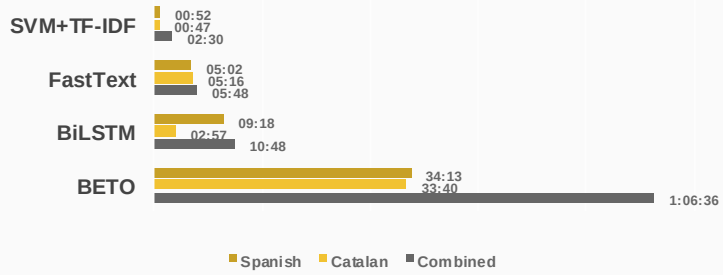
2-classes Macro F1



3-classes Macro F1

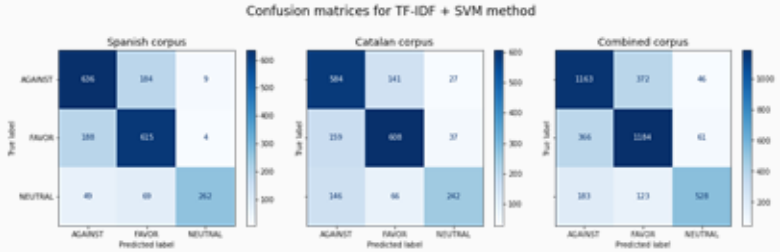


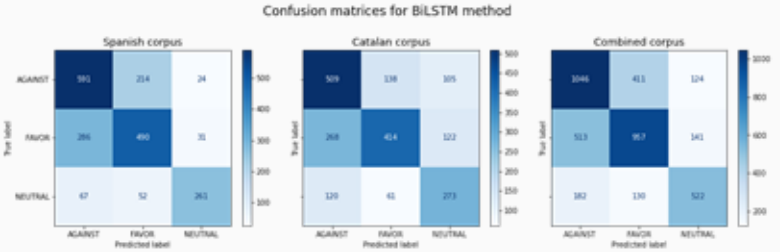
Training + Evaluation Times

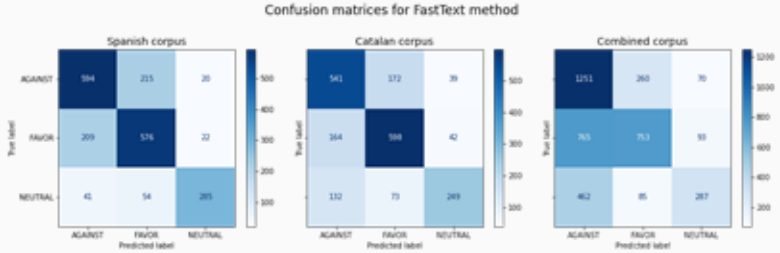


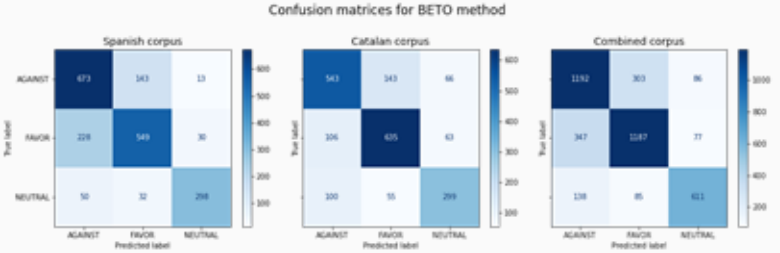
$$F1_{2classes} = \frac{F1_{favour} + F1_{against}}{2} \quad \text{and} \quad F1_{3classes} = \frac{F1_{favour} + F1_{against} + F1_{neutral}}{3}$$

CONFUSION MATRICES

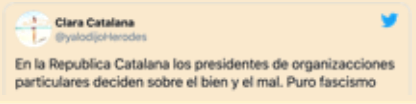
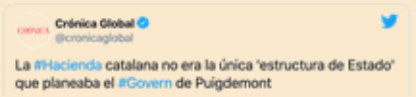




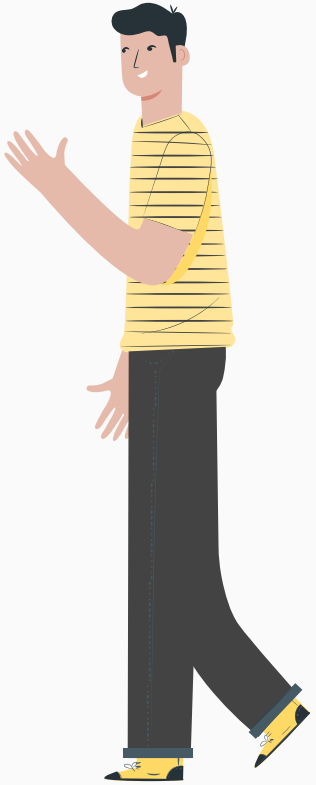




ERROR ANALYSIS

Tweet	Predicted Stance	Annotated Stance
	FAVOR	AGAINST
	AGAINST	FAVOR
	NEUTRAL	AGAINST
	AGAINST	NEUTRAL
	NEUTRAL	AGAINST

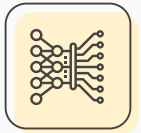
04.
CONCLUSIONS AND FUTURE WORK



CONCLUSIONS

CATALAN INDEPENDENCE CORPUS

Having a good dataset makes a difference



BETO

New State-of-the-Art performance.
First study that uses it in this task

COMBINED DATASET

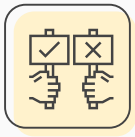
Slightly lower scores than individual languages - no big difference

CONTEXT

Would help determining stance

PRE-PROCESSING

Would need to improved



STANCE DETECTION

Significant role in measuring public opinions

FUTURE WORK

PREPARE

- Improve pre-process pipeline
- Feature engineering: Named Entity Extraction

UNDERSTAND

- Train models from scratch
- Debug BETO model's performance
- Tune BETO's parameters

TRY NEW THINGS

- Train other Transformers
- Combine Spanish + Catalan word embeddings

THANKS

Does anyone have any questions?

QUESTION 1

Answer 1

QUESTION 2

Answer 2

QUESTION 3

Answer 3

Fonts & colors used

This presentation has been made using the following fonts:

Staatliches
(https://fonts.google.com/specimen/Staatliches)

Anaheim
(https://fonts.google.com/specimen/Anaheim)

#434343 #f3f3f3 #fafafa #fff2cc #ffe599 #ffe599 #ffd966