Conspiracy Detection

Andrei Ancuța and Andrei Daniel Tavă
Mentorship: Florin Brad* and Marius Drăgoi*

University of Bucharest, Romania
*Bitdefender, Romania
andrei.ancuta@gmail.com, tavaandrei@gmail.com, fbrad@bitdefender.com, mdragoi@bitdefender.com

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

1. Introduction

The studied problem is analyzing oppositional thinking in both English and Spanish text, and it involves two tasks:

Task 1: perform binary classification to detect conspiratorial thinking.

Task 2: identify and label smaller text sequences under six different labels related to oppositional thinking

Identifying conspirational texts and differentiating them from critical texts that oppose mainstream views can be a considerable challenge and it is prone to false positives.

2. Dataset and Preprocessing

Dataset Description:
For each language we used a dataset consisting of messages from the **Telegram** platform during the **Covid-19 Pandemic**. The messages are classified and have a list of span labels associated.

Statistic	EN	ES
CONSPIRATORIAL : CRITICAL RATIO	1:1.9	1:1.73
MEDIAN TOKENS PER CONSPIRATORIAL	102	159
MEDIAN TOKENS PER CRITICAL	71	92
MEDIAN LABELED SPANS PER MESSAGE	4	3

Data preprocessing: For the sequence labelling task, the span lists had to be converted to tensors of token classes, for each label.

3. Models

Tested encoders:

1. **Baseline**: base BERT
2. BERT multilingual
3. Spanish BERT (BETO)

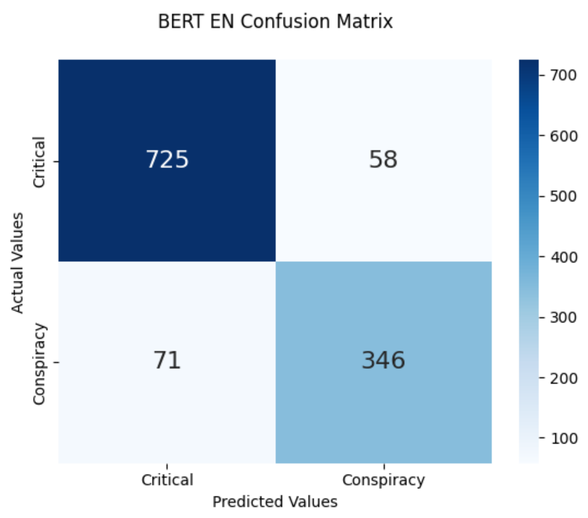
Parameters and Training:

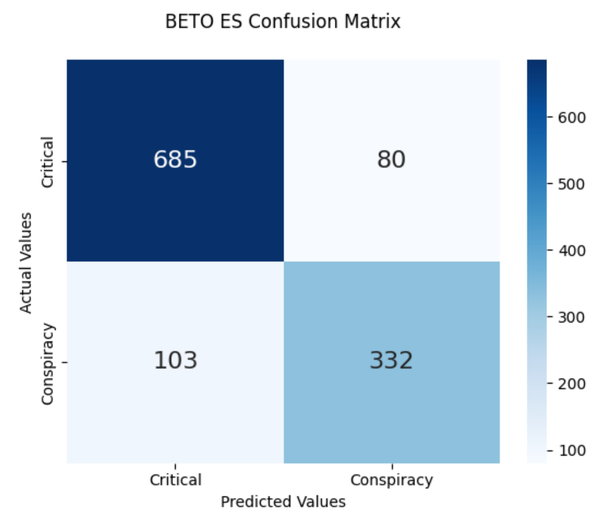
- Classification used a linear classifier with sizes **(768, 64, 2)**.
- Sequence Labelling was done by token classification using 6 different classifiers with sizes **(768, 3)**. The classes are: **not part**, **beginning** and **continuation** (of span).
- The encoder had a maximum sequence length of **256**.
- Training was done with **Adam** optimizer, learning rate **2e-5** and batch size **16** for **3-4** epochs.

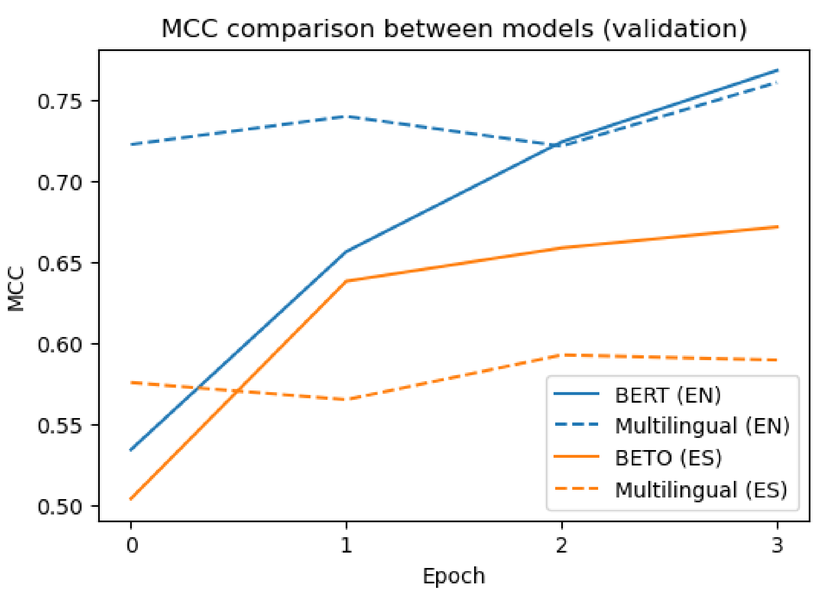
4. Results

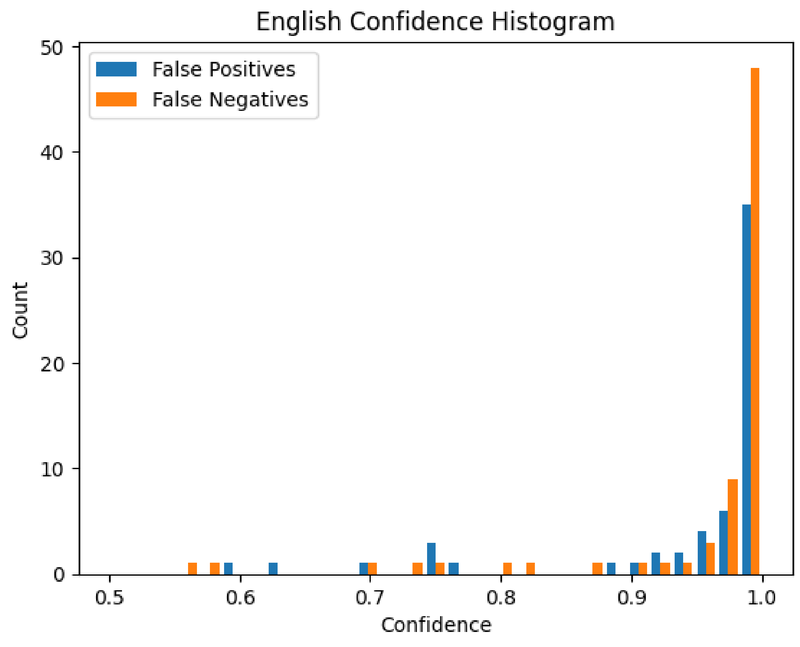
Classification results

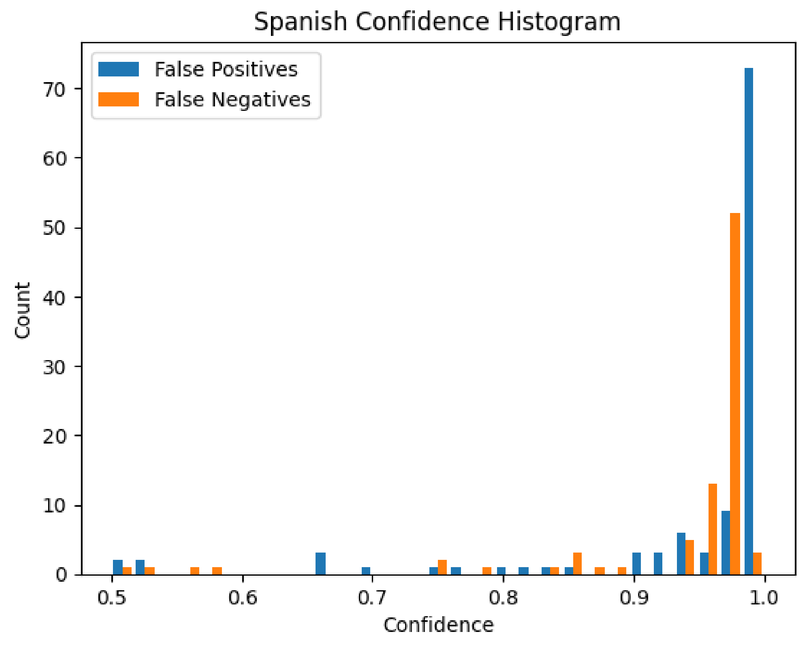
Language	Model	MCC	Precision	F1	Recall
ENGLISH	BERT BASE	**0.76**	**0.91**	**0.91**	**0.92**
ENGLISH	BERT MULTILINGUAL	0.75	0.93	0.90	0.88
SPANISH	BERT MULTILINGUAL	0.59	0.88	0.83	0.78
SPANISH	BETO	**0.66**	**0.86**	**0.88**	**0.89**











BERT (EN) most confident FP sample: The Deep State continues to be a WASP, Northeast good ole boy's club of snooty (insider trading) rich guys who think they know better. [...] If you actually drive around America, you'll see that other than strip malls everywhere, the country is basically very different than it is depicted.

BERT (EN) most confident FN sample: The developer of an implanted microchip that is linked to a COVID vaccine passport says that the mass chipping of humans as a means of verifying compliance is happening "whether we like it or not." [LINK]

Sequence labeling span-F1 scores

Language	Model	**Overall**	Agent	Campaigner	Facilitator	Negative Effect	Objective	Victim
EN	BERT BASE	**0.44**	**0.58**	**0.51**	0.33	**0.59**	**0.46**	**0.59**
EN	BERT MULTI.	0.44	0.57	0.50	**0.34**	0.55	0.41	0.59
EN	BETO	0.38	0.52	0.44	0.29	0.50	0.39	0.53
ES	BERT BASE	0.32	0.35	0.39	0.23	0.50	0.20	0.45
ES	BERT MULTI.	0.41	0.40	0.43	0.32	0.61	0.31	0.52
ES	BETO	**0.45**	**0.44**	**0.47**	**0.35**	**0.63**	**0.40**	**0.55**

Median tokens per sample

	EN	ES
TP	104	177
FP	110.5	169
TN	72	84
FN	69	121

5. Conclusion

- Language specific models perform best
- Multilingual BERT performs very well overall
- Models tend to associate length with conspiracy
- There is considerable variation between sequence labels