Conspiracy Detection PAN 2024

Scarlat Marius-Stefan, Dinu Matei-Alexandru, coordinators: Florin Brad*, Ioana Pintilie*

University of Bucharest, Romania
*Bitdefender, Romania

1. Introduction

Conspiracy theories pose significant challenges to social cohesion and public health. Detecting conspiracy content automatically is crucial for moderating social media platforms. This project tackles conspiracy detection as part of the PAN 2024 shared task on oppositional thinking analysis. We train several BERT-based models with different strategies for handling English and Spanish conspiracy content. We compare monolingual models (BERT for English, BETO for Spanish) with multilingual BERT on individual languages and a combined dataset. Our results show that language-specific models outperform multilingual models in single-language settings, while multilingual models offer impressive cross-lingual transfer capabilities.

2. Dataset

We use the PAN 2024 oppositional thinking dataset, which contains social media posts in English and Spanish labeled as conspiracy or non-conspiracy. Based on our analysis, the dataset has the following characteristics:

English Characteristic	**Value**
Total samples	4,000
Conspiracy samples	1,379 (34.5%)
Non-conspiracy samples	2,621 (65.5%)
Average token count	132.1

Spanish Characteristic	**Value**
Total samples	4,000
Conspiracy samples	1,462 (36.6%)
Non-conspiracy samples	2,538 (63.5%)
Average token count	190.3

The token length distribution shows Spanish texts tend to be longer than English texts, with both languages exhibiting a right-skewed distribution. We found that some texts exceed the 512 token limit for BERT models, requiring truncation.

3. Models

We trained and evaluated five BERT-based models:

- **English BERT**: Pretrained BERT-base-uncased fine-tuned on English data
- **Spanish BERT (BETO)**: BETO model fine-tuned on Spanish data
- **Multilingual BERT on English**: mBERT fine-tuned on English data
- **Multilingual BERT on Spanish**: mBERT fine-tuned on Spanish data
- **Multilingual BERT on Both**: mBERT jointly trained on English and Spanish data

Each model consists of a pretrained transformer with a dropout layer (rate=0.3) and a classification head. For all models, we experimented with two approaches: full fine-tuning and partial fine-tuning where only the final layer was trainable (require_grad=True) with all other layers frozen. The partial fine-tuning approach showed approximately 6-7% worse performance across all metrics. For instance, English-BERT accuracy dropped from 94.8% to 90.3% when using frozen layers. All models were trained using the AdamW optimizer (learning rate=2e-5) for 5 epochs.

4. Results

Our experimental results show that language-specific models outperform multilingual models when tested on the same language, but multilingual models offer better cross-lingual transfer.

Model	Accuracy	Precision	Recall	F1	AUC	Avg. Prec.
English-English	**0.948**	**0.937**	0.892	**0.914**	**0.977**	**0.963**
Spanish-Spanish	0.884	0.927	0.752	0.831	0.938	0.929
Multilingual-English	0.846	0.719	0.831	0.771	0.932	0.881
Multilingual-Spanish	0.745	0.628	0.802	0.704	0.867	0.831
Multilingual-Both	0.878	0.822	0.826	0.824	0.946	0.921

English BERT

	PN	**PP**
TN	179	5
TP	9	74

Spanish BERT

	PN	**PP**
TN	160	6
TP	25	76

mBERT-Eng

	PN	**PP**
TN	157	27
TP	14	69

mBERT-Spa

	PN	**PP**
TN	118	48
TP	20	81

mBERT-Both

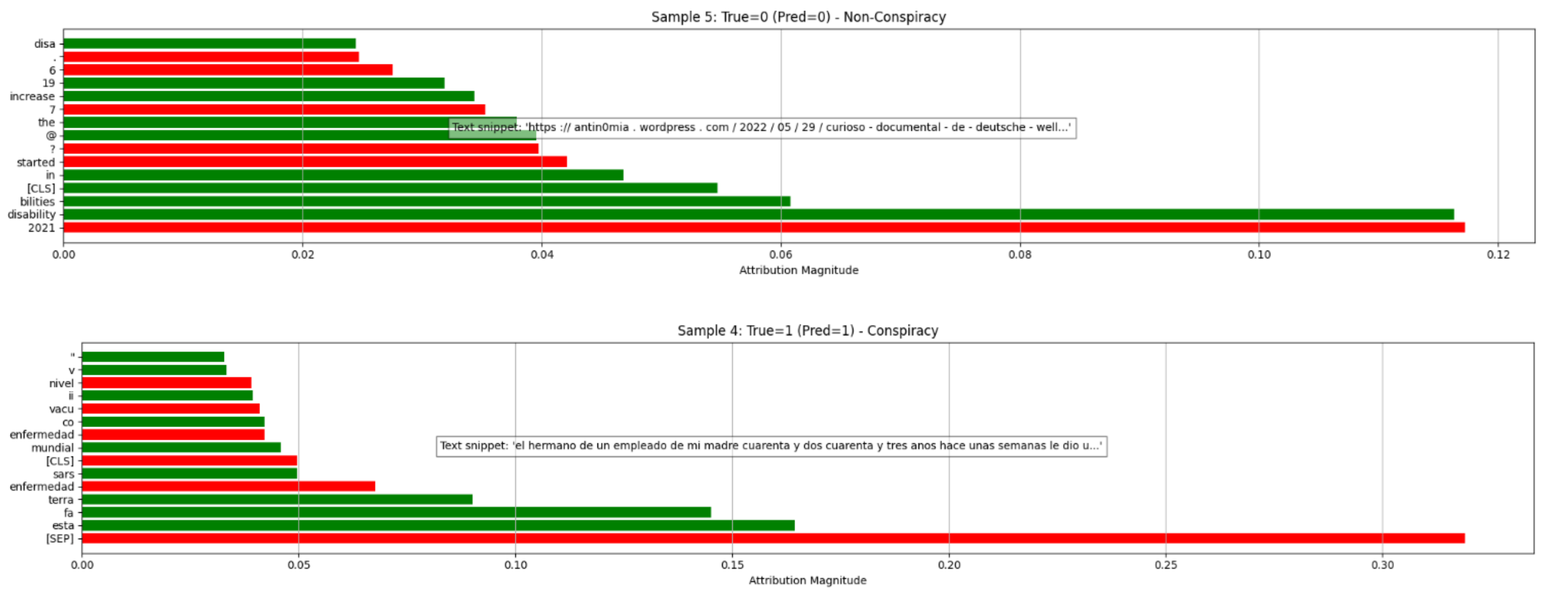
	PN	**PP**
TN	317	33
TP	32	152

Cross-Language Transfer

Model	**Accuracy**
Multilingual BERT (English → Spanish)	0.42
Multilingual BERT (Spanish → English)	0.65

The results demonstrate that the monolingual models perform best within their respective languages, with English BERT achieving the highest overall performance (94.8% accuracy). However, the multilingual model trained on both languages shows strong performance (87.8% accuracy) while offering the advantage of handling both languages with a single model.

5. Integrated Gradients Analysis



Our analysis reveals that:

- The models often misclassify posts containing strong language or insults as conspiracy content, even when no conspiratorial claims are made.
- Subtle conspiracies expressed in neutral language are frequently missed.
- Our observations show that longer texts tend to be classified more accurately than shorter ones, with accuracy generally increasing with text length.

The token importance analysis reveals distinctive patterns: conspiracy content typically features terms related to "terra", diseases, viruses, and mythological references like dragons, while non-conspiracy content is characterized by numbers, @ symbols, and modern neologisms - reflecting the linguistic differences between these content types.

6. Conclusions

- Language-specific models (English BERT, BETO) outperform multilingual models in single-language settings
- Multilingual BERT trained on both languages offers strong performance with the advantage of handling multiple languages
- Cross-lingual transfer shows promise, with Spanish→English transfer being more effective than English→Spanish
- Despite the class imbalance (35.5% conspiracy vs. 64.5% non-conspiracy), models achieved strong performance through proper training and evaluation techniques
- Integrated gradients shows conspiracy content contains terms about "terra" and diseases, while non-conspiracy features numbers and modern slang.