Conspiracy Detection PAN 2024

Cosmin Colceru, Cosmin-Ionuț Moarcăs, Florin Brad* and Marius Drăgoi*
University of Bucharest, Romania
*Bitdefender, Romania

cosmincolceru@gmail.com, cosminmoarcas01@gmail.com, fbrad@bitdefender.com, bdragoi@bitdefender.com

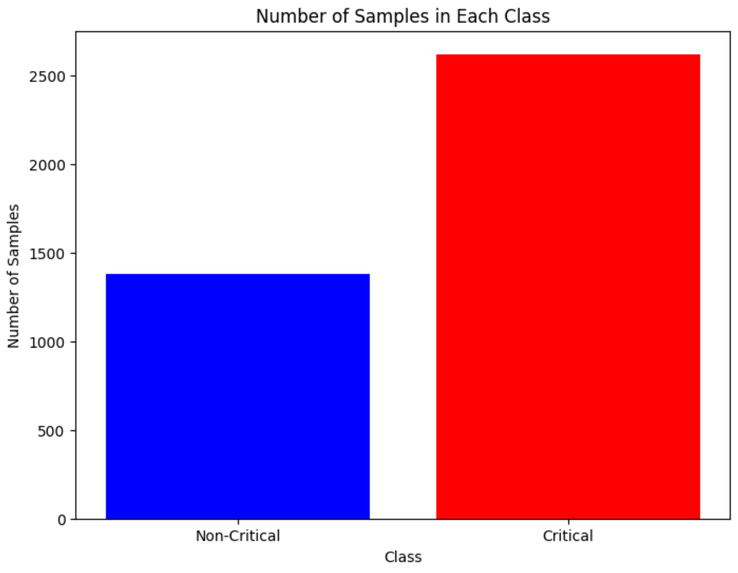
1. Introduction

Our task is too differentiate between public messaging and conspiracy theories based on text data. To achieve this we train models such as BERT and then more specific models.

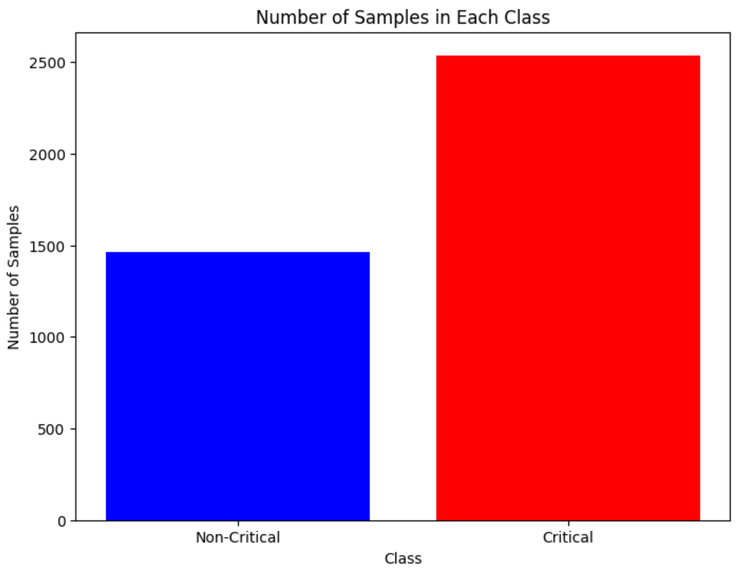
One challenge is the multilingual dataset, which includes both English and Spanish texts. Another challenge is the inherent subjectivity and ambiguity present in distinguishing between public messaging and conspiracy theories, witch requires the model to comprehend subtle linguistic cues, context and other factors.

2. Dataset and Preprocessing

Dataset Description:
Our datasets is composed of 4000 English and 4000 Spanish texts that are annotated as either "CONSPIRACY" or "CRITICAL". The distribution is mostly balanced between the two classes, slightly biased towards "CRITICAL".
Number of samples in English dataset:



Number of samples in Spanish dataset:



Preprocessing included two main steps:
1. Elimination of Punctuation: Punctuation marks, such as commas, periods, and quotation marks, were removed using standard Python string manipulation techniques.
2. Elimination of Stopwords: Stopwords are common words that often do not carry significant meaning in the context of natural language processing tasks. Examples of stopwords include "the", "is", "and", and "in". We eliminated stopwords from the text data using the NLTK (Natural Language Toolkit) library in Python. By removing stopwords, we aimed to reduce noise and improve the quality of the text data.
These preprocessing steps were essential for cleaning and preparing the text data, ensuring that it is suitable for training machine learning models and extracting meaningful insights.

3. Models

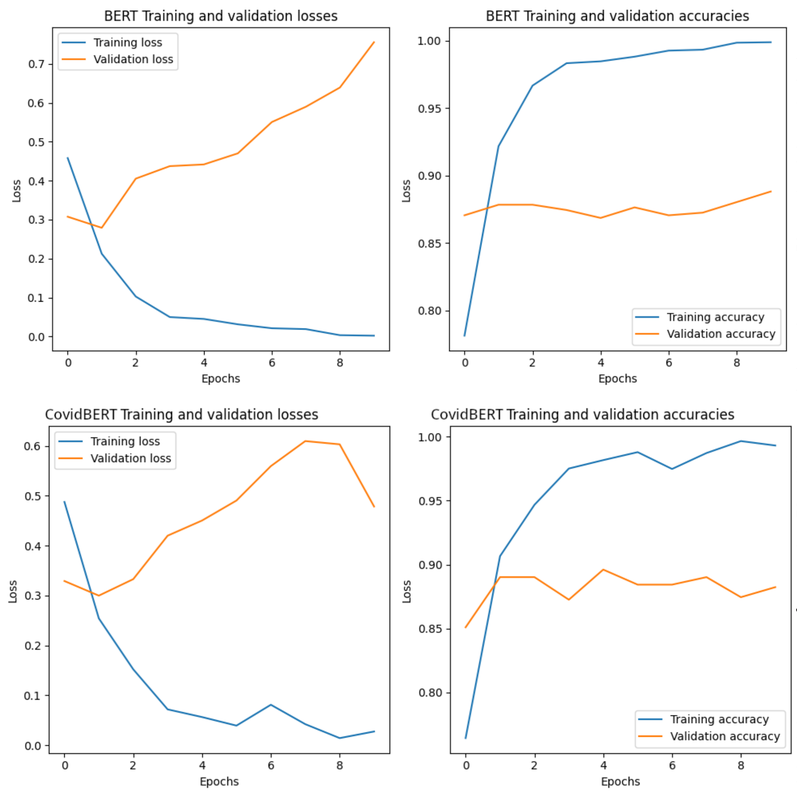
Baseline: We trained a BERT model with AdamW optimizer for our baseline. We experimented with multiple values for the learning rate, and for the number of epochs and achieved the best results for 10 epochs and a learning rate of 5e-5.
Domain specific model: We trained CovidBERT, a BERT model pretrained on tweets about Covid, also using the AdamW optimizer with a learning rate of 5e-5 and for 10 epochs.
Language specific models: Since our dataset is in both English and Spanish, we also trained Multilingual BERT and BETO (Spanish BERT) for 30 epochs and a learning rate of 1e-5.

4. Results

We trained BERT and CovidBERT on the English texts and achieved the following results:

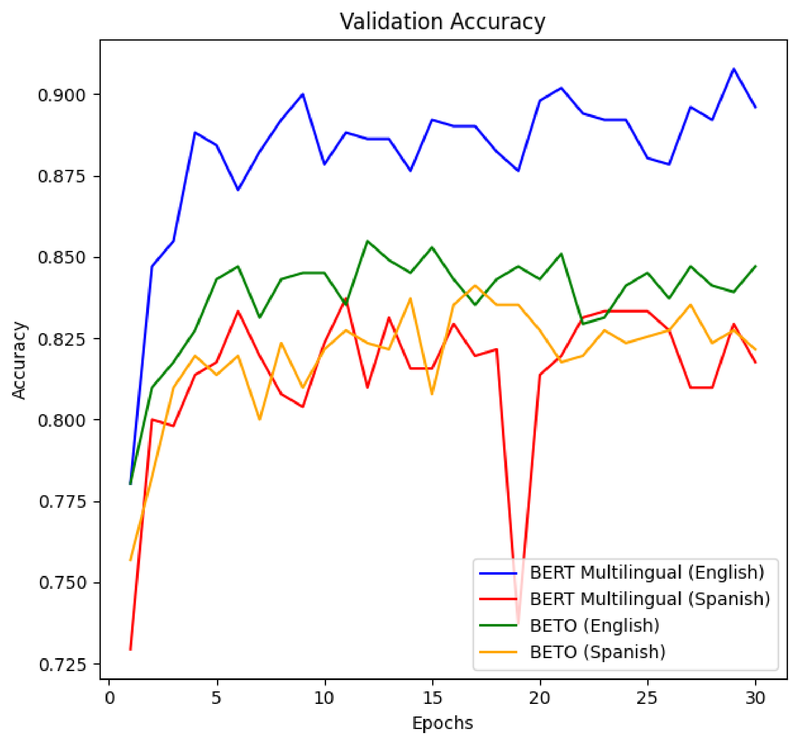
Architecture	Accuracy	F1	Precision	Recall	MCC
BERT	0.89	0.88	0.89	0.89	0.76
CovidBERT	0.89	0.89	0.89	0.89	0.74

As you can see, both BERT and CovidBERT perform comparably, yielding nearly identical results. This suggests that while CovidBERT is specifically tailored for COVID-19 related text, its performance doesn't significantly outpace the general-purpose BERT model.



Predictions examples from BERT:
Critical text predicted as critical
professor sir andrew pollard, who helped develop the astrazeneca covid vaccine, warns that endless rounds of booster shots are not the answer and that "we can't vaccinate the planet every four to six months." https://summit.news/2022/jab-chief
Critical text predicted as conspiracy
us and them: global elites drive demand for private jets to new highs during pandemic
Conspiracy text predicted as conspiracy
no vehicle that runs on gas will be allowed on the streets in 15 years. most likely way before that. the "green totalitarianism by the nwo" underway. join us at: agentsoftrutht.me/agentsoftruthchat
Conspiracy text predicted as critical
aaron russo | the american people are living in a matrix clip taken from interview with info wars in 2009 "democracy = new world order, democracy = slavery"

We also trained BERT multilingual and BETO on the English and Spanish texts and achieved the following results:



We trained each model for 30 epochs and assessed its performance using standard evaluation metrics, including the F1 score and Matthews correlation coefficient (MCC).
We calculated the confusion matrix for the best-performing model and found 66 false positives (FP) and 26 false negatives (FN). These instances highlight areas for potential improvement and emphasize the need for further model refinement.

Architecture	accuracy	MCC	F1
BERT mul. en	0.89	0.66	0.88
BERT mul. es	0.81	0.61	0.8
BETO en	0.84	0.63	0.84
BETO es	0.82	0.62	0.81

5. Conclusion

Both BERT and CovidBERT performed almost equally well. While CovidBERT was tailored for COVID-19 content, BERT showed comparable performance, showcasing the versatility of general-purpose models. In the multilingual task, Multilingual BERT performed the best. Also, while we anticipated discrepancies between BETO's performance on English and Spanish texts, no significant difference was observed.