Fake News Detection

Radu-Mihai Gheorghe, Iulita Alpetri, coordinator: Florin Brad*
University of Bucharest, Romania
*Bitdefender, Romania

radu2002ro@gmail.com, iulitaalpetri66@gmail.com, fbrad@bitdefender.com

1. Introduction

In this project, we address the pervasive issue of fake news detection by classifying news articles into four categories of truthfulness: true, partially true, false, or other. Our approach involves training robust models like BERT and RoBERTa to manage texts effectively within the 512-token limit imposed by BERT, despite the lengthy nature of typical news articles. We also devise strategies to distinguish between nuanced truth categories, achieving our best results with a fine-tuned RoBERTa model on the dataset provided by the CLEF2021-CheckThat! competition.

2. Dataset and Preprocessing

Dataset Description:
The *CLEF2021-CheckThat!* dataset is split into three CSV files designated for training, validation, and testing. After merging these files into a single dataset, we divided it into training and testing sets, allocating 20% for testing. The dataset consists of 1,876 articles each labeled as *true*, *partially true*, *false*, or *other*. The *title*, *text*, and *our rating* columns form the core components of each article record. Notably, a significant number of articles were labeled as *false*, and some articles lacked titles.

Preprocessing Steps:
To prepare the data for analysis, we concatenated the *title* and *text* fields when titles were present. The preprocessing step included removing stopwords, emojis, and URLs, and converting all text to lowercase to ensure uniformity. A linguistic analysis revealed that the most common words across labels were *said* and *will*, as highlighted by word cloud visualizations.

Tokenization:
For tokenization, we used the RoBERTaTokenizer, which leverages byte-level Byte-Pair-Encoding derived from the GPT-2 tokenizer. This tokenizer treats spaces as part of the tokens, meaning the encoding of a word varies if it appears at the beginning of a sentence. For BERT, we used the BertTokenizer, which is built on a WordPiece model. It optimizes vocabulary size and allows for more efficient model training by breaking words down into the most frequently occurring subwords.

6. References

Text Data Augmentation Techniques for Fake News Detection in the Romanian Language - https://www.mdpi.com/2076-3417/13/13/7389

3. Models

Baseline:

- Trained a BERT model using the pretrained BERT model.
- Used a weighted Cross Entropy Loss function.

BERT Fine-Tuning:

- We used the sliding-window technique and text preprocessing.
- For creating the model, we used as core the pre-trained BERT model, added two dropout layers with a 0.2 rate to reduce overfitting. Between this two, we added a fully connected layer and finally, the classifier layer (linear layer).
- We used GeLU activation function and a learning rate of 1e-05.

RoBERTa:

- We used a RoBERTa pretrained model (RobertaForSequenceClassification) and trimmed data (512- token limit)
- We extended the pre-trained RoBERTa model by replacing its classifier head with a custom classifier composed of two linear layers. The first linear layer reduces the dimensionality of the RoBERTa model's output to the number of classes, followed by a ReLU activation function and dropout regularization. Finally, the output is passed through another linear layer to produce logits.

4. Results

Architecture	Accuracy	Precision	F1	Recall
BERT - Baseline	0.34	0.29	0.29	0.30
BERT	0.50	0.31	0.31	0.32
RoBERTa	0.71	0.60	0.59	0.59

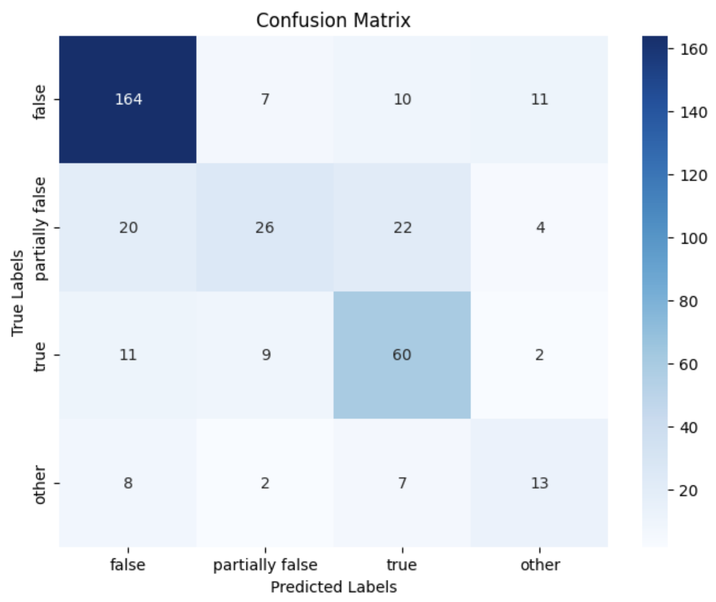

Figure 1: RoBERTa Confusion Matrix

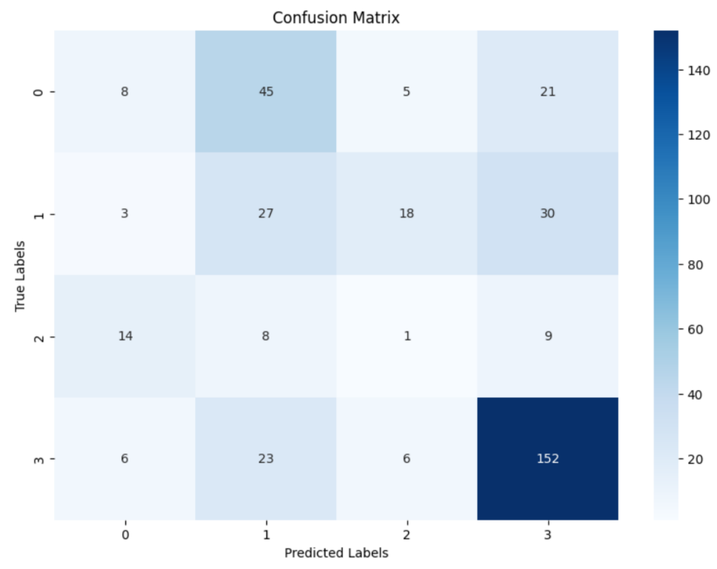

Figure 2: BERT Confusion Matrix

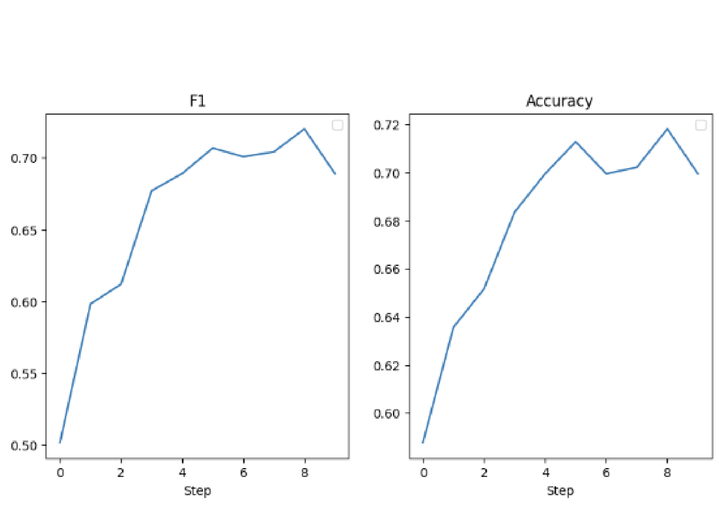

Figure 3: F1 Score and Accuracy RoBERTa

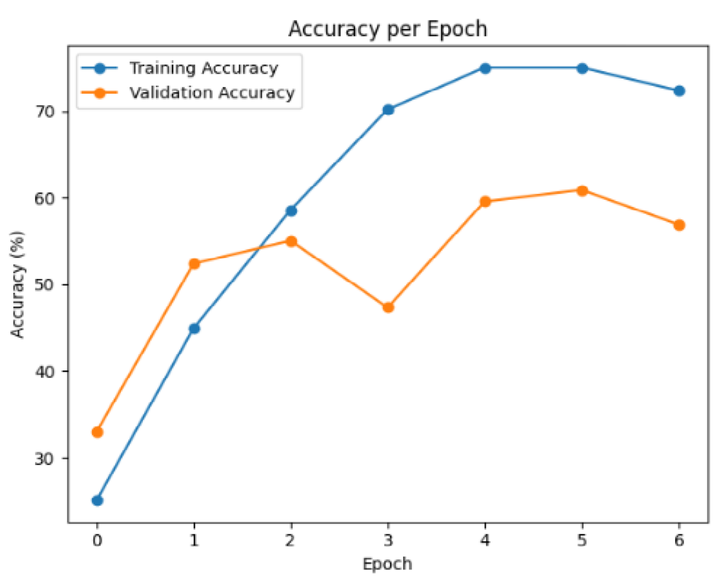

Figure 4: BERT train and validation Accuracy

5. Conclusion

In conclusion, the RoBERTa model demonstrated the strongest performance in detecting the veracity of news articles when compared to BERT. While both models faced challenges related to overfitting and learning stability, RoBERTa's fine-tuning on a trimmed dataset exhibited more promising results in terms of accuracy and reliability across different metrics.