# Fake News Detection

**Moroșan Eric-Alexandru** eric-alexandru.morosan@s.unibuc.ro
**Gheorghe Radu-Mihai** radu-mihai.gheorghe@s.unibuc.ro

## Abstract

In this project, we address the 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.

## 1 Introduction

There has been a massive growth of concern in information integrity in the past few years as misinformation has the potential to spread extremely fast on social media networks and news channels, with so many people misusing the internet. Our research focuses to address this problem by analysing and developing a new approach at classifying the accuracy of a piece of content in a more sophisticated way. This is where fake or real news is assessed as a multi-class generalised classification system instead of a single binary classification system oversimplifying the context. This multi-class system also fits better as legions of fabricated articles are scattered all around the internet and each article has some mix of truths, some more than others.

Research within this domain has significant consequences on the automated fact checking systems as well as the content moderation platforms. In particular, the ability to consistently judge the existence of partially true content, provides important subtlety to the content moderators and to the fact checkers, allowing them to better address the issue of misinformation.

## 2 Related Work

Determining claim credibility has become a significant research problem in recent years (Li et al., 2012). While check-worthy claims originate from various sources, research has predominantly focused on social media content (Gupta et al., 2014). News article verification has received some attention but has typically been treated as a binary classification problem. Research has shown that false news propagates faster than partially false news on social media platforms (Shahi et al., 2021a).

Several evaluation campaigns have addressed different aspects of claim verification. The CheckThat! lab at CLEF, RumourEval (Derczynski et al., 2017), (Shahi et al., 2021b), and initiatives at SemEval have explored various facets including stance detection (Sobhani et al., 2016) and propaganda identification (Martino et al., 2020). Other notable efforts include the FEVER task (Thorne et al., 2018), which focuses on fact extraction using Wikipedia evidence, and the FakeNews task at MediaEval (Pogorelov et al., 2021), which specifically addresses coronavirus-related misinformation.

## 3 Method

### 3.1 Dataset

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.

The dataset analysis continues with the content that allows us to identify several other features. This content presented as social media data has is-

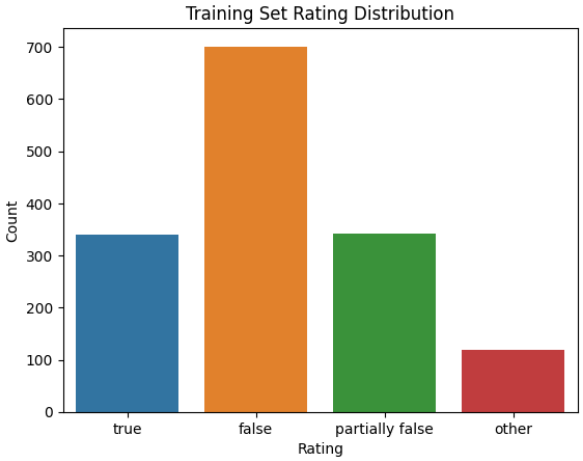


Figure 1: Training set distribution

sues related to class distribution which was skewed with misinformation being a lot of it. The length of articles in the set ranged from one or two paragraphs to lengthy reports on certain events. The issues related to the length of the articles posed a unique problem for transformer models due to their token limits.

### 3.2 Preprocessing

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**. [2 and 3]

For our article, we added some actions into our preprocessing pipeline seeking to improve data quality as well as the model accuracy. We also integrated text normalization techniques which included the elimination of special symbols, unification of quotation marks, and addressing whitespace errors.

The linguistic analysis included more than just word count calculations. Sentence structuring, quote frequency, and numerical claims across truth categories were areas of interest. The mentions of words such as “said” and “will” were overheard across all the categories indicating that quoting sources and making predictions about the future are staples of news articles regardless of the truth. This result emphasized the value of concentrating on more in-depth linguistic characteristics in the classification process.

For tokenization, we used the RoBERTa-

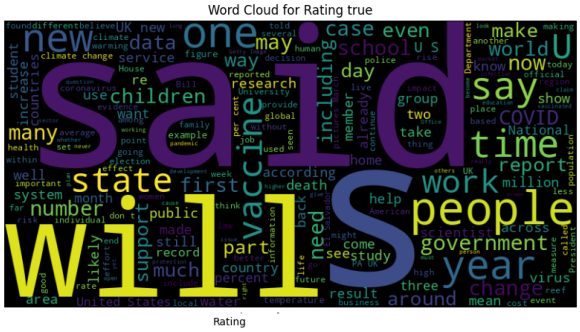


Figure 2: Word Cloud True Label

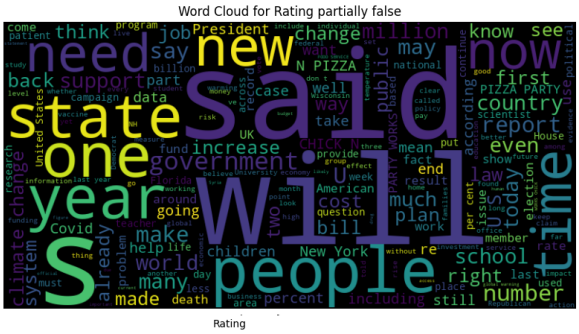


Figure 3: Word Cloud Partially False Label

Tokenizer, 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.

### 3.3 Models

Our experimental approach encompassed three main components: a baseline BERT implementation, an enhanced BERT fine-tuning strategy, and a modified RoBERTa architecture.

For the baseline model, we utilized a **pretrained BERT model** complemented by a weighted Cross Entropy Loss function to address class imbalance in the dataset. The model architecture consists of a standard BERT base model (bert-base-uncased) followed by a linear classification layer, with gradient accumulation through sub-batching to manage memory constraints effectively. We implemented early stopping with a patience of 3 epochs to prevent overfitting, while using the Adam optimizer with a learning rate of 1e-5 for stable training.

The confusion matrix [4] reveals that the model performs best at identifying "true" cases, with 110

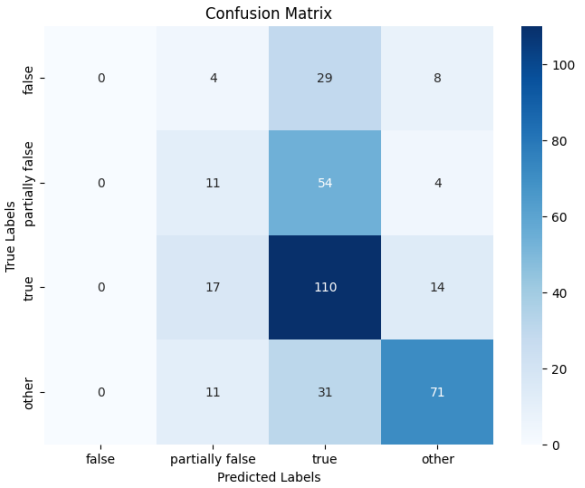


Figure 4: Bert Baseline Confusion Matrix

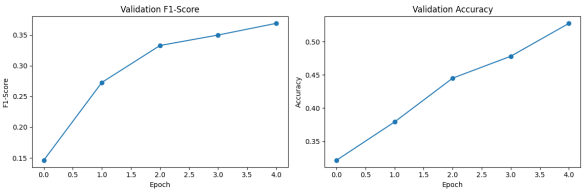


Figure 5: Bert Baseline Metrics

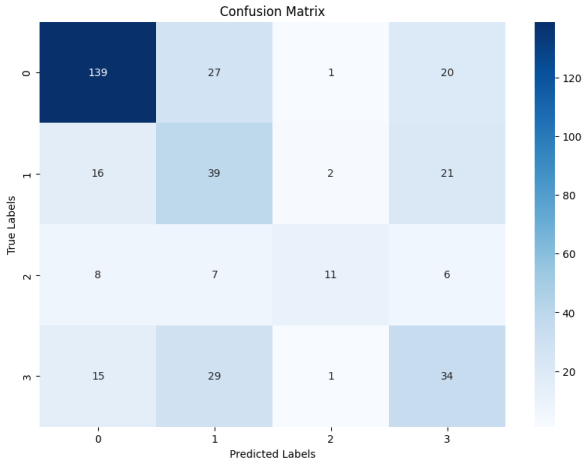


Figure 6: Bert Finetuned Confusion Matrix

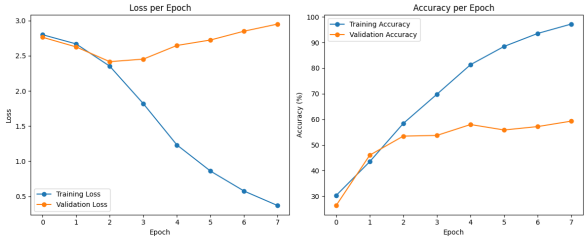


Figure 7: Bert Finetuned Metrics

correct predictions. However, there's notable confusion between "true" and "partially false" categories, with 54 instances of overlap, suggesting that the model sometimes struggles with fine-grained distinctions between completely true statements and those containing partial inaccuracies. The "other" category also shows significant mixing with both "true" and "partially false" predictions.

The validation metrics [5] demonstrate steady improvement across epochs, with the F1-score growing from 0.15 to 0.37 and accuracy increasing from 0.32 to 0.52. This consistent growth without significant plateauing suggests that the model might benefit from additional training epochs, though the improvement rate does begin to slow in later epochs. The smooth learning curves indicate that our training process is stable, likely due to the conservative learning rate choice and sub-batching strategy.

In our **enhanced BERT implementation**, we developed a more sophisticated BERT-based architecture with additional features to handle the task complexity. The model incorporates a sliding-window approach to process long articles, along with a dual dropout strategy (rate 0.2) and an intermediate dense layer for better feature extraction. Training utilized the Adam optimizer with a learning rate of 5e-5.

The confusion matrix [6] for this enhanced model shows significant improvement in class prediction accuracy, particularly for class 0 with 139 correct predictions. The model maintains strong performance across other classes as well, though there's still some confusion between adjacent classes, which is expected given the nuanced nature of the classification task.

Looking at the training curves [7], we observe an interesting pattern where the training loss steadily decreases from about 2.75 to 0.4, indicating good optimization progress. However, the validation loss shows signs of divergence after epoch 2, suggesting some overfitting despite our dropout measures. The accuracy plots reveal that while training accuracy climbs to nearly 0.97, validation accuracy plateaus around 0.60, further confirming the overfitting behavior. This suggests that additional regularization techniques might be worth exploring in future iterations.

For our **RoBERTa-based approach**, we employed the RobertaForSequenceClassification pretrained model, adapting the input data to conform to the 512-token limitation. We enhanced the base architecture by replacing the standard clas-

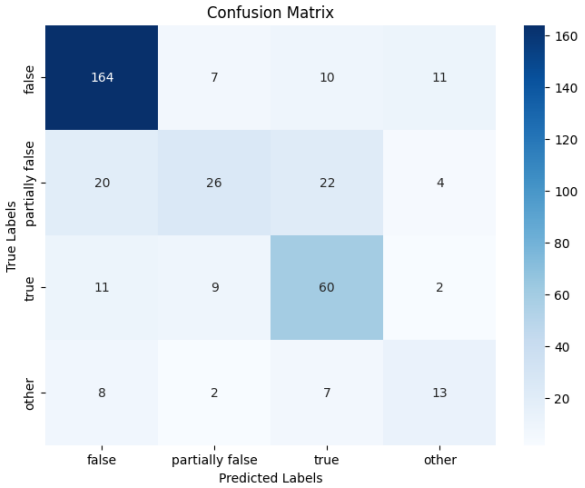


Figure 8: RoBERTa Confusion Matrix

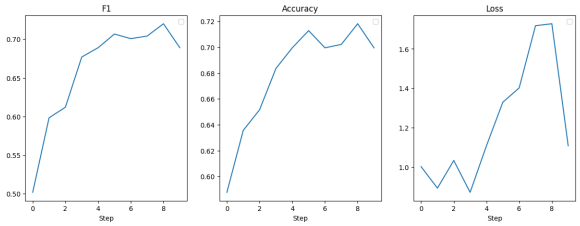


Figure 9: RoBERTa Metrics

sifier head with a custom classification module. This custom classifier consisted of two linear layers: the first layer reduced the dimensionality of RoBERTa's output to match the number of target classes, followed by a ReLU activation function and dropout regularization. The final linear layer generated the classification logits. This architectural modification was designed to better capture the nuanced differences between truth categories while maintaining the robust feature extraction capabilities of the base RoBERTa model.

The confusion matrix [8] demonstrates particularly strong performance in identifying false claims, with 164 correct predictions in this category. The model also shows good accuracy in identifying true statements (60 correct predictions), though there's some confusion in the partially false category where predictions are more evenly distributed among the classes. This suggests the model is most confident in identifying clear cases of false information but shows more uncertainty with partially false claims.

Looking at the training metrics [9], we observe steady improvement in both F1-score and accuracy, reaching peaks of approximately 0.72 and 0.71 respectively. The loss curve shows some interesting dynamics, with an initial decrease followed by fluctuations and a final sharp decline, indicating that the model eventually found a better optimization direction after some exploration. Despite these fluctuations, the consistent improvements in F1-score and accuracy suggest that the model successfully learned to distinguish between different truth categories while maintaining stability in its predictions.

| Architecture | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| BERT - BASE | 0.53 | 0.38 | 0.37 | 0.36 |
| BERT | 0.60 | 0.53 | 0.52 | 0.51 |
| ROBERTA | 0.71 | 0.69 | 0.72 | 0.68 |

## 4 Conclusion

The proliferation of fake news introduces a formidable threat and hinders societal advancement by losing faith in the medians, manipulating public perception and demonizing democratic systems. In this regard, AI has appeared to be an important instrument in the fight against disinformation. Artificial intelligence can quickly detect and label manipulative information through algorithmic technologies, natural language processing, and machine learning, thus contributing to safeguarding information ecosystems.

AI also helps to increase the accuracy in news interpretation allowing people, institutions, and countries to access relevant information as well as empowering the anti-fake news campaigns. In addition to that, it combats media bias by equipping the audience with resources that detect an agenda or a lie, thus promoting better understanding and use of the news. Furthermore, AI enabled means can keep pace with the changing strategies of the perpetrators ensuring that false information spread throughout society does not cause pool of resources across the adaptive area.

In the end, AI based means for detecting fake news will help to build a better society promoting accessibility of information and transparency of speech. It provides guarantees to the public that basic tenets such as democracy, accountability, and critical reasoning are preserved and remain necessary for sustaining material and moral progress in a highly competitive world.

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 promis-

ing results in terms of accuracy and reliability across different metrics.

## 5 Future Work

There are a number of areas worth pursuing in the field of AI-based fake news detection. First of all, there is the challenge of creating more advanced text preprocessing methods appropriate for long-form articles while taking care not to lose any contextual information. While our sliding window approach of text segmentation had some success, another approach such as hierarchical attention mechanisms or document level transformers may put the model in a better incomprehensible or follow-up setting and tend to achieve better results in the understanding that it has a greater context.

Another important challenge that must be addressed is the analysis of multiple media content. Since fake news is often disseminated in the form of text, images, and videos, designing models to jointly assess these types of content may help significantly increase the level of accuracy of the text. This may involve the use of vision transformers in conjunction with language models in order to examine the correspondence of the visual component to the text.

So is the problem of robustness of the model against new disinformation strategies changes remain. The future work may include development of new adaptable and learning based methods that automatically update and retrain the models whenever new trends and fakes are made available. Another approach could be looking into effective ways of dealing with edge cases and vague content, especially regarding the "partially false" cases where our predictive models were not performing that well.

## Limitations

- **Language Dependency**: The proposed method may exhibit language dependency, performing better on languages with limited morphology, such as English, compared to languages with more complex grammatical structures or multilingual content.

- **Cultural Sensitivity**: The effectiveness of the fake news detection model may vary across different cultural contexts. Cultural nuances, regional expressions, and varying journalistic practices might not be adequately represented in the training data, impacting the model's performance.

- **Ambiguity Handling**: The model may struggle to accurately classify content when the fake news is subtle or incorporates ambiguous cues, such as satirical news, opinion pieces, or borderline misinformation. Distinguishing between such cases and clear fake news remains a challenge.

- **Limited Generalization**: The performance of the proposed method may be limited to specific domains or genres of text that closely resemble the training data. Extending the model's capabilities to handle diverse sources, such as social media, blogs, and traditional news outlets, requires further adaptation and evaluation.

- **Scalability to Long Text**: While the proposed method shows promising results on shorter or medium-length text samples, its ability to scale effectively to analyze lengthy documents or multi-article discourse is limited. Techniques for processing long-form content efficiently are crucial for addressing this issue.

- **Resource Intensiveness**: The computational resources required for training and inference, particularly for large-scale datasets, can pose a barrier to the widespread adoption of the proposed method. Developing resource-efficient models and optimizing algorithms for deployment on hardware-constrained environments is an area for improvement.

- **Limited Robustness to Noisy Data**: The model's performance may degrade when dealing with noisy or incomplete input data, such as user-generated content with misspellings, grammatical errors, or informal language. Enhancing robustness through data augmentation or noise-resilient training strategies is essential.

- **Interpretability**: Despite high accuracy in detecting fake news, the inner workings of the model may lack interpretability, making it difficult to understand the rationale behind its predictions. Developing methods to improve model interpretability and provide clear explanations of decision-making processes is important for building trust with users.

By acknowledging these limitations, we aim to provide a balanced assessment of the proposed method's strengths and weaknesses, guiding future research and development in fake news detection to improve its effectiveness and applicability.

## Ethical Statement

When conducting research on fake news detection, as with any field, it is essential to be aware of potential ethical considerations and biases. Below are some unethical uses and biases that researchers in this domain should be mindful of:

- **Manipulation and Censorship**: Fake news detection models could be misused to suppress free speech or censor legitimate information under the guise of combating misinformation. Malicious actors might exploit these tools to target dissenting voices or manipulate online discourse for political or economic gain.
- **Sampling Bias**: The datasets used to train fake news detection models may not be representative of the diversity of language use, news sources, and perspectives across different demographics, cultures, and regions. This could result in models that unfairly flag or overlook content from certain groups, perpetuating inequalities or biases.
- **Labeling Bias**: The process of labeling data for fake news detection is inherently subjective and may be influenced by human biases, differing cultural norms, or political ideologies. Inconsistent or biased labeling can lead to unreliable models that reflect the annotators' preconceptions rather than objective assessments.
- **Overreliance on Automation**: Excessive dependence on automated systems for fake news detection can lead to ethical concerns if these systems lack transparency or accountability. Misclassifications by the models might have severe consequences, such as unjustly harming reputations or undermining trust in legitimate sources.

To address these ethical concerns, researchers in fake news detection should prioritize responsible data collection and annotation practices, ensure diversity and inclusivity in their datasets, and adopt transparent methods for model development and evaluation. They should also consider the broader social and ethical implications of their work and collaborate with experts from ethics, journalism, sociology, and other relevant fields to develop solutions that are fair, unbiased, and accountable.

## References


Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. pages 69–76.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. pages 228–243.

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth finding on the deep web: is the problem solved? *Proc. VLDB Endow.*, 6(2):97–108.

Giovanni Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. pages 1377–1414.

Konstantin Pogorelov, Daniel Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkuková, and Johannes Langguth. 2021. Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020.

Gautam Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021a. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104.

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021b. Overview of the clef-2021 checkthat! lab: Task 3 on fake news detection. In *CLEF (Working Notes)*, pages 406–423.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. pages 159–169.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.