Fake News Detection

Moroșan Eric-Alexandru & Gheorghe Radu-Mihai

eric-alexandru.morosan@s.unibuc.ro radu-mihai.gheorghe@s.unibuc.ro

1. Abstract & Introduction

This project focuses on developing an advanced fake news detection system that categorizes news articles into four truthfulness levels: true, partially true, false, or other. The approach utilizes **BERT and RoBERTa** models to handle text analysis within BERT's 512-token limit, with the best results achieved using a **fine-tuned RoBERTa mode**l on CLEF2021-CheckThat! competition data.

With the growing concern over misinformation's rapid spread across social media and news channels, this research introduces a sophisticated **multi-class classification system** instead of a simple binary approach. This nuanced method is particularly valuable for content moderation platforms and fact-checking systems, as it can better handle the complex reality of online articles that often contain varying degrees of truth, ultimately helping moderators and fact-checkers address misinformation more effectively.

2. Related work

Previous research in claim credibility has mainly focused on social media content (Gupta et al., 2014), with evaluation campaigns like CLEF CheckThat! and FEVER (Thorne et al., 2018) exploring **various aspects of fact-checking**, including stance detection (Sobhani et al., 2016) and propaganda identification (Martino et al., 2020). Studies by Shahi et al. (2021a) have shown that **completely false news spreads faster than partially false information on social platforms**, while recent initiatives like MediaEval's FakeNews task (Pogorelov et al., 2021) have specifically addressed coronavirus-related misinformation.

3. Method

In our study, we developed a systematic approach to tackle the challenge of multi-class fake news detection, focusing on classifying news articles across different levels of truthfulness. We adopted a comprehensive methodology that leverages transformer-based architectures to effectively process and analyze news content. Our experimental framework was built around three main components: **a baseline BERT** implementation, an **enhanced BERT** fine-tuning strategy, and a **modified RoBERTa architecture**.

The methodology was designed to address key challenges in news content classification, particularly the processing of lengthy articles within transformer model constraints and the nuanced nature of truthfulness categories. By implementing both BERT and RoBERTa architectures with strategic modifications, our approach aimed to optimize the balance between model capacity and computational constraints while maintaining robust classification performance. This comprehensive framework provides a foundation for accurate multi-class fake news detection while addressing the practical challenges of processing news content.

3.1 Dataset

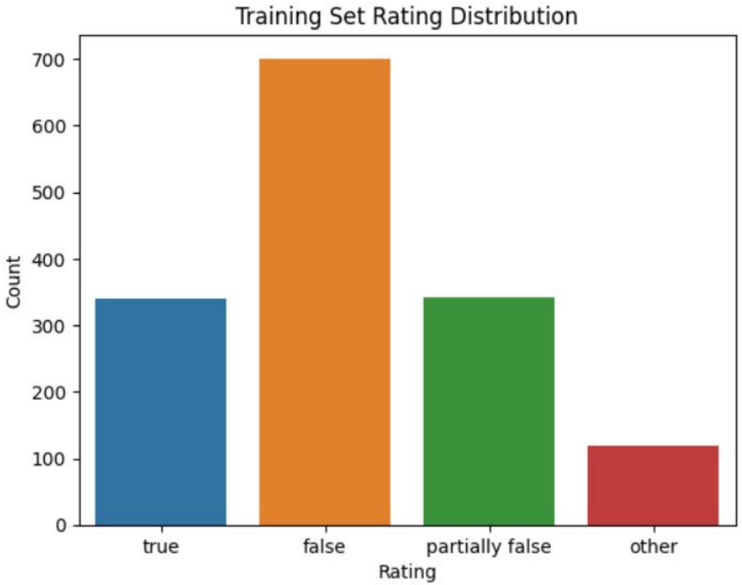
3.2 Preprocessing

3.3 Models

3.4 Experiments

3.1 Dataset

The **CLEF2021-CheckThat!** dataset contains 1,876 articles labeled as **true, partially true, false, or other**, split into training (80%) and testing (20%) sets. Each article contains title, text, and truthfulness rating, with some entries missing titles. The dataset exhibits key characteristics including a skewed class distribution towards misinformation, varying article lengths from short paragraphs to lengthy reports, and token length constraints that pose challenges for transformer models.



3.2 Preprocessing

Our data preparation pipeline began by concatenating article titles and text, followed by comprehensive cleaning procedures including stopword removal, emoji and URL elimination, and lowercase conversion. The enhanced preprocessing incorporated text normalization techniques to ensure data consistency and quality. Linguistic analysis revealed interesting patterns across truth categories, such as the consistent use of source quotations and future predictions, highlighting the importance of deeper linguistic features in classification.

For text processing, we implemented **dual tokenization approaches** using **RoBERTa and BERT architectures**, each offering unique advantages in handling news content. These preprocessing decisions were driven by the specific challenges of news article analysis, where preserving linguistic nuances is crucial for truthfulness classification while maintaining computational efficiency.

RoBERTa-Tokenizer is an advanced tokenization method that uses byte-level **Byte-Pair-Encoding** derived from GPT-2. It treats spaces as part of the tokens, meaning the same word can be encoded differently based on its position in a sentence. This approach allows for **more nuanced text representation** by considering contextual positioning of words, which is particularly valuable for analyzing news article structure and style.

BertTokenizer utilizes the **WordPiece** model, a subword tokenization algorithm that breaks down words into **commonly occurring pieces**. This approach optimizes vocabulary size by finding the most frequent subword units, enabling efficient handling of rare words and morphological variations. For example, "playing" might be broken into "play" and "##ing", allowing the model to understand related word forms while maintaining a manageable vocabulary size. This is particularly useful when processing news articles that may contain specialized or uncommon terms.

Text Normalization encompasses various techniques to standardize text format, including special symbol elimination, quotation mark unification, and whitespace correction. This process ensures consistency across the dataset and reduces noise that could interfere with model learning, while preserving meaningful linguistic features that are crucial for truthfulness classification.

3.3 Models

BERT (Baseline Implementation) serves as our foundation model, utilizing its bidirectional training approach to understand news article context from both directions. This transformer-based architecture excels at capturing complex relationships in text, making it particularly suitable for detecting subtle variations in truthfulness across news content.

BERT (Enhanced Fine-tuning) builds upon the baseline by implementing specialized fine-tuning strategies optimized for multi-class truthfulness detection. This enhanced approach focuses on adapting BERT's pre-trained knowledge to the specific nuances of news article verification while maintaining computational efficiency.

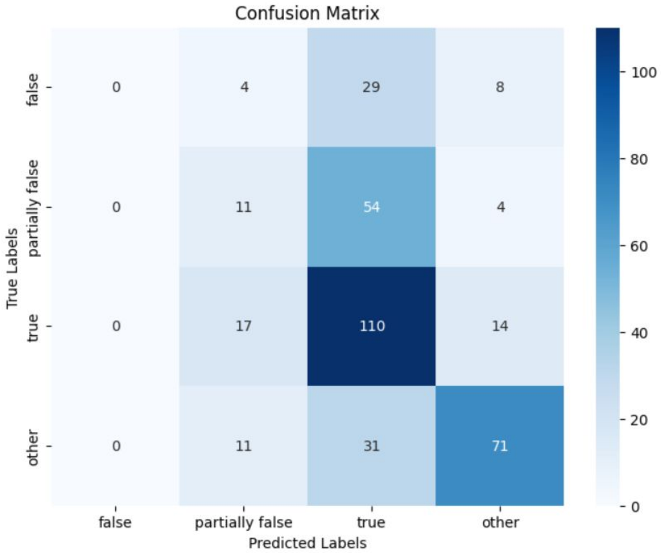
RoBERTa (Modified Architecture) represents an optimized version of BERT with enhanced training methodology and modified hyperparameters. Our implementation includes specific adaptations for handling news article length and preserving crucial linguistic features that distinguish between different levels of truthfulness. RoBERTa's robust architecture proves particularly effective in capturing the subtle distinctions between partially true and false content.

3.4 Experiments

a. BERT Baseline Model

The foundation of our approach utilizes a pre-trained BERT model (bert-base-uncased) with a linear classification layer, optimized using **weighted Cross Entropy Loss** to address class imbalance. Training was conducted using Adam optimizer with a 1e-5 learning rate, incorporating gradient accumulation through sub-batching to manage memory constraints. Early stopping with 3 epochs patience was employed to prevent overfitting.

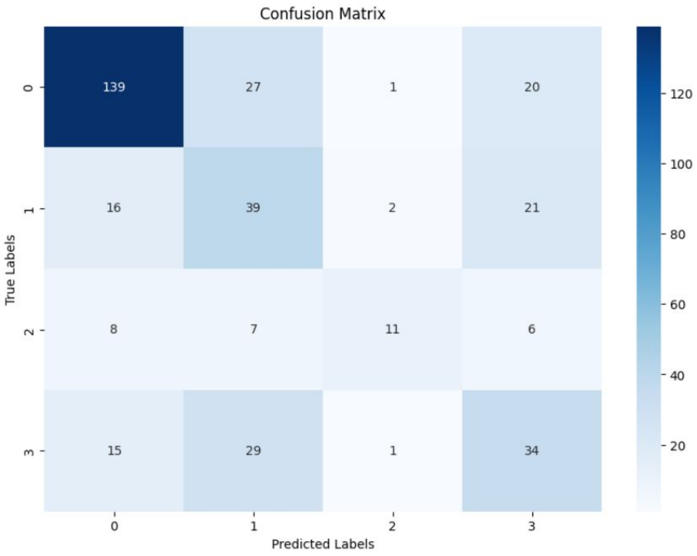
Performance analysis revealed strong accuracy in identifying "true" content (110 correct predictions), though the model showed some difficulty distinguishing between "true" and "partially false" categories (54 overlap instances). The validation metrics demonstrated steady improvement, with F1-score increasing from 0.15 to 0.37 and accuracy from 0.32 to **0.52**, suggesting **potential for further optimization with extended training**. The stable learning curves validated the effectiveness of the chosen conservative learning rate and sub-batching strategy.



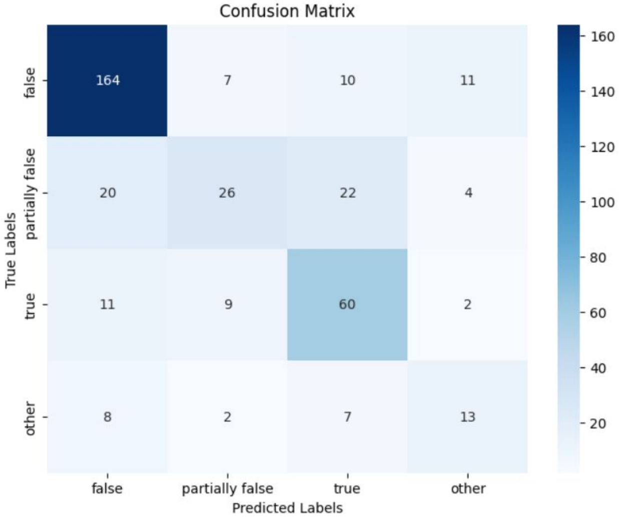
b. BERT Enhanced Model

The **enhanced BERT architecture** introduces a sliding-window mechanism to process lengthy articles, coupled with a dual dropout strategy (rate 0.2) and an intermediate dense layer for improved feature extraction. The training process employed Adam optimizer with a 5e-5 learning rate, resulting in **notable performance improvements** across all classes, particularly for class 0 with 139 correct predictions.

Analysis of model performance revealed interesting patterns: while the training loss showed steady improvement (decreasing from 2.75 to 0.4), the validation metrics indicated potential overfitting after epoch 2. Training accuracy reached 0.97, but validation accuracy plateaued around **0.60**, suggesting the need for additional regularization techniques in future iterations. Despite some remaining confusion between adjacent classes, this enhanced architecture demonstrated significant improvement over the baseline in handling the nuanced nature of truthfulness classification.



c. RoBERTa Model



	false	partially false	true	other
false	164	7	10	11
partially false	20	26	22	4
true	11	9	60	2
other	8	2	7	13

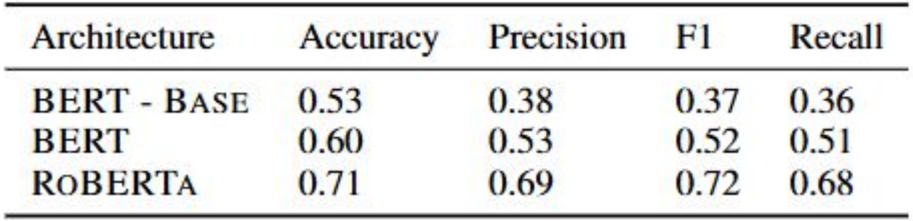
The final architecture leverages **RobertaForSequenceClassification** with a customized classification module, designed to handle the 512-token limitation while maintaining optimal performance. The **modified structure replaces the standard classifier head with two linear layers**: one reducing RoBERTa's output dimensionality to match target classes (with ReLU activation and dropout), and another generating classification logits. This design specifically targets the nuanced distinctions between truth categories.

Performance metrics demonstrated exceptional accuracy in identifying false claims (164 correct predictions) and true statements (60 correct predictions), though partially false categories showed more distributed predictions. The training curves revealed steady improvement, with F1-score and accuracy peaking at approximately **0.72 and 0.71 respectively**. Despite some loss curve fluctuations, the model exhibited consistent performance improvements, indicating successful adaptation to the complex task of multi-class truthfulness classification.

Results

Analysis of model performance across architectures demonstrates a clear progression in classification capability. The baseline BERT model achieved moderate performance with 0.53 accuracy and 0.37 F1-score, establishing a foundation for comparison. The enhanced BERT implementation showed notable improvement, reaching 0.60 accuracy and 0.52 F1-score through its sophisticated architecture and fine-tuning strategies.

RoBERTa emerged as the superior architecture, significantly outperforming both BERT implementations across all metrics. With **0.71 accuracy, 0.69 precision, and 0.72 F1-score**, RoBERTa's custom classification head and architectural modifications proved particularly effective in handling the nuanced nature of truthfulness classification in news articles. The substantial improvement in recall (**0.68**) also indicates RoBERTa's enhanced ability to identify relevant instances across all truth categories.



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 study demonstrates the effectiveness of transformer-based architectures in multi-class fake news detection, with RoBERTa emerging as the most successful model. The progression from baseline BERT to enhanced BERT, and finally to RoBERTa, showcases the importance of architectural refinements in handling the nuanced nature of truthfulness classification. Beyond technical achievements, this research contributes to the broader goal of combating misinformation through AI.

The implications of this work extend to practical applications in safeguarding information ecosystems. By developing more accurate classification systems, we enhance **the ability to detect and label manipulative information, supporting critical reasoning and transparency in news consumption**. This technology serves as a crucial tool in maintaining information integrity and promoting democratic values in an increasingly complex media landscape.

5. Future work

Looking ahead, our research in multi-class fake news detection opens several promising avenues for advancement. While our current work establishes a foundation for distinguishing between different levels of truthfulness in news content, there are numerous opportunities for enhancement and expansion.

The **integration of advanced preprocessing techniques** and **cross-modal analysis** could significantly improve our models' performance. By incorporating both textual and visual elements of news articles, along with better context preservation methods, we could develop more robust systems capable of handling the complex nature of modern misinformation.

Future applications could extend to **real-time news verification systems, social media content moderation, and automated fact-checking tools**. These implementations would contribute to creating more reliable information ecosystems and supporting critical media literacy. By building upon our current findings with **RoBERTa** and exploring new architectural approaches, we can continue to advance the field of automated fake news detection while addressing the evolving challenges of misinformation in digital media.

Limitations

Language and Cultural Dependency: Limited effectiveness across different languages and cultural contexts

Content Ambiguity: Challenges in classifying subtle fake news, satire, and opinion pieces

Generalization Issues: Performance may vary across different text sources and genres

Technical Constraints: Scalability challenges with long-form content and resource-intensive processing requirements

Data Quality Sensitivity: Reduced performance with noisy or informal content

Limited Interpretability: Difficulty in explaining the model's decision-making process

These limitations highlight areas for future improvement in developing more robust and widely applicable fake news detection systems.

Ethical statement

Research in fake news detection must carefully balance combating misinformation with protecting free speech. Key concerns include **potential misuse** for censorship, **dataset biases** in cultural representation, **subjective labeling influences**, and **risks of over-automation**.

The development of these systems requires responsible data practices, transparent methodology, and cross-disciplinary collaboration to develop fair and accountable solutions while maintaining information integrity.

References

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.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Realtime 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.