Rezumat
„Advancements in Fake News Automatic Detection”,

Ștefan Emil Repede

The unchecked proliferation of misinformation in the digital age has profoundly eroded public trust in institutions, skewed democratic processes, and even endangered public health. **Fake news**, broadly **defined in this work** as *“information that is verifiably false and deliberately spread to mislead”*, represents a pressing socio-technical challenge. Despite increased efforts in manual fact-checking and content moderation, the sheer volume and velocity of misleading content online highlight the urgent need for advanced automated detection systems. This doctoral thesis addresses that need by advancing the field of fake news detection through a multifaceted approach that spans theory, dataset creation, model development, and extensive evaluation.

To ground the research, we first establish a clear operational definition and scope for fake news. Building on this definition, we use established datasets but also curate and **annotate two bilingual datasets (English and Romanian)** to support both binary classification (real vs. fake) and nuanced multi-class classification. The proposed datasets incorporate **a rich labeling taxonomy** beyond the simplistic true/false dichotomy, including intermediate labels for partially false or misleading content, to reflect real-world complexity. In our example, news items are tagged not only as “true” or “false,” but also as “mostly true,” “mostly false,” or “misleading”.

Drawing on a comprehensive literature review, we identify four principal perspectives in automated misinformation detection – **fact-checking**, **stylometric/textual analysis**, **propagation dynamics**, and **source credibility**. Insights from these perspectives inform a **refined taxonomy of fake news types**. We differentiate between related concepts such as disinformation (deliberate falsehoods), misinformation (unintentional inaccuracies), MALinformation (genuine information used maliciously), satire/parody, propaganda, clickbait, and manipulated media. By categorizing fake news along dimensions of intent (e.g. malicious vs. accidental), style (e.g. news report vs. satire), and motivation/impact (political, financial, or humorous), the thesis contributes to a structured labeling methodology that clarifies the boundaries of what constitutes “fake news” under our definition. This not only aids in more granular annotation of the datasets, but also guides feature selection and model architecture, ensuring that detection methods are attuned to the diverse **forms of deceptive content**, from outright hoaxes and fabricated stories to subtly misleading half-truths. We also highlight the **multifaceted role of Large Language Models (LLMs)** in the fake news ecosystem, demonstrating how the same advances in AI that enable fluent text generation can **paradoxically both exacerbate and help combat fake news**. On one hand, cutting-edge LLMs can produce highly persuasive false narratives at scale; on the other, their deep linguistic understanding can be harnessed to identify inconsistencies and semantic cues indicative of deception.

Using the latest advances in natural language processing, the thesis introduces two **transformer-based detection models** that form the core technical contribution of this work. **FakeLuke**, a novel model we developed, builds upon the LUKE architecture (Language Understanding with Knowledge Embeddings) and is **purpose-built for binary fake news classification**. It incorporates an **entity-aware self-attention mechanism**, allowing it to recognize and cross-reference key entities (people, places, organizations) within text. The result is a classifier that achieved **state-of-the-art performance on standard benchmarks** (over 99% accuracy on the widely used ISOT fake news dataset), substantially outperforming classic models like BERT or logistic regression in terms of precision and recall. Complementing FakeLuke, we present a **fine-tuned version of Meta's LLaMA 3**, an advanced LLM, adapted specifically for the task of fake news detection across both English and Romanian. This model, extended with a sequence classification head, is trained to handle the more **nuanced, multi-class labeling** scheme of our bilingual dataset. We employ **Parameter-Efficient Fine-Tuning (PEFT)** techniques, notably **Low-Rank Adaptation (LoRA)**, to efficiently adapt LLaMA 3's billions of parameters to our domain without prohibitive computational cost. These **architectural innovations** (including mixed-precision training and language-specific tokenization optimizations for Romanian) enable the model to retain the rich semantic knowledge of its pre-training while **focusing on domain-specific cues of misinformation**. Notably, by updating only small weight matrices via LoRA, the fine-tuning process becomes memory-efficient and scalable, making it feasible to redeploy the detector for other languages or evolving news topics in the future. Together, FakeLuke and the enhanced LLaMA 3 model represent a two-pronged modeling strategy: one focused on lean, targeted binary classification with interpretable entity focus, and one using massive pre-trained knowledge for both binary and multi-label multilingual classification.

We validate our approaches through **experiments and comparative evaluations** against a wide range of baseline methods. Traditional machine learning classifiers (such as SVM, Naïve Bayes, and Random Forest), deep neural networks (CNNs, LSTMs, and transformer models like BERT/RoBERTa), and other state-of-the-art LLMs (including OpenAI's GPT-4 and Google's Gemini, where possible) are benchmarked on our datasets. We report performance using standard evaluation metrics – **accuracy, precision, recall, F1-score**, and other appropriate measures of agreement and overlap, including **Agreement rates**, to ensure rigorous assessment. The results demonstrate the efficacy of our proposed models. **FakeLuke** consistently outperforms earlier approaches on binary classification tasks, achieving high precision and a low false-alarm rate, which is crucial for maintaining trust in a real-world deployment. The fine-tuned **LLaMA 3** model proves even more powerful on the bilingual, multi-class tasks: it not only handles English and Romanian news with equal adeptness, but also correctly distinguishes fine-grained truthfulness categories better than other competing LLMs. Notably, our LLaMA-based detector showed improvements of several percentage points in F1-score over vanilla LLaMA and GPT-4 baselines on the nuanced classification task, setting a new state-of-the-art for that evaluation. We analyze misclassification cases and find that the entity-aware attention in FakeLuke helps catch subtle misinformation (e.g. misattributed quotes or dates), while the broad semantic coverage of LLaMA 3 helps in understanding idiomatic or context-heavy false claims in Romanian that stump simpler models. Our experiments also highlight

the importance of **data curation**: by balancing class distributions and removing extraneous cues (like author names or location tags that models might latch onto), we ensure the models learn genuine patterns of deception rather than spurious correlations. This careful dataset preparation contributed to a notable increase in cross-domain robustness, as evidenced when testing the models on unseen news articles outside the training set.

Finally, the thesis **acknowledges current limitations** and charts several promising avenues for **future research**. One limitation is that our work focuses primarily on textual news articles; however, misinformation today is increasingly **multimodal**, often combining text with fabricated images, deepfake videos, or synthetic audio. To address this, we outline plans for extending fake news detection to **mixed-media content**, for instance, developing models that can jointly analyze an article's text and a related video or image to detect inconsistencies (an approach critical for platforms like YouTube or TikTok where audio-visual misinformation thrives). Another challenge is the **interpretability of LLM-based detectors**. While large models like LLaMA 3 are highly accurate, they can behave as "black boxes," making it difficult for journalists and policymakers to trust and act on their outputs. We discuss techniques to enhance **LLM interpretability and transparency**, such as integrating explainable AI methods that highlight which phrases or patterns influenced a model's decision, or using knowledge graphs to trace claims back to sources. Additionally, the ethical dimension of automated detection is explored: we advocate for **ethics-first system design** to ensure that these AI tools are used responsibly, do not infringe on free expression, and incorporate human oversight. **Interdisciplinary collaboration**, bringing together AI experts, social scientists, and media scholars, is emphasized as crucial for keeping detection methods aligned with real-world needs and values.

In summary, this doctoral research delivers a comprehensive contribution to the fight against online misinformation. We provide: (1) a clear definitional framework and taxonomy that sharpen the conceptual understanding of fake news; (2) enriched multilingual datasets and labeling schemes that capture the full spectrum of veracity in news content; (3) advanced detection models (FakeLuke and a LLaMA 3 variant) that significantly advance the state-of-the-art in accuracy and cross-lingual capability; and (4) thorough empirical evidence and analysis validating these innovations against existing methods. These contributions not only improve current detection performance but also lay a versatile foundation for future developments. By bridging technical advances in transformer-based AI with insights from journalism and cognitive science, the thesis sets the stage for next-generation **multimodal misinformation detection systems** that are more accurate, interpretable, and resilient. Ultimately, our work aims to strengthen the integrity of digital information ecosystems, helping society **detect and deter fake news more effectively** while updating a roadmap of ongoing research to outpace the evolving tactics of misinformation spreaders.

Student-doctorand,