Final Report

Xinze Yu

Misinformation Detection

Abstract

Misinformation detection is a crucial area of research that focuses on identifying and mitigating the spread of false information, particularly on social media and online platforms. This report provides an in-depth overview of misinformation detection techniques based on three papers. It examines the role of large language models (LLMs), machine learning-based approaches, and deep learning models in detecting misinformation. The report highlights the effectiveness of various methodologies, the datasets used, and the challenges associated with misinformation detection. Additionally, it discusses emerging trends and areas for future research to improve the accuracy and adaptability of misinformation detection systems.

1 Introduction

Misinformation refers to false or misleading information that is disseminated with or without malicious intent. The increasing prevalence of misinformation has severe implications for politics, health, and society at large. Misinformation can spread rapidly through social media platforms, where it is often amplified by algorithms that prioritize engagement. This has led to the development of misinformation detection systems that utilize advanced artificial intelligence techniques to analyze textual content, user interactions, and propagation patterns. With the advancement of artificial intelligence, researchers have developed sophisticated algorithms to detect misinformation by leveraging both linguistic and contextual features of online content.

2 Methods of Misinformation Detection

2.1 Large Language Models(LLMs) in Misinformation Detection

The study "Explore the Potential of LLMs in Misinformation Detection" (arXiv:2311.12699) evaluates the effectiveness of LLMs such as GPT-3.5, GLM, and Mistral. The study categorizes misinformation detection into two approaches:

1. LLM-based Detectors**:** Direct application of LLMs with task-specific prompts to detect misinformation. These models are tested using various prompting strategies, such as chain-of-thought(CoT) reasoning, which allows them to generate more context-aware responses.
2. LLM-enhanced Detectors**:** Integration of LLM-generated embeddings and synthetic data to enhance traditional machine learning models. This method helps improve feature extraction, making traditional classifiers more robust against misinformation.

Findings indicate that while LLMs can perform comparably to smaller fine-tuned models in text-based misinformation detection, their ability to analyze propagation structures remains limited. However, LLMs contribute significantly to data augmentation and feature enhancement, which can improve misinformation detection pipelines.

2.2 Machine Learning-Based Approaches

The literature review "Misinformation Detection: Datasets, Models, and Performance"(https://www.emerald.com/insight/content/doi/10.1108/oir-06-2024-0388/full/html) identifies commonly used machine learning and deep learning models for misinformation detection. The study categorizes computational approaches as follows:

- **Classic Machine Learning:** Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT) for basic classification tasks. These models rely on handcrafted features such as word frequency, sentiment analysis, and linguistic markers to distinguish misinformation.
- **Deep Learning Models:** Transformer-based models like BERT, as well as CNN, LSTM, and BiGRU, which use contextual embeddings to classify misinformation with high accuracy. These models leverage massive pre-trained datasets to improve their ability to identify subtle differences between real and false news.

The study finds that approximately 65% of misinformation detection models achieve an accuracy of 90% or higher, demonstrating the effectiveness of these methods. However, deep learning models require large labeled datasets and significant computational resources, which may limit their accessibility.

2.3 Fake News Detection from a New Perspective

The paper "An Overview of Fake News Detection"(https://doi.org/10.1016/j.fmre.2024.01.017) introduces a novel classification of misinformation detection approaches based on three key aspects:

- **Intentional Creation:** Identifying textual and stylistic features that indicate deliberate fabrication. This approach examines how fake news articles use exaggerated claims, emotionally charged language, and misleading formatting to appear credible.
- **Heteromorphic Transmission:** Analyzing the way fake news spreads differently from genuine news. Researchers have found that fake news tends to be shared more rapidly and widely than factual information, often involving bots and coordinated campaigns.
- **Controversial Reception:** Investigating public reactions and sentiment discrepancies toward fake news. Fake news tends to generate polarized opinions, with some users endorsing it while others actively debunk it.

This perspective emphasizes the importance of incorporating social dynamics and propagation behavior in misinformation detection models. The study suggests that combining text-based and propagation-based approaches can significantly improve detection accuracy.

3 Datasets Used in Misinformation Detection

The reviewed studies identify several benchmark datasets frequently used in misinformation detection research:

- **FakeNewsNet**: Aggregates fake and real news articles labeled by fact-checking organizations. It is widely used to train and test machine learning models.
- **LIAR Dataset**: Contains short political statements classified into six truthfulness categories ranging from "true" to "pants-on-fire false."
- **CoAID Dataset**: Focuses on COVID-19 misinformation, incorporating news articles, tweets, and user engagements.
- **PHEME Dataset**: Includes rumors and non-rumors from social media platforms, with annotations about their credibility.
- **Twitter-based Datasets (Twitter15, Twitter16)**: Provide misinformation samples from social media, often used for studying propagation patterns.

These datasets serve as critical resources for training and evaluating misinformation detection algorithms. However, researchers acknowledge the limitations of these datasets, including potential biases and the need for continuous updates to reflect evolving misinformation tactics.

4 Challenges in Misinformation Detection

Despite the advancements in misinformation detection, several challenges persist:

- **Adaptability:** Misinformation evolves over time, requiring models to continuously update and adapt. New forms of misinformation, such as deepfake videos, require innovative detection methods.
- **Context Understanding:** Nuanced misinformation, such as satire or misleading headlines, is difficult to classify. Models must be able to differentiate between intentional misinformation and humor or opinion.
- **Propagation Analysis:** LLMs struggle to process graph-based propagation structures effectively. Current models lack the ability to fully capture the relationships between users and how misinformation spreads through networks.
- **Data Limitations:** High-quality labeled misinformation datasets are limited and require significant human effort for annotation. Crowdsourcing and semi-supervised learning approaches are being explored to address this issue.

5 Conclusion

Misinformation detection has made significant progress through the use of LLMs, deep learning models, and machine learning techniques. While LLMs provide promising enhancements, traditional machine learning and deep learning models remain more reliable for structured misinformation detection. The integration of textual analysis, social context, and propagation patterns offers a comprehensive approach to addressing misinformation challenges. Future research should focus on improving the adaptability of models, enhancing dataset quality, and refining methodologies to detect misinformation more accurately. Additionally, interdisciplinary collaboration between AI researchers, social scientists, and policymakers is necessary to develop more robust misinformation detection frameworks.

LLM Agents

Abstract

Large Language Model (LLM) agents have emerged as a transformative technology for decision-making, planning, and multi-step reasoning. These agents utilize pre-trained LLMs to process and generate information, facilitating applications in problem-solving, automation, and decision-making. This report presents an extensive review of LLM agent frameworks based on four recent academic papers. It examines various methodologies, including experiential learning, modular benchmarking, multi-agent collaboration, and advanced planning mechanisms. The report highlights the effectiveness of these approaches, the datasets and benchmarks used, and the challenges associated with deploying LLM agents in real-world applications. Furthermore, it discusses potential future research directions aimed at improving adaptability, efficiency, and robustness of LLM agents.

6 Introduction

The rapid advancements in LLMs have led to their adoption in autonomous agent-based systems that require sophisticated reasoning, adaptation, and planning capabilities. These agents leverage LLMs to process textual data, generate responses, and interact dynamically with environments. The ability of LLM agents to execute complex tasks has led to significant interest in their applications in fields such as robotics, customer support, code generation, and research automation.

Recent research efforts have focused on improving LLM agent adaptability, learning efficiency, and evaluation methodologies. However, despite their potential, LLM agents still face challenges such as maintaining long-term memory, handling complex planning tasks, and ensuring consistency in performance evaluation. This report explores cutting-edge methodologies that seek to address these challenges and improve the overall efficiency and effectiveness of LLM-based agents.

7 Methods of LLM Agents

7.1 Experiential Learning for LLM Agents(ExpeL)

The Experiential Learning (**ExpeL**) framework introduces an experiential learning mechanism where LLM agents autonomously gather and refine experiences to improve decision-making without parameter updates.(https://arxiv.org/abs/2308.10144)

Key Components of ExpeL:

- **Experience Collection** - The agent interacts with multiple tasks and documents both successful and failed attempts.
- **Knowledge Abstraction** - Insights from experiences are stored in a structured format for future retrieval.
- **Application of Learned Insights** - The agent recalls and applies past experiences to enhance decision-making in new tasks.

Findings and Contributions:

- **ExpeL** operates without requiring fine-tuning of LLM weights, making it adaptable to proprietary models like GPT-4 and Claude.
- It outperforms baseline decision-making agents across multiple domains without needing extensive human supervision.
- The study demonstrates the agent's ability to generalize across tasks through the abstraction of prior experiences, resembling human-like learning processes

7.2 Benchmarking and Evaluation Frameworks(AgentQuest)

AgentQuest is a modular benchmarking framework designed to assess LLM agents using structured, multi-faceted evaluation metrics.(https://arxiv.org/abs/2404.06411)

Key Features:

- **Modular API Design** - Allows integration with various benchmarking tools and datasets.

- **Progress Rate (PR)** - Measures the incremental advancement of an agent toward completing a task.
- **Repetition Rate (RR)** - Tracks redundant steps to help identify inefficiencies in agent workflows.
- **Debugging Capabilities** - Identifies specific failure points in LLM agent architectures and suggests refinements.

Findings and Contributions:

- **AgentQuest** provides a detailed, metric-driven analysis of agent performance, enabling researchers to refine LLM-based architectures effectively.
- It has been successfully applied in structured problem-solving environments such as ALFWorld and Sudoku, where it has highlighted key failure patterns in multi-step reasoning tasks.
- By providing a standardized benchmarking framework, AgentQuest enables reproducibility in LLM agent research and comparison across different architectures.

7.3 Multi-Agent Collaboration with AutoGen

AutoGen is a multi-agent framework that enhances LLM collaboration through structured dialogues and flexible interaction protocols.(https://arxiv.org/abs/2308.08155)

Key Features:

- **Conversable Agents** - Enables multiple LLMs, external tools, and human inputs to work together seamlessly.
- **Flexible Conversation Programming** - Supports both static predefined workflows and dynamic adaptive interactions.
- **Hierarchical and Joint Task Execution** - Allows agents to function independently or coordinate tasks through structured dialogues.

Findings and Contributions:

- Empirical evaluations demonstrate that AutoGen outperforms single-agent approaches by effectively distributing workloads among specialized agents.
- By structuring LLM interactions, AutoGen reduces task redundancy and enhances the reliability of complex decision-making processes.

7.4 Planning Strategies for LLM Agents

A systematic survey on **LLM-based planning** categorizes existing methodologies into five key strategies.(https://arxiv.org/abs/2402.02716)

Key Features:

- **Task Decomposition** – Breaking down complex tasks into structured, manageable sub-problems.
- **Plan Selection** – Generating multiple plans and choosing the most optimal strategy based on evaluation metrics.
- **External Module Assistance** – Integrating third-party tools,APIs and databases to refine decision-making processes.
- **Reflection and Refinement**- Allowing agents to self-evaluate, learn from mistakes and iteratively improve.
- **Memory-Augmented Planning** – Storing previous decision-making paths and retrieved knowledge to enhance long-term adaptability

Findings and Contributions:

- The study identifies **task decomposition** as one of the most effective methods for improving LLM planning efficiency.
- **Reflection and refinement techniques** allow LLMs to correct mistakes and iteratively improve performance.
- **Memory-augmented planning** enables long-term retention of past strategies, leading to more consistent decision-making

8 Challenges and Future Research Directions

Despite significant advancements, LLM agents still face several challenges:

- **Contextual Retention** - Maintaining memory consistency across multi-turn interactions remains an open problem.
- **Efficiency Trade-offs** - Balancing computational cost with real-time adaptability is critical for scalable deployment.
- Generalization - Ensuring agents can generalize their problem-solving approaches across diverse environments without overfitting to specific datasets.
- **Benchmarking Consistency** - Standardizing evaluation metrics across different LLM architectures remains a research priority.

Future research should focus on:

- Enhancing agent autonomy by improving self-reflection and iterative learning.
- Reducing computational overhead to enable real-time applications in resource-constrained environments.
- Developing unified benchmarking methodologies to standardize performance assessment across LLM architectures.
- Improving collaborative planning by enabling more sophisticated interactions between multiple agents.

Combination LLM Agents with Misinformation Detection

Large Language Model Agent for Fake News Detection

Abstract

This paper introduces FactAgent, an agentic LLM-based framework for misinformation detection that systematically verifies news claims without requiring additional training. Unlike traditional LLM fact-checking approaches that classify claims in a single step, FactAgent follows a structured workflow, breaking down the verification process into multiple sub-tasks. These sub-tasks leverage both internal LLM knowledge and external tools, such as search engines and credibility assessments, to ensure comprehensive claim verification. Experimental results demonstrate that FactAgent outperforms supervised models and non-agentic LLM approaches on three benchmark datasets, proving its effectiveness in misinformation detection. (https://arxiv.org/abs/2405.01593)

Methodology

FactAgent's workflow consists of the following steps:

- **Phrase Analysis** – Detects sensationalist and emotionally charged language.
- **Linguistic Analysis** – Identifies grammatical and style inconsistencies.
- **Commonsense Verification** – Checks claims against general knowledge.
- **Political Standing Analysis** – Identifies partisan bias in claims (if applicable).
- **Search-Based Evidence Retrieval** – Uses search engines to find conflicting reports.
- **URL Credibility Check** – Evaluates the source's reliability.

At the final stage, FactAgent aggregates findings from all sub-steps and makes a **transparent, explainable decision** regarding claim veracity.

Experiments

The model was evaluated on PolitiFact(political fact-checking dataset),GossipCop(entertainment and celebrity news),Snopes(general misinformation dataset) and compared with traditional NLP models(LSTM,TextCNN, BERT) and LLM-based approaches(Zero-shot promting,CoT,HiSS(Hierarchical Step-by-Step prompting).

Result

- Consistently outperformed all baseline models:
- Higher accuracy and F1 scores
- Better interpretability
- No need for training on labeled datasets.

Web Retrieval Agents for Evidence-based Misinformation Detection

Abstract

This paper proposes a retrieval-augmented LLM agent for misinformation detection, combining LLM reasoning with external web searches**.** Instead of relying solely on LLMs' internal knowledge, the system dynamically generates search queries**,** retrieves supporting evidence, and refines its decision-making through an iterative process. Experiments show that integrating web retrieval increases misinformation detection accuracy by up to 20%, significantly reducing LLM hallucinations.(https://arxiv.org/abs/2409.00009)

Methodology

- The system consists of two primary agents:
- LLM Query Generator: Decomposes a claim and formulates search queries.
- Web Search Agent: Fetches relevant sources from DuckDuckGo, Cohere RAG, Wikipedia or other external sources.
- The LLM integrates retrieved evidence before making a final factuality judgement.

Experiments

The model was tested on LIAR-New(fact-checking database from PolitiFact) and FEVER-v2(Wikipedia-based fact verification) and compared with baseline models:LLM without search(GPT-4),HiSS,WikiChat, BERT.

Result

- Web retrieval improves LLM accuracy by up to 20%
- More sources lead to better misinformation detection
- PolitiFact is often the most-used evidence source, but the system still performs well without it.

Future Directions

- **Multi-agent Collaboration**: Using multiple LLM agents specialize in different aspects of misinformation detection.
- **Adaptive Reasoning Agents:**Using reinforcement learning techniques to allow agents to refine their fact-checking process based on the feedback.

- **Live Fact-Checking Agents**:Using LLM agents monitoring trending social media content for detecting the early stage of viral misinformation.
- **Network-based propagation Analysis:** Use graph-based models to track the spread pattern of misinformation across digital platforms.

- **Human Experts Guided LLM Agent Training**:
- Incorporate human-in-the-loop methods to finetuning the LLM agents for better factual consistency.

References

Bo Hu, Zhendong Mao,Yongdong Zhang, An overview of fake news detection: From a new perspective. *ScienceDirect,* Volume 5, Issue 1, Jan 2025, Page 332-346. https://doi.org/10.1016/j.fmre.2024.01.017

Mengyang Chen, Lingwei Wei, Han Cao, Wei Zhou and Songlin Hu.Explore the Potential of LLMs in Misinformation Detection: An Empirical Study. Dec 25,2024.https://arxiv.org/abs/2311.12699.

Hsin-Hsuan Chung,Jiangping Chen, Misinformation detection: datasets, models and performance. *Emerald insight.*Jan6,2025. https://www.emerald.com/insight/content/doi/10.1108/oir-06-2024-0388/full/html

Andrew Zhao, Daniel Huang, etc. ExpeL:LLM Agents Are Experiential Learners.*AAAI-24.*Dec 20,2024.https://doi.org/10.48550/arXiv.2308.10144

Luca Gioacchini, Guiseppe Siracusano, etc. AgentQuest: A modular Benchmark. Framework to Measure Progress and Improve LLM Agents. *NAACL-HLT2024.*Apr.9,2024. https://doi.org/10.48550/arXiv.2404.06411

Qingyun Wu, Gagan Bansal, etc. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.Aug16,2023. https://doi.org/10.48550/arXiv.2308.08155

Xu Huang, Weiwen Liu,etc. Understanding the planning of LLM agents: A survey. Feb5,2024. https://doi.org/10.48550/arXiv.2402.02716

Xinyi Li, Yongfeng Zhang, Edward C. Malthouse. Large Language Model Agent for Fake News Detection.Apr30,2024. https://doi.org/10.48550/arXiv.2405.01593

Jacob-Junqi Tian, Hao Yu, etc.Web Retrieval Agents for Evidence-Based Misinformation Detection. Oct9,2024. https://doi.org/10.48550/arXiv.2409.00009