Data Auditing System Using Machine Learning

Poj Netsiri
Immatriculation number: 12402153
e124021538@student.tuwien.ac.at

Main Supervisors
Assoz.Prof. PD Dr. Jakob Müllner, Wirtschaftsuniversität Wien
Asst.Prof. Dr. Harald Puhr, Universität Innsbruck
Co-Supervisor
Undecided, Technische Universität Wien

1 Abstract

Unethical data manipulation involves the deliberate alteration or misrepresentation of data to support a specific hypothesis, resulting in fraudulent research outcomes. The consequences of such practices include paper retraction, loss of credibility, withdrawal of funding, and potential harm to public health and policy. This project aims to develop a data auditing system utilizing machine learning algorithms to detect and classify instances of data fraud. By training models on both authentic and manipulated datasets, the system will be capable of distinguishing between legitimate and fraudulent data, there by promoting integrity in scientific research.

2 Introduction

Unethical data manipulation refers to the intentional misrepresentation of research data to produce predetermined or favorable conclusions. Such misconduct may be driven by personal, financial, or institutional interests and manifests in various forms, including:

- **Data fabrication**: Generating entirely fictitious data that was never collected.
- **Data falsification**: Altering or selectively modifying data to fit a specific narrative.
- **Selective reporting**: Publishing only favorable results while disregarding contradictory evidence.
- **P-hacking**: Adjusting statistical analyses or repeatedly testing data until significant p-values are obtained.
- **Textual manipulation**: The use of repetitive phrases, AI-generated content, or other linguistic inconsistencies in scientific papers, which may indicate fraudulent activity.

The negative impacts of unethical data manipulation are severe, leading to the dissemination of false information, erosion of public trust in scientific research, and policy decisions based on inaccurate findings. The ability to detect such manipulations is crucial to uphold the credibility and integrity of academic and scientific research.

3 Methodology

This research will leverage machine learning techniques to develop a robust data auditing system. The methodology involves the following steps:

1. **Data Collection**: Gathering datasets containing both verified legitimate research data and known fraudulent cases.
2. **Feature Engineering**: Identifying key indicators of data manipulation, including statistical anomalies, linguistic inconsistencies, and irregular data distributions.
3. **Model Selection and Training**: Evaluating various machine learning models, such as logistic regression, random forests, and neural networks, for their ability to classify manipulated data.
4. **Comparison with Traditional Methods**: The machine learning model's performance will be compared with conventional statistical techniques to assess their relative effectiveness.
5. **Evaluation Metrics**: Using evaluation such as F1-score to measure model performance, ensuring a balanced trade-off between precision and recall.

4 Advantages of Machine Learning

Machine learning-based fraud detection provides several advantages over conventional statistical approaches:

- **Automated Pattern Recognition**: ML models can learn complex patterns and detect subtle anomalies that may escape manual review.
- **Scalability**: ML-based systems are capable of auditing large volumes of data efficiently, making them well-suited for large-scale institutional or journal-level applications.
- **Adaptive Learning**: ML algorithms can evolve with new data, continuously improving their ability to identify emerging patterns of fraud.
- **Higher Accuracy**: Deep learning models, particularly those leveraging structured data and natural language, can outperform traditional methods in classification tasks.
- **Text-Based Fraud Detection**: Natural Language Processing (NLP) techniques can uncover suspicious textual patterns—such as unnatural repetition, or AI-generated content—which are often overlooked by statistical checks.

5 Expected Results

The proposed system is expected to produce machine learning models capable of accurately detecting fraudulent data across multiple domains. This tool will aid researchers, peer reviewers, and regulatory bodies in identifying unethical data manipulation, thereby strengthening the reliability of published research.

By advancing automated techniques for data auditing, this project will contribute to broader efforts in research ethics, transparency, and accountability. Ultimately, it aims to offer a practical, scalable solution to support the long-term integrity of scientific inquiry.

6 References

1. Data Falsificad (Part 1):"Clusterfake". `https://datacolada.org/109`. Accessed: 2025-04-01.
2. Reinhart and Rogoff are wrong about austerity. `https://peri.umass.edu/wp-content/uploads/joomla/images/wp322FT_Pollin_Ash.pdf`. Accessed: 2025-04-01.
3. Stapel report finds faked data in at least 30 papers, possibly more. `https://retractionwatch.com/2011/10/31/stapel-report-finds-faked-data-in-at-least-30-papers-possibly-more/`