UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

MIPRO 2021

Methods for Automatic Sensitive Data Detection in Large Datasets: a Review

V. Kužina*, E. Vušak* and A. Jović*

* University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10 000 Zagreb, Croatia

vjeko.kuzina@fer.hr

As part of:
"Digital platform for ensuring data privacy and prevention of malicious manipulation of the personal data – AIPD2"

Overview

Introduction

Approaches

- Rule-based
- Machine learning

Comparison and best performance

Conclusion

Introduction

- What is sensitive data?
- Presence and requirements of different domains.
- Automatization

Approaches to sensitive data detection

Rule-based

Machine learning

Rule-based approaches

LOOKUP TABLE

REGULAR EXPRESSIONS

IDENTIFYING METADATA

Rule-based approaches

Pros:

- No samples
- Adding/changing rules

Cons:

- Knowing and writing all rules
- Low generalizability

Machine learning approaches

- Hidden Markov model (HMM)
- Conditional random fields (CNN)
- Recurrent neural networks (RNN)
- Long short-term memory (LSTM)
- Bidirectional Encoder Representations from Transformers (BERT)

Machine learning approaches

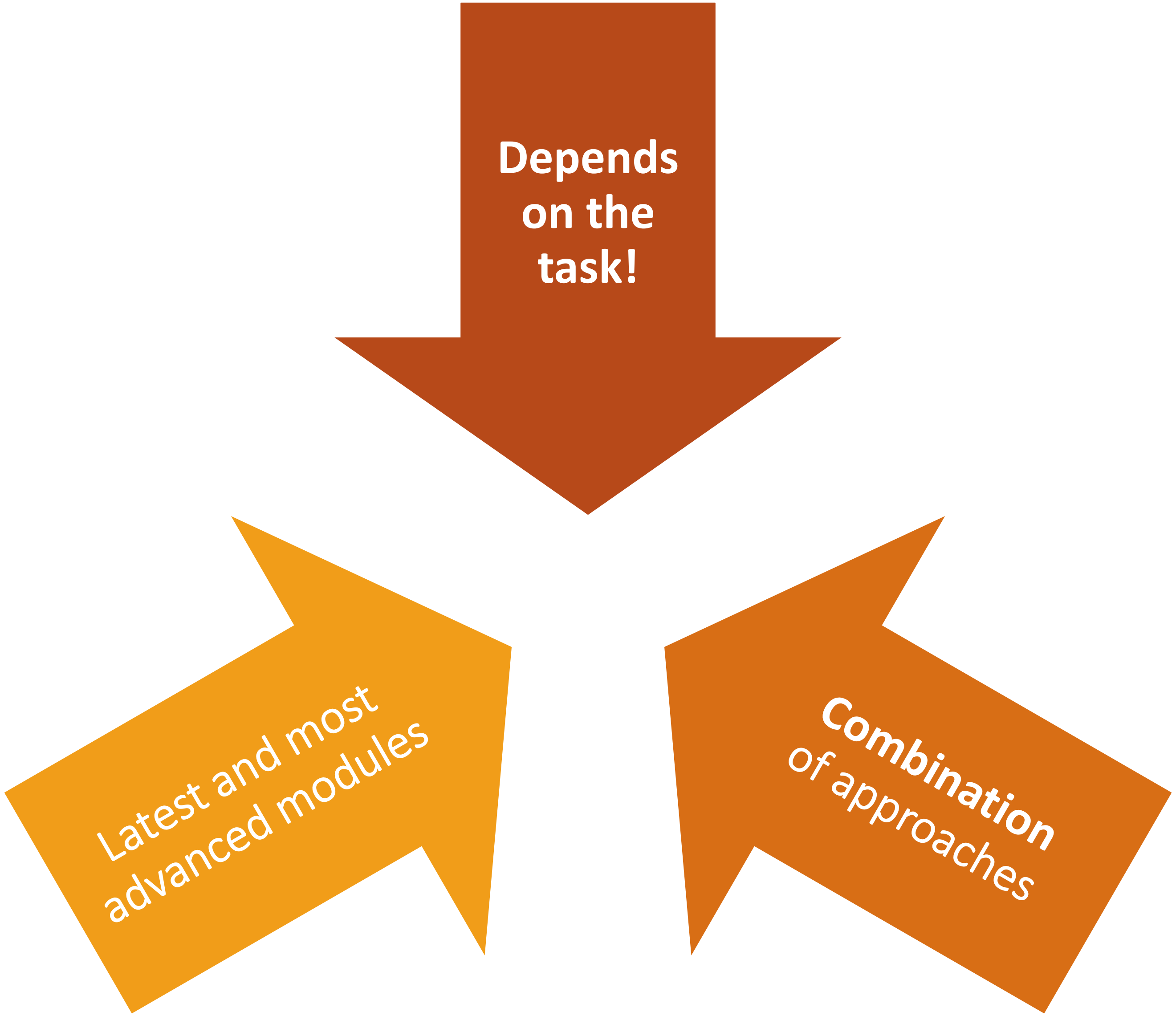
Pros:

- Generalizable
- Learns edge cases by iself

Cons:

- Large amounts of data
- Hard to add/change edge cases later

Best performance



Conclusion

Automation detection processes

Comparison of various detection approaches

Need for more research

Questions?