Open WebSearch .eu

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

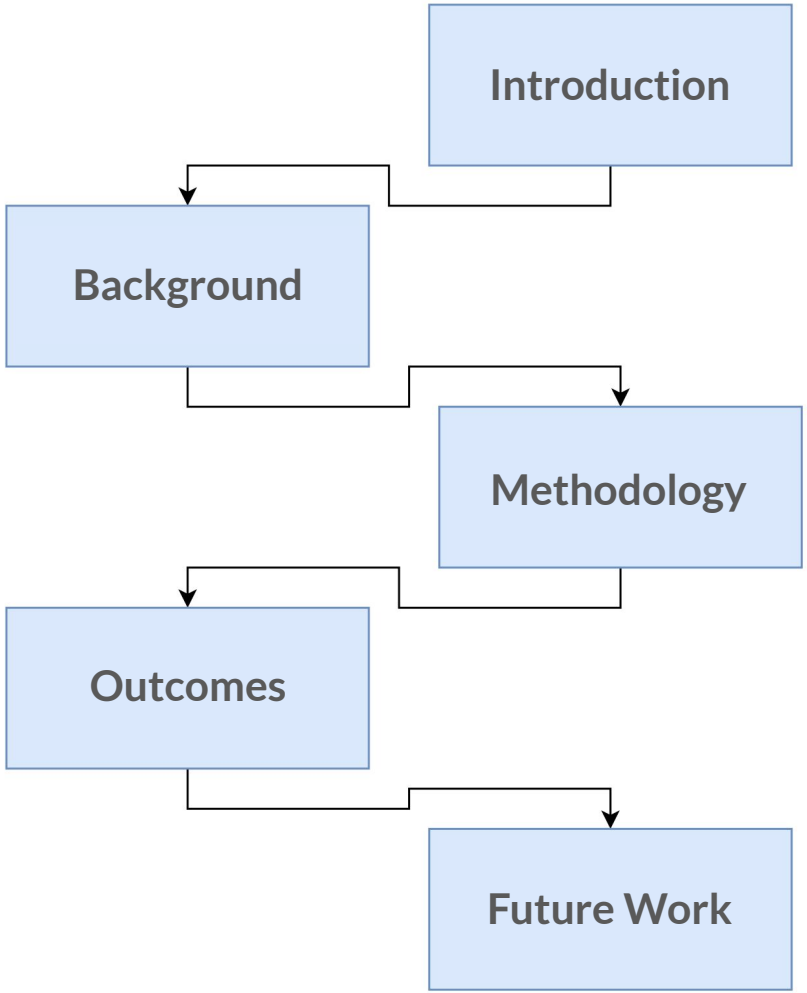
A Comprehensive Dataset for Webpage Classification

M. Al-Maamari, M. Istaiti, S. Zerhoudi,
M. Dinzinger, M. Granitzer, J. Mitrovic

October 4th 2023

Agenda

- Introduction
- Background
- Methodology
- Outcomes
- Future Work



Research Objectives

- **Present a comprehensive dataset:**
  - **Size:** 116,000 URLs
  - **Task:** Webpage classification.
- **Establish two levels of labels:**
  - **Broad categorization**: [Benign, Malicious, Adult]
  - **Nuanced labeling**: 20 subclasses
- **Test and compare machine learning model performance.**

Importance of Webpage Classification

- **Why Webpage Classification Matters?**
  - Exponential growth of the web.
  - Challenges for search engines and web crawlers.
  - Need for improved crawling efficiency and targeted content indexing.

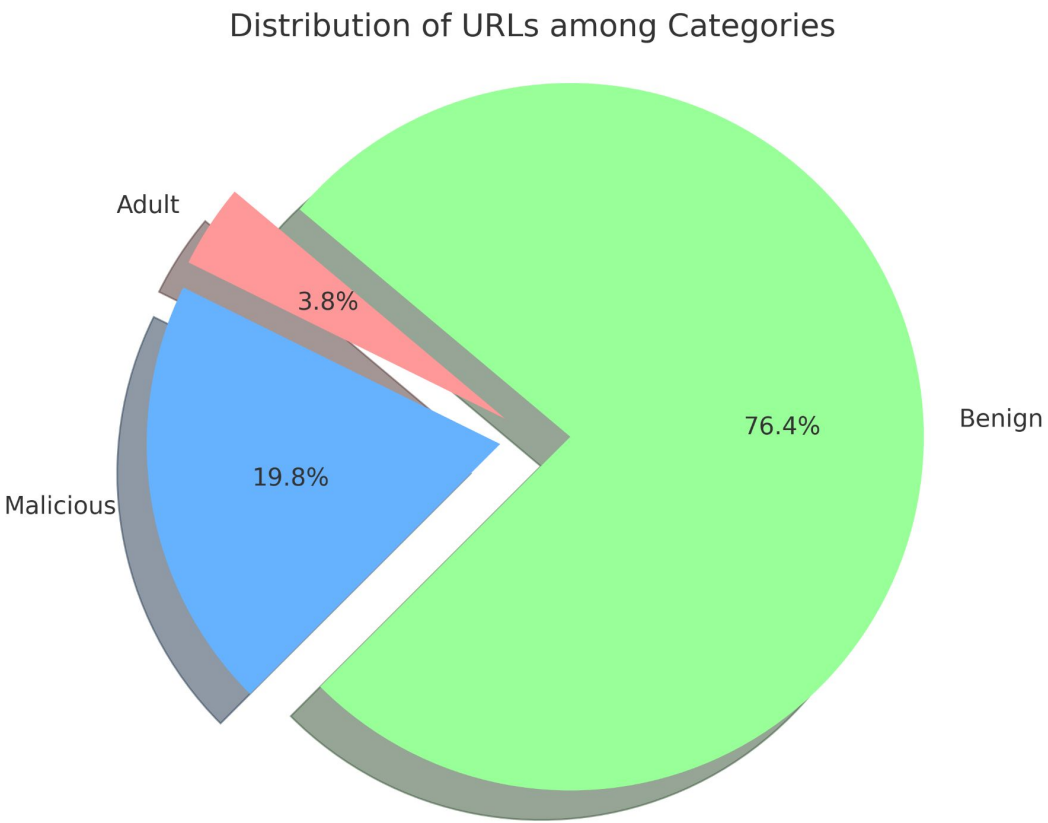
AI generated image: “Smart search engine crawling websites”

Dataset Overview: Data Collection

- **Dataset curated from multiple online sources:**
  - URL Classification Dataset [DMOZ] for benign URLs
  - URLhaus for malware URLs.
- **Crawling process used "OWler" to gather raw HTML content.**
  - Non-working URLs were ignored.
- **Raw HTML parsed to extract structured content**
- **Each URL labeled with:**
  - **Main label** [Benign, Malicious, Adult]
  - **Subclass label:** 20 subclasses (e.g. News, Spam, …, etc)

Dataset Overview: Dataset Construction

- **Dataset cleaning**
  - Removed around 2,000 duplicate URLs, mostly malicious.
  - 23 URLs with no content were removed.


Distribution of URLs among Categories
Adult
3.8%
Malicious
19.8%
76.4%
Benign

Our Approach to Webpage Classification

- Data collection from diverse sources ✅
- Use of the "OWler" crawler for content extraction ✅
- Dataset cleaning ✅
- **Application of machine learning models: SVC and SGD.**
- **Evaluation metrics: precision, recall, F1, and F2 scores.**

Feature Representation

- **Three types of input.**
- **Each input type offers unique strengths and challenges for classification:**
  - URLs only → Fast classification **But** Less information
  - Raw HTML content → More information to be used **But** slower performance
  - Parsed HTML content → More information + natural language **But** also slow
- **Input is tokenized and vectorized first.**

Experimented Variables

- **Tokenization:**
  - Different techniques [n-grams, Byte-Pair Encoding BPE]
  - Token levels [char, word, subword]

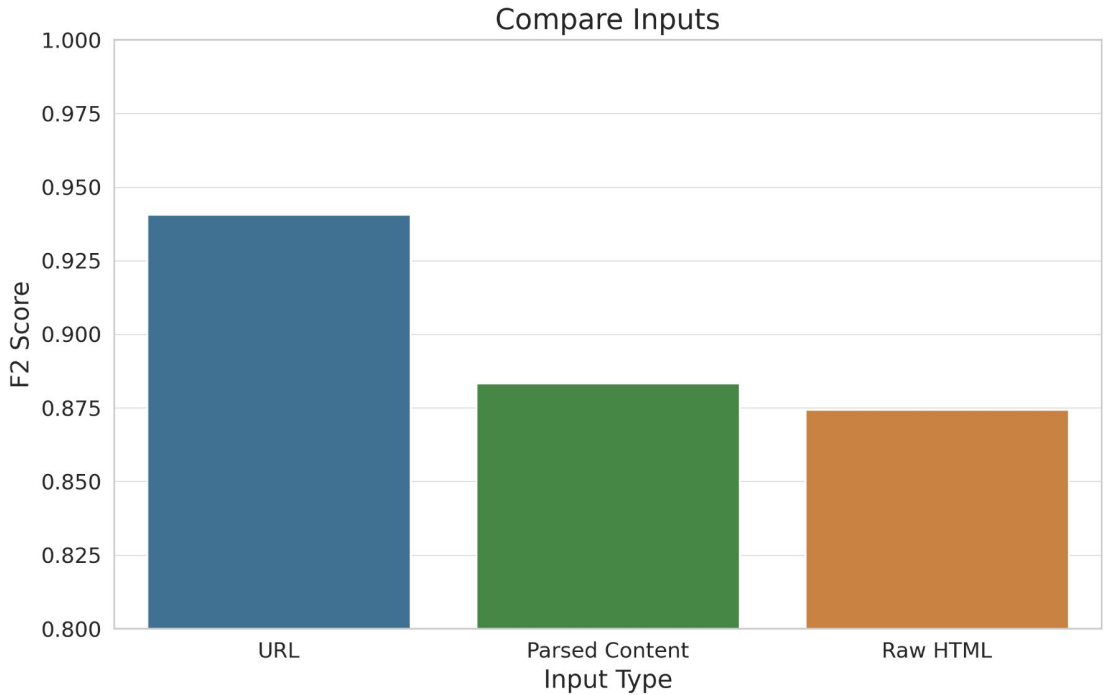
- **Vectorized using TF-IDF.**

Experimented Variables

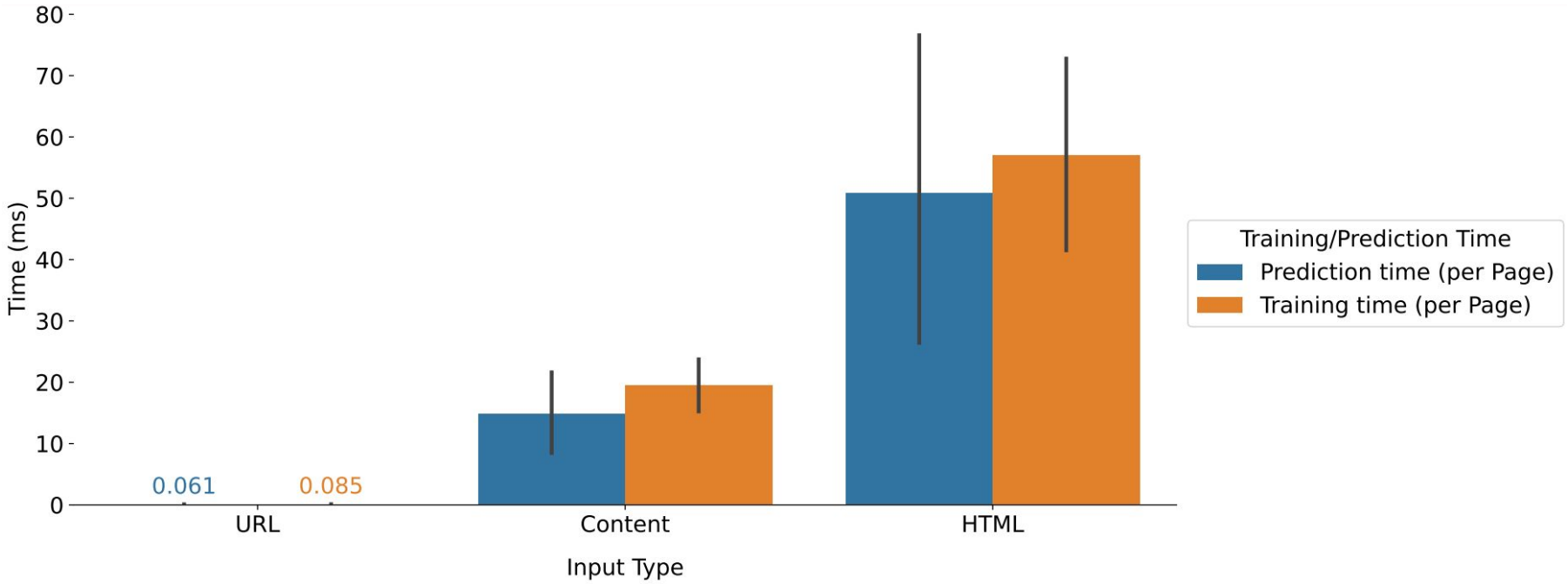
- **Different types of output:**
  - Classify the input into one of the 3 main labels [Benign, Malicious, Adult]
  - Classify the input into one of the 20 sublabels (e.g. News, Spam, …, etc)

- **Models:**
  - Support Vector Classifier (SVC)
  - Linear models with Stochastic Gradient Descent (SGD)

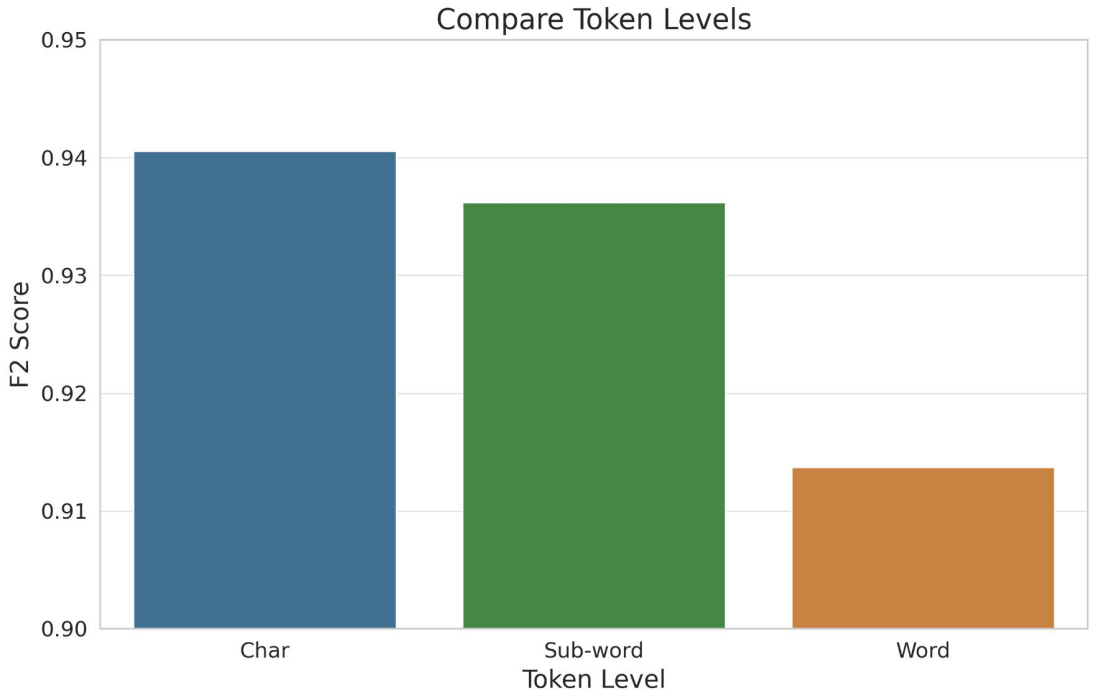
Results & Evaluation



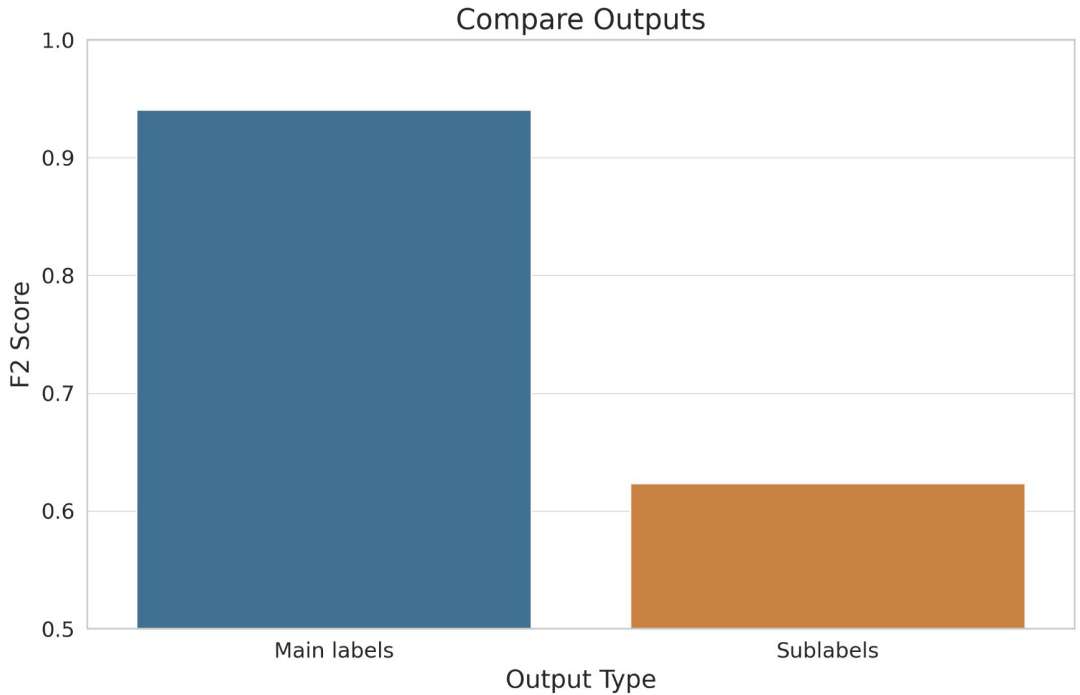
Results & Evaluation



Results & Evaluation



Results & Evaluation



Results & Evaluation

Future Work & Recommendations

- **Expanding the dataset:**
  - More data points
  - More Categories
- **More Experiments:**
  - URL augmentation
  - Try different vectorization techniques

Thank you for listening!

Happy to answer
any question