Open WebSearch.eu

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

Impact of Tokenization Techniques on URL Classification

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

October 10th 2024

Agenda

- Introduction
- Background
- Data
- Results

Research Objectives

- **Study the effect of different tokenization methods on URL classification:**
- **Compare tokenization methods:**
  - Byte Pair Encoding (BPE)
  - Enhanced BPE with GPT-4 generated keywords
  - Punctuation-based splitting
  - Character-level n-grams
- **Assess effects on:**
  - Classification accuracy
  - Computational efficiency

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.

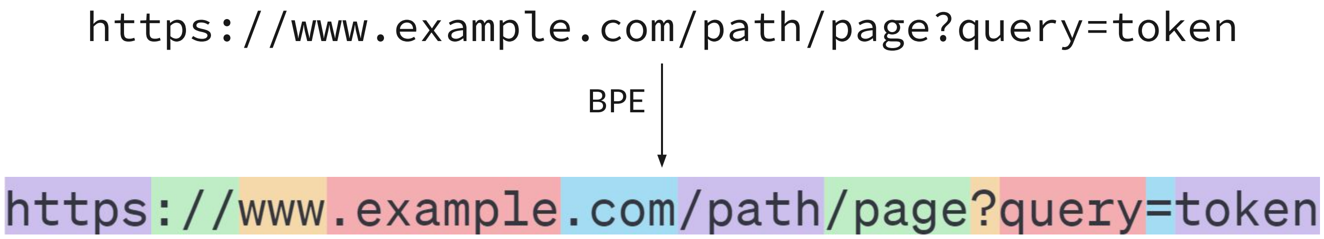
AI generated image: “Smart and cute spider-like robot crawling websites, behind it are floating web pages”

Tokenization

- **What is Tokenization?**
  - Breaking down text into **smaller units** called **tokens**.
  - Tokens can be words, subwords, characters, or symbols
- **Problem: Lack of Whitespace in URLs**
  - URLs are continuous strings without spaces, unlike regular text.
  - Word-based tokenizers are ineffective for URLs.

Byte Pair Encoding (BPE)

- **What is BPE?**
  - Merges frequent pairs of characters or sequences.
  - Reduces vocabulary size.
  - Captures **subword** units.



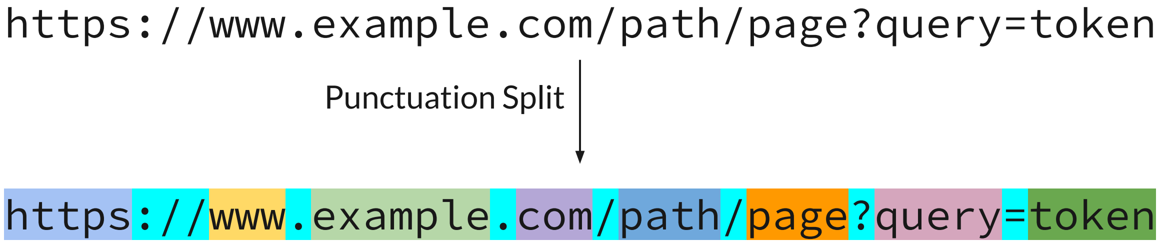
https://www.example.com/path/page?query=token

BPE

https://www.example.com/path/page?query=token

Punctuation Split

- **How it works?**
  - Tokenize URLs at punctuation marks.
  - Splits the URL based on punctuations like:
    - Slashes `/` or `//`
    - Question marks ` ? `
    - Dots ` . `
    - other marks such as `=`or `:`



```
https://www.example.com/path/page?query=token
```

Punctuation Split

```
https://www.example.com/path/page?query=token
```

Character-level n-grams

- **How it works?**
  - Captures sequences of characters.
  - Explored n-gram ranges:
    - 1-gram
    - (1-3)-grams
    - (3-6)-grams

```
https://www.example.com/path/page?query=token
```

1-grams

```
['h', 't', 't', 'p', 's', ':', '/', '/', 'w', 'w', 'w', '.', 'e', 'x', 'a', 'm', 'p', 'l', 'e',
     '.', 'c', 'o', 'm', ' ... '?', 'q', 'u', 'e', 'r', 'y', '=', 't', 'o', 'k', 'e', 'n']
```

Character-level n-grams

```
https://www.example.com/path/page?query=token
```

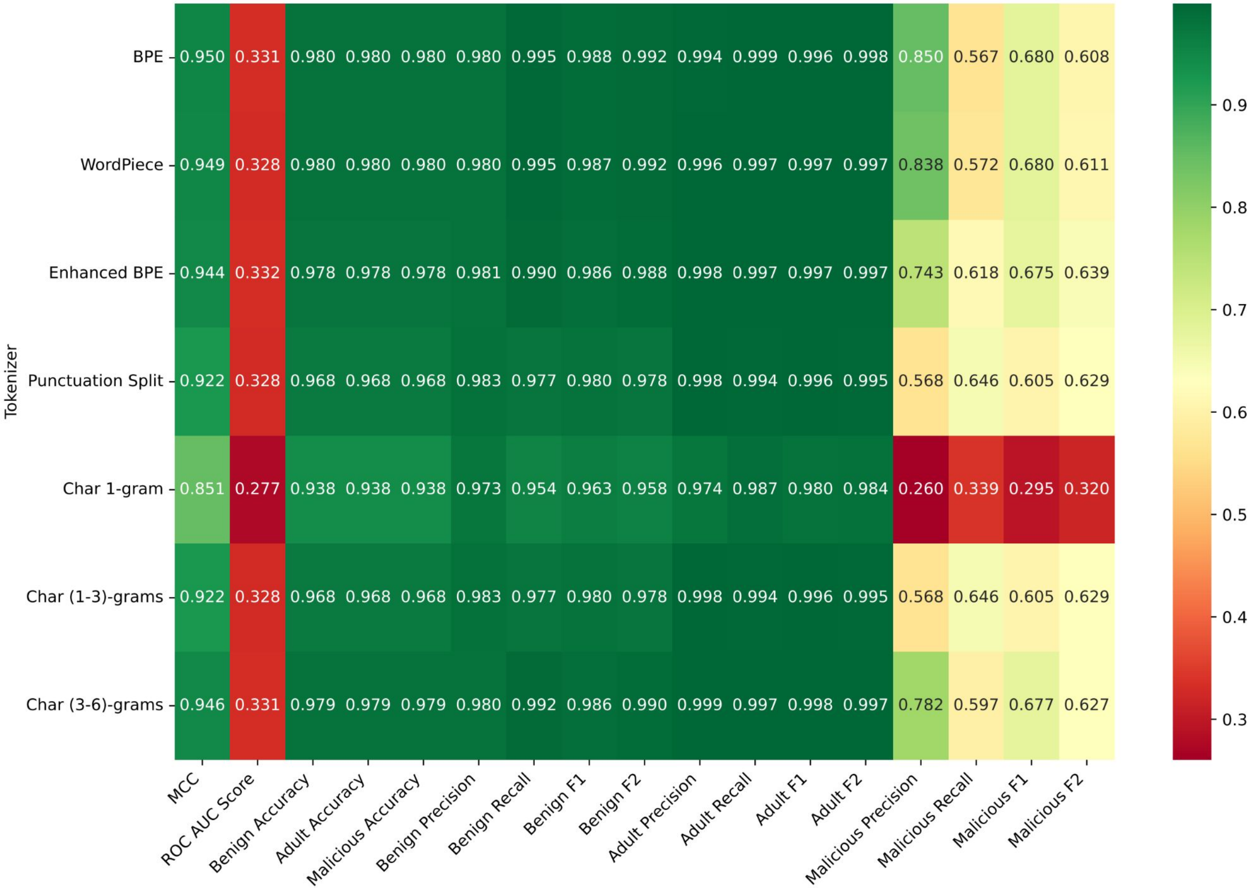
3 grams ↓

```
 ['htt', 'ttp', 'tps', 'ps:', 's:/', '://', '//w', 'www', 'ww.', 'w.e', '.ex', 'exa', 'xam',
'amp', 'mpl', 'ple', 'le.', 'e.c', '.co', 'com', 'om/', 'm/p', '/pa', 'pat', 'ath', 'th/',
'h/p', '/pa', 'pag', 'age', 'ge?', 'e?q', '?qu', 'que', 'uer', 'ery', 'ry=', 'y=t', '=to',
'tok', 'oke', 'ken']
```

Dataset Overview:

- **Over 1 million labeled URLs analyzed.**
- **Categories:**
  - Malicious
  - Benign
  - Adult

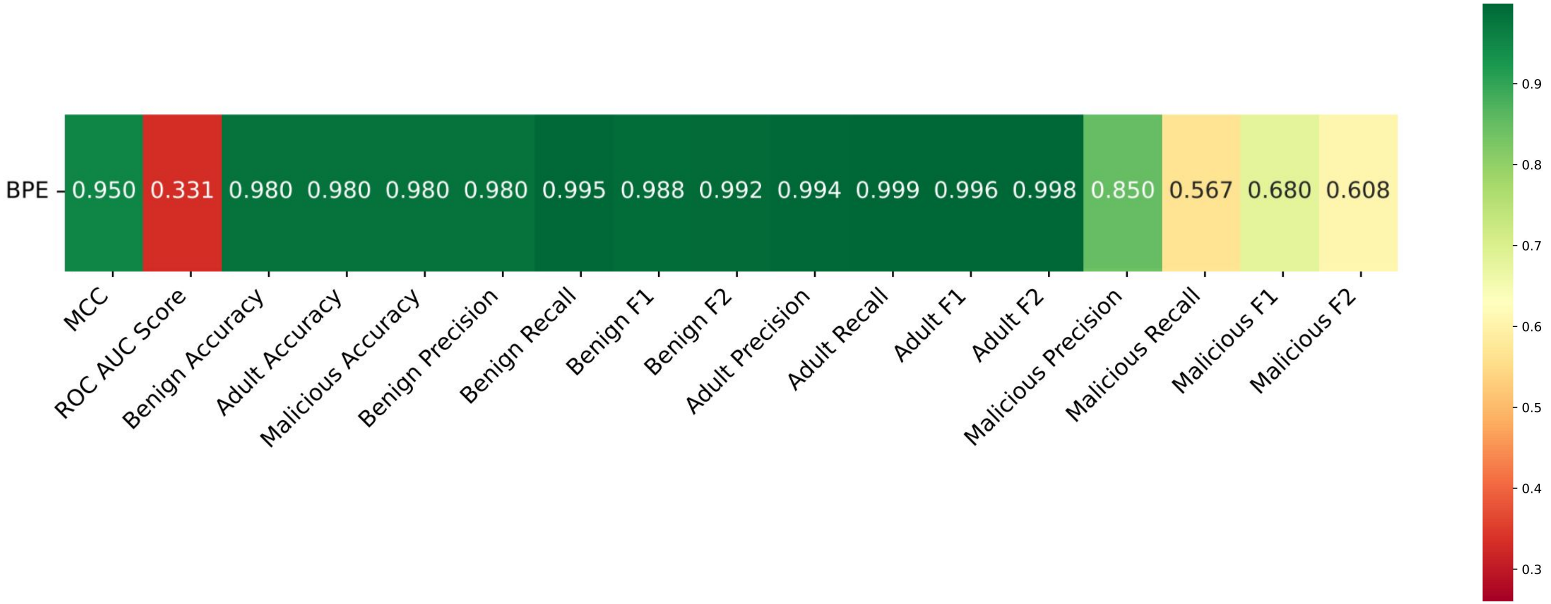
Results: Classification accuracy



Tokenizer	MCC	ROC AUC Score	Benign Accuracy	Adult Accuracy	Malicious Accuracy	Benign Precision	Benign Recall	Benign F1	Benign F2	Adult Precision	Adult Recall	Adult F1	Adult F2	Malicious Precision	Malicious Recall	Malicious F1	Malicious F2
BPE	0.950	0.331	0.980	0.980	0.980	0.980	0.995	0.988	0.992	0.994	0.999	0.996	0.998	0.850	0.567	0.680	0.608
WordPiece	0.949	0.328	0.980	0.980	0.980	0.980	0.995	0.987	0.992	0.996	0.997	0.997	0.997	0.838	0.572	0.680	0.611
Enhanced BPE	0.944	0.332	0.978	0.978	0.978	0.981	0.990	0.986	0.988	0.998	0.997	0.997	0.997	0.743	0.618	0.675	0.639
Punctuation Split	0.922	0.328	0.968	0.968	0.968	0.983	0.977	0.980	0.978	0.998	0.994	0.996	0.995	0.568	0.646	0.605	0.629
Char 1-gram	0.851	0.277	0.938	0.938	0.938	0.973	0.954	0.963	0.958	0.974	0.987	0.980	0.984	0.260	0.339	0.295	0.320
Char (1-3)-grams	0.922	0.328	0.968	0.968	0.968	0.983	0.977	0.980	0.978	0.998	0.994	0.996	0.995	0.568	0.646	0.605	0.629
Char (3-6)-grams	0.946	0.331	0.979	0.979	0.979	0.980	0.992	0.986	0.990	0.999	0.997	0.998	0.997	0.782	0.597	0.677	0.627

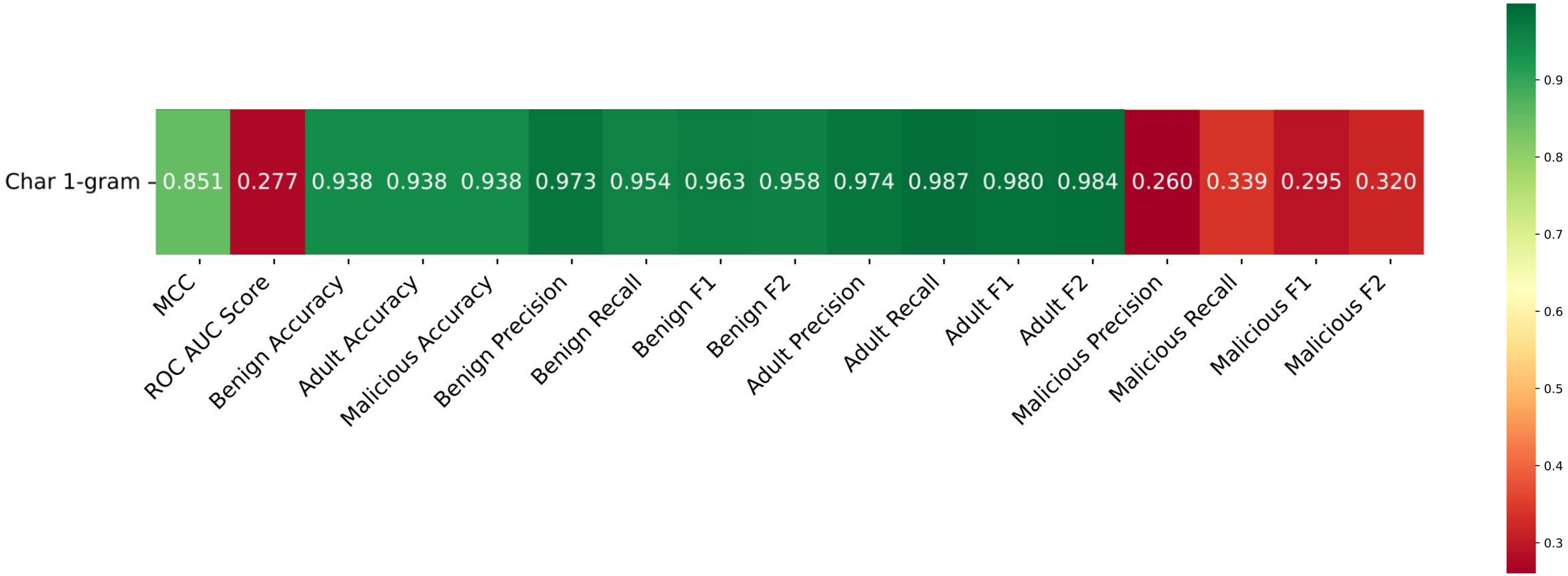
Results: Classification accuracy

Best Method: BPE



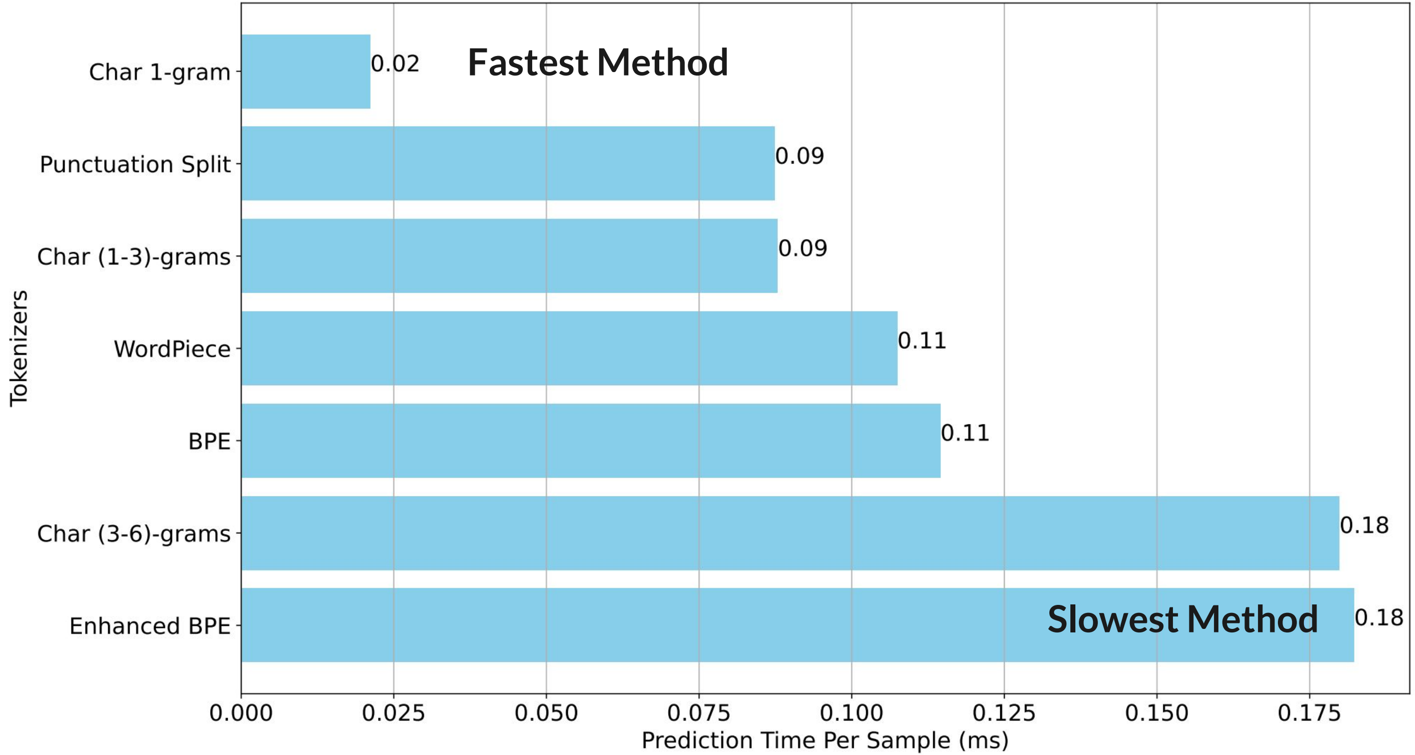
Results: Classification accuracy

Worst Method: BPE



	MCC	ROC AUC Score	Benign Accuracy	Adult Accuracy	Malicious Accuracy	Benign Precision	Benign Recall	Benign F1	Benign F2	Adult Precision	Adult Recall	Adult F1	Adult F2	Malicious Precision	Malicious Recall	Malicious F1	Malicious F2
Char 1-gram	0.851	0.277	0.938	0.938	0.938	0.973	0.954	0.963	0.958	0.974	0.987	0.980	0.984	0.260	0.339	0.295	0.320

Results: Computational efficiency



Key Findings

- **Simple tokenizers are faster but less accurate:**
  - They process data quickly due to their simplicity and limited scope.
  - They lack the depth needed to recognize complex patterns, reducing accuracy.

- **Advanced tokenizers perform better overall:**
  - These methods capture meaningful subword structures, improving understanding.
  - Their complexity leads to better accuracy but increases computational time.

Key Findings

- **All methods struggle with 'Malicious' URLs:**
  - Malicious URLs mimic benign ones, making them harder to classify accurately.
  - Tokenization alone may not detect subtle differences that indicate malicious intent.

Thank you for listening!

Happy to answer any question

?