IMPACT OF TOKENIZATION TECHNIQUES ON URL CLASSIFICATION

M. Al-Maamari, M. Istaiti, S. Zerhoudi,
M. Dinzinger, M. Granitzer, J. Mitrovic
University of Passau, 94032 Passau, Germany

Abstract

Web crawling can be improved by the accurate classification of URLs to ensure relevant content is indexed and harmful content is filtered out. In this study, we examined the impact of various tokenization techniques on URL classification, a task integral to the development of intelligent web crawlers. Our investigation was conducted using a large-scale dataset of over one million URLs, categorized into 'Malicious', 'Benign', and 'Adult' classes, with detailed sublabels for in-depth analysis [1]. We explored a range of tokenization methods, including Byte Pair Encoding (BPE), Enhanced BPE with a GPT-4 generated keyword dictionary, punctuation-based splitting, and character-level n-grams, to assess their effect on the classification accuracy and computational efficiency [2, 3]. The results indicated that while simple tokenization methods like Char 1-gram offered rapid prediction times, they were inadequate in correctly identifying more complex 'Malicious' URLs. More sophisticated techniques such as BPE and WordPiece achieved a better balance of precision and recall for 'Benign' and 'Adult' content, yet they, along with other methods, struggled with the 'Malicious' category. The findings highlight the nuanced challenges of URL classification and underscore the need for advanced tokenization approaches that can compete with the nature of malicious content while maintaining computational efficiency. Future work should focus on integrating diverse tokenization strategies and enhancing semantic comprehension within the tokenization process to improve classification performance, particularly for detecting malicious content within the vast and dynamic landscape of the web.

INTRODUCTION

Web crawlers are fundamental tools used by search engines to collect data from the Internet, which demands the classification of URLs to improve efficiency and filter out irrelevant or harmful content. Efficient web crawling is contingent upon the avoidance of resource expense on unneeded or harmful URLs, such as those that are malicious, spam, or not relevant to the crawler's purpose. The incentive for developing robust URL classification systems is to support these intelligent crawling strategies.

The process of URL classification is a form of text classification, which involves categorizing text into organized groups. In this domain, tokenization plays a important role by breaking down text into smaller units, or tokens, that serve as input for machine learning algorithms. The choice of tokenization technique is a critical decision that can significantly influence the effectiveness of a classification model. [4] [5] Tokenization affects not only the granularity of the data but also the ability of the model to recognize patterns and make accurate classifications.

This paper aims to clarify the impact of various tokenization techniques on the task of URL classification. Given the diverse nature of URLs, which may include various structures and subcomponents, selecting an appropriate tokenization method is not trivial. We compare several tokenization methods, including Byte Pair Encoding (BPE), an enhanced version of BPE supplemented with a initial keyword dictionary, character-level n-grams, and a method based on splitting at punctuation marks.

Our investigation is grounded in the analysis of a comprehensive dataset of approximately one million URLs [1]. We employ a suite of evaluation metrics to assess the efficacy of each tokenization strategy, with a focus on accuracy, precision, recall, and the F1 score.

By concentrating on tokenization as a fundamental aspect of the URL classification process, our study provides granular insights into the influence of different tokenization approaches. The main objective is to identify the most effective tokenization technique, balancing high classification performance while being mindful of computational efficiency. Such insights are invaluable for the development of web crawlers that are more selective, sparing resources by avoiding the retrieval and indexing of unwanted URLs.

In the context of creating a more open web search ecosystem, this paper also contributes to the larger project aimed at developing an Open Web Index (OWI). As outlined in the recent work of [6], an OWI would promote a more open search ecosystem, offering genuine choice among alternative search engines and fostering a fair and collaborative information space. Our research supports this vision by enhancing the technology that supports web crawling, an essential component of search engine infrastructure. The classification of URLs based on reliable tokenization methods is a step towards enriching the open index with quality data, thereby enabling the development of declarative search engines and innovative web data products.

BACKGROUND AND RELATED WORK

The classification of URLs has emerged as a task for enabling web crawlers to efficiently process the growing data on the World Wide Web. A web crawler, by definition, systematically navigates the web to index content for search engines and data retrieval applications [7]. With the sheer volume of web pages, it is essential to employ intelligent crawling strategies, such as focused crawlers, which aim to selectively retrieve pages relevant to specific topics or areas. URL classification facilitates this selective approach by identifying and filtering out URLs likely to lead to irrelevant or malicious content, thus optimizing the crawling process [8].

Tokenization, as the process of segmenting text into tokens, represents the first and a foundational step in any Natural Language Processing (NLP) pipeline. While the simplest approach to tokenization is to use whitespace-separated words, this can result in an inordinately large vocabulary, especially in the context of extensive corpora such as the web, additionally, this method does not work for URLs since there are no whitespaces in them. To address the inefficiencies associated with large vocabularies, subword tokenization algorithms have been developed. These algorithms, including Byte Pair Encoding (BPE), create subwords or tokens that can significantly limit the vocabulary size while retaining meaningful linguistic units, it also works with text that does not contain whitespace (e.g. URLs) [3,4]. Tokenization strategies can significantly alter linguistic understanding and, thus, are crucial in the composition of input features for machine learning models, particularly in languages with rich morphology [5].

Previous studies have studied the impact of tokenization on machine learning model performance. In the context of text classification, various tokenization algorithms have been evaluated, demonstrating that the performance of these algorithms is contingent on multiple factors. These factors include the size and nature of the dataset, the specific classification task at hand, and the morphological complexity inherent to the language of the dataset [4]. Tokenization has also been shown to play a significant role in the context of named entity recognition (NER), where the choice of tokenization strategy can either enhance or impair model performance based on how it copes with the linguistic challenges posed by the target language [5]. In the domain of web page classification, character n-gram based features extracted from URLs have been successfully employed, showcasing the utility of tokenization techniques that do not rely on the actual content or the hyperlink structure of the pages [8]. This approach highlights the influence of tokenization in addressing the challenges associated with URL classification.

Collectively, these studies form the background against which we examine the effectiveness of various tokenization methods, with a particular emphasis on their application in URL classification for web crawlers. This exploration aims to contribute to the ongoing discussion on the optimal integration of tokenization techniques within machine learning frameworks for the enhancement of web crawling and indexing efficiency.

METHODOLOGY

Data

Our investigation utilized a comprehensive dataset comprising 1,069,715 URLs, each annotated with labels denoting its classification into 'Malicious', 'Benign', or 'Adult' categories, and further specified into 20 sublabels for detailed analysis [1]. This dataset is constructed to facilitate the development and comparative assessment of machine learning models. The dataset was curated to enable research in enhancing webpage classification, one component in optimizing web crawling and content filtering systems.

Tokenization Techniques

The tokenization methods explored in this paper include:

- **Byte Pair Encoding (BPE):** BPE is a hybrid between character-level and word-level tokenization. It iteratively merges the most frequent pair of characters or character sequences, thereby reducing vocabulary size and capturing more information than individual characters [3]. We applied BPE to URLs to examine its effect on capturing token patterns significant for classification tasks.
- **Enhanced BPE:** This method extends BPE by integrating an initial dictionary of keywords generated by GPT-4 for each class [2]. The keywords enrich the BPE token dictionary, expected to refine the granularity with which URLs are tokenized and enhance classification performance.
- **Punctuation Split:** Utilizing regular expressions, specifically the pattern "(\w+|\S)", we tokenize on punctuation. This approach recognizes the structural nuances of URLs, which often contain meaningful delimiters such as periods and slashes.
- **Character-level N-grams:** We analyzed the performance of various n-gram levels, ranging from unigrams to longer spans of characters (1-gram, (1 to 3)-grams, and (3 to 6)-grams). This analysis aims to understand the impact of n-gram granularity on model performance, examining the trade-offs between the specificity of longer n-grams and the broader context captured by sequences.

Machine Learning Model

Given the scope of this paper is to examine the impact of tokenization on URL classification, we selected the SGDClassifier from SKLearn as our machine learning model [9]. The choice of SGDClassifier is motivated by its computational efficiency and moderate performance across various text classification tasks. The SGDClassifier is well-suited for handling large-scale data and provides a consistent benchmark to evaluate the influence of different tokenization methods. By fixing the variable of the machine learning model, we isolate the effects of tokenization techniques on classification outcomes, thereby ensuring the focus of this study remains on the comparative analysis of the tokenization strategies employed.

RESULTS

The heatmap visualization in Figure 1 shows the comparative performance of various tokenization techniques utilized for URL classification across three primary content categories. A key observation is the uniform struggle among

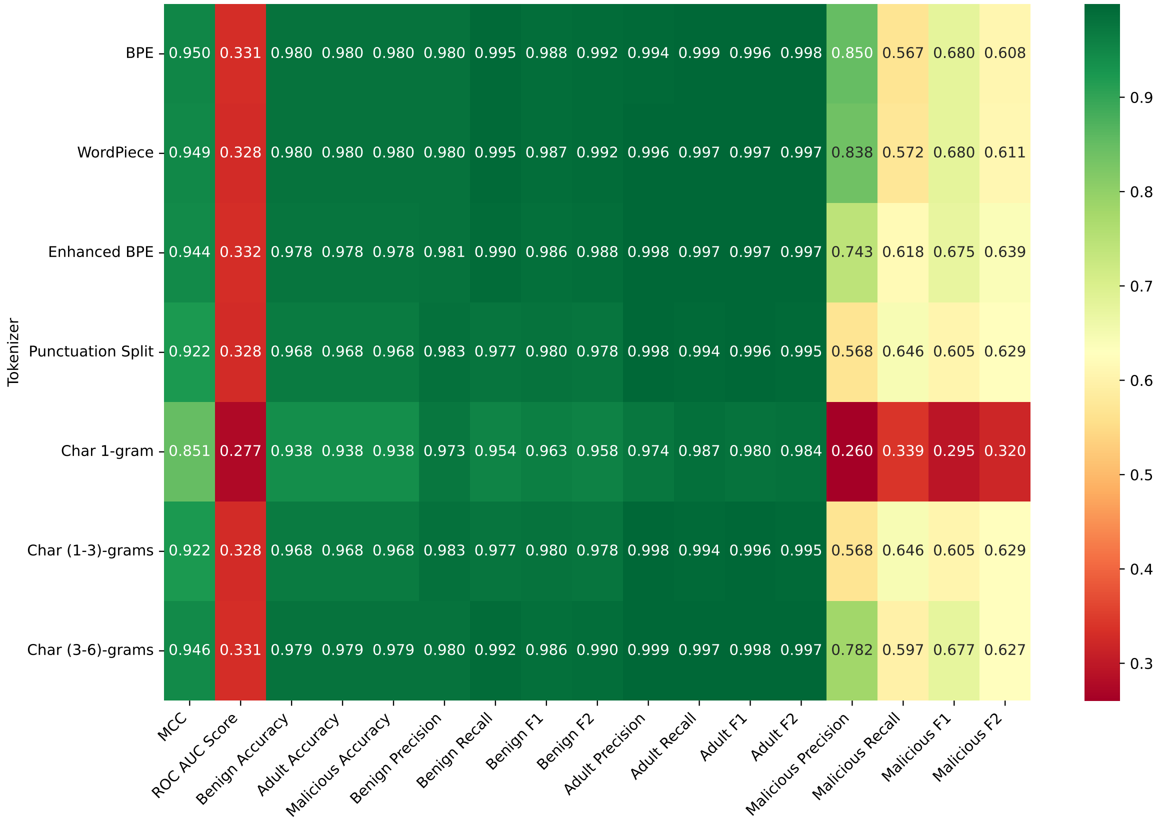

Figure 1: Performance of the tokenizers

all tokenization strategies to accurately classify ’Malicious’ URLs. Despite this common challenge, certain tokenizers emerged with relatively superior performance in the ’Malicious’ class, with **Byte Pair Encoding (BPE)**, **WordPiece**, and **Char (3-6)-grams** positioned as the frontrunners, respectively. Their ability to capture longer subword structures or sequences may attribute to their marginally better performance, suggesting a nuanced but high impact of token granularity on classification outcomes.

On the other hand, using **Char 1-gram** tokenization manifests as the least effective, particularly pronounced in its inability to classify ’Malicious’ URLs. The results signify the insufficiency of singular characters to encapsulate the contextual complexity required for the identification of malicious content.

Furthermore, the ROC AUC Score, a probabilistic measure indicating a model’s capability to discriminate between classes, is markedly low for all tokenization techniques. This uniform underperformance emphasizes a broader issue in the classification model’s capacity to distinguish ’Malicious’ URLs from others, reflecting a pivotal limitation within the current scope of tokenization approaches.

In contrast to the ’Malicious’ class, tokenization techniques exhibit an excellent performance in classifying ’Benign’ and ’Adult’ URLs. This great performance indicates that the nature of tokens common in these categories is well-captured by the tokenizers, facilitating reliable classification. The differential success across the content categories underscores a key conclusion: while tokenization methods adeptly handle general content, they stumble in reliably identifying content with potentially harmful intent, where context and semantic complexity play an instrumental role, additionally, it is known that malicious URLs usually try to be similar to benign URLs to avoid being detected.

In light of the findings, it is obvious that the pursuit of enhanced tokenization strategies remains necessary. The quest entails refining the balance between token granularity and the semantic richness essential for the robust classification of web content, particularly for ensuring web crawlers’ efficacy and safety in their navigational endeavors.

DISCUSSION

The comparative analysis reveals significant insights into the performance landscape of various tokenization techniques in URL classification. Notably, the **Char 1-gram** tokenizer, despite its operational speed at a mere 0.02 milliseconds per URL Figure 2, demonstrates suboptimal performance metrics, with MCC values and F2 scores for the ’Malicious’ class indicating insufficient precision and recall balance. This finding highlights the trade-off between prediction speed and classification robustness, particularly underlining the tokenizer’s insufficiency in complex URL categorization that demands a richer contextual understanding.

Meanwhile, the **Punctuation Split** tokenizer exhibits improvement in critical areas, including ROC AUC and MCC scores, over the Char 1-gram. At 0.09 milliseconds per URL, it encapsulates meaningful URL delimiters, hinting at the value of structural tokens in distinguishing between content categories. Similarly, the **Char (1-3)-grams** tokenizer maintains the same prediction time but advances in balancing precision and recall, except in the classification of 'Malicious' URLs, suggesting a need for an enhanced tokenization strategy to address URLs with malicious intent.

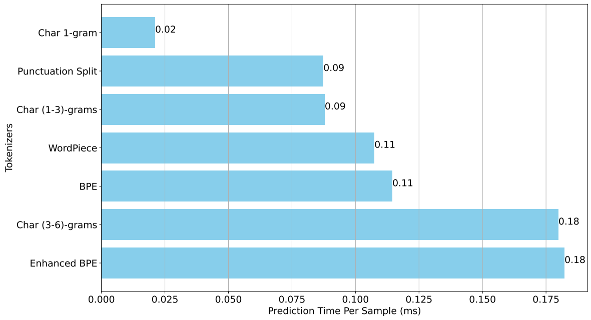

Figure 2: Comparison of Prediction Time Per URL Across Tokenizers

The **WordPiece** and **BPE** tokenizers, both clocking prediction times at 0.11 milliseconds per URL, achieve a admirable balance across evaluation metrics. However, their limitations become apparent in the 'Malicious' class, showing a challenge in detecting URLs of harmful web pages.

With a prediction time of 0.18 milliseconds per URL, the **Char (3-6)-grams** tokenizer shows potential in classifying 'Adult' and 'Benign' URLs but experiences a decline in performance when it comes to 'Malicious' URLs. This pattern suggests that while extended n-gram ranges might improve context capture, they may also result in overly specific tokens that lack generalizability.

Lastly, the **Enhanced BPE** tokenizer, also with a prediction time of 0.18 milliseconds, reveals a nuanced performance. It slightly improves upon BPE in the 'Adult' precision metric yet falls behind in critical areas such as 'Malicious' recall and F2 scores. The addition of GPT-4 generated keywords does not seem to uniformly enhance classification, particularly of 'Malicious' URLs, which remain challenging for all tokenizer models under study.

The practical application of these tokenization techniques within web crawlers has far-reaching implications. The efficiency of web crawlers is pivotal, as is their capability to sieve through the vast web content accurately. In this light, the findings of our study point to the necessity for carefully calibrated tokenizers that can adeptly handle URLs across varying content types without costing prohibitive computational costs.

In real-world applications, the decision to employ a particular tokenizer must be informed by the specific requirements of the web crawling task. The analysis highlights the need for a tokenizer that not only provides computational efficiency but also maintains high classification accuracy, especially for detecting 'Malicious' URLs. As web content continues to expand, the advancement of tokenization strategies will remain an essential area of research, with the objective of refining web crawlers to operate with enhanced precision and efficiency.

CONCLUSION

Our comprehensive evaluation of tokenization techniques in URL classification has yielded several key findings. The study confirms that while faster tokenizers like **Char 1-gram** offer computational expediency, they fall short in effectively classifying URLs, particularly those that are malicious. In contrast, more complex tokenization strategies such as **BPE** and **WordPiece** demonstrate a commendable balance of speed and accuracy for 'Benign' and 'Adult' classes but exhibit limitations in discerning 'Malicious' URLs. Enhanced tokenizers like **Enhanced BPE**, despite incorporating domain-specific keywords, do not consistently improve classification outcomes, indicating the complex challenge of URL classification.

The pursuit of an optimal tokenization technique is complex and context-dependent. Our findings suggest that there is no one-size-fits-all solution; the choice of tokenizer must be tailored to the specific nuances of the classification task, with considerations for both computational efficiency and accuracy. For instance, while **Char (3-6)-grams** and **Enhanced BPE** offer detailed token representations, their slower prediction times may not be suitable for all web crawling contexts.

Suggestions for Future Research Directions

Future research should explore the integration of multiple tokenization techniques, potentially leveraging the strengths of each to improve classification performance, especially for the elusive 'Malicious' class. Additionally, investigating the incorporation of semantic analysis and contextual understanding into the tokenization process could yield significant advancements. Another promising direction is the application of deep learning models that could learn optimal token representations in an end-to-end manner, potentially overcoming the limitations of predetermined tokenization schemes.

Continued exploration in tokenization techniques is critical as web content evolves. The development of more adaptive, context-aware models could greatly enhance the precision of web crawlers and their utility in navigating the ever-growing expanse of the internet.

ACKNOWLEDGEMENTS

Federal Ministry
of Education
and Research

This work is part of the OpenWebSearch.eu project, funded by the EU under the GA 101070014, and part of the CAROLL project, funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049.

REFERENCES

[1] Mohammed Al-Maamari, Mahmoud Istaiti, Saber Zerhoudi, Michael Dinzinger, Michael Granitzer, and Jelena Mitrovic. A Comprehensive Dataset for Webpage Classification (Part 1: Adult & Malicious), March 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

[4] Zaid Alyafeai, Maged S Al-shaibani, Mustafa Ghaleb, and Irfan Ahmad. Evaluating various tokenizers for arabic text classification. *Neural Processing Letters*, 55(3):2911–2933, 2023.

[5] Gyeongmin Kim, Junyoung Son, Jinsung Kim, Hyunhee Lee, and Heuiseok Lim. Enhancing korean named entity recognition with linguistic tokenization strategies. *IEEE Access*, 9:151814–151823, 2021.

[6] Michael Granitzer, Stefan Voigt, Noor Afshan Fathima, Martin Golasowski, Christian Guetl, Tobias Hecking, Gijs Hendriksen, Djoerd Hiemstra, Jan Martinovič, Jelena Mitrović, et al. Impact and development of an open web index for open web search. *Journal of the Association for Information Science and Technology*, 2023.

[7] Duygu Taylan, Mitat Poyraz, Selim Akyokuş, and Murat Can Ganiz. Intelligent focused crawler: Learning which links to crawl. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 504–508. IEEE, 2011.

[8] R Rajalakshmi and Chandrabose Aravindan. Web page classification using n-gram based url features. In *2013 fifth international conference on advanced computing (ICoAC)*, pages 15–21. IEEE, 2013.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.