# SoAC and SoACer: A Sector-Based Corpus and LLM-Based Framework for Sectoral Website Classification

Shahriar Shayesteh[1], Mukund Srinath[1], Lee Matheson[2], Lu Xian[3], Sinjoy Saha[1], C. Lee Giles[1], Shomir Wilson[1]
[1]Pennsylvania State University [2]Future of Privacy Forum [3]University of Michigan
{sxs7285,mus824,sks7620,clg20,shomir}@psu.edu,lmatheson@fpf.org,xianl@umich.edu

## Abstract

One approach to understanding the vastness and complexity of the web is to categorize websites into sectors that reflect the specific industries or domains in which they operate. However, existing website classification approaches often struggle to handle the noisy, unstructured, and lengthy nature of web content, and current datasets lack a universal sector classification labeling system specifically designed for the web. To address these issues, we introduce SoAC (Sector of Activity Corpus), a large-scale corpus comprising 195,495 websites categorized into 10 broad sectors tailored for web content, which serves as the benchmark for evaluating our proposed classification framework, SoACer (Sector of Activity Classifier). Building on this resource, SoACer is a novel end-to-end classification framework that first fetches website information, then incorporates extractive summarization to condense noisy and lengthy content into a concise representation, and finally employs large language model (LLM) embeddings (Llama3-8B) combined with a classification head to achieve accurate sectoral prediction. Through extensive experiments, including ablation studies and detailed error analysis, we demonstrate that SoACer achieves an overall accuracy of 72.6% on our proposed SoAC dataset. Our ablation study confirms that extractive summarization not only reduces computational overhead but also enhances classification performance, while our error analysis reveals meaningful sector overlaps that underscore the need for multi-label and hierarchical classification frameworks. These findings provide a robust foundation for future exploration of advanced classification techniques that better capture the complex nature of modern website content. [1]

## Keywords

Website Classification, Web Content Analysis, Multi-Class Text Classification, Sector-Based Corpus, LLM-Based Text Classification.

**ACM Reference Format:**
Shahriar Shayesteh[1], Mukund Srinath[1], Lee Matheson[2], Lu Xian[3], Sinjoy Saha[1], C. Lee Giles[1], Shomir Wilson[1]. 2025. SoAC and SoACer: A Sector-Based Corpus and LLM-Based Framework for Sectoral Website Classification. In *Proceedings of The 25th ACM Symposium on Document Engineering (DocEng'25).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3704268.3742691

## 1 Introduction

Automatically classifying website content by sector of activity offers a practical way to organize and interpret the web's information landscape and is beneficial for a range of applications, including cybersecurity, where it enables the identification of vulnerabilities specific to particular industries, and targeted advertising, where aligning ad content with a site's sector enhances user engagement and click-through rates [12, 21, 33]. Additionally, sector classification is pivotal for enforcing sector-specific privacy regulations in the U.S., which adopts a sectoral approach to data privacy legislation. For example, the Health Insurance Portability and Accountability Act (HIPAA) mandates rigorous standards for healthcare websites handling protected health information. Thus, accurately assigning websites to sectors such as finance, healthcare, or education is essential for effective regulatory compliance.

However, the inherent complexity and lack of central governance on the Web present significant challenges in developing a universal sector-based classification labeling system tailored specifically for web content. Existing classification labeling systems, such as the North American Industry Classification System (NAICS) and the Statistical Classification of Economic Activities in the European Community (NACE), were initially designed for broader economic reporting [27] and often fail to capture the dynamic, and multifaceted nature of contemporary website content [15].

To address these limitations, we introduce the SoAC, a novel dataset specifically designed for sector-based website classification. Our dataset employs a simplified coarse-grained labeling system of 10 sectors, derived from the PrivaSeer framework [31]. This system consolidates 148 detailed industry categories from the fine-grained labeling system originally developed by People Data Labs (PDL),[2] resulting in broader sector categories that are better suited for web-based classification tasks. Throughout this paper, we refer to the 10-sector classification developed by PrivaSeer as the coarse-grained labeling system, and to the original 148-category classification from PDL as the fine-grained labeling system. The mapping between these two labeling systems is provided in Table D.

To operationalize this dataset, we also introduce SoACer, a novel framework for automated multi-class website classification that leverages large language models (LLMs) to address the challenges of noisy, unstructured web content. SoACer is built on the SoAC and designed to push beyond the limitations of existing website

[1] The code is available at https://privaseer.ist.psu.edu/data.

[2] People Data Labs (PDL) provides comprehensive data on companies and industries; see https://docs.peopledatalabs.com/docs/industries for details.

*DocEng'25, Nottingham, UK*

ACM ISBN 979-8-4007-1351-4/25/09
https://doi.org/10.1145/3704268.3742691

classification methods through a three-stage pipeline. The framework begins by applying extractive summarization with LexRank to condense lengthy and often inconsistent website content[3] into concise representations. This approach assumes that the most frequently discussed content in a document reflects the website's core activity. Since longer texts demand higher memory and computation, we systematically evaluate multiple summary lengths and select the optimal configuration based on validation performance. These optimized summaries are then transformed into embeddings using a frozen Llama3-8B model, and passed through a fine-tuned classification head to predict website sectors.

Our thorough analysis shows that the effectiveness of SoACer in leveraging lightweight LLMs (Llama3-8B) for accurate sectoral classification. Furthermore, the ablation study shows concise summaries not only reduce computational overhead but also yield superior classification performance compared to using the full text. Although state-of-the-art LLMs are capable of handling long contexts, we find that summary-based inputs offer a more efficient alternative without sacrificing accuracy for this task. Furthermore, we demonstrate that adopting the coarse-grained labeling system for the classification task, rather than the initial 148-category fine-grained labeling system, improves the classification performance of SoACer. An error analysis reveals semantic overlaps resulting from shared vocabulary among sectors with similar activities. Through sector-based evaluation, we find that sectors characterized by blurred thematic boundaries or broad content scopes experience lower classification accuracy, motivating the need for multi-label or hierarchical classification in future work.

The main contributions of this work are as follows:

- **SoAC Corpus:** A dataset of 195,495 websites categorized into 10 universal website sectors.
- **LLM-based Website Classification Framework:** SoACer, a novel pipeline introducing multi-class website classification using LLMs for the first time in website classification.

The following sections review related work, describe the SoAC Corpus and SoACer framework in detail, present comprehensive experimental results highlighting our framework's performance improvements and practical benefits, and finally discuss the broader implications, limitations, and future directions for sector-based website classification research.

## 2 Related Works

Web content classification, encompassing both page and website categorization, is crucial for online information retrieval and management [22, 29]. Web page classification focuses on individual pages, whereas website classification addresses entire sites holistically.

### 2.1 Classification of Web Content

Traditional web page classification has utilized methods like Naive Bayes, KNN [5, 20], and SVMs [28, 29]. Deep learning approaches (CNNs [1], LSTMs [22]) improved performance but still lag behind transformer-based models (e.g., Llama, BERT, RoBERTa) in capturing complex semantics and richer contexts [34]. Nevertheless, model performance heavily relies on dataset quality. Prior research typically employed public directories like Yahoo! and dmoz ODP, as well as datasets such as WebKB [11] and 20 Newsgroups [29], which have become outdated and lack sector-specific categories. Domain-specific sectoral classification (e.g., IndustrySector [4]) exists, but typically focuses on structured data and targeted domains. Unlike these approaches, SoAC addresses broader sector-based classification by directly modeling noisy and unstructured website content.

### 2.2 Website Classification

Website classification requires holistic analysis of entire sites and has received comparatively less attention [10]. Earlier studies primarily focused on topic-based categorization using manual or single-label approaches [25, 30]. Automated sector-based methods often utilized industry classification systems such as NAICS, originally not designed for web categorization [12, 21, 33]. Addressing this, the PrivaSeer labeling system consolidates 148 industries into 10 web-specific sectors (Table 1), enabling practical and regulatory applicability. For example, medical sites must adhere to HIPAA regulations, while finance or education sectors follow different privacy standards tailored to their specific data.

*2.2.1 Challenges in Website Classification.* Lengthy and noisy content significantly challenges website classification. Traditional models struggle with the extensive, multi-page, and structurally diverse nature of websites [7, 16]. Commercial sites often include non-informative elements such as advertisements and navigation panels, which must be effectively removed to avoid distorting thematic signals critical for accurate categorization [2, 36].

### 2.3 Text Classification with LLMs

Text classification has been significantly advanced by Large Language Models (LLMs), starting with transformer-based models such as BERT [8] and RoBERTa [24]. Although these models demonstrated strong performance, their limited context length (512 tokens) restricted effectiveness on longer documents. Recent research transitioned toward embedding-based methods using decoder-only LLMs like GPT to better capture semantic relationships [32]. Innovations like lightweight LLMEmbed [23] optimize computational efficiency by combining embeddings from multiple layers, achieving performance comparable to larger models with reduced overhead.

Prompt-based classification leverages pre-trained LLMs by framing tasks as natural language prompts, enabling zero-shot classification without extensive labeled data [26]. However, these methods face sensitivity to prompt wording, inconsistent outputs, and biases, limiting reliability for large-scale tasks [6, 18]. Hybrid approaches, such as PTEC [4], address these issues by integrating prompt tuning with embedding-based classification, enhancing scalability and accuracy. Selecting an appropriate approach depends on specific application requirements, balancing accuracy, computational constraints, and task nature.

In summary, previous web classification research using traditional and topic-based methods struggled with the complexities and sectoral nature of modern web content. To overcome these, we introduce the SoAC corpus, an up-to-date, large-scale dataset structured around the web-specific PrivaSeer sector classification

[3] Noisy website content refers to the inconsistency and lack of coherence often found across different pages of a website.

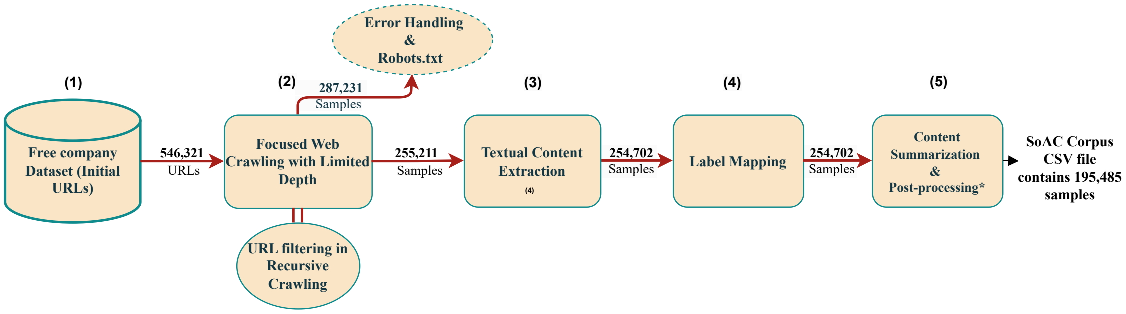


**Figure 1: Data collection and processing pipeline for the SoAC Corpus.**

system, offering enhanced utility for this task compared to the more granular 148-category PDL labeling system. Complementing this, our novel SoACer framework leverages extractive summarization combined with lightweight LLM embeddings, effectively addressing computational overhead, document length, and content noise. This integration demonstrates significant improvements in sector-based website classification performance.

## 3 SoAC Corpus: Collection and Processing

The SoAC Corpus comprises 195,495 websites categorized by their primary sector of activity, collected in 2024. This comprehensive dataset was developed through a systematic process designed to accurately represent the current web ecosystem.

### 3.1 Document Collection

The data collection process, as depicted in Figure 1, involved multiple systematic steps designed to ensure dataset comprehensiveness and quality:

**(1) Initial URLs**. The corpus foundation was established using the People Data Labs platform[4], which provided **546,321** company profiles collected in 2018. Each profile included a URL linking to the company's landing webpage and industry labels based on 148 fine-grain PDL's [5] classification labeling system.

**(2) Focused Web Crawling**. A custom web crawling framework was developed to initiate data collection from each company's landing page. Using a breadth-first search with a depth limit of one, the crawler parsed not only the landing page but also all internal hyperlinks on the same domain (i.e., links pointing to other webpages within the same website). During this process, it systematically excluded universal legal documents, such as Terms and Conditions or Privacy Policies, that typically contain limited sector-specific information.

**(3) Content Extraction and Storage**. Initially, the dataset included raw HTML content from 254,702 websites. To enhance usability, boilerplate content (universal, non-informative sections) was subsequently removed to extract the textual content of each website. This refinement resulted in the exclusion of 509 instances where raw HTML content could not be parsed by the Boilerpipe framework.

**(4) Labeling Methodology**. The initial company data were annotated with 148 self-declared industry categories from People Data Labs (PDL). To construct the SoAC, we systematically mapped these fine-grained labels into 10 broader, coarse-grained sectors defined by the PrivaSeer classification labeling system (see Table 6). This transformation reduces data sparsity and semantic overlap, enhancing interpretability. We empirically validate the superiority of this coarse-grained framework over the original 148 fine-grain labels in Section 5.5.

**(5) Content Summarization and Post-processing**. To optimize the corpus for classification tasks, we applied content-length filtering to exclude documents exceeding 100,000 words, thereby preventing length-induced bias in the training data [19]. We also filtered out overly short documents, which often consisted of noisy or non-informative content. In particular, we excluded website content with fewer than 70, a threshold derived from empirical observation. We found that such short entries frequently lacked meaningful semantic structure and made the LexRank algorithm ineffective in producing coherent summaries. While the 70-word cutoff was determined heuristically, it was informed by manual inspection and practical performance considerations during the preprocessing stage.

### 3.2 Dataset Statistics

The SoAC Corpus, available in the HuggingFace (HF) dataset repository [6], comprises 195,495 unique websites collected as of 2024, serving as a robust resource for website content classification research. The dataset is systematically divided into training (56%), validation (14%), and test (30%) sets (see Table 5). On average, each website contains 6,544 words, with a median length of 3,212 words. Table 1 summarizes the distribution of websites across the 10 defined sectors.

Notably, the dataset exhibits inherent class imbalance, mirroring real-world sector distribution patterns where certain sectors dominate the digital landscape. Recognizing and possibly addressing this imbalance is crucial for accurate model evaluation and real-world applicability.

[4] https://docs.peopledatalabs.com/docs/free-company-dataset
[5] https://docs.peopledatalabs.com/docs/industries

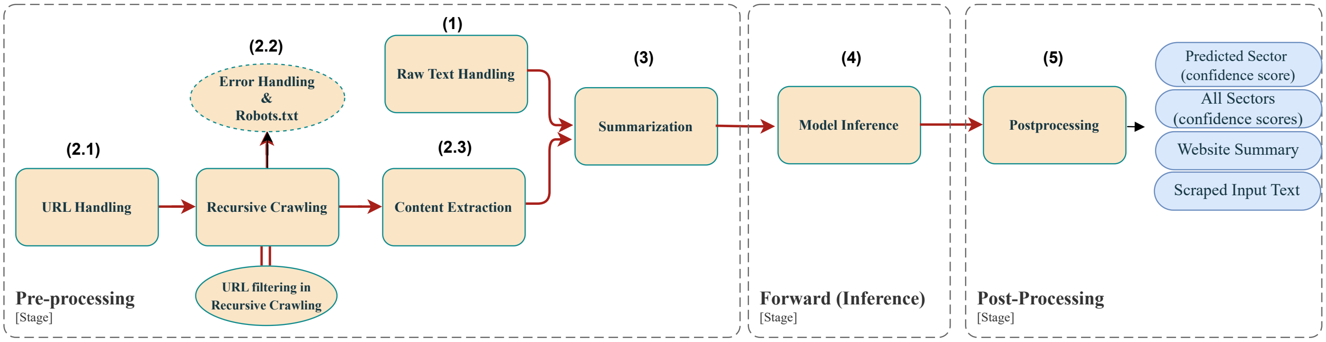


**Figure 2: SoACer Framework a three-stage pipeline design: Pre-processing, Forward (Inference), and Post-Processing**

**Table 1: Distribution of Websites Across Sectors of Activity in SoAC Corpus.**

| Category (Acronym) | Count | % |
|---|---|---|
| Finance, Marketing & Human Resources (**FMHR**) | 38,331 | 19.6% |
| Information Technology & Electronics (**ITE**) | 29,588 | 15.1% |
| Consumer & Supply Chain (**CSC**) | 27,030 | 13.8% |
| Civil, Mechanical & Electrical (**CME**) | 25,460 | 13.0% |
| Medical (**MED**) | 17,393 | 8.9% |
| Sports, Media & Entertainment (**SME**) | 15,808 | 8.1% |
| Education (**EDU**) | 13,247 | 6.8% |
| Government, Defense & Legal (**GDL**) | 11,124 | 5.7% |
| Travel, Food & Hospitality (**TFH**) | 10,281 | 5.3% |
| Non-Profit (**NP**) | 7,233 | 3.7% |

**Table 2: Partition of the SoAC Corpus into training, validation, and test sets for supervised classification.**

| Set | Count | Percentage |
|---|---|---|
| Training | 109,476 | 56% |
| Validation | 27,370 | 14% |
| Test | 58,649 | 30% |

# 4 SoACer: Website Classification Framework

The SoACer framework, illustrated in Figure 2, enables high-level sectoral classification by integrating content summarization, transformer-based embeddings, and a multi-class linear classification head. This section details the training and inference procedures of the framework.

## 4.1 Training Architecture Overview

The training procedure consists of three main components: content summarization, model architecture, and training strategy for multi-class classification.

*4.1.1 Content Summarization.* To handle lengthy web content, we employ LexRank [9], an extractive summarization technique, to generate concise summaries for each website. Summarization improves the efficiency of processing lengthy website content with transformer models like Meta-Llama-3-8B by reducing input length while aiming to retain key information about the service that a website offers. The summary for each website content $w_i$ where $i \in [1, 2, ..., n]$ is computed as:

$$s_i = \text{LexRank}(w_i, \text{sentences\_count}) \tag{1}$$

where sentences_count specifies the target number of sentences in the summary. LexRank uses a graph-based approach with a PageRank-derived algorithm to ensure that the most pertinent information is retained for classification purposes. Further technical details on LexRank are provided in Appendix A.

*4.1.2 Multi-class Classification.* We frame sector prediction as a multi-class classification problem. For each website summary $s_i$, we first obtain contextualized token embeddings using Meta-LLaMA-3-8B:

$$\mathbf{H}^{(L)} = \text{LlamaEmbed}(s_i) = \left[\mathbf{h}_1^{(L)}, \mathbf{h}_2^{(L)}, \ldots, \mathbf{h}_T^{(L)}\right], \tag{2}$$

where $\mathbf{h}_t^{(L)} \in \mathbb{R}^d$ is the embedding of token $t$ from the last hidden layer $L$, and $T$ is the total number of tokens in the summary.

The embedding vector $\mathbf{x}$ is then computed by applying mean pooling over the transformer outputs:

$$\mathbf{x} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t^{(L)}. \tag{3}$$

The embedding $x$ is then processed through a multi-layer perceptron (MLP) to produce logits, defined as:

$$\mathbf{h}^{(l)} = \text{Drop}\left(\text{LReLU}\left(\text{BN}\left(\mathbf{W}_l \mathbf{x}^{(l-1)} + \mathbf{b}_l\right)\right)\right), \quad l = 1, 2 \tag{4}$$

$$\mathbf{z} = \mathbf{W}_3 \mathbf{h}^{(2)} + \mathbf{b}_3 \tag{5}$$

where Drop(·), LReLU(·), and BN(·) denote Dropout, LeakyReLU, and Batch Normalization, respectively.

The model predicts a sector class $y \in \{1, ..., C\}$, where $C$ is the number of sectors. We use the Cross-Entropy Loss function defined as:

$$\mathcal{L}(\mathbf{x}, y) = -\sum_{c=1}^{C} y_c \log\left(\frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}\right) \tag{6}$$

where $z$ represents the logits produced by the model's final linear layer, and $y$ denotes the ground truth label represented as a one-hot

[6] https://huggingface.co/datasets/Shahriar/SoAC_Corpus

encoded vector. During training, this loss is minimized to optimize the model parameters.

*4.1.3 Training Procedure.* We divide the SoAC Corpus into training, validation, and test splits (Table 5). Model training is conducted in epochs, optimizing model parameters via Adam optimizer with a learning rate of 2e-4. After each epoch, the model's performance is evaluated on the validation loss. The best-performing model, based on the lowest validation loss, is selected for testing. This approach balances overfitting prevention and computational efficiency, ensuring robust model performance.

## 4.2 Inference

The inference process in the SoACer Framework is designed to categorize new, unseen web content into the most relevant sector designed similarly to Hugging Face pipelines and consists of three main components: Pre-processing, Forward (Inference), and Post-Processing.

*Pre-processing Stage.* This stage handles two types of input: raw text and website URLs. The process varies based on the input type:

- **Raw Text (1)**: Directly forwarded to text summarization, assuming it represents the website content.
- **Website URL (2.1 and 2.2)**: If the input is a URL, the framework initiates a recursive crawling process by parsing the main webpage of the given website URL, followed by depth-limited crawling to efficiently gather relevant content while adhering to `robots.txt` rules. The extracted content is then processed using Boilerpipe to retrieve the main textual content.
- **Text Summarization (3)**: Both crawled and directly submitted texts undergo summarization using LexRank to produce concise summaries.

*Forward (Inference) Stage (4).* The summarized content is then passed through the fine-tuned classifier to generate the final sector prediction.

*Postprocessing Stage (5).* The final prediction, along with all sectors with their confidence score, website-generated summary, and original scraped input text from the website, are the output of SoACer framework in the inference time.

# 5 Experiments

This section outlines the experiments conducted to evaluate the SoAC and SoACer framework. We assess our website classfiers performance through a series of evaluations, focusing on four key aspects: (1) the impact of LexRank summary length on both classification accuracy and computational efficiency, (2) an analysis of classifiers' performance on the website classification task on the optimal summary length selected in step 1 (3) a comparative analysis of full-text versus summary- based classification as an ablation study, and (4) an evaluation of PrivaSeer labeling system applied to the SoAC dataset, highlighting its strengths and limitations.

In this section, we summarize the experimental evaluations conducted to assess the SoACer framework. A detailed description of the baseline model architectures, as well as SoACer hyperparameters, training setup, and classification head design, can be found in Appendix B.

## 5.1 Evaluation Metrics

We evaluate classification performance using five standard metrics: Accuracy, Weighted Accuracy, Weighted Precision, Weighted Recall, and Weighted F1-score. These weighted metrics reflect the relative frequency of each class by assigning a higher weight to classes with more true instances. This provides an overall performance estimate that aligns with the data distribution, ensuring that the majority classes are proportionally represented in the evaluation. Formal definitions are provided in Appendix C.

## 5.2 Impact of Summary Length on Classification Performance

In this subsection, we systematically evaluated the impact of varying the number of sentences extracted by LexRank summarization on the classification performance (using Meta-Llama-3-8B embeddings) of the SoACer framework. Table 3 summarizes the results obtained from different summary lengths, quantified by the number of extracted sentences and their corresponding token counts[7]. Our findings indicate a clear trend in model performance relative to summary length.

Increasing the summary length consistently enhances classification accuracy, balanced accuracy, precision, recall, and F1-score, with substantial improvements observed between 2-sentence summaries (sc2) with an average of 107 tokens and 20-sentence summaries (sc20) with an average of 765 tokens. Specifically, the accuracy increased from 66.3% to 72.3%, balanced accuracy from 64.0% to 70.1%, and weighted precision, recall, and F1-score similarly improved.

The peak performance is achieved at 20-sentence summaries (average 765 tokens), achieving the highest accuracy of 72.3%. Beyond 20 sentences, we observed small fluctuations with a slight performance drop. For instance, summaries consisting of 25 and 30 sentences showed marginal variations in performance (accuracy at 72.0% and 72.1%, respectively).

From a computational efficiency perspective, shorter summaries significantly reduce memory and processing power requirements, enabling faster inference times and lower resource usage. Therefore, selecting a summary length of around 20 sentences (765 tokens) provides the optimal balance between high classification performance and computational efficiency. This experiment serves as a preliminary step in determining the optimal configuration for the SoACer framework.

## 5.3 Classifiers' Performance on the SoAC

In this subsection, we present a comprehensive comparative analysis of different classifiers' performance using various Large Language Model (LLMs) embeddings in the proposed SoACer framework. Table 4 details the performance metrics for different models evaluated on the website classification task.

[7] The number of token counts is computed using a rule of thumb, approximated as the number of words multiplied by 1.5.

**Table 3: Performance comparison across different summary lengths on the validation set. Sent. Count (Tok.) represents the number of top n extracted sentences for extractive summarization and the corresponding average token count. Acc. = Accuracy, W Acc. = Weighted Accuracy, W Prec. = Weighted Precision, W Rec. = Weighted Recall, and W F1 = Weighted F1-score.**

| Sent. Count (Tok.) | Acc. | B Acc. | W Prec. | W Rec. | W F1 |
|---|---|---|---|---|---|
| sc2 (107 tok.) | 66.3% | 64.0% | 66.1% | 66.3% | 66.0% |
| sc4 (195 tok.) | 69.3% | 68.2% | 69.4% | 69.3% | 69.3% |
| sc10 (427 tok.) | 71.1% | 69.6% | 71.1% | 71.1% | 71.0% |
| sc15 (601 tok.) | 71.7% | 69.9% | 71.6% | 71.7% | 71.5% |
| **sc20 (765 tok.)** | **72.3%** | **70.1%** | **72.2%** | **72.3%** | **72.1%** |
| sc25 (920 tok.) | 72.0% | 69.7% | 72.1% | 72.0% | 71.9% |
| sc30 (1067 tok.) | 72.1% | 70.2% | 72.1% | 72.1% | 71.9% |

*Optimal Model Selection for SoACer.* Among all models evaluated, **LLaMA3-8B** demonstrated the best performance, achieving an overall accuracy of 72.6%, balanced accuracy of 70.6%, and weighted precision, recall, and F1-scores of approximately 72.7%, 72.6%, and 72.4%, respectively. Consequently, we selected LLaMA3-8B as the primary model for the SoACer framework due to its superior predictive performance. In addition, interestingly utilizing the embeddings from reasoning language models [3], such as DS-LLaMA-8B (DeepSeek-R1-Distill-Llama-8B), does not improve classification performance compared to LLaMA3-8B.

*Performance Comparison Based on Model Size.* We observe that model performance generally correlates positively with model size. For instance, larger models such as LLaMA3-8B (8 billion parameters) and DS-LLaMA-8B achieve higher performance compared to smaller models like LLaMA-3.2-1B (1 billion parameters). Specifically, LLaMA3-8B outperforms LLaMA-3.2-1B by approximately 1.6% in overall accuracy, suggesting that larger models, due to increased parameter capacity, better capture semantic nuances critical for accurate classification.

*Auto-regressive vs. Encoder-based Models.* In this paper, we compare light-weight auto-regressive models (e.g., LLaMA3-8B, DS-LLaMA-8B, LLaMA-3.2-3B) with encoder-based architectures such as ModernBERT [35]. The results reveal significant differences. ModernBERT, representing a recent advancement in encoder-based architectures, achieved slightly lower accuracy (70.0%) compared to auto-regressive LLM like LLaMA3-8B (72.6%). Unlike previous encoder-based models, such as BERT and RoBERTa, ModernBERT incorporates several architectural improvements, making it a more competitive alternative for document-level understanding [35]. These enhancements allow ModernBERT to capture long-range dependencies more effectively while maintaining efficiency, making it a strong representative of encoder-based models in this task. Despite ModernBERT's strength in bi-directional context comprehension, the larger auto-regressive LLMs lead to more robust performance on website classification tasks that require nuanced contextual understanding and long-range dependencies.

*Traditional Methods vs. Advanced Models.* Evaluating traditional recurrent neural network models, particularly the Long Short-Term Memory (LSTM), we note a notable performance gap compared to modern transformer-based architectures. The LSTM model achieved significantly lower accuracy (66.0%) compared to advanced transformer models, such as LLaMA3-8B (72.6%) and DS-LLaMA-8B (72.1%). This underscores the transformational shift in natural language processing capabilities offered by transformer-based architectures which is primarily due to their superior handling of long-range contextual dependencies and providing a reach contextual embedding representation.

In summary, our analysis demonstrates that LLaMA3-8B consistently outperform smaller models and encoder-based architectures. This shows a critical role of model size in capturing the semantic nuances of web content. These results highlight that advanced transformer-based approaches offer a significant advantage over traditional methods such as LSTM by effectively handling long-range dependencies. Ultimately, our findings confirm that leveraging state-of-the-art LLM embeddings is key to achieving robust performance in website classification tasks.

**Table 4: Performance comparison of various model architectures on website classification. Metrics include Overall Accuracy (Acc.), Balanced Accuracy (B Acc.), Weighted Precision (W Prec.), Weighted Recall (W Rec.), and Weighted F1-score (W F1).**

| Model | Acc. | B Acc. | W Prec. | W Rec. | W F1 |
|---|---|---|---|---|---|
| DS-LLaMA-8B | 72.1% | 70.3% | 72.1% | 72.1% | 72.0% |
| LLaMA-3.2-1B | 71.0% | 69.0% | 71.0% | 71.0% | 70.9% |
| LLaMA-3.2-3B | 72.1% | 70.3% | 72.1% | 72.1% | 71.9% |
| **LLaMA3-8B** | **72.6%** | **70.6%** | **72.7%** | **72.6%** | **72.4%** |
| ModernBERT | 70.0% | 69.6% | 70.2% | 70.0% | 70.0% |
| LSTM | 66.0% | 66.0% | 65.8% | 66.0% | 65.7% |

## 5.4 Ablation Study: Comparative Analysis of Full-text vs. Summary-based Classification using LLM Embeddings

To evaluate the effectiveness of extractive summarization in handling lengthy and noisy website content, we conducted an ablation study comparing the classification performance of full-text versus summary-based inputs using the LLaMA-3.2-1B model. Given the significant computational resources required for processing full-text inputs, we strategically subsampled websites containing 7,000 tokens or fewer—far above the median website content length (4,878 tokens)—to reduce the computational footprint while preserving the essence of the dataset. This approach ensured that both models were trained on the same subsampled dataset, therefore enabling a fair and accurate comparison. Table 5 details the resulting subsampled dataset.

Figure 3 compares the classification performance metrics between summary-based and full-text inputs. Our findings demonstrate that summary-based classification consistently outperforms full-text classification across all metrics, with improvements of 3.5% in overall accuracy, 3.2% in balanced accuracy, 2.50% in weighted precision, 3.5% in weighted recall, and 3.8% in weighted F1-score.

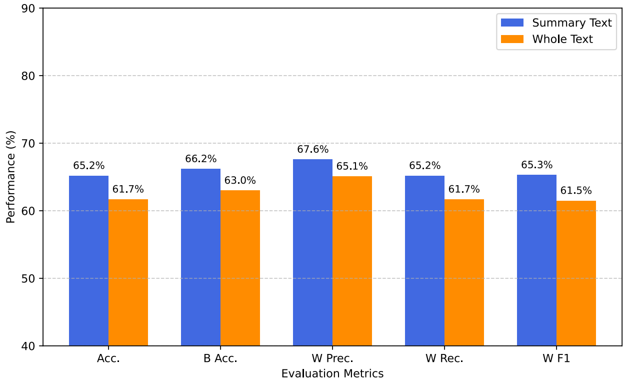


**Figure 3: Performance comparison of summary-based vs. full-text classification for Llama-3.2-1B. The bar chart illustrates Accuracy (Acc.), Balanced Accuracy (B Acc.), Weighted Precision (W Prec.), Weighted Recall (W Rec.), and Weighted F1-score (W F1). across both settings.**

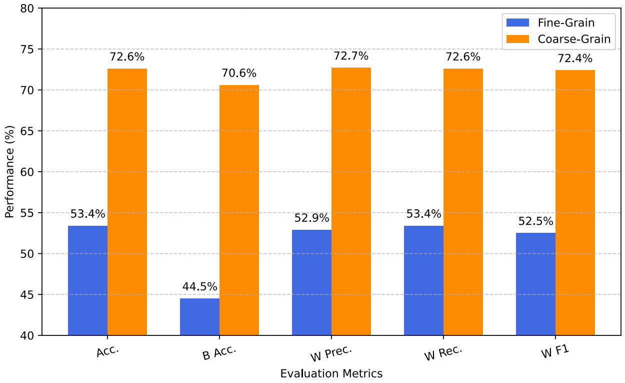


**Figure 4: Performance comparison of the sector classification task using fine-grained (148 detailed industry categories) labels versus coarse-grained (10 broad sectors from the PrivaSeer labeling system) labels.**

**Table 5: Dataset split and subsampling percentages for the ablation study.**

| Set | Count | Subsampled | Reduction (%) |
|---|---|---|---|
| Training | 109,476 | 20,000 | 81.74% |
| Validation | 27,370 | 14,021 | 48.77% |
| Test | 58,649 | 18,790 | 67.96% |

These results clearly indicate that extractive summarization not only significantly reduces computational overhead but also enhances model performance by distilling noisy and long website content into more coherent and contextually meaningful summaries. Additionally, another interesting observation aligns with our initial hypothesis that the most frequently discussed sections reflect a website's core activity, is supported by the performance gap between summary-based and full-text classification.

## 5.5 Evaluation of PrivaSeer Labeling System

This section investigates the effect of using the PrivaSeer labeling system in SoAC as classification labels on the website classification task performance. We do this by comparing the advantages of utilizing the PrivaSeer labeling system over the original fine-grained PDL labelling system. Then, to analyze the limitations of the selected coarse-grain (PrivaSeer) labeling system, we perform an error analysis to pinpoint common inter-sector misclassification and provide insights and recommendations for future improvements.

*5.5.1 Advantages of Coarse-Grain Categories over Fine-Grain Categories .* As shown in Figure 4, adopting the coarse-grained PrivaSeer labeling system as classification labels in the SoAC dataset yields significant improvements in SoACer performance compared to the original 148 fine-grained labels from People Data Labs (PDL). Specifically, we observe a notable increase in the weighted F1-score from approximately 52.5% with the fine-grained labeling system to 72.4% using the coarse-grained labeling system. Similar improvements are observed in accuracy, precision, and recall metrics. This improvement is attributed to the coarse-grained labels' ability to reduce data sparsity (less sectors with low number of samples in PrivaSeer ) and semantic overlap (less categories with similar activities in PrivaSeer). For instance, "Information Technology and Services" and "Computer Software" are distinct fine-grained categories that fall under the broader "IT & Electronics" coarse-grained category, as illustrated in the first row of Table 6.

*5.5.2 Error Analysis and Sector Overlaps.* While the previous section demonstrated a significant improvement in sectoral classification performance by adopting the coarse-grained labeling system in the SoAC, this section provides a detailed error analysis of the SoACer classification results. Our focus is on the confusion between coarse-grained sectors, to identify the underlying causes of misclassifications and discuss their implications for future work. To achieve this, we analyze the confusion matrix (Figure 5) and class-based performance metrics (Table 6) to pinpoint specific error patterns.

Figure 5, a normalized confusion matrix with the diagonal removed, visually represents misclassification percentages between the ten PrivaSeer sectors, where darker cells indicate higher misclassification percentages. Several noteworthy sectoral overlaps are evident:

*Education (EDU) ⟶ Finance, Marketing & HR (FMHR) (13.94%).* This represents the most frequent confusion. The high misclassification rate can be attributed to the inherent semantic overlap between the EDU and FMHR coarse-grained sectors, as defined by the underlying fine-grained categories (Table 6). Specifically, the EDU sector includes fine-grained subsectors such as 'Professional Training,' 'Fund-Raising,' and 'Market Research.' These subsectors naturally involve activities closely related to finance and marketing, and human resources. Therefore, these fine-grain sectoral overlaps create textual and conceptual similarities with the FMHR sector, which contains fine-grain sectors like 'Marketing/Advertising,' 'Financial Services,' 'Fund-Raising,' and 'Management Consulting.'

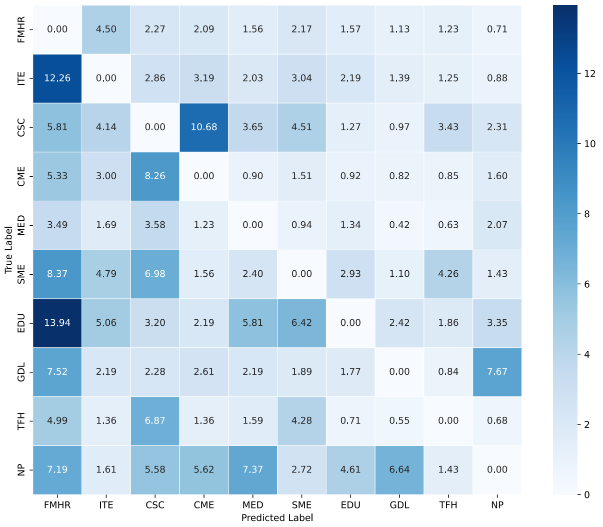
*Information Technology & Electronics (ITE) ⟶ Finance, Marketing & HR (FMHR) (12.26%).* The misclassification of 12.26% of websites labeled as ITE into the FMHR sector is associated to possible semantic overlap between the two coarse-grain sectors (Table 6). The ITE sector includes fine-grain categories like 'Information Technology and Services' and 'Program Development'. Therefore, websites that are related to the mentioned fine-grain sectors in ITE contain similar vocabulary to several FMHR fine-grain categories, such as 'Financial Services', 'Marketing/Advertising', and 'Human Resources'; therefore, SoACer may incorrectly classify them as FMHR. Essentially, the classifier detects the mention of FMHR-related activities within ITE websites.

*Consumer & Supply Chain (CSC) ⟶ Civil, Mechanical & Electrical (CME) (10.68%).* The misclassification between CSC and CME is attributed to the semantic overlap in their focus on physical goods and associated systems. While CSC encompasses categories like 'Consumer Goods,' 'Packaging,' and 'Transportation' (addressing products and their movement), CME includes categories such as 'Machinery,' 'Building Materials,' and 'Automotive' (addressing the creation and engineering of those products and the systems that utilize them). Consequently, websites related to CSC may discuss the materials used in packaging or the vehicles used in transportation, resulting in shared vocabulary between the sectors and their misclassification as CME, which centers on design and production aspects.

In summary, the error analysis reveals a consistent trend: semantic overlap between coarse-grained sectors, as defined by their constituent fine-grained categories (Table 6), contributes to misclassification errors. This suggests that the inherent ambiguity in assigning websites to a single sector, particularly when they exhibit characteristics of multiple sectors, limits the accuracy of the classification. The three most frequent misclassification patterns, discussed above, exemplify this issue. This analysis provides valuable insights into the limitations of a single-label classification approach for websites and highlights the need for future research to address these challenges.

*5.5.3 Sector-Based Performance.* Figure 6 and the accompanying metrics reveal that SoACer's accuracy differs substantially across the ten sectors. While unequal class sizes also contribute to these performance gaps, here we focus on differences arising from how narrowly or broadly each coarse-grained sector is defined by its underlying fine-grained subsectors (see Table 6).

Sectors that achieve higher classification accuracies, such as Medical (MED) at 84.6% and Finance, Marketing & Human Resources (FMHR) at 82.8%, are characterized by fine-grained subsectors that represent relatively focused and less overlapping domains. For example, the Medical sector includes subsectors specifically related to 'Hospital Care', 'Medical Practice', and 'Pharmaceuticals'. The concentrated nature of activities within these fine-grained subsectors contributes to more distinct textual characteristics, which leads to more accurate classification. However, FMHR also benefits from being the largest sector in the dataset (Table 1), with 38,331 samples (19.6%). This considerable sample size provides SoACer with a more extensive body of text to learn the characteristic patterns associated with FMHR, likely helping to counterbalance some of the potential ambiguity introduced by the diversity of its fine-grained components and contributing significantly to its high accuracy.



| | FMHR | ITE | CSC | CME | MED | SME | EDU | GDL | TFH | NP |
|---|---|---|---|---|---|---|---|---|---|---|
| FMHR | 0.00 | 4.50 | 2.27 | 2.09 | 1.56 | 2.17 | 1.57 | 1.13 | 1.23 | 0.71 |
| ITE | 12.26 | 0.00 | 2.86 | 3.19 | 2.03 | 3.04 | 2.19 | 1.39 | 1.25 | 0.88 |
| CSC | 5.81 | 4.14 | 0.00 | 10.68 | 3.65 | 4.51 | 1.27 | 0.97 | 3.43 | 2.31 |
| CME | 5.33 | 3.00 | 8.26 | 0.00 | 0.90 | 1.51 | 0.92 | 0.82 | 0.85 | 1.60 |
| MED | 3.49 | 1.69 | 3.58 | 1.23 | 0.00 | 0.94 | 1.34 | 0.42 | 0.63 | 2.07 |
| SME | 8.37 | 4.79 | 6.98 | 1.56 | 2.40 | 0.00 | 2.93 | 1.10 | 4.26 | 1.43 |
| EDU | 13.94 | 5.06 | 3.20 | 2.19 | 5.81 | 6.42 | 0.00 | 2.42 | 1.86 | 3.35 |
| GDL | 7.52 | 2.19 | 2.28 | 2.61 | 2.19 | 1.89 | 1.77 | 0.00 | 0.84 | 7.67 |
| TFH | 4.99 | 1.36 | 6.87 | 1.36 | 1.59 | 4.28 | 0.71 | 0.55 | 0.00 | 0.68 |
| NP | 7.19 | 1.61 | 5.58 | 5.62 | 7.37 | 2.72 | 4.61 | 6.64 | 1.43 | 0.00 |

**Figure 5: Presents a row-normalized confusion matrix with the diagonal (correct predictions) blanked out so that only misclassification rates remain. Each row corresponds to the true sector, and each column to the predicted sector; darker cells indicate a higher fraction of websites from the true sector being assigned to the predicted sector. This makes it easy to spot which sector pairs SoACer most frequently confuses, for example, a dark cell in the row for *Education* (EDU) under the column for *Finance, Marketing & HR* (FMHR) shows that 13.94% of Education websites are misclassified as Finance, Marketing & HR.**

In contrast, sectors with lower accuracies, including Education (EDU) at 55.8%, Non-Profit (NP) at 57.2%, and Consumer & Supply Chain (CSC) at 63.2%, are defined by a more diverse and extensive set of fine-grained subsectors. This broader scope introduces greater potential for ambiguity and overlap with other sectors. For instance, the Education sector incorporates a wide range of subsectors, from 'Education Management' and 'Higher Education' to 'Publishing' and 'Market Research'. Similarly, the Non-Profit sector includes different areas such as 'Non-Profit Management', 'Environmental Services', and 'Public Safety'. The Consumer & Supply Chain (CSC) sector is also broadly defined. The CSC encompasses 'Retail', 'Consumer Goods', and 'Logistics', among others. The inclusion of such varied fine-grained activities within these coarse-grained sectors can lead to less distinct textual patterns and increased semantic overlap, which presents greater challenges for accurate classification. This analysis indicates that the degree of focus and diversity within the fine-grained subsectors comprising each coarse-grained sector influences the clarity of sector boundaries and contributes to the observed differences in classification performance.

Based on the error analysis and the class-based performance evaluation, the misclassification errors observed are closely related to the semantic overlaps between coarse-grained sectors, which arise from the diverse and sometimes overlapping nature of their

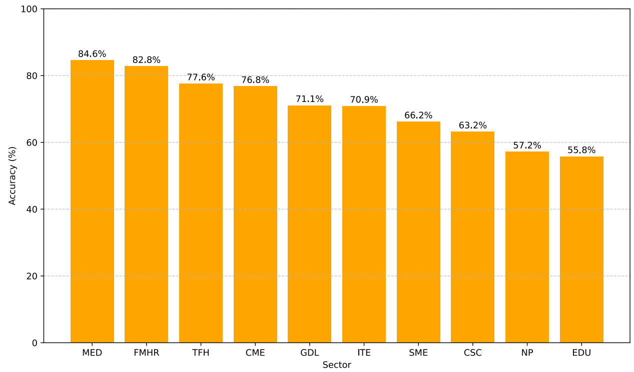


**Figure 6: Accuracy performance metrics across sectors.**

constituent fine-grained categories as defined in Table 6. The variation in sector-based classification performance directly reflects this, where sectors composed of more focused, fine-grained categories achieve higher accuracy, while those encompassing broader and more interconnected fine-grained categories present greater challenges for accurate classification by SoACer.

For future work, two promising directions emerge from these findings. First, the PrivaSeer labeling system could be refined for specific applications. For instance, adapting the coarse-grained sector definitions to align more closely with sector-specific privacy laws in the U.S. could enhance the framework's utility for regulatory compliance checks. Second, exploring multi-label or hierarchical classification approaches could better capture the multifaceted nature of websites that span multiple sectors. Such methods could not only improve overall classification accuracy by resolving ambiguities caused by sector overlaps but also enrich downstream applications, such as targeted advertising, where identifying multiple sector associations can lead to more effective ad placement strategies.

## 6 Disscussion

Our experimental investigations reveal several key insights into the efficacy of the SoACer framework for sector-based website classification. First, our analysis of the LexRank summarization parameters demonstrates that carefully tuning the summary length specifically, extracting approximately 20 sentences per document, yields a performance boost in classification. This experiment shows that an optimal level of content distillation can reduce computational overhead while preserving the important information necessary for the accurate classification of websites.

Secondly, the comparative study of various classifier architectures indicates that **LLaMA3-8B** outperforms both lighter autoregressive models and encoder-based counterparts on this task. The superior performance of larger models suggests that increased parameter capacity is crucial for capturing the nuanced semantic features of web content. Moreover, our experiments show that reasoning model embeddings (DeepSeek-R1-Distill-Llama-8B), despite sharing the same parameter size as LLaMA3-8B, do not improve classification accuracy. This suggests that reasoning-focused fine-tuning may not necessarily enhance the contextual representations needed for sector-based website classification tasks.

In the ablation study comparing full-text with summary-based classification, we observed that summaries not only enable faster inference times but also improve overall accuracy and other evaluation metrics. This finding supports our hypothesis that extractive summarization can effectively distill the core thematic elements of a website, which are often overshadowed by extraneous or redundant content scraped from the websites. It also opens a discussion on the benefits of refining summarization techniques and exploring controllable text summarization to extract more task-relevant information for website classification.

Finally, the in-depth analysis of the PrivaSeer labeling system highlights the advantages of adopting a coarse-grained sectoral categorization over fine-grained industry classifications. While the coarse-grain labeling approach substantially improves performance metrics by alleviating issues of data sparsity for underrepresented sectors and semantic overlap compared to the 148 fine-grain labeling approach, the error analysis reveals persistent challenges for the PrivaSeer labeling system. Notably, significant inter-sector misclassification (Figure 5), such as overlaps observed between Education and FMHR, ITE and FMHR, or CSC and CME, stems particularly from the inherent complexity and diversity of the fine-grained subsectors that compose certain coarse-grained categories (Table 6). These findings motivate the exploration of multi-label or hierarchical approaches that could better accommodate the multifaceted nature of modern websites and address the limitations of rigid single-label assignments.

## 7 Conclusion and Future Work

We introduced the SoAC corpus, a large-scale dataset of websites categorized into 10 distinct sectors, and SoACer, a novel LLM-based framework for sector-based website classification. Our experiments demonstrated the effectiveness of using 20-sentence extractive summarization and showed that classifiers leveraging LLaMA3-8B embeddings outperform tested alternatives. While the coarse-grained PrivaSeer labeling system offers advantages over fine-grained classification, our analysis revealed that misclassification errors and varying sector performance are linked to the overlaps and diversity of constituent fine-grained subsectors.

Moving forward, we identify two key directions for future research. First, the PrivaSeer labeling system could be refined for specific applications, such as aligning with U.S. sector-based privacy laws to enhance regulatory compliance checks. Second, exploring multi-label or hierarchical classification paradigms is essential to handle sector overlaps and more accurately represent multi-sector websites, which can be beneficial for applications like targeted advertising by identifying relevant sectoral associations.

## 8 Limitations and Ethical Considerations

Our analysis is based on self-declared industry categories, which may introduce selection bias because websites typically report only their primary sector. This practice can lead to the underrepresentation of secondary sectors and an overemphasis on dominant industries. Additionally, although our PrivaSeer labeling system aggregates 148 fine-grained categories into 10 broad sectors, this coarse-grained approach may overlook subtle distinctions that are crucial for a nuanced understanding of certain industries. These limitations could affect the fairness and generalizability of our findings and may have unintended implications for regulatory or practical applications.

## References

[1] Siti Hawa Apandi, Jamaludin Sallim, Rozlina Mohamed, and Araby Madbouly. 2021. Web Page Classification Using Convolutional Neural Network (CNN) Towards Eliminating Internet Addiction. In *2021 International Conference on Software Engineering  Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*. 149–154. doi:10.1109/ICSECS52883.2021.00034

[2] Hiteshwar Kumar Azad, Rahul Raj, Rahul Kumar, Harshit Ranjan, Kumar Abhishek, and M. P. Singh. 2014. Removal of Noisy Information in Web Pages *(ICTCS '14)*. Association for Computing Machinery, New York, NY, USA, Article 88, 5 pages. doi:10.1145/2677855.2677943

[3] Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, et al. 2025. Reasoning Language Models: A Blueprint. *arXiv preprint arXiv:2501.11223* (2025).

[4] Valentin Buchner, Lele Cao, Jan-Christoph Kalo, and Vilhelm Von Ehrenheim. 2024. Prompt Tuned Embedding Classification for Industry Sector Allocation. In *Proc. NAACL 2024 (Industry Track)*, Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 108–118. doi:10.18653/v1/2024.naacl-industry.10

[5] Pável Calado, Marco Cristo, Edleno Moura, Nivio Ziviani, Berthier Ribeiro-Neto, and Marcos André Gonçalves. 2003. Combining link-based and content-based methods for web document classification. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (New Orleans, LA, USA) *(CIKM '03)*. Association for Computing Machinery, New York, NY, USA, 394–401. doi:10.1145/956863.956938

[6] Arnav Chavan, Raghav Magazine, Shubham Kushwaha, Mérouane Debbah, and Deepak Gupta. 2024. Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward. *arXiv preprint arXiv:2402.01799* (2024).

[7] Célia D'Cruz, Jean-Marc Bereder, Frédéric Precioso, and Michel Riveill. 2024. Domain-specific long text classification from sparse relevant information. *arXiv preprint arXiv:2408.13253* (2024).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[9] Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.* 22, 1 (dec 2004), 457–479.

[10] Martin Ester, Hans-Peter Kriegel, and Matthias Schubert. 2002. Web site mining: a new way to spot competitors, customers and suppliers in the world wide web. In *Proc. KDD'02* (Edmonton, Alberta, Canada) *(KDD '02)*. Association for Computing Machinery, New York, NY, USA, 249–258. doi:10.1145/775047.775084

[11] Johannes Fürnkranz. 1999. Exploiting Structural Information for Text Classification on the WWW. In *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis (IDA '99)*. Springer-Verlag, Berlin, Heidelberg, 487–498.

[12] Kanwarpartap Singh Gill, Vatsala Anand, and Rupesh Gupta. 2023. Website Classification Through Exploratory Data Analysis Using Naive Bayes, Random Forest, and Support Vector Machine Classifier. In *Proc. CONIT 2023*. 1–5. doi:10.1109/CONIT59222.2023.10205766

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[15] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. 2008. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM* 51, 7 (jul 2008), 60–69. doi:10.1145/1364782.1364798

[16] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi. 2013. Overview of web content mining tools. *arXiv preprint arXiv:1307.1024* (2013).

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[18] Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. 2023. Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's GPT-4 with Self-Hosted Open Source SLMs in Production. *arXiv preprint arXiv:2312.14972* (2023).

[19] Lan Jiang, Tianshu Lyu, Yankai Lin, Meng Chong, Xiaoyong Lyu, and Dawei Yin. 2022. On Length Divergence Bias in Textual Matching Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. 4187–4193.

[20] Oh-Woog Kwon and Jong-Hyeok Lee. 2000. Web page classification based on k-nearest neighbor approach. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages* (Hong Kong, China) *(IRAL '00)*. Association for Computing Machinery, New York, NY, USA, 9–15. doi:10.1145/355214.355216

[21] Heidi Kühnemann, Arnout Delden, and Dick Windmeijer. 2020. Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. *Statistical Journal of the IAOS* 36 (09 2020), 807–821. doi:10.3233/SJI-200675

[22] Huaxin Li, Zhaoxin Zhang, and Yongdong Xu. 2019. Web Page Classification Method Based on Semantics and Structure. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. 238–243. doi:10.1109/ICAIBD.2019.8837027

[23] Chun Liu, Hongguang Zhang, Kainan Zhao, Xinghai Ju, and Lin Yang. 2024. LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7994–8004. doi:10.18653/v1/2024.acl-long.433

[24] Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[25] George Moiseev. 2016. Classification of E-commerce Websites by Product Categories. In *International Joint Conference on the Analysis of Images, Social Networks and Texts*. https://api.semanticscholar.org/CorpusID:17659714

[26] Simon Münker, Kai Kugler, and Achim Rettinger. 2024. Zero-shot prompt-based classification: topic labeling in times of foundation models in German Tweets. *arXiv preprint arXiv:2406.18239* (2024).

[27] Ryan Lee Phillips and Rita Ormsby. 2016. Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship* 21 (2016), 1 – 25. https://api.semanticscholar.org/CorpusID:61708603

[28] Xiaoguang Qi and Brian D. Davison. 2006. Knowing a web page by the company it keeps. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (Arlington, Virginia, USA) *(CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 228–237. doi:10.1145/1183614.1183650

[29] Xiaoguang Qi and Brian D. Davison. 2009. Web page classification: Features and algorithms. *ACM Comput. Surv.* 41, 2 (2009), 1–31. doi:10.1145/1459352.1459357

[30] Galuh Tunggadewi Sahid, Rahmad Mahendra, and Indra Budi. 2019. E-Commerce Merchant Classification using Website Information. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics* (Seoul, Republic of Korea) *(WIMS2019)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. doi:10.1145/3326467.3326486

[31] Mukund Srinath, Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson. 2021. PrivaSeer: A Privacy Policy Search Engine. In *Web Engineering*, Marco Brambilla, Richard Chbeir, Flavius Frasincar, and Ioana Manolescu (Eds.). Springer International Publishing, Cham, 286–301.

[32] Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Zhengwei Tao, and Shuai Ma. 2024. LLMs are Also Effective Embedding Models: An In-depth Overview. *arXiv preprint arXiv:2412.12591* (2024).

[33] Chaman Thapa, Osmar Zaiane, Davood Rafiei, and Arya M. Sharma. 2012. Classifying websites into non-topical categories. In *Proceedings of the 14th International Conference on Data Warehousing and Knowledge Discovery* (Vienna, Austria) *(DaWaK'12)*. Springer-Verlag, Berlin, Heidelberg, 364–377. doi:10.1007/978-3-642-32584-7_30

[34] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[35] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663* (2024).

[36] Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating noisy information in Web pages for data mining. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, D.C.) *(KDD '03)*. Association for Computing Machinery, New York, NY, USA, 296–305. doi:10.1145/956750.956785