GenCrawl: A Generative Multimedia Focused Crawler for Web Pages Classification

Domenico Benfenati[a], Antonio Maria Rinaldi[b], Cristiano Russo[c] and Cristian Tommasino[d]
Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Naples, Italy
{domenico.benfenati, antoniomaria.rinaldi, cristiano.russo, cristian.tommasino}@unina.it

Keywords: Web Crawling, Web Pages Classification, Generative AI, Web Topic Analysis.

Abstract: The unprecedented expansion of the internet necessitates the development of increasingly efficient techniques for systematic data categorization and organization. However, contemporary state-of-the-art techniques often need help with the complex nature of heterogeneous multimedia content within web pages. These challenges, which are becoming more pressing with the rapid growth of the internet, highlight the urgent need for advancements in information retrieval methods to improve classification accuracy and relevance in the context of varied and dynamic web content. In this work, we propose GenCrawl, a generative multimedia-focused crawler designed to enhance web document classification by integrating textual and visual content analysis. Our approach combines the most relevant topics extracted from textual and visual content, using innovative generative techniques to create a visual topic. The reported findings demonstrate significant improvements and a paradigm shift in classification efficiency and accuracy over traditional methods. GenCrawl represents a substantial advancement in web page classification, offering a promising solution for systematically organizing web content. Its practical benefits are immense, paving the way for more efficient and accurate information retrieval in the era of the expanding internet.

1 INTRODUCTION

The rapid growth of the web has led to an overwhelming amount of information, with over 5 billion web pages available online (Kunder, 2018; Bergman, 2001). Effective web page classification is crucial for various applications, including information retrieval, content recommendation, and search engine optimization. However, web content's diverse and dynamic nature, including textual and multimedia elements, presents significant challenges for traditional classification methods (Chakrabarti et al., 1999). Previous approaches to web page classification have primarily focused on analyzing textual content, often neglecting the valuable information embedded in visual elements (Rinaldi et al., 2021c; Rinaldi et al., 2021b). While some methods have attempted to incorporate multimedia data, they typically treat textual and visual content separately, failing to leverage the synergistic potential of combining these modalities (Ahmed and Singh, 2019). Furthermore, the complexity and resource-intensive nature of processing large volumes of multimedia content necessitate the development of more efficient and scalable solutions (Fernàndez-Cañellas et al., 2020). Introducing GenCrawl, a genuinely innovative, generative, multimedia-focused crawler, in response to these challenges. This novel approach, which integrates textual and visual content analysis, promises to enhance web page classification accuracy and revolutionize the field. Our work makes some primary contributions: we introduce a novel interdisciplinary approach that combines textual and visual content analysis, significantly improving the classification performance of multimedia-focused web crawlers. This comprehensive approach ensures that no aspect of web content is overlooked, enhancing the accuracy of our classification system. A conventional focused crawler relies on link-based navigation, which may not prioritize systematic data acquisition and processing. This problem limits discerning meaningful patterns, extracting valuable insights, and adapting to dynamic web content evolution. The data generated may lack refinement, potentially leading to lower data quality and hindering analysis accuracy (Kumar and Aggarwal, 2023). Ethical

[a] https://orcid.org/0009-0008-5825-8043
[b] https://orcid.org/0000-0001-7003-4781
[c] https://orcid.org/0000-0002-8732-1733
[d] https://orcid.org/0000-0001-9763-8745

Benfenati, D., Rinaldi, A., Russo, C. and Tommasino, C.
GenCrawl: A Generative Multimedia Focused Crawler for Web Pages Classification.
DOI: 10.5220/0012998900003838
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 91-101
ISBN: 978-989-758-716-0; ISSN: 2184-3228
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

and privacy considerations may also pose challenges. Adopting a more adaptive framework for web crawling strategies could enhance the effectiveness of conventional focused crawlers. The synergistic integration of deep learning techniques can enhance crawler proficiency in analyzing and categorizing web pages, ensuring seamless incorporation of diverse instances into the evolving web page classification task (K et al., 2023). This approach addresses web page classification intricacies and advances information representation and retrieval methodologies in the expansive and dynamic web-based knowledge dissemination era.

The article is organized as follows: in Section 2 a literature review is presented and discussed, putting in evidence the novelties of our approach; the system architecture and the proposed methodology for crawling strategy and web pages classification methodology are discussed in Section 3; a use case of our crawler together with experimental results are in Section 4; eventually, conclusions and future works are presented in Section 5.

2 RELATED WORKS

General-purpose and special-purpose web agents are the categories into which web crawlers fall (Bhatt et al., 2015). Instead of providing a thorough analysis of general-purpose crawlers, this section highlights relevant research on focused crawlers within the framework of web page classification. Since our framework is based on ontologies, we carefully evaluate works that utilize and conform to this approach. On the other hand, we present a summary of alternative approaches, emphasizing studies that show how well Convolutional Neural Network (CNN) features operate as general feature extractors for multimedia content retrieval tasks.

A focused crawler filters millions of pages and finds relevant resources distant from the initial batch. Machine learning approaches are used to train focused crawlers, which can be extracted from online taxonomies or manually classified datasets (Pant and Srinivasan, 2005). The seminal work on focused crawlers is (Chakrabarti et al., 1999), where the authors developed two hypertext mining software: a classifier for document relevance assessment and a distiller for identifying hypertext nodes providing multiple access points to relevant pages. Moreover, genetic algorithms show intriguing results when it comes to concentrated crawling.

Ontology-based crawlers employ unsupervised ontology learning and domain-based ontology with multi-objective optimization for improved crawling performance and selection of weighted coefficients for web pages (Hassan et al., 2017; Liu et al., 2022; Russo et al., 2020), or for improvement of visualization and document summarization (Rinaldi and Russo, 2021). Event-based crawlers utilize event models and temporal intent recognition methods, including Google Trends data, to capture and prioritize event-related information (Farag et al., 2018; Wu and Hou, 2023). Phishing detection crawlers use isomorphic graph techniques to detect phishing content by identifying subgraph similarities between web pages (Tchakounte et al., 2022). Machine learning crawlers implement LSTM and CNN for word embeddings and classification, and Attention Enhanced Siamese LSTM Networks for predicting web page relevance in specific domains like biomedical information (Shrivastava et al., 2023; Mary et al., 2022). Rule-based and specialized domain crawlers leverage rule-based approaches for obfuscating audio file crawlers in the AIR domain and incremental crawling systems for the Dark Web (Benfenati et al., 2023; Fu et al., 2010). Genetic algorithm-based crawlers utilize modified Genetic Algorithms for web page classification based on keyword feature sets (Fatima et al., 2023). Additionally, an improved genetic algorithm can increase the focused crawler's memory and precision while broadening its search area, focusing on the direct influence of textual content and topic on user information retrieval (Yan and Pan, 2018).

The strategy proposed in this article introduces a multimedia-focused crawler for web page classification, which combines textual and visual topics from text and images as in (Rinaldi et al., 2021a). It uses a supervised learning algorithm to classify web pages, including those with text and images. In the latter case, a generative model extracts and creates images, improving visual topic extraction and crawling performance. The crawler's flexibility and adaptability to different domains make it more adaptable to predefined keywords or exemplary documents. Our study compares a web page classification method using text or images with existing methods, revealing superior accuracy, recall, and greater resilience to noisy or irrelevant content. We also discuss its benefits and drawbacks and suggest future enhancement directions.

3 PROPOSED FRAMEWORK

This section describes in detail how the proposed framework is composed, indicating specifically what the components are and how they function.

Figure 1 shows a sketch of the proposed system architecture. Our system involves multiple modules for crawler tasks. It starts with an online document repository, generating crawler threads. These threads retrieve web pages, analyze structures, classify text and images, and use a model for synthetic image generation if direct comparison isn't possible. Extracted topics from text and images are combined for document classification. The crawling process then selects the next page from the repository. The crawler initialization phase begins by gathering seed URLs from an online document repository and setting up a structured workflow for the subsequent stages of the crawling process. Following this, the web page retrieval and hyperlink extraction phase constructs a network of interconnected pages, creating a comprehensive dataset that mirrors the web's interconnected nature. In the parsing phase, the crawler differentiates between main content (relevant text) and auxiliary content (unrelated text), focusing on essential elements. It also identifies visual information, adding a multimedia dimension to the understanding of web pages.

In the web page classification phase, machine learning models categorize textual and visual elements based on their topics. This classification process automates the evaluation of content relevance, enhancing system efficiency and accuracy. When discrepancies arise between textual and visual classifications, the generative phase harmonizes these interpretations. Using a Latent Diffusion Model (LDM) for image generation, it synthesizes images that align with textual descriptions, ensuring cohesive content representation. The combined topics of textual content and generated images are then used to calculate a priority value for subsequent actions, indicating each web page's significance in the crawler's exploration.

The preprocessing module is crucial for analyzing textual and visual information on web pages, using the DMOZ web collection as a data source. Despite its official closure in 2017, DMOZ remains valuable for classification tasks due to its heterogeneous nature. A screenshot of this repository is available on the Kaggle Dataset platform[1].

The text preprocessing pipeline involves HTML parsing, normalization, tokenization, stopword removal, lemmatization, removal of special characters and symbols, spell-checking, feature extraction, and vectorization. These steps ensure consistent representation, enhance topic extraction efficiency, and prepare the text for analysis.

For encoding the text of web pages, we used SBERT (Cheng et al., 2023) to create vector representations capturing semantic meanings. Clustering techniques grouped sentences into topics based on semantic similarity, extracting representative keywords and generating labels for each topic using a rule-based method. This approach allowed the identification of main themes or concepts in the web pages.

The focused crawler's image processing pipeline involves a sequential process of extracting and analyzing images from web pages, starting with HTML parsing, downloading and preprocessing images for quality enhancement, and extracting relevant features as feature vectors. The critical phase involves applying advanced computer vision techniques, explicitly utilizing the VGG19 model (Simonyan and Zisserman, 2014). VGG19 is a mighty Convolutional Neural Network (CNN) consisting of 19 layers. It has been trained on a large and diverse dataset featuring complex image classification tasks, such as ImageNet (Ridnik et al., 2021), a large image dataset consisting of 14,197,122 images, each tagged in a single-label fashion by one of 21,841 possible classes. VGG19's adaptability makes it well-suited for the task of visual topic extraction in our work, as it excels at discerning patterns and objects in the analyzed images. Figure 2 represents the described textual and visual pipelines.

3.1 Textual Topic Detection

We employ a pipeline to identify representative topics in text using Sentence-BERT (SBERT) embeddings and WordNet synsets. Text is preprocessed through lowercasing, sentence tokenization, and removal of common linguistic artifacts, such as stopwords, special characters, etc, for uniformity. SBERT embeddings encode each sentence into a high-dimensional vector space, capturing contextual relationships. The SBERT application to the preprocessed text ensure that the vector created by the model doesn't effect the noisy informations in the long text of the web pages of the dataset. K-means clustering identifies distinct topics, while WordNet synsets enrich semantic interpretation.

This method's effectiveness relies on the input text's language richness and diversity, facilitating granular content exploration.

3.2 Visual Topic Detection

The visual topic detection task utilizes multimedia components, particularly images, to determine a document's principal topic, enhancing the overall framework's performance. Our research currently focuses on recognizing multimedia descriptors to measure the

[1] https://www.kaggle.com/datasets/shawon10/url-classification-dataset-dmoz?select=URL+Classification.csv

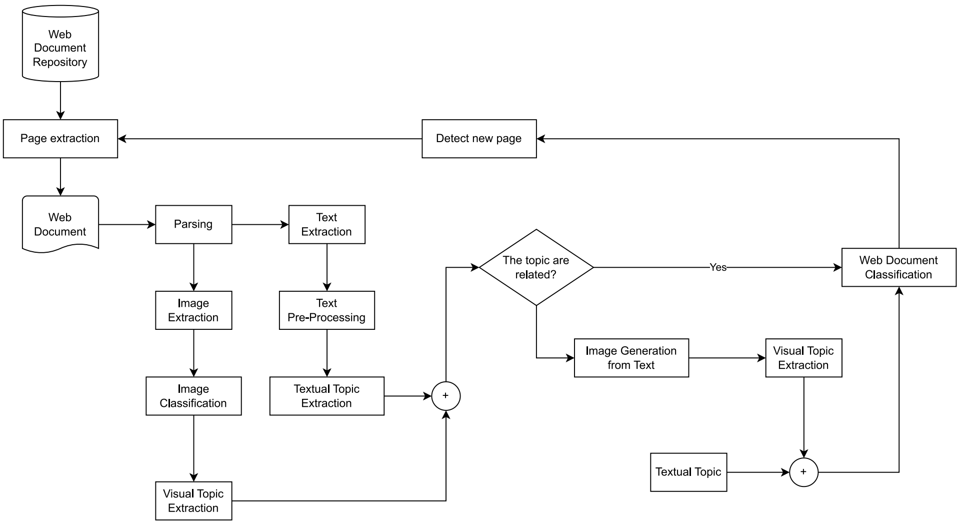

Figure 1: System architecture overview.

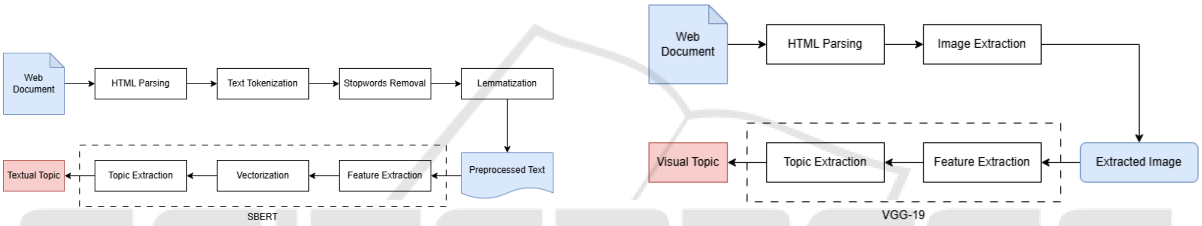

Figure 2: A detailed pipeline for textual (on the left) and visual (on the right) topic extraction and detection task.

similarity between document images and our multimedia knowledge base. Diverse descriptors, including local, global, and deep features, are evaluated (see Section 3). We employ the pre-trained VGG19 model on ImageNet, applied to the complete image, to identify and visualize relevant image regions associated with specific predictions, facilitating efficient transfer learning. This model's straightforward architecture promotes ease of interpretability and implementation.

If more than one image are detected from the web page, we need to select only the images that are more compliant with the textual topic.

3.3 Text-to-Image Generation

We incorporated a Latent Diffusion Model (LDM) into the crawling process to address the challenge of web pages lacking relevant multimedia content. This model generates high-quality images consistent with the textual description of the web page, even if no multimedia data is present initially. The Stable Diffusion (Rombach et al., 2022) serves as the latent model for translating text into images. Based on a latent diffusion approach, it is tailored for generating and manipulating images from textual prompts. The model utilizes the pre-trained CLIP ViT-L/14 text encoder (Radford et al., 2021), following Imagen's methodology (Saharia et al., 2022). The choice of Stable Diffusion over other generation techniques is justified by its superior performance, as demonstrated in Section 3.

To show that the visual topic extraction procedure works in the same way, using generated images, we chose to show the predictions made by the VGG19 model on the example Figure 3 and extract the probabilities for the first three classes predicted by the model. As shown in Table 1, the top 3 predicted classes fully reflect the content of the image, including the topics indicated within the textual prompt used to generate the image.

Table 1: Top 3 Prediction probabilities of generated image using the prompt ”a parrot that rides a bicycle”.

Top Prediction	Probability
macaw	**0.919**
bicycle-built-for-two	0.027
lorikeet	0.017

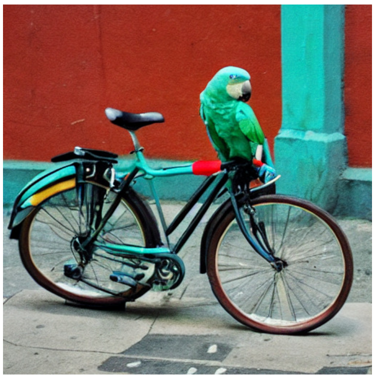

Figure 3: Image generated using the prompt ”a parrot that rides a bicycle”.

3.4 Combined Topic Extraction

This paper aims to improve classification performance by combining two classifiers: one for text-based topic detection and another for visual-based topic detection. Existing research suggests that different classifiers may provide complementary information models based on the specific patterns requiring classification (Mohandes et al., 2018; Clinchant et al., 2011).

The textual and visual classifications are combined by normalizing scores from different topic detection methods and scaling them to fit within the $[0,1]$ interval. The fusion of textual and visual classifications adopts various schemes, and in line with (Rinaldi, 2014), this study opts for the SUM operator and the Ordered Weighted Averaging (OWA) operators proposed by (Yager and Kacprzyk, 2012). These operators provide a systematic way to aggregate the results, allowing for a more robust and comprehensive integration of information from both textual and visual modalities. The SUM function stands out as one of the widely adopted techniques for linear combinations of classifiers, offering various versions such as weighted sum and average. Its prevalence in ensemble methods is attributed to its superior noise tolerance, contributing to enhanced overall performance compared to other elementary functions (Kittler et al., 1998). The versatility of the SUM function makes it particularly effective in capturing the collective decision-making power of diverse classifiers, making it a popular choice in ensemble learning approaches.

Formally an OWA operator of size n is a function $F : R_n \rightarrow R$ with a collection of associated weights $W = [w_1,...,w_n]$ whose elements are in the unit range such that $\sum_{i=1}^{n} w_i = 1$. The function is defined as:

$$F(a_1,...,a_2) = \sum_{j=1}^{n} w_j b_j \qquad (1)$$

where b_j represent the jth value of the $\vec{a}$ vector ordered.

4 EXPERIMENTAL RESULTS

This section details the experiments conducted to evaluate the performance of key components within the proposed framework, specifically focusing on textual, visual, and combined topic detection strategies. The system outlined in this study exhibits a high degree of generalization owing to the inherent versatility of the developed modules.

4.1 Dataset Description

This study employs a parsed and pre-elaborated multimedia dataset derived from DMOZ (Sood, 2016), a renowned and extensive multilingual web directory recognized for its popularity and open-content richness. DMOZ was selected as the experimental framework to establish a real-world scenario, offering a publicly accessible and widely recognized repository for result comparisons against baseline measures.

The Scraper module compiles URLs for download, using a “*text-only*” parameter to acquire only textual components of each document. The complete DMOZ repository subset is used based on web-scraping policies and link prevalence. It is crucial to map DMOZ categories on WordNet synset and definitions and if a corresponding mapping exists for all categories, because we have to obtain a compliant representation of the synsets and the categories that we have in the dataset. In Table 2, we provide a simple association between the category extracted and the respective WordNet synset tag and the description of it. We use English documents for our experiments, automatically handled with the help of a Python library porting of the Google algorithm for language detection and for the scraping of the pages (Danilak, 2017; Hajba, 2018). Out of a total corpus comprising 12,120 documents, 10,910 were allocated for creating topic modeling models. The remaining 1,210 documents were designated as test sets for the comprehensive evaluation of the entire system. A practical test set for the system requires documents that undergo textual and visual analyses. Specifically, efforts were directed towards selecting random documents from the web directory, ensuring they possess a substantial textual component and a minimum of three images. A DOM Parser algorithm was used to identify and retain a “valid” multimedia document aligned with the system’s objectives.

Table 2: Categories with WordNet Synsets and Definitions.

Category	Synset	Definition	Offset
Arts	art.n.01	The products of human creativity; works of art collectively	2,743,547
Business	commercial_enterprise .n.02	The activity of providing goods and services involving financial and commercial and industrial aspects	1,094,725
Computers	computer.n.01	A machine for performing calculations automatically	3,082,979
Games	game.n.01	A contest with rules to determine a winner	455,599
Health	health.n.01	A healthy state of well-being free from disease	14,447,908
News	news.n.01	Information about recent and important events	6,642,138
Science	science.n.01	A particular branch of scientific knowledge	5,999,797
Shopping	shopping.n.01	Searching for or buying goods or services	81,836
Society	society.n.01	An extended social group having a distinctive cultural and economic organization	7,966,140
Sports	sport.n.01	An active diversion requiring physical exertion and competition	523,513

4.2 Textual Topic Detection

The process of annotating the DMOZ category shown in Table 2 makes the classification using the selected algorithms easier for comparison of the resulted topic detection task because they return no information about the DMOZ category but only about the number of topics that represent the main topic of the text corpus of the web page.

To ensure a comprehensive evaluation of our system's performance, we compare two established reference algorithms widely employed in topic detection research: Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Latent Semantic Analysis (LSA) (Landauer et al., 1998) employs a vectorial representation approach to capture the essence of a document using the bag-of-words model. Meanwhile, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a text-mining model rooted in statistical methodologies. By benchmarking our system against these reference algorithms, we aim to provide a robust assessment of its efficacy in comparison to established techniques in the realm of topic detection.

The three selected strategy performance for textual topic detection, compared with SBERT, are presented in Figure 4 and summarized in Table 3. The SBERT model yielded the most favorable results regarding accuracy for textual topic detection, followed by LDA, while the LSA algorithm demonstrated comparatively lower performance. This discrepancy may be attributed to the outcomes being contingent on the feasibility of mapping DMOZ categories onto the concepts within the proposed ontology, as made in Table 2. Notably, in the case of LSA, the diminished accuracy appears linked to challenges associating specific topics generated by the model with their corresponding WordNet synsets. It can be postulated that SBERT exhibits superior generalization in concept recognition and is adept at mitigating noise inherent in specific datasets. Algorithm 1 shows the pseudo-code of the procedure adopted for detecting the textual topic from the full text of the web page using SBERT.

Table 3: Accuracy score detail for textual topic detection.

Algorithm	Accuracy	Num. Correct
LSA	0.1	117
LDA	0.34	407
SBERT	**0.53**	**620**

Algorithm 1: Detect Topics.

```
function PREPROCESS_TEXT(text)
    return Tokenize, lemmatize, and remove stop words from text
end function
function DETECT_TOPICS(text, num_clusters)
    processed_text ← PREPROCESS_TEXT(text)
    embeddings ← Generate SBERT embeddings for processed_text
    clusters ← Apply k-means clustering with num_clusters to embeddings
    for each cluster do
        synset_map ← Map words to WordNet synsets in processed_text
        top_synset ← Identify most common synset in synset_map
        Associate top_synset with the cluster as the representative topic
    end
    return (detected topics, associated synsets)
end function
```

4.3 Visual Topic Detection

The task of visual topic detection harnesses multimedia components, particularly images, to identify a document's primary topic, enhancing the overall performance of our framework. Our methodology takes a comprehensive approach to visual topic detection, employing three distinct configurations based on different feature extraction methods. This thorough exploration allows us to understand the strengths and limitations of each method.

PHOG (Bosch et al., 2007) is a global feature representation method that ensures comprehensive feature extraction and accuracy through dense grid computation and local contrast normalization with overlapping regions, making it suitable for precise visual analysis tasks.

SIFT (Lowe, 2004) is a robust local feature extraction technique that identifies key points of interest in gray-scale images, providing invariant descriptors for translation, rotation, and scaling, and is widely used in computer vision tasks.

VGG19 (Simonyan and Zisserman, 2014) is a deep convolutional neural network (CNN) designed for image classification tasks, with 19 layers, including convolutional and fully connected layers. It extracts visual content representation from intermediate layers, particularly the last max pooling layer, yielding deep features suitable for topic detection.

Evaluation of the three configurations based on utilized features (presented in Figure 4 and summarized in Table 4) reveals VGG19 features exhibit the highest accuracy, followed by PHOG and SIFT. These results align with expectations, as VGG19 and PHOG are promising candidates for precise feature matching, albeit with VGG19's computational time tradeoff due to its high dimensionality.

PHOG's superior accuracy over SIFT is attributed to its global nature. However, it can also be interpreted as a local feature due to its processing of images at various scales and resolutions. Selection of the optimal feature depends on a comprehensive analysis of the combined strategy and specific application requirements.

Table 4: Accuracy score detail for visual topic detection.

Algorithm	Accuracy	Num. Correct
SIFT	0.29	471
PHOG	0.39	348
VGG19	**0.82**	**1331**

4.4 Text-to-Image Generation

In our proposed strategy, an additional step involves generating multimedia data, prompting us to identify the most suitable model for this task. We evaluate three different models: StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018), and Stable Diffusion (Rombach et al., 2022). StackGAN employs a hierarchical Generative Adversarial Network for multistage image synthesis, progressively generating high-resolution images. AttnGAN incorporates attention mechanisms to enhance fine-grained image generation by focusing selectively on relevant regions. Stable Diffusion utilizes stable diffusion processes, controlling the gradual evolution of generated images to achieve high-quality synthesis. To assess fidelity in generated images from textual descriptions, we consider several metrics addressing different aspects of image fidelity. The Fréchet Inception Distance (FID) quantifies dissimilarity between the distributions of real and generated images, offering a comprehensive assessment. Other metrics, such as R-precision, Semantic Object Accuracy (SOA), and CLIP score, evaluate image-text matching. R-precision measures visual-semantic similarity by ranking retrieval results based on image and text features. SOA assesses object detection in images, providing class (SOA-C) and image (SOA-I) averages to gauge model alignment with textual descriptions. As highlighted in (Hinz et al., 2020), not all metrics are equally suited for evaluating generative models. Metrics like FID, SOA, and CLIP score better align with human visual assessment than IS and R-precision. Thus, we focus on the FID score, SOA metric, and CLIP score as crucial metrics for evaluating the generative model's performance. In Table 5, we have reported some metrics comparisons obtained during the evaluation process. The results show that Stable Diffusion is the model that best generates visually correct images that represent the textual prompt used for generation purposes.

4.5 Combined Topic Detection

The objective of the combined topic detection is to allocate a singular score based on weights assigned to individual textual and visual topic detection classifiers. The integration employs SUM and Ordered Weighted Averaging (OWA) operators. Various weight combinations have been systematically tested for each proposed scheme, excluding the OWA schema that employs OWA operators, providing a fuzzy logic approach. This comprehensive evaluation aims to identify optimal weight configurations contributing to an effective and nuanced integration of textual and visual topic detection outputs.

We decided to define four types of combination between the textual and visual strategy: the *A* combination pertains to the tests detailed earlier, wherein visual topic detection demonstrated superior performance. The *B* combination represents an average of computed scores, while the *C* combination assigns

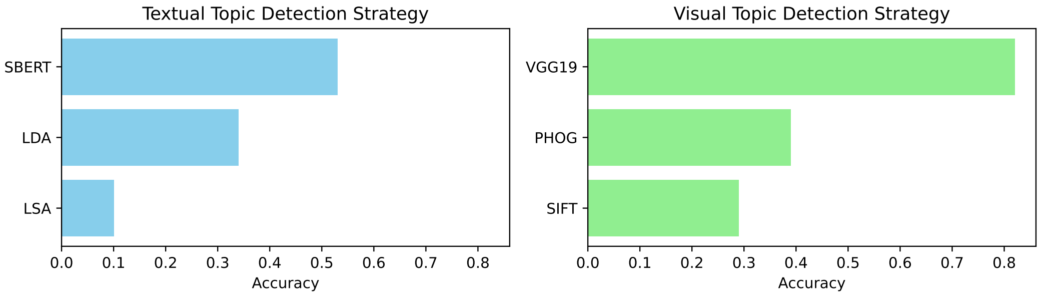

Figure 4: Accuracy comparison between the selected methods of textual topic detection (on the left) and visual topic detection (on the right).

Table 5: Metrics value comparison between GAN generative models and Stable Diffusion.

Model	**FID**	**CLIP**	**SOA-C**	**SOA-I**	**R-precision**
StackGAN (Zhang et al., 2017)	12.50	23.18	25.88	39.01	68.40
AttnGAN (Xu et al., 2018)	9.14	28.39	31.70	47.78	83.79
Stable Diffusion (Rombach et al., 2022)	**7.31**	**32.03**	**33.27**	**51.81**	**92.53**

greater importance to textual topic detection. Additionally, a combination was tested using OWA operators, denoted as the D scheme, with a weight vector $\vec{w} = (0.65, 0.35)$. This diverse set of combinations aims to thoroughly explore the interplay between textual and visual topic detection and identify optimal configurations for comprehensive evaluation. The testing procedure encompasses 36 combinations, employing schemes outlined in Table 6. In Figure 5, all results in classification accuracy among the described combinations are presented. Combinations incorporating visual topic detection and a fuzzy logic approach (combination D) are the most effective, particularly with deep feature-based schemas achieving high accuracy. However, textual topic detection schemes exhibit lower or similar accuracy. Combinations with equal weights for both classifiers can sometimes yield suboptimal classification accuracy.

Table 6: Weight configuration mapping for combined topic detection.

Combination	**Text topic weight**	**Image topic weight**
A	0.4	0.6
B	0.5	0.5
C	0.6	0.4
D	0.65	0.35

Visual classifiers with high accuracy significantly contribute to topic detection due to their enhanced discriminatory capabilities, particularly when images better represent specific concepts. However, terms with multiple meanings introduce uncertainty in the task. Despite their high computational demands, the SBERT model approach and VGG19 network-based visual topic detection yield the best results. The categorization process is an offline task prioritizing accuracy within the specific domain of interest.

Table 7 only summarizes the most significant experiments. The decision to choose the strategy for extracting topics from text and images and the method of combination used for classification was based on minimizing noise introduced by the image generation step, ensuring its inclusion as a variable in the evaluation step.

5 CONCLUSIONS AND FUTURE WORKS

In this work, we have proposed an innovative way to classify web documents using a generative approach and combined topic detection algorithm. We have unveiled promising directions for advancing the capabilities of an ontology-driven focused crawler. Exploring its extension to diverse conceptual domains, accompanied by a comparative assessment across various ontologies and knowledge bases, offers valuable insights into its adaptability and performance in capturing domain-specific information. Additionally, the investigation into alternative deep learning architectures, encompassing attention mechanisms, generative models, and self-supervised learning, holds the potential for augmenting the efficacy of image classification and retrieval tasks.

By evaluating different strategies and weighting schemes for merging textual and visual classifications, we have refined the model's performance. This

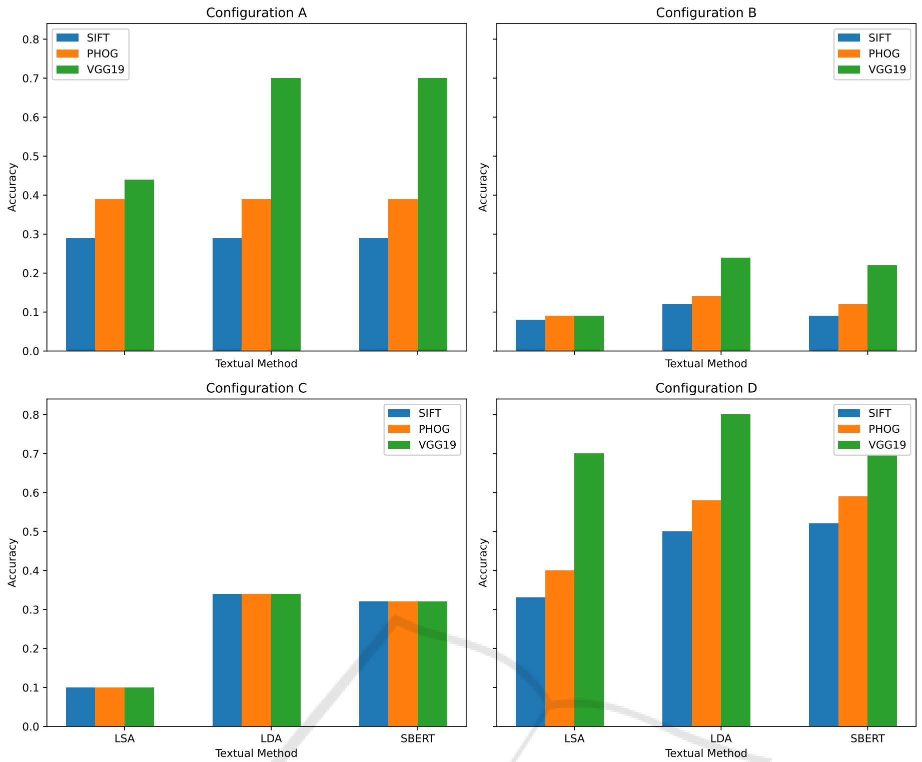

Figure 5: Accuracy performance among all the combinations selected, with all the visual and textual topic detection strategies descripted.

Table 7: Most relevant test details for combined topic detection.

Combination	Accuracy	Num. Correct	Combination	Accuracy	Num. Correct	Combination	Accuracy	Num. Correct
LSA-SIFT-A	0.29	348	LDA-SIFT-A	0.29	348	SBERT-SIFT-A	0.29	348
LSA-SIFT-B	0.08	100	LDA-SIFT-B	0.12	147	SBERT-SIFT-B	0.09	108
LSA-SIFT-C	0.10	117	LDA-SIFT-C	0.34	408	SBERT-SIFT-C	0.32	389
LSA-SIFT-D	0.33	402	LDA-SIFT-D	0.50	609	SBERT-SIFT-D	0.52	629
LSA-PHOG-A	0.39	471	LDA-PHOG-A	0.39	471	SBERT-PHOG-A	0.39	471
LSA-PHOG-B	0.09	108	LDA-PHOG-B	0.14	171	SBERT-PHOG-B	0.12	149
LSA-PHOG-C	0.10	117	LDA-PHOG-C	0.34	408	SBERT-PHOG-C	0.32	389
LSA-PHOG-D	0.40	480	LDA-PHOG-D	0.58	708	SBERT-PHOG-D	0.59	711
LSA-VGG19-A	0.44	539	LDA-VGG19-A	**0.70**	**846**	SBERT-VGG19-A	**0.70**	**846**
LSA-VGG19-B	0.09	111	LDA-VGG19-B	0.24	288	SBERT-VGG19-B	0.22	270
LSA-VGG19-C	0.10	117	LDA-VGG19-C	0.34	408	SBERT-VGG19-C	0.32	389
LSA-VGG19-D	**0.70**	**852**	LDA-VGG19-D	**0.80**	**966**	SBERT-VGG19-D	**0.80**	**965**

process has also enhanced our understanding of the robustness and sensitivity of these strategies under various conditions. Moreover, the expansion of the focused crawler framework to include multimedia content such as audio, video, and 3D models suggests a comprehensive approach to web page analysis. This expansion has the potential to significantly enhance our understanding of multimedia-rich online content, thereby improving the overall web document classification process.

ACKNOWLEDGEMENTS

This research has been partially supported by the Spoke 9 - Digital Society & Smart Cities - Centro Nazionale di Ricerca in High-Performance-Computing, Big Data and Quantum Computing, funded by European Union - NextGenerationEU - CUP: E63C22000980007. We also acknowledge financial support from the PNRR MUR project PE0000013-FAIR.

REFERENCES

Ahmed, Z. and Singh, H. (2019). Text extraction and clustering for multimedia: A review on techniques and challenges. In *2019 International Conference on Digitization (ICD)*, pages 38–43. IEEE.

Benfenati, D., Montanaro, M., Rinaldi, A. M., Russo, C., and Tommasino, C. (2023). Using focused crawlers with obfuscation techniques in the audio retrieval domain. In *International Conference on Management of Digital*, pages 3–17. Springer.

Bergman, M. K. (2001). White paper: the deep web: surfacing hidden value. *Journal of electronic publishing*, 7(1).

Bhatt, D., Vyas, D. A., and Pandya, S. (2015). Focused web crawler. *algorithms*, 5:18.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408.

Chakrabarti, S., Van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer networks*, 31(11-16):1623–1640.

Cheng, H., Liu, S., Sun, W., and Sun, Q. (2023). A neural topic modeling study integrating sbert and data augmentation. *Applied Sciences*, 13(7).

Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, pages 1–8.

Danilak, M. (2017). Langdetect 1.0. 7. *Python Package Index*.

Farag, M. M., Lee, S., and Fox, E. A. (2018). Focused crawler for events. *International Journal on Digital Libraries*, 19:3–19.

Fatima, N., Faheem, M., and Dar, M. Z. N. (2023). Optimized focused crawling for web page classification. In *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)*, pages 1–6.

Fernàndez-Cañellas, D., Marco Rimmek, J., Espadaler, J., Garolera, B., Barja, A., Codina, M., Sastre, M., Giro-i Nieto, X., Riveiro, J. C., and Bou-Balust, E. (2020). Enhancing online knowledge graph population with semantic knowledge. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19*, pages 183–200. Springer.

Fu, T., Abbasi, A., and Chen, H. (2010). A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology*, 61(6):1213–1231.

Hajba, G. L. (2018). Website scraping with python. *Berkeley: Apress*.

Hassan, T., Cruz, C., and Bertaux, A. (2017). Ontology-based approach for unsupervised and adaptive focused crawling. In *Proceedings of The International Workshop on Semantic Big Data*, pages 1–6.

Hinz, T., Heinrich, S., and Wermter, S. (2020). Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1552–1565.

K, N. T., S, C., G, B., Dharani, C., and Karishma, M. S. (2023). Comparative analysis of various web crawler algorithms.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.

Kumar, N. and Aggarwal, D. (2023). Learning-based focused web crawler. *IETE Journal of Research*, 69(4):2037–2045.

Kunder, M. d. (2018). The size of the world wide web (the internet). *Pobrano z: http://www. world widewebsize. com/(19.01. 2017)*.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Liu, J., Li, X., Zhang, Q., and Zhong, G. (2022). A novel focused crawler combining web space evolution and domain ontology. *Knowledge-based systems*, 243:108495.

Lowe, G. (2004). Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2.

Mary, J. D. P. N. R., Balasubramanian, S., and Raj, R. S. P. (2022). An enhanced focused web crawler for biomedical topics using attention enhanced siamese long short term memory networks. *Brazilian Archives of Biology and Technology*, 64:e21210163.

Mohandes, M., Deriche, M., and Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639.

Pant, G. and Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems (TOIS)*, 23(4):430–462.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Rinaldi, A. M. (2014). Using multimedia ontologies for automatic image annotation and classification. In *2014 IEEE International Congress on Big Data*, pages 242–249. IEEE.

Rinaldi, A. M. and Russo, C. (2021). Using a multimedia semantic graph for web document visualization and summarization. *Multimedia Tools and Applications*, 80(3):3885–3925.

Rinaldi, A. M., Russo, C., and Tommasino, C. (2021a). A semantic approach for document classification using deep neural networks and multimedia knowledge graph. *Expert Systems with Applications*, 169:114320.

Rinaldi, A. M., Russo, C., and Tommasino, C. (2021b). Visual query posing in multimedia web document retrieval. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 415–420. IEEE.

Rinaldi, A. M., Russo, C., and Tommasino, C. (2021c). Web document categorization using knowledge graph and semantic textual topic detection. In *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part III 21*, pages 40–51. Springer.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Russo, C., Madani, K., and Rinaldi, A. M. (2020). An unsupervised approach for knowledge construction applied to personal robots. *IEEE Transactions on Cognitive and Developmental Systems*, 13(1):6–15.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Shrivastava, G. K., Pateriya, R. K., and Kaushik, P. (2023). An efficient focused crawler using lstm-cnn based deep learning. *International Journal of System Assurance Engineering and Management*, 14(1):391–407.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sood, G. (2016). Parsed DMOZ data.

Tchakounte, F., Ngnintedem, J. C. T., Damakoa, I., Ahmadou, F., and Fotso, F. A. K. (2022). Crawlshing: A focused crawler for fetching phishing contents based on graph isomorphism. *Journal of King Saud University-Computer and Information Sciences*, 34(10):8888–8898.

Wu, H. and Hou, D. (2023). A focused event crawler with temporal intent. *Applied Sciences*, 13(7).

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Yager, R. R. and Kacprzyk, J. (2012). *The ordered weighted averaging operators: theory and applications*. Springer Science & Business Media.

Yan, W. and Pan, L. (2018). Designing focused crawler based on improved genetic algorithm. In *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, pages 319–323. IEEE.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.