Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 8.1 (2026) 1–29
HTTPS://DOI.ORG/10.5117/CCR2026.1.1.HAGA

Topic Classification of News Articles from URLs Alone

Nick Hagar
Department of Communication Studies, Northwestern University, USA

Abstract
This paper presents a novel approach to classifying news articles by topic using only their URLs, addressing growing challenges in accessing article text due to paywalls and scraping restrictions. By fine-tuning a DistilBERT transformer model on URL data alone, I demonstrate topic classification performance that matches or exceeds traditional approaches requiring article text. Across three benchmark datasets spanning multiple languages and over 660,000 articles from more than 11,000 news domains, this URL-based topic classifier achieved superior F1 scores compared to both conventional machine learning methods and existing URL-based techniques. While this method requires more computational resources than simpler topic classification approaches, it dramatically reduces data collection requirements, offering researchers a practical alternative when text access is limited. These findings suggest that news article URLs contain richer semantic information than previously recognized, opening new possibilities for large-scale news content analysis in increasingly restrictive digital environments.

Keywords: URL classification, news classification, natural language processing, machine learning, data access

Introduction

Classifying news articles is a common research task, enabling segmentation by topic, news type, domain relevance, or writing style (Bakshy et al., 2015; de León & Trilling, 2021; de León et al., 2023a; Flaxman et al., 2016). While machine learning has greatly expanded classification capacity, it presents notable challenges around computational resources and data access.

Compute difficulties center around the demand for computational resources that machine learning models introduce. Whereas simple statistical approaches, heuristic rules, and hand labeling have negligible resource demands, training and running a machine learning model can be beyond the capabilities of typical consumer-grade computer hardware. The need for specialized equipment adds a barrier to entry that may prevent researchers

from being able to apply state-of-the-art methods to their topic classification problems.

But even when using a sufficiently optimized model, or sufficiently powerful hardware, access to *data* is a necessary and challenging prerequisite for classifying news articles at scale. Amassing a large dataset of news articles presents a range of technical and policy obstacles. At a technical level, researchers must develop a web scraping approach that accurately extracts the information of interest (e.g., headline, article text, publication date) from an HTML document, across multiple types of documents, potentially from multiple websites with their own structure and document layout. The need to identify, download, and process articles across hundreds to thousands of web pages is cumbersome, even for researchers with significant technical expertise (Freelon, 2018).

Given the challenges of data access, less resource-intensive topic classification methods are desirable. Building on prior work that has leveraged URLs' semantic information (de León & Vermeer, 2023; Flaxman et al., 2016; Guess, 2021), this paper proposes classifying news articles using only their URLs. Specifically, by fine-tuning a DistilBERT model on URL data, I demonstrate performance that is not only comparable to methods using full article text but can even outperform models trained on headlines or article excerpts. This data-efficient approach offers a practical solution for news topic classification amid increasing digital access constraints.

Background

News article classification underpins a wide range of common analyses in communication research. For researchers, the ability to sort news coverage (e.g., by topic, or by whether or not articles represent political coverage) enables them to make comparisons across groups, filter down to a salient subset of articles for analysis, or quantify the representation of a particular kind of coverage in a a larger corpus. But while machine learning approaches to classification can be applied at scale to large corpora, they generally require text from each article to make reliable predictions. This text requirement is cumbersome for researchers, both from the technical perspective of large-scale web scraping, and the perspective of gaining access to news articles in an increasingly restrictive data environment. To address these challenges, I draw from research in machine learning techniques that classify web pages solely based on their URLs, and explore a potential application to news articles.

Classifying news articles

Machine learning classifiers have permeated quantitative study of news coverage. Because these algorithms allow researchers to annotate or filter text at scale, they offer a way to process data without requiring human review of every record in a sample. They are also extremely versatile, capable of classifying text along many dimensions.

In one common use case, researchers use machine learning to classify articles into a set of predetermined topics (R. Singh et al., 2020). Those topics then provide the basis for further quantitative analysis. For example, de León et al. (2023a) train a classifier to label news articles across six topics, to gauge the relative shift in engagement across coverage areas around elections. Similarly, Chuang and Larochelle (2014) apply topic modeling to the news coverage database Media Cloud, in an effort to analyze coverage areas by news outlet and over time. And Kaiser et al. (2019) apply a topic model to the news coverage produced by far-right outlets in the U.S., to gauge the relative presence or absence of coverage areas across this subset of the media. In these cases, an algorithm allows the researchers to aggregate individual articles into broader coverage areas, either as a way to explore the contents of that coverage or to monitor changes in theoretically interesting types of coverage over time.

For many empirical studies, only a subset of news coverage is relevant to the research question. In these cases, binary classification often provides an important filtering technique to identify a salient sample within a larger news corpus. For instance, researchers might use binary classifiers to limit a corpus to only hard news about politics (Budak et al., 2016; Flaxman et al., 2016). Similarly, de León and Vermeer (2023) employed binary classification to identify political news content across multiple languages and contexts, and Bakshy et al. (2015) utilized binary classification in their study of ideological exposure on Facebook, to identify hard news content. These studies illustrate how binary classification serves as a crucial preliminary step in large-scale news analysis, allowing researchers to efficiently identify relevant subsets of content for more detailed examination.

Finally, algorithmic classification can be useful for identifying more abstract concepts in news coverage. For example, researchers have leveraged a range of algorithmic classification techniques to operationalize news values in articles (Burggraaff & Trilling, 2020; Hagar et al., 2021; Trilling et al., 2017) and to explore automated techniques for measuring news agendas (Korenčić et al., 2015). And while sentiment analysis often attempts to place text on a continuous measure from negative to positive (e.g., Hutto & Gilbert,

2014), some researchers also operationalize discrete emotional labels via supervised classifiers (de León & Trilling, 2021).

Together, these approaches highlight the ubiquity of algorithmic classification as an analysis tool in the study of news coverage. Machine learning classifiers are a widely-accepted tool for quantitative communication researchers, often leveraged to label or filter news corpora at a scale that would be infeasible for human review.

Studies that leverage classification use a range of algorithmic approaches. Many researchers take a supervised learning approach, using model architectures such as support vector machines (Bakshy et al., 2015; de León & Trilling, 2021), random forests (Hagar et al., 2021), and transformers (De Clercq et al., 2020). These approaches are appropriate for cases where the topics of interest are known ahead of time. In more exploratory cases, where the goal of the analysis is to identify the types of coverage represented in a corpus, researchers rely on unsupervised topic modeling approaches like LDA (Chuang & Larochelle, 2014; Korenčić et al., 2015) and STM (Kaiser et al., 2019). Each of these techniques generally require some form of text preprocessing (e.g., counting the frequency of each word) to turn documents into numerical vectors.

These algorithms vary in computational complexity and predictive performance, but typically rely on text from news articles as input data. Many researchers use the full article text as features for classifiers (De Clercq et al., 2020; Trilling et al., 2017), while others find headlines provide a more useful signal (Hagar et al., 2021; Kuiken et al., 2017). In all cases, large-scale algorithmic classification requires some form of news text to determine group membership.

This requirement presents a challenge for researchers. While much news coverage can be collected via web scraping, an increasing number of restrictions limit the kinds of text that are programmatically accessible. For news websites with paywalls, web scrapers often cannot access the full text of articles. And as news publishers have updated their scraping policies in response to the large-scale collection of training data for large language models, web scraping for research that was formerly permissible may no longer be so (Welsh, 2024). The articles that researchers are able to classify are therefore limited by what is accessible—through, for example, archival databases like the Internet Archive—which may not be a representative sample (Thelwall & Vaughan, 2004). In the absence of reliable access to primary-source records of news coverage, researchers require alternative methodological approaches that still enable the kinds of large-scale sorting

and filtering that text classification provides.

URL-based web page classification

In some domains, researchers have demonstrated the efficacy of methods that classify web pages without accessing the contents of the pages themselves. These approaches use the same supervised machine learning algorithms, text preprocessing pipelines, and evaluation metrics as the classifiers explored above. But rather than applying them to the text of a web page, these approaches apply them to the contents of the URL.

The URL (uniform resource locator) of a web page is its address on the Internet, which allows computers to retrieve its contents. URLs often contain human-readable descriptive information about their contents. For example, the URL huffpost.com/static/about-us likely has information about the Huffington Post. This characteristic makes URLs potentially informative for downstream natural language tasks.

URL-based classification is prevalent in domains where accessing web pages directly is either impossible or inadvisable. For example, when identifying malicious websites, detecting harmful pages without visiting them is crucial (Ma et al., 2009; Vanhoenshoven et al., 2016). URLs may also be the only available textual information for pages containing solely multimedia content (Baykan et al., 2011; J. Zhang et al., 2006). In addition, URL-based classifiers can provide faster inference by avoiding the download and processing overhead of full web pages (e.g., in personalized web readers— Baykan et al., 2011; Hernández et al., 2014; Kan & Thi, 2005).

Of course, URL-based classification creates trade-offs. Researchers often leverage such approaches because of their efficiency and practical application, not because of their predictive performance (Hernández et al., 2014). URLs tend to be extremely short, and so the information they convey is often limited. In some cases, it is not possible to glean any information about the topic of a web page solely from its URL (e.g., the domain name of a small business without context— Baykan et al., 2009). And URLs can be complex and nonstandard, creating challenges from a language processing perspective: Words can be abbreviated, capitalization is irregular or not present, and there is no punctuation to indicate structure (Baykan et al., 2011; Kan, 2004).

Still, URL-based classification approaches have proven effective on real-world datasets. Numerous researchers report strong predictive performance on datasets ranging from news to malicious web pages to academic websites (Kan, 2004; Ma et al., 2009; Rajalakshmi & Aravindan, 2018; N. Singh et al.,

2012). These approaches use a combination of natural language processing techniques and supervised machine learning—many create input features from URLs by splitting them into n-grams of various sizes (e.g., Baykan et al., 2011) and by applying token weighting approaches (Hernández et al., 2012; Rajalakshmi & Aravindan, 2018). They then train a range of classification models—including Naive Bayes classifiers, support vector machines, and random forests—on these features.

While combining token-based featurization with supervised learning algorithms has been the dominant approach, it may not be optimal for URL classification. The field's focus on tokenization strategies and reliance on support vector machines underscores a fundamental challenge: sparsity in limited text data. This limitation suggests the need for more sophisticated approaches to text representation and modeling.

Transformers offer a promising solution to this challenge. This architecture has demonstrated exceptional performance across diverse sequence modeling tasks, from basic classification to sophisticated generation in large language models (Brown et al., 2020; Vaswani et al., 2017). While transformers still require text tokenization, their use of dense vector representations directly addresses the sparsity problem that has constrained traditional approaches (Vaswani et al., 2017). This characteristic makes them particularly well-suited for URL classification, where the limited text available demands more efficient representation.

One effective transformer for natural language processing tasks is BERT. BERT is a transformer model, which generates a context-aware representation of text sequences (Devlin et al., 2019). It has proven effective as a classification model, including for news articles, with performance that surpasses other supervised machine learning methods (De Clercq et al., 2020). BERT has also shown promise as a URL-based classifier, in the context of detecting malicious web pages (Chang et al., 2021).

Combining these threads of research, then, BERT-based text classification has proven effective for labeling news articles. A range of classification techniques, including BERT, have proven effective when trained solely on URLs, across a range of domains. And there have even been some efforts to incorporate news article URLs into classification tasks for communication research. Multiple studies have taken a "distant labeling" approach to classifying news articles, which partially relies on information from the URL (de León & Vermeer, 2023; de León et al., 2023b). In these studies, researchers map the section contained in a URL to a topic of interest (e.g., a URL with /politics/ would be a political story, while one with /sports/ would not), as

a first step in a classification pipeline. This approach augments the classification pipeline, producing more accurate inference than a supervised machine learning model on its own.

Given the efficacy of URLs as predictive signals, in news and other domains, these data provide an opportunity to develop less data intensive training sets for news article classifications. By leveraging a state-of-the-art transformer architecture, I argue that it is possible to train a BERT-based model to assign topics to articles based solely on their URLs, without needing access to the articles themselves.

H1: A BERT-based URL topic classification model will achieve equal or higher F1 scores compared to traditional supervised machine learning approaches that use news article text.

Data

To evaluate each topic classification approach, I used three publicly-available news topic classification benchmark datasets. These datasets were selected by searching for data on popular repository archives Kaggle, HuggingFace, and the Harvard Dataverse, for projects that 1) consisted of news articles and associated topics, and 2) contained a URL for each article. I evaluated each method on multiple datasets to get a sense of their effectiveness across a range of topic labeling approaches, news domains, and languages. Below is a description of each dataset. Table 1 contains summary statistics for each.

Table 1: Descriptive statistics for the filtered subset of the three benchmarking datasets used in this analysis.

Dataset	Rows	Unique domains	Topics	Gini coefficient (URLs per domain)	Topic label entropy
HuffPost	144,167	1	14	N/A	3.47
News Aggregator	421,890	10,925	4	0.82	1.89
RecognaSumm	98,023	14	7	0.61	2.66

News Category (HuffPost) Dataset: This is a computational linguistics benchmarking dataset, which consists of about 210,000 news articles from the Huffington Post (Misra, 2022). The dataset includes 14 topics, with labels originating from the section in which each article was published (i.e., applied by a human editor or journalist) (Misra, 2022). Because this dataset focuses on a single website and categorizes articles into relatively

specific groups, it presents an opportunity to evaluate the ability of topic classification approaches to make granular distinctions.

News Aggregator Dataset: This is a widely-used machine learning benchmark dataset, maintained as part of a dataset repository at the University of California, Irvine (Gasparetti, 2017). It contains approximately 400,000 news articles, from multiple websites, across 4 topics (business, science and technology, health, and entertainment). The process used to generate the topic labels in this dataset is not specified in its documentation. These data complement the Huffington Post corpus well: While the topics here are much broader, they are applied across many websites, which is valuable for assessing the generalizability of topic classification methods.

RecognaSumm: This dataset captures news articles in Portuguese, from Brazilian news outlets (Paiola et al., 2024). While the purpose of the dataset is to train machine learning models for text summarization, each record also has a topic label across 7 topics. These topic labels come from the metadata of each news website (Paiola et al., 2024). Given the bias present toward English text in many transformer pre-training datasets (Gouvert et al., 2025), this dataset is useful for evaluating topic classification approaches for their effectiveness in Portuguese.

To offer a comprehensive view of the many ways that article text and URL text might be represented, I trained models on a wide variety of input features, described in Section 3.1.

For the HuffPost and RecognaSumm datasets, there are a large number of unique topics with ambiguous overlap (e.g., "WORLDPOST", "THE WORLDPOST", and "WORLD NEWS"). To account for this, I removed any records with labels that applied to less than 2% of total articles. The resulting number of records and topics is shown in Table 1.

In addition, it is worth noting that the volume of URLs across domains in the RecognaSumm and News Aggregator datasets is not equal, as indicated by their Gini coefficients. This skew raises the possibility that certain domain names may appear disproportionately in the training dataset and may have imbalanced labels (e.g., in.reuters.com accounts for 0.9% of records in the News Aggregator dataset, and 79.5% of its articles have the "business" topic label), potentially biasing results if the classification approaches learn from the domain–topic association. The text input variants described below—in particular, the URL text conditions that remove domain names—are an effort to measure the amount of performance that might be attributed to these associations.

I trained a separate instance of each model type for each dataset and

input feature, for a total of 139 unique classifiers. I trained each model on 80% of records. I used 10% as a validation set, and held out the remaining 10% as a test set, from which I report performance metrics. Each model used the same splits.

Input Feature Conditions

To comprehensively evaluate the performance of article text versus URL text for topic classification, I trained models on seven distinct input feature conditions: **Article Text Conditions (baseline comparisons):**

1. **Title only:** Article headline text
2. **Title and subtitle:** Combined headline and subtitle text
3. **Snippet/description:** Excerpt from the article body (not available in the News Aggregator dataset)

URL Text Conditions (primary focus):

4. **URL path (cleaned):** URL path with query parameters removed and special characters replaced with spaces, in line with established practices in URL classification research (Baykan et al., 2011)
5. **URL path (raw):** URL path in its original form, excluding the domain
6. **URL (raw):** Complete URL including domain and path

Blended Condition:

7. **URL + Title + Subtitle:** Combination of full URL with article title and subtitle

The "URL path (raw)" condition (#5) represents the primary contribution of this work, demonstrating that topic classification can be achieved using only the path portion of URLs without any article text, domain information, or text cleaning.

Methods

Proposed URL-based Classification Approach

In this work, I propose a news article topic classification pipeline that relies on a BERT-based transformer model, trained on a labeled dataset of URLs[1].

[1]All code and data available at https://osf.io/qfrzh/overview

This approach builds on empirical work demonstrating both the effectiveness of BERT as a classifier for news text (De Clercq et al., 2020; R. Singh et al., 2020) and the value of the text of news article URLs as input features for machine learning models (de León et al., 2023b).

While I evaluate multiple URL encoding strategies to understand their relative merits, my primary contribution is demonstrating that news article classification can be achieved using only the raw URL path—the portion of the URL after the domain name—as input to a topic classification model. This "URL path (raw)" condition most directly tests whether the semantic information in URL paths alone can enable accurate topic classification.

My analysis pipeline used DistilBERT, a compressed version of BERT that achieves nearly identical performance while using only 60% of the original parameters (Sanh et al., 2020). This compression makes the model more efficient, reducing computational resources needed for both fine-tuning and inference. I applied the standard DistilBERT to English-language datasets and DistilBERT-multilingual for other languages (Sanh et al., 2020).

To fine-tune these models for topic classification, I used a training dataset of labeled URLs from the benchmarking corpora described above. I used HuggingFace's Transformers library for fine tuning (Wolf et al., 2020).

Comparison Models

I compared my topic classification algorithm against a range of other approaches. First, where possible, I implemented the distant labeling algorithm from de León et al. (2023b). While both approaches classify news articles using URL information, they make different trade-offs. My approach uses a more complex model on a smaller dataset, whereas de León et al. (2023b) employ a computationally efficient model supplemented by larger datasets and manual annotations. This comparison helps quantify how different approaches balance model complexity, dataset size, and prediction accuracy.

In addition, I implement a standard machine learning classification approach that is often leveraged for processing text data: embedding text, then training an XGBoost classifier on those embeddings (e.g., Jahnavi et al., 2024). This approach, and related approaches with other kinds of tree-based ensemble models, encodes the text information from an article into a dense vector. The tree-based models often used for classification are robust to collinearity and capable of capturing non-linear relationships, making them well-suited to use dense vectors as input features (Breiman, 2001). To embed article text, I used the gte-Qwen2-1.5B-instruct embedding model (Li et al.,

2023). At the time of writing, this is a relatively efficient model that performs well on the MTEB text embedding benchmark (Muennighoff et al., 2023). It is also capable of embedding text across multiple languages.

Finally, to provide a comprehensive set of baselines, I also implemented four additional traditional machine learning models. In contrast to the embedding-based XGBoost approach, these classifiers were trained on features generated using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a standard technique for converting text into numerical features. The models include:

- Logistic Regression: A widely-used linear model that is a common baseline for text classification tasks.
- Support Vector Machine (SVM): A linear SVM classifier, which is often highly effective for high-dimensional, sparse data like TF-IDF vectors.
- Gradient Boosting: A tree-based ensemble technique that builds models sequentially, with each new model correcting the errors of the previous ones.
- Tree Ensemble: An ensemble model using a random forest classifier, which constructs a multitude of decision trees during training.

These models represent a range of established and effective techniques for text classification and serve as robust comparisons for the URL path-based DistilBERT approach. The parameters used to train each model are included in the replication package.

Section-to-Topic Mapping Procedure

To reproduce the distant-labeling approach of de León et al. (2023b), I created a site section to topic lookup table for the News Aggregator and RecognaSumm datasets described above (the HuffPost dataset URLs lack sections). For both datasets, the site section corresponded to the first element of the URL path. I extracted these elements, ranked them by frequency, then retained the top 250. This provided enough data to cover the long-tail of possible site sections while keeping the mapping task tractable.

Then, I manually assigned each section token to one of the dataset's high-level topics by assessing the closest fit, based on the section name and a sample of its URLs. The resulting mapping tables covered 16.8% of all URLs in the News Aggregator dataset and 61.1% in RecognaSumm. This discrepancy stems from a difference in section specificity—while common RecognaSumm sections mapped cleanly onto high-level topics (e.g., politica,

economia), the most common in News Aggregator were more generic (e.g., news, article, story, content). Unmapped URLs received no pseudo-label and were therefore excluded from distant-labeler training, mirroring the procedure in de León et al. (2023b). The mapping tables and generation code are included in the replication package.

Evaluation Framework

To comprehensively assess the proposed URL-based classification approach, I leveraged a multi-faceted evaluation framework examining predictive performance, computational efficiency, and data requirements.

Performance Metrics

I evaluated all models using F1 scores to account for potential class imbalances in the datasets. For each model, I computed both overall F1 scores and per-topic F1 scores to identify performance variations across different topics.

To investigate potential biases from unequal domain representation in the training data, I also conducted domain-level analyses. For each domain in the test set, I calculated model F1 scores and examined their relationship to domain frequency using Pearson correlation. To see if the model was memorizing domain-topic associations, I computed the accuracy lift over a naive majority-topic baseline (i.e., the accuracy rate that would be achieved by always predicting each domain's modal topic) for each domain and correlated this with the domain's topic entropy.

Computational Efficiency Evaluation

Given the practical compute constraints researchers face (de León et al., 2023b), I measured the computational efficiency of each approach through throughput analysis. I defined throughput as predictions per second on the test set, providing a standardized measure of inference speed across models. All models were evaluated on identical hardware: a server equipped with 8 Intel Xeon 6338 CPU cores and an NVIDIA A100 GPU. For each model type, I used the input features that achieved the highest F1 score on the test set, ensuring I compared the models at their optimal configurations.

Training Data Requirements Analysis

To understand the relationship between training data volume and classification performance, I conducted systematic ablation studies. I created stratified samples of 1,000 and 3,000 training records from each dataset, maintaining the original distribution of topic labels. These sample sizes were chosen to represent scenarios with limited annotation resources—depending on the dataset, they comprised between 1% and 4% of the full training data.

For each sample size, I trained new DistilBERT classifiers using identical hyperparameters to the full models, focusing on raw URL path as the input feature. This allowed me to identify the minimum data requirements for achieving competitive performance relative to both the full DistilBERT models and the traditional machine learning baselines.

URL Structure Analysis

To investigate whether the effectiveness of URL-based classification depends on specific URL formats or generalizes across diverse structures, I conducted a systematic analysis of URL composition and its relationship to model performance.

URL Feature Extraction

I programmatically analyzed all URLs in the News Aggregator dataset ($n = 421,890$), extracting and parsing each URL path to identify key structural features. For each URL, I computed:

- **Path length:** Total character count in the URL path
- **Segment count:** Number of path components separated by forward slashes
- **Numeric density:** Percentage of numeric characters in the path
- **Date indicators:** Presence of recognizable date patterns (e.g., YYYY/MM/DD, YYYY-MM-DD)
- **Semantic keywords:** Presence of topic-indicative terms (e.g., "politics," "business," "health") or structural markers (e.g., "article," "story," "news")

URL Categorization Schema

Based on these extracted features, I developed a classification schema to categorize each URL into one of six mutually exclusive structural types:

- **Semantic Structured:** URLs containing topic keywords within a clear hierarchical structure (e.g., /politics/article/biden-announcement-2024)
- **Semantic Unstructured:** URLs with topic keywords but lacking clear hierarchical organization (e.g., /biden-politics-announcement.html)
- **Date-Based:** URLs where the primary organizational structure uses date formatting (e.g., /2024/03/15/)
- **ID-Based:** URLs composed primarily ($> 50\%$) of numeric or alphanumeric identifiers without semantic keywords (e.g., /article/837g4qw9f2)
- **Simple:** Short URLs containing two or fewer path segments
- **Mixed:** URLs not fitting cleanly into the above categories, typically combining multiple organizational schemes

Performance Analysis by Structure Type

To assess how URL structure impacts classification accuracy, I merged the structural categorizations with the domain-level performance results from the raw URL path DistilBERT classifier. This allowed me to calculate aggregate F1 scores for each URL type and examine correlations between specific structural characteristics (e.g., path length, keyword presence) and model performance. I used Pearson correlation to quantify the relationship between continuous structural features and F1 scores.

Results

Table 2 presents F1 scores across all seven input conditions described in Section 3.1. The URL path (raw) condition represents the core contribution—classification using only URL paths—while other conditions provide comparative baselines (article text), alternative URL-based approaches, and upper bounds (blended features). These metrics demonstrate two key high-level findings. First, the fine-tuned DistilBERT classifier achieves the highest topic classification performance for every dataset. This holds true for every input feature, emphasizing the general utility of DistilBERT as a classifier, regardless of available text. Second, URLs offer strong topic classification

performance—in all cases, the raw URL path-based DistilBERT classifier outperforms every non-DistilBERT model, confirming H1. In addition, the URL-based DistilBERT models trail their highest-performing counterparts by an average F1 of only 0.04 across datasets. And those highest-performing models benefit from the inclusion of URL text, as well. When comparing DistilBERT models using only titles and subtitles, versus those that also leverage URLs, the latter gain an average F1 of 0.04 across datasets.

The results from the URL-based models also help to mitigate concerns around domain-level bias. While it was possible that models would glean associations between domain names themselves and their most common topic labels, in practice, there is virtually no change in performance when domain names are added. This indicates that the most predictive text lies in the URL paths, gleaned from section and article information.

Looking at the other models, there is not a consistent second-best approach. However, the logistic regression, XGBoost, and SVM approaches all offer comparable performance across datasets. The distant labeling algorithm consistently has the lowest performance. In particular, the distant labeling pipeline struggles with the RecognaSumm dataset, with a maximum F1 score of just 0.25. Digging into the section-to-label mapping more closely, this performance drop appears to stem from an ambiguous relationship between site sections and topics. For example, "Mundo" is one of the most common topics in the dataset, containing 11,000 URLs. There is a corresponding "mundo" site section, attributed to this label in the manual mapping step. However, only 15% of URLs with this section are actually labeled with the corresponding category—rather, many articles fall under topics including "Internacional" (71%), "Saúde" (9%), and "Ciência e Tecnologia" (0.1%). Because there is not a clear one-to-one mapping in this case, the manual labeling process works against accurate topic classification—as evidenced by the much higher performance on this dataset from other traditional machine learning methods without a manual labeling step.

It's worth noting that the section attribution step, which is key to the distant labeling algorithm, could not be applied for the HuffPost dataset. The URLs in this dataset do not have section labels, so the reported performance solely relies on the count vectorization and Naive Bayes classification steps of the algorithm. In addition, the article snippets in the HuffPost dataset were too long for the context window of the embedding model used with the XGBoost classifier, so no result is reported for that condition.

Table 2: F1 scores for all combinations of models and input features across datasets. The URL Path (raw)* column represents the primary contribution of this work (URL-only classification). Bold values indicate the highest-performing model for each feature–dataset combination.

	Article Text Conditions			**URL Conditions**			**Blended**
Model	**Snippet/ Desc.**	**Title**	**Title+ Subtitle**	**URL Path (cleaned)**	**URL Path (raw)***	**URL (raw)**	**URL Title+ Subtitle**
				HuffPost Dataset			
Distant labeling	0.37	0.54	0.57	0.53	0.53	0.52	0.61
DistilBERT	**0.59**	**0.74**	**0.80**	**0.79**	**0.80**	**0.80**	**0.87**
Gradient Boosting	0.47	0.63	0.67	0.56	0.56	0.56	0.72
Logistic Regression	0.47	0.64	0.68	0.58	0.58	0.59	0.72
SVM	0.46	0.63	0.67	0.57	0.57	0.57	0.71
Tree Ensemble	0.30	0.54	0.54	0.52	0.52	0.49	0.57
XGBoost	–	0.65	0.71	0.64	0.61	0.55	0.72
				News Aggregator Dataset			
Distant labeling	–	0.67	0.67	0.62	0.62	0.62	0.70
DistilBERT	–	**0.96**	**0.96**	**0.93**	**0.93**	**0.93**	**0.97**
Gradient Boosting	–	0.88	0.88	0.84	0.84	0.84	0.89
Logistic Regression	–	0.92	0.92	0.86	0.86	0.87	0.92
SVM	–	0.91	0.91	0.85	0.85	0.86	0.92
Tree Ensemble	–	0.90	0.90	0.84	0.84	0.83	0.88
XGBoost	–	0.90	0.90	0.85	0.85	0.86	0.91
				RecognaSumm Dataset			
Distant labeling	0.24	0.24	0.25	0.23	0.15	0.15	0.15
DistilBERT	**0.94**	**0.92**	**0.94**	**0.96**	**0.96**	**0.96**	**0.98**
Gradient Boosting	0.76	0.73	0.76	0.93	0.93	0.94	0.95
Logistic Regression	0.77	0.74	0.77	0.91	0.91	0.92	0.92
SVM	0.77	0.74	0.77	0.91	0.91	0.92	0.92
Tree Ensemble	0.71	0.67	0.71	0.93	0.93	0.93	0.94
XGBoost	0.87	0.82	0.86	0.93	0.95	0.94	0.88

*Primary contribution: Classification using only URL paths without article text or domain information

Throughput

Figure 1 shows the throughput of each approach (higher is better). Several of the traditional machine learning models—logistic regression, support vector machines, and tree ensembles—offer the highest throughput, processing between 20,000 and 50,000 records per second. These models therefore offer a clear option for cases where researchers cannot sacrifice efficiency for predictive performance, or do not have the computational resources required to run a more demanding model.

DistilBERT, on the other hand, sacrifices throughput for the strongest performance, processing 200-600 records per second. Because it generates compute-intensive embeddings before running inference, the XGBoost approach has the lowest throughput.

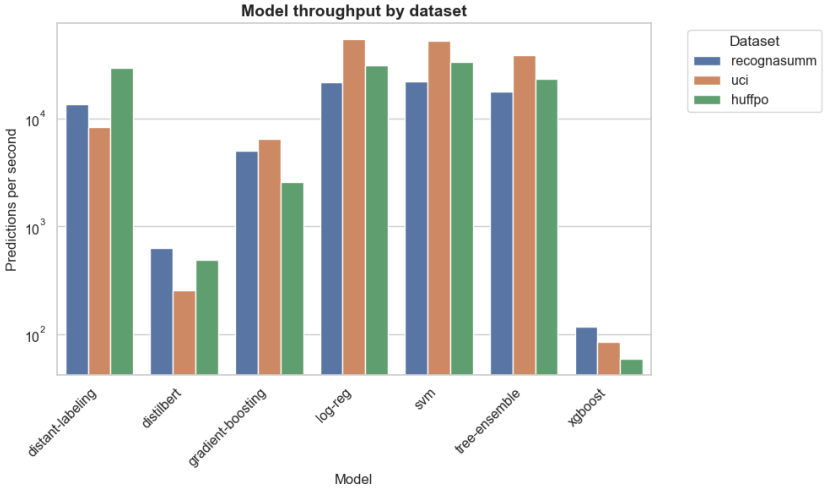

Figure 1: Prediction throughput (predictions per second) for each model. The logistic regression, SVM, and tree ensemble approaches offer the highest throughput, with comparable predictive performance, while DistilBERT trades off throughput for leading performance.

Training Data Volume Ablation

Figure 2 shows how training data volume affects DistilBERT classification performance. The impact varied substantially across datasets. The HuffPost dataset showed the steepest performance degradation: F1 scores dropped from 0.80 with full training data to 0.60 with 3,000 samples and 0.18 with 1,000 samples. This sensitivity likely stems from the dataset's 14 topic classes

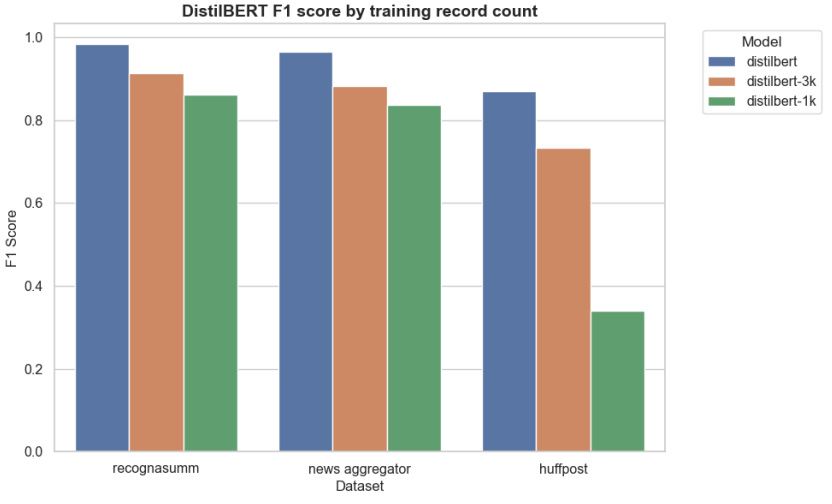

Figure 2: Effect of training data ablation on URL-based DistilBERT classifier performance. Performance takes the largest hit for the dataset with the most topic labels.

requiring more examples per category.

The News Aggregator and RecognaSumm datasets proved more robust to reduced training data. With 3,000 training samples, F1 scores decreased by an average of only 0.09 points. Even with just 1,000 samples, these models maintained F1 scores within 0.20 points of their full-data counterparts. Notably, the 3,000-sample models achieved performance comparable to traditional machine learning approaches trained on complete datasets, suggesting DistilBERT can be viable even with limited labeled data.

Topic and Domain Performance Variance

Table A1 in Appendix A reveals substantial heterogeneity in topic-level classification performance. The HuffPost dataset exhibited the widest performance range, with F1 scores spanning from 0.65 (Black Voices) to 0.89 (Politics) for the raw URL path model. Political content and lifestyle topics (Style & Beauty: 0.87, Wellness: 0.87) showed the strongest URL-based discriminability. Identity-focused categories (Black Voices: 0.65, Queer Voices: 0.74) proved more challenging, suggesting these topics may appear across more diverse URL structures.

In contrast, the News Aggregator and RecognaSumm datasets showed remarkably consistent performance across topics, with F1 ranges of only

0.06 points each.

Domain-level analysis revealed no meaningful relationship between domain representation and performance (Pearson $r = 0.02$), indicating that aggregate results were not inflated by a few heavily represented sources. The accuracy lift analysis showed that 35.1% of domains gained at least 5 percentage points over the naive baseline, while only 8.9% experienced a comparable drop. Domains with higher topic diversity showed greater model improvement ($r = 0.63$ between topic entropy and accuracy lift), confirming the classifier leverages URL content rather than memorizing domain-topic associations.

URL Structure

The URL structure analysis revealed both diversity and consistent patterns across the News Aggregator dataset. Over half (53.6%) of URLs contained explicit semantic keywords, with "semantic unstructured" (27.6%) and "semantic structured" (25.9%) being the most common types. Pure ID-based URLs were rare, comprising only 1.1% of the dataset.

Classification performance varied meaningfully by URL type. Date-based URLs achieved the highest F1 score (0.93), followed by semantic structured (0.87) and semantic unstructured (0.84) URLs. The model could not accurately classify ID-based URLs ($F1 = 0.41$). URL path length showed a moderate positive correlation with performance ($r = 0.32$), suggesting longer, more descriptive paths provide richer classification signals. To verify that the strong performance of date-based URLs stems from semantic content rather than spurious temporal patterns, I conducted an ablation study removing all date components from URL paths. Across all three datasets, logistic regression models showed minimal performance change when dates were removed, with an average F1 decrease of only 0.008 (see Appendix B). These findings demonstrate that while URL structures vary considerably, the vast majority contain meaningful semantic information that the classifier successfully exploits. Performance degradation occurs primarily in the small fraction of opaque, ID-based URLs, validating the general applicability of URL-based classification for news articles when semantically meaningful text is present.

Discussion

This research demonstrates the potential for URL-based topic classification of news articles. Across three commonly-used benchmark datasets, a BERT-

based classifier trained solely on the paths of article URLs more accurately assigns news articles to topics than other machine learning approaches that leverage vectorization of article text. In addition, URLs offer a performance gain for BERT-based topic classification compared to article text, whether in isolation or in combination with that text. This suggests that the extremely compact text contained in the path of a news article URL may offer a stronger signal than the text of the article itself for topic classification tasks.

These findings highlight a valuable approach for news article topic classification: URLs alone, when analyzed with sophisticated models, can provide highly accurate classification while requiring minimal data. This is particularly relevant given emerging challenges in accessing news content. While URL-based classification has long been necessary in domains where accessing web pages is impossible (Baykan et al., 2011; Ma et al., 2009), news researchers now face similar constraints. Recent increases in robots.txt restrictions, driven by publishers' responses to AI scraping, combined with the growing adoption of paywalls, are making it increasingly difficult to collect large-scale news text data (Longpre et al., 2024). My results demonstrate that effective classification remains possible in these cases, requiring only article URLs rather than article text access.

A key consideration for this supervised learning approach is the need for labeled training data, which researchers can procure in several ways. First, the data ablation study showed that competitive performance is possible with just a few thousand training examples, making manual annotation on URLs or articles a feasible option. Second, because the model's performance is not dependent on any particular domain within a diverse dataset, a model trained on a wide-ranging corpus may generalize to new data if the URL structures are comparable. Finally, for cases where manual annotation is impractical, researchers could explore emerging techniques that use large language models (LLMs) for automated labeling (Kuzman & Ljubešić, 2025; Y. Zhang et al., 2025).

This research also illustrates an important methodological trade-off in computational communication research: data efficiency versus compute efficiency. Many existing approaches prioritize compute efficiency. For instance, de León et al. (2023b) deliberately designed their distant labeling approach to be computationally lightweight, making it accessible to researchers with limited computational resources. However, there are increasingly common scenarios where data—rather than compute—is the limiting factor. Researchers may face restricted access to news content, limited availability of digitized historical records, or technical challenges in

parsing certain types of information at scale. In these cases, traditional approaches requiring large training datasets may be impractical. My findings demonstrate an alternative: Using more sophisticated model architectures, especially those pretrained on large text corpora, can achieve strong performance even with limited domain-specific data. This approach enables researchers to maximize the value of scarce data, making previously infeasible large-scale annotation and analysis tasks possible.

These findings point to an important direction for computational communication research: exploring the balance between data and compute efficiency. While leading general-purpose models like the GPT series demand both massive datasets and intensive computing resources (Brown et al., 2020), more focused applications may offer practical alternatives. Sophisticated models applied to specific research tasks can achieve strong performance with smaller datasets and less manual annotation. This data-efficient approach complements existing work on compute efficiency, with each strategy addressing different barriers to research. Just as compute-efficient methods make computational analysis more accessible to researchers with limited processing power, data-efficient approaches can enable research that would otherwise be blocked by the costs of web scraping, data licensing, or human annotation. Together, these approaches can democratize computational methods across a broader range of research contexts.

There are several important limitations to this work. First, the methodology described here requires meaningful information encoded in the URLs used for training. As demonstrated in the URL structure analysis in Section 5.4, longer URLs with more semantic information generally yield better results. This method does not perform well in cases where, as on some news websites, the URL path is a random sequence of letters and numbers. Researchers should therefore consider whether the URLs in their data encode meaningful information to map to topic labels before using this method. In addition, while the benchmark datasets used to evaluate the approaches in this study contain a wide range of news websites and multiple languages, they are not comprehensive. The results here may not transfer to any particular news website, label set, or language. Finally, the model used for URL-based topic classification here is based on BERT, which was trained on a corpus of books and Wikipedia data (Devlin et al., 2019). While fine-tuning pretrained models can yield strong performance on specific classification tasks, the model's initial pretraining fundamentally shapes its behavior. When research priorities include model interpretability or precise control over model behavior, training custom models from scratch may be

more appropriate.

Looking ahead, this work opens several promising avenues for future research in computational communication studies. The demonstrated effectiveness of URL-based topic classification suggests opportunities to analyze news content in contexts where article text access is limited or impossible. Researchers might explore applying these methods to historical news archives where only URLs or headlines survived, or to contemporary paywalled content that cannot be scraped at scale. Future work should also investigate how URL-based topic classification performs across different languages, cultures, and digital media contexts, particularly as URL conventions evolve. More broadly, this research highlights the need to critically examine assumptions about what constitutes "rich" versus "sparse" data for machine learning applications in communication research. While the field has often prioritized collecting comprehensive text corpora, there may be many cases where strategic use of limited but information-dense features, combined with sophisticated modeling approaches, could yield comparable or superior results with fewer resources. Understanding these tradeoffs between data requirements, computational costs, and model performance will be crucial as computational methods continue to expand within communication research.

References

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Baykan, E., Henzinger, M., Marian, L., & Weber, I. (2009). Purely URL-based topic classification. *Proceedings of the 18th international conference on World wide web*, 1109–1110. https://doi.org/10.1145/1526709.1526880

Baykan, E., Henzinger, M., Marian, L., & Weber, I. (2011). A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification. *ACM Transactions on the Web*, *5*(3), 1–29. https://doi.org/10.1145/1993053.1993057

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly*, *80*(S1), 250–271.

Burggraaff, C., & Trilling, D. (2020). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, *21*(1), 112–129. https://doi.org/10.1177/1464884917716699

Chang, W., Du, F., & Wang, Y. (2021). Research on Malicious URL Detection Technology Based on BERT Model. *2021 IEEE 9th International Conference on Information, Communication and Networks (ICICN)*, 340–345. https://doi.org/10.1109/ICICN52636.2021.9673860

Chuang, J., & Larochelle, D. (2014). Large-Scale Topical Analysis of Multiple Online News Sources with Media Cloud. *NewsKDD: Data Science for News Publishing*.

De Clercq, O., de Bruyne, L., & Hoste, V. (2020). News Topic Classification as a First Step Towards Diverse News Recommendation. *Computational Linguistics in the Netherlands Journal*, *10*, 37–55.

de León, E., & Trilling, D. (2021). A Sadness Bias in Political News Sharing? The Role of Discrete Emotions in the Engagement and Dissemination of Political News on Facebook. *Social Media + Society*, *7*(4), 205630512110597. https://doi.org/10.1177/20563051211059710

de León, E., & Vermeer, S. (2023). The News Sharing Gap: Divergence in Online Political News Publication and Dissemination Patterns across Elections and Countries. *Digital Journalism*, *11*(2), 343–362. https://doi.org/10.1080/21670811.2022.2099920

de León, E., Vermeer, S., & Trilling, D. (2023a). Electoral news sharing: A study of changes in news coverage and Facebook sharing behaviour during the 2018 Mexican elections. *Information, Communication & Society*, *26*(6), 1193–1209. https://doi.org/10.1080/1369118X.2021.1994629

de León, E., Vermeer, S., & Trilling, D. (2023b). URLs Can Facilitate Machine Learning Classification of News Stories Across Languages and Contexts. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.4.DELE

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. Retrieved March 10, 2020, from http://arxiv.org/abs/1810.04805

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gasparetti, F. (2017). News Aggregator [Published: UCI Machine Learning Repository] DOI: https://doi.org/10.24432/C5F61C.

Gouvert, O., Hunter, J., Louradour, J., Cerisara, C., Dufraisse, E., Sy, Y., Rivière, L., Lorré, J.-P., & community, O.-F. (2025). The Lucie-7B LLM and the Lucie Training

Dataset: Open resources for multilingual language generation. https://doi.org/10.48550/arXiv.2503.12294

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science, 65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

Hagar, N., Diakopoulos, N., & DeWilde, B. (2021). Anticipating Attention: On the Predictability of News Headline Tests. *Digital Journalism*. https://doi.org/10.1080/21670811.2021.1984266

Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R. (2012). A statistical approach to URL-based web page clustering. *Proceedings of the 21st International Conference on World Wide Web*, 525–526. https://doi.org/10.1145/2187980.2188109

Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R. (2014). CALA: An unsupervised URL-based web page classification system. *Knowledge-Based Systems, 57*, 168–180. https://doi.org/10.1016/j.knosys.2013.12.019

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Jahnavi, M., Chandana, K., Nair, P. C., & Dheeraj, K. (2024). Classification of News Category Using Contextual Features. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), 1*, 1–7. https://doi.org/10.1109/ICKECS61492.2024.10616859

Kaiser, J., Rauchfleisch, A., & Bourassa, N. (2019). Connecting the (Far-)Right Dots: A Topic Modeling and Hyperlink Analysis of (Far-)Right Media Coverage during the US Elections 2016. *Digital Journalism, 8*(3), 422–441. https://doi.org/10.1080/21670811.2019.1682629

Kan, M.-Y. (2004). Web Page Categorization without the Web Page. *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, 262–263.

Kan, M.-Y., & Thi, H. O. N. (2005). Fast webpage classification using URL features. *Proceedings of the 14th ACM international conference on Information and knowledge management*, 325–326. https://doi.org/10.1145/1099554.1099649

Korenčić, D., Ristov, S., & Šnajder, J. (2015). Getting the Agenda Right: Measuring Media Agenda using Topic Models. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, 61–66.

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective Headlines of Newspaper Articles in a Digital Environment. *Digital Journalism, 5*(10), 1300–1314. https://doi.org/10.1080/21670811.2017.1279978

Kuzman, T., & Ljubešić, N. (2025). LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access, 13*, 35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards General Text Embeddings with Multi-stage Contrastive Learning. Retrieved October 28, 2024, from http://arxiv.org/abs/2308.03281

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, *6*(8), 975–987. https://doi.org/10.1038/s42256-024-00878-8

Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1245–1254. https://doi.org/10.1145/1557019.1557153

Misra, R. (2022). News Category Dataset. Retrieved October 30, 2024, from http://arxiv.org/abs/2209.11429

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive Text Embedding Benchmark. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Afonso, L. C. S., & Papa, J. P. (2024). RecognaSumm: A Novel Brazilian Summarization Dataset. *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, *1*, 575–579.

Rajalakshmi, R., & Aravindan, C. (2018). A Naive Bayes approach for URL classification with supervised feature selection and rejection framework. *Computational Intelligence*, *34*(1), 363–396. https://doi.org/10.1111/coin.12158

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. Retrieved October 28, 2024, from http://arxiv.org/abs/1910.01108

Singh, N., Sandhawalia, H., Monet, N., Poirier, H., & Coursimault, J.-M. (2012). Large Scale URL-based Classification Using Online Incremental Learning. *2012 11th International Conference on Machine Learning and Applications*, 402–409. https://doi.org/10.1109/ICMLA.2012.199

Singh, R., Chun, S. A., & Atluri, V. (2020). Developing Machine Learning Models to Automate News Classification. *The 21st Annual International Conference on Digital Government Research*, 354–355. https://doi.org/10.1145/3396956.3397001

Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, *26*, 162–176.

Trilling, D., Tolochko, P., & Burscher, B. (2017). From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics. *Journalism & Mass Communication Quarterly*, *94*(1), 38–60. https://doi.org/10.1177/1077699016654682

Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., & Koppen, M. (2016). Detecting malicious URLs using machine learning techniques. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI.2016.7850079

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. Retrieved October 29, 2024, from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Welsh, B. (2024). Who blocks OpenAI, Google AI and Common Crawl? — News Homepages documentation. Retrieved October 30, 2024, from https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. v., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. Retrieved October 29, 2024, from http://arxiv.org/abs/1910.03771

Zhang, J., Qin, J., & Yan, Q. (2006). The Role of URLs in Objectionable Web Content Categorization. *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, 277–283. https://doi.org/10.1109/WI.2006.170

Zhang, Y., Wang, M., Li, Q., Tiwari, P., & Qin, J. (2025). Pushing The Limit of LLM Capacity for Text Classification. *Companion Proceedings of the ACM on Web Conference 2025*, 1524–1528. https://doi.org/10.1145/3701716.3715528

Appendix A Topic-Level F1 Scores

Table A1 presents detailed F1 scores for each topic across all model input feature combinations for the three benchmark datasets. These granular results complement the aggregate performance metrics reported in the main text, revealing how classification accuracy varies by both topic and input feature type. For the HuffPost dataset with its 14 fine-grained topics, performance variations are most pronounced, while the News Aggregator and RecognaSumm datasets show more consistent performance across their broader topic categories. The table further demonstrates that URL-based features achieve competitive or superior performance compared to article text features across topic categories.

Table A1: F1 scores for all combinations of topics and input features across datasets, with the highest F1 scores in bold.

Topic	Snippet & Desc.	Title	Title+ Subtitle	URL Path (cleaned)	URL Path (raw)	URL (raw)	URL Title+ Subtitle
			HuffPost Dataset				
Black Voices	0.40	0.58	0.63	0.66	0.65	0.64	**0.72**
Business	0.51	0.65	0.70	0.72	0.72	0.72	**0.82**
Comedy	0.31	0.60	0.65	0.71	0.72	0.71	**0.75**
Entertainment	0.59	0.82	0.85	0.84	0.85	0.85	**0.88**
Food & Drink	0.71	0.82	0.87	0.82	0.83	0.83	**0.91**
Healthy Living	0.32	0.43	0.63	0.74	0.74	0.75	**0.84**
Home & Living	0.64	0.84	0.87	0.84	0.83	0.84	**0.90**
Parenting	0.70	0.75	0.83	0.79	0.78	0.78	**0.90**
Politics	0.76	0.89	0.91	0.90	0.89	0.89	**0.93**
Queer Voices	0.54	0.77	0.81	0.73	0.74	0.75	**0.86**
Sports	0.60	0.82	0.87	0.81	0.82	0.82	**0.89**
Style & Beauty	0.76	0.86	0.90	0.87	0.87	0.88	**0.93**
Travel	0.73	0.87	0.90	0.80	0.80	0.80	**0.92**
Wellness	0.75	0.73	0.82	0.87	0.87	0.88	**0.93**
			News Aggregator Dataset				
b (Business)	–	0.95	0.95	0.91	0.91	0.92	**0.95**
e (Entertainment)	–	0.98	0.98	0.97	0.97	0.97	**0.99**
m (Health)	–	0.96	0.96	0.91	0.91	0.91	**0.96**
t (Science & Technology)	–	0.95	0.95	0.92	0.92	0.93	**0.96**
			RecognaSumm Dataset				
Brasil	0.90	0.87	0.92	0.94	0.94	0.95	**0.98**
Ciência e Tecnologia	0.92	0.89	0.92	0.96	0.96	0.96	**0.98**
Economia	0.92	0.89	0.92	0.96	0.95	0.95	**0.98**
Esporte	0.97	0.98	0.98	1.00	1.00	1.00	**1.00**
Internacional	0.96	0.95	0.96	0.97	0.97	0.97	**0.99**
Política	0.93	0.90	0.94	0.96	0.96	0.96	**0.98**
Saúde	0.95	0.93	0.95	0.96	0.95	0.96	**0.98**

Appendix B Date Ablation Analysis

To assess whether models were learning spurious associations between publication dates and topics, I conducted an ablation study removing all date components (e.g., YYYY/MM/DD patterns) from URL paths, then training and evaluating performance of logistic regression models on the same train and test splits. Table A2 compares model performance with and without dates across all three datasets. The minimal performance differences confirm that semantic content in URL paths, rather than temporal patterns, drives classification accuracy.

Table A2: Impact of date removal on logistic regression classifier performance using raw URL paths. Performance remains largely stable across datasets when dates are excluded, indicating minimal reliance on temporal patterns.

Dataset	**F1 (With Dates)**	**F1 (No Dates)**	**Δ F1**
HuffPost	0.584	0.576	-0.009
News Aggregator	0.861	0.861	$+0.000$
RecognaSumm	0.926	0.909	-0.016
Average	*0.790*	*0.782*	-0.008