RESEARCH-ARTICLE

Language-Agnostic Modeling of Source Reliability on Wikipedia

JACOPO D'IGNAZI, Pompeu Fabra University Barcelona, Barcelona, Barcelona, Spain

ANDREAS KALTENBRUNNER, Open University of Catalonia, Barcelona, Barcelona, Spain

YELENA MEJOVA, Institute for Scientific Interchange Foundation, Turin, TO, Italy

MICHELE TIZZANI, Institute for Scientific Interchange Foundation, Turin, TO, Italy

KYRIAKI KALIMERI, Institute for Scientific Interchange Foundation, Turin, TO, Italy

MARIANO G. BEIRÓ, University of San Andrés, Victoria, Provincia de Buenos Aires, Argentina

View all

Open Access Support provided by:

Institute for Scientific Interchange Foundation

Pompeu Fabra University Barcelona

Open University of Catalonia

University of San Andrés

Published: 19 February 2026
Online AM: 18 November 2025
Accepted: 26 October 2025
Revised: 18 July 2025
Received: 11 November 2024

Language-Agnostic Modeling of Source Reliability on Wikipedia

JACOPO D'IGNAZI, Universitat Pompeu Fabra, Barcelona, Spain and ISI Foundation, Torino, Spain
ANDREAS KALTENBRUNNER, Universitat Oberta de Catalunya, Barcelona, Spain, ISI Foundation, Torino, Italy, and Universitat Pompeu Fabra, Barcelona, Spain
YELENA MEJOVA, ISI Foundation, Torino, Italy
MICHELE TIZZANI, ISI Foundation, Torino, Italy
KYRIAKI KALIMERI, ISI Foundation, Torino, Italy
MARIANO G. BEIRÓ, Universidad de San Andrés, Buenos Aires, Argentina and CONICET, Buenos Aires, Argentina
PABLO ARAGÓN, Universitat Pompeu Fabra, Barcelona, Spain

Over the last few years, verifying the credibility of information sources has become a fundamental need to combat disinformation. Here, we present a language-agnostic model designed to assess the reliability of web domains as sources in references across multiple language editions of Wikipedia. Utilizing editing activity data, the model evaluates domain reliability within different articles of varying controversiality, such as Climate Change, COVID-19, History, Media, and Biology topics. Crafting features that express domain usage across articles, the model effectively predicts domain reliability, achieving an F1 Macro score of approximately 0.80 for English and other high-resource languages. For mid-resource languages, we achieve 0.65, while the performance of low-resource languages varies. In all cases, the time the domain remains present in the articles (which we dub as *permanence*) is one of the most predictive features. We highlight the challenge of maintaining consistent model performance across languages of varying resource levels and demonstrate that adapting models from higher-resource languages can improve performance. We believe these findings can assist Wikipedia editors in their ongoing efforts to verify citations and may offer useful insights for other user-generated content communities.

CCS Concepts: • **Information systems → Wikis**; • **Computing methodologies → Machine learning algorithms;**

This work is partially supported by a grant from the Credibility Coalition, a project of Meedan and Hacks/Hackers. Authors also acknowledge support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation), funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT). Earlier versions of this article benefited from valuable feedback from Diego Sáez Trumper, Francisco Navas, Aitolkyn Baigutanova, and attendees of Wiki Workshop 2024.
Authors' Contact Information: Jacopo D'Ignazi, Universitat Pompeu Fabra, Barcelona, Catalunya, Spain and ISI Foundation, Torino, Piemonte, Spain; e-mail: jacopo.dignazi@upf.edu; Andreas Kaltenbrunner (corresponding author), Universitat Oberta de Catalunya, Barcelona, Barcelona, Spain, ISI Foundation, Torino, Piemonte, Italy, and Universitat Pompeu Fabra, Barcelona, Barcelona, Spain; e-mail: kaltenbrunner@gmail.com; Yelena Mejova, ISI Foundation, Torino, Piemonte, Italy; e-mail: yelena.mejova@acm.org; Michele Tizzani, ISI Foundation, Torino, Piemonte, Italy; e-mail: mt.tizzani@isi.it; Kyriaki Kalimeri, ISI Foundation, Torino, Piemonte, Italy; e-mail: Kyriaki.Kalimeri@isi.it; Mariano G. Beiró, Universidad de San Andrés, Buenos Aires, Argentina and CONICET, Buenos Aires, Buenos Aires, Argentina; e-mail: mbeiro@fi.uba.ar; Pablo Aragón, Universitat Pompeu Fabra, Barcelona, Catalunya, Spain; e-mail: elaragon@gmail.com.

ACM 1559-1131/2026/02-ART2
https://doi.org/10.1145/3777444

Additional Key Words and Phrases: Wikipedia, knowledge equity, knowledge integrity, information reliability

ACM Reference Format:
Jacopo D'Ignazi, Andreas Kaltenbrunner, Yelena Mejova, Michele Tizzani, Kyriaki Kalimeri, Mariano G. Beiró, and Pablo Aragón. 2026. Language-Agnostic Modeling of Source Reliability on Wikipedia. *ACM Trans. Web* 20, 1, Article 2 (February 2026), 29 pages. https://doi.org/10.1145/3777444

1 Introduction

The proliferation of fake news in recent years is deteriorating the reliability and trustworthiness of the online information ecosystem [15, 45]. To mitigate their spread and impact in society, researchers in web and data mining are devoting a great deal of effort to generating resources to detect untrue content on social media platforms [3, 20, 41, 49–52].

In popular web platforms like Wikipedia,[1] disinformation is also a main threat to knowledge integrity [1]. Wikipedia's conception of knowledge as a service requires content to be verifiable to preserve its reliability and integrity.[2] Thus, the platform's policy on verifiability implies that readers should be able to check that all the information in Wikipedia articles comes from reliable sources.[3] While a classical patrolling technique in this platform is to detect and remove statements that violate basic core content policies [31], some Wikipedia editors have proven more effective in tackling disinformation by first identifying unreliable sources [12]. Recent research revealed the positive impact of the community-curated list of perennial sources in English Wikipedia [4]. However, that list is limited or even non-existent in other language editions [5]. A great deal of manual effort would be necessary to create such community-curated resources in all of the languages on Wikipedia.

To address this challenge, we propose a language-agnostic approach to reveal source reliability that leverages the implicit signals in the edit activity data from multiple Wikipedia language editions. We aim to support Wikipedia editors, including those in medium and small projects that often miss advanced tools, in identifying sources with patterns associated with low reliability, and in monitoring their prevalence across language editions. To meet this goal, we address the following research questions:

- **RQ1.** What are the most predictive language-agnostic features when modeling source reliability in Wikipedia, considering the Climate Change topic in English as a case study?
- **RQ2.** To what extent can model performance be related to the topic and size of a language edition of Wikipedia?
- **RQ3.** Is it possible to adapt these models across topics and/or languages?

Our methods to answer these questions are inspired by existing language-agnostic approaches to measure the controversiality of linked elements in a given Wikipedia article [8]. Results with datasets from different topics and language editions suggest a promising scenario. Given the several limitations outlined in the following sections, we pay attention to model performance not only for the largest editions but also in mid- and low-resource settings, where adaptation of the models may provide the most benefit. To our knowledge, this is the first work modeling source reliability on Wikipedia using only language-agnostic features.

2 Related Work

This work approaches the challenge of measuring source reliability on Wikipedia through the lens of controversiality metrics. This approach builds on the methodology of the Contropedia project [8],

[1]Wikipedia is among the top visited websites worldwide: https://www.semrush.com/trending-websites/global/all
[2]https://research.wikimedia.org/knowledge-integrity.html
[3]https://en.wikipedia.org/wiki/Wikipedia:Verifiability

aimed at identifying and visualizing controversial links between Wikipedia articles, i.e., links that point to other articles in the platform and whose collaborative edit history shows signs of dispute between editors before reaching consensus. This is not the only effort in this space. Systems like Edit-History Vis [21] have also included interactive visualizations to illustrate the path to consensus for a Wikipedia article. Cultural differences in the controversiality of specific topics have been shown to exist (e.g., [18, 38]). Unlike these previous approaches primarily focusing on wikilinks within article text, our work shifts attention to citations, aiming to explore the reliability of sources in Wikipedia.

Previous research has examined assessed the verifiability of Wikipedia content [7, 9, 24, 25] and the reliability of the sources [33, 53]. Baigutanova et al. [4] addressed both goals by analyzing reference quality on English Wikipedia over time using two key metrics: the proportion of sentences lacking necessary citations, as identified by machine learning models [10, 36], and the proportion of citations linking to non-authoritative sources. For the latter, non-authoritative sources were determined using the *perennial source list* [46], a curated list of sources whose reliability and use on Wikipedia are frequently discussed. The study found the share of references to non-authoritative sources has remained below 1%, with a notable decline following the introduction of the list in 2018.

Although the *perennial source list* proved effective in decreasing the presence of non-authoritative sources on English Wikipedia, later research revealed significant limitations when applying this concept to other language editions [5]. A few other language communities have developed their own versions of the *perennial source list*, but such cases are rare among the hundreds of Wikipedia editions, and these lists typically contain only limited data. Furthermore, sources flagged as unreliable in one language edition are often still used in others, highlighting the variability in how source reliability is perceived across cultural and linguistic contexts. These discrepancies pose a major challenge to the development of universally applicable models for assessing source reliability across Wikipedia language editions.

The multilingual nature of Wikipedia calls for specific modeling techniques [28]. Models such as multilingual BERT have been used to detect vandalism [44] or political bias [42], while other works have built rankings of relative quality in multiple language versions of the platform [32, 33]. To build a multilingual model, such as one recently proposed for entity insertion [17], one needs a large dataset of training data in all the languages of interest. For some other tasks, and due to the limitations found in languages with smaller Wikipedia datasets, *language-agnostic* approaches. These approaches have proven effective across a wide range of tasks, e.g., the assessment of article content quality [13], reliability [48], automatic entity-linking [19], orphan article detection [2], and topic modeling [35] and classification [26].

Language-agnostic models using features extracted from the network structure or user behavior have also been used in the more general context of detecting misinformation or low-quality information in social media. In this line, Shu et al. [40] constructed embeddings of users and news and used diffusion models to trace the paths through which fake news propagates, while Zhao et al. [58] extracted features from the user interaction network to help identify misinformation in online health communities. As shown later, user-behavior related language-agnostic signals are useful for modeling source reliability, and can be adapted across languages to improve performance in low-resource settings.

3 Data and Methods

3.1 Perennial Sources

We use the English Wikipedia perennial source list [46] as ground truth for modeling source reliability. As only a few other languages have perennial source lists, in this work we focus on the list from English Wikipedia, as it is the most complete one, and concerns the largest language Wikipedia. We discuss this lack of ground truth data in the Limitations section below. It was collected on 30 March, 2023. Although the list exists for 12 other languages, we only use the English version as it is by far the most extensive (see [5] for details on coverage). In English, it consists of a collection of web domains in five categories: blacklisted (569), deprecated (125), generally unreliable (178), no consensus (137), and generally reliable (184). To build a binary classifier, we use the generally reliable domains as positive labels, merge the domains blacklisted, deprecated, and generally unreliable into a single unreliable category (i.e., the negative label), and ignore the no consensus domains.

3.2 Language Groups

As we aim to test the proposed model on a variety of languages, we collect data on all 326 available language editions of Wikipedia. We then discard all languages that have fewer than two reliable and unreliable sources according to the English perennial source list. Among the remaining ones, we distinguish between low, middle, and high-resource languages by examining the number of "active users", defined as "registered users who have made at least one edit in the last 30 days."[4] We consider the top 5% of the languages having the most active users as high-resource, the subsequent 25% as mid-resource, and the remaining 70% as low-resource languages.

3.3 Article Collection

3.3.1 Topic Selection. In this study, we have focused on articles from the following five topic datasets: Climate Change, COVID-19, Biology, History, and Media. Climate Change and COVID-19 were selected as they are both controversial, but for different time durations. COVID-19 appeared at the end of 2019, while Climate Change has been a controversial topic for decades [30]. Both have a Wikiproject community of editors[5] who label articles as relevant to the topic with special guidelines. For comparison, we also selected three topics that are not as controversial, but which varied in terms of reference quality based on the "reference risk" metric. This metric was introduced by Baigutanova et al. [4] and estimates the likelihood of articles citing unreliable sources. Specifically, reference risk is calculated as the proportion of citations in an article that point to domains categorized as generally unreliable, deprecated, or blacklisted in the English Wikipedia perennial source list. Higher reference risk indicates a greater potential threat to verifiability and information quality.

This stratification allowed us to evaluate our model's performance across topics with varying levels of citation quality and editorial scrutiny.

3.3.2 Data Download. For the first two topics, we use the articles listed by the corresponding topic-based WikiProjects. For the latter three, we use the topics identified using the Wikipedia **Objective Revision Evaluation Service** (**ORES**) machine learning system [23]. We use this method rather than later approaches [26], since it is the one employed in Baigutanova et al. [4], which inspires our comparison among topics. To keep the sizes of datasets comparable and manageable, we sample the articles identified by ORES to be in Biology and History topics by 10% (each), and in the Media topic by 2%. These two approaches allow us to find articles in the English Wikipedia for each of the five topics, resulting in a total of 84 027 articles. To find the versions of these articles in other languages, we use the MediaWiki API, which results in another 335 809 of articles in all of the 326 languages. We collect the articles and their complete edit history using the MediaWiki API in between December 2023 and February 2024.

[4]https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed on 01-03-2024).
[5]https://en.wikipedia.org/wiki/Wikipedia:WikiProject

Table 1. Statistics of the Topic-Language Datasets used in the Study Grouped by High, Mid, and Low Resource Languages

	Topic	Langs	Articles	Revs	Revs/Ar	URLs	URLs/Ar	Domains	P. Domains
High	Climate Change	7	1, 378	372, 133	269	58, 671	49	15, 363	203
	COVID-19	7	898	233, 905	260	60, 168	89	8, 289	206
	Biology Sample	7	13, 607	586, 171	43	59, 763	6	11, 713	148
	History Sample	7	5, 710	624, 428	109	52, 702	12	12, 121	166
	Media Sample	7	5, 577	819, 541	146	108, 583	22	21, 290	259
Mid	Climate Change	36	395	28, 364	71	6, 224	19	2, 599	94
	COVID-19	37	272	18, 523	68	8, 984	39	1, 788	102
	Biology	37	4, 153	83, 599	20	8, 691	3	2, 022	61
	History	37	1, 680	83, 258	49	6, 274	7	2, 054	67
	Media	37	1, 018	46, 957	46	7, 932	10	2, 416	123
Low	Climate Change	56	115	3, 430	29	1, 079	12	570	35
	COVID-19	75	32	916	28	568	22	242	32
	Biology	41	593	21, 096	35	1, 159	4	460	16
	History	55	511	13, 501	26	952	6	406	22
	Media	70	159	2, 262	14	528	7	264	30

All numbers are averages (apart from the number of languages). Perennial Domains (P. Domains) refers to the number of domains for the perennial source list in English found in the different datasets. See Table A1 in Appendix B for the corresponding median values.

3.3.3 Extraction of Source Domains. Citations may appear in the body of the article text and as references at the bottom of the article, each of which has a standard template. Each revision that involves a citation contains the author of the revision, position within the document of the citation, text of the citation, and meta information, such as whether it is a DOI, URL, ISBN, ISSN, and so on. or a raw URL. We standardize these revisions to create a list of "source edits" containing the metadata about each insertion, deletion, or edit that involves a source. We resolve redirects, apply rules to clean and standardize the URLs (including DOI references redirecting to a resource), and merge revisions when the same user edits an article in succession (regardless of the time). Finally, we extract the domains from the URLs, obtaining 376 566 unique source domains.

3.3.4 Dataset Statistics. The summary statistics about the topic-language datasets (or simply "datasets" throughout this article), for the five topics grouped into low-, mid-, and high-resource languages can be seen in Table 1. The medians of these values can be found in Appendix B, while more in-depth statistics for each individual language are in Appendix C. Recall that we only use datasets that have passed the threshold of having at least 2 reliable and 2 unreliable perennial sources.

The smallest topic in terms of the number of articles is COVID-19, while the highest numbers of revisions per article are found around Climate Change and COVID-19 in high-resource languages. On the contrary, biology has the largest number of articles, but these are revised on average the fewest number of times in high- and mid-resource languages. High-resource languages have much better coverage of the English perennial sources, with an average of 150 to 260 sources per topic. These values decrease to an average of 15 to 35 sources for low-resource languages, making the training and evaluation more challenging.

3.4 Feature Definition

Unlike previous studies measuring Wikipedia edit quality that did not consider certain types of edits and reverts [43], we propose to focus on the entire edit history concerning each domain. Using edit metadata, we define 52 features capturing the popularity of the domain, the "permanence" of the domain across edits, and the number and type of users involved in edits adding or removing that domain. As the features are computed for each domain within each dataset separately (recall, by "dataset" we mean a collection of articles on a certain topic, in a certain language) and the datasets are of varying sizes and editorial frequencies, we explore different ways of normalizing these statistics w.r.t. the first appearance of the domain, age of the dataset, and in terms of time duration or number of revisions. Below, we give a short overview of the features we have used. A more detailed description can be found in Appendix A.

3.4.1 Popularity Features.

- N_{articles}, $\overline{N_{articles}}$: Number of articles a domain has appeared in (sensitive to the size of the dataset), and its normalized version (by the number of articles in the dataset).
- $CurrN_{articles}$, $\overline{CurrN_{articles}}$: Number of articles a domain is used in at collection time, and normalized by the total number of articles in the dataset.

3.4.2 Permanence Features.

- $\Sigma Perm_d$, $\Sigma Perm_r$: Permanence, how long a domain has been used in an article (not necessarily consecutively), summed across all articles and measured in days (subscript d) or by this number of revisions (subscript r).
- $\Sigma CurrPerm_d$, $\Sigma CurrPerm_r$: Same as above, but considering only the articles where the domain is currently used.
- $\overline{\Sigma Perm_d}$, $\overline{\Sigma Perm_r}$: Sum of all permanences for a domain in all articles, normalized by the sum of the ages of all articles (measured by the number of days or the number of revisions that an article existed).
- $\overline{\Sigma CurrPerm_d}$, $\overline{\Sigma CurrPerm_r}$: Same as above, but considering only the articles where the domain is currently used.
- $\langle Perm_d \rangle$, $\langle Perm_r \rangle$: The average permanence of a domain (over all articles where it was used).
- $\langle SelfPerm_d \rangle$, $\langle SelfPerm_r \rangle$: Self-permanence: permanence divided by the number of days (or revisions) since the first time the domains were added to the article, then averaged article-wise.
- Σage_d, Σage_r: The sum of ages of the domain over all the articles in a dataset, where the age of a domain in an article is the number of days or revisions since a URL from that domain has been first added to the article.
- $\langle age_d \rangle$, $\langle age_r \rangle$: The average age of a domain over all the articles it appears at least once.

3.4.3 User-based Features.

- U_{add}, U_{rem}: Number of users that have added or removed a domain.
- R_{add}, R_{rem}: Same as above, but only counting registered users.
- $\overline{U_{add}}$, $\overline{U_{rem}}$, $\overline{R_{add}}$, $\overline{R_{rem}}$: The above four features, normalized by the total number of unique users in the dataset.
- $\langle U_{add} \rangle$, $\langle U_{rem} \rangle$, $\langle R_{add} \rangle$, $\langle R_{rem} \rangle$: The above four features, measured as an average per article.
- $Ratio(R_{add}/U_{add})$, $Ratio(R_{rem}/U_{rem})$: The proportion of registered vs. all users that ever added or removed a domain.
- $Proba(R_{add})$, $Proba(R_{rem})$: Probability that, when a domain is added or removed, this is done by the registered users. Contrary to the ratio, here we take into account the number of revisions per user.
- For all user-based features listed above, we compute a version wherein instead of counting any instance of adding a domain, we only count instances of adding it for the first time on an article ("starting"), and similarly, counting only the instances that the domain was removed for the last time ("ending"). For example, $Proba(R_{add})$, and $Proba(R_{rem})$ would have a corresponding version $Proba(R_{start})$ and $Proba(R_{end})$ (similarly for the related features).

Table 2. Leave-one-out Validation F1 Macro, and Precision and Recall Metrics for Both Classes for the Model Trained and Tested in the English Wikipedia on Each Topic, and for All Topics Combined (Standard Deviations Specified with ±)

Topic	F1 Macro	Precision (rel)	Recall (rel)	Precision (unr)	Recall (unr)	# rel	# unr
Climate Change	0.81 ± 0.02	0.80 ± 0.03	0.80 ± 0.04	0.83 ± 0.04	0.83 ± 0.03	129	152
COVID-19	0.88 ± 0.02	0.87 ± 0.03	0.91 ± 0.03	0.89 ± 0.03	0.85 ± 0.03	136	125
Biology	0.80 ± 0.03	0.81 ± 0.03	0.81 ± 0.03	0.80 ± 0.03	0.80 ± 0.03	124	120
History	0.75 ± 0.03	0.70 ± 0.04	0.83 ± 0.04	0.82 ± 0.03	0.69 ± 0.05	120	139
Media	0.83 ± 0.02	0.81 ± 0.03	0.82 ± 0.03	0.85 ± 0.03	0.84 ± 0.03	146	176
All topics	0.83 ± 0.02	0.76 ± 0.04	0.85 ± 0.03	0.89 ± 0.02	0.82 ± 0.03	159	234

Class balance in terms of the number of reliable (rel) and unreliable (unr) domains.

We remark that, to aggregate the various statistics over articles for a domain, we have considered both the average and the median and found the best performance with the average, which we use in the models in the following sections.[6]

3.5 Model Training and Evaluation

Using the above features, we train an XGBoost[7] model to classify the domains as reliable or unreliable. To account for class imbalance, we use weights (e.g., the weight for the positive class is the fraction of negative/positive). To limit overfitting, we set the maximum depth of the classifier to 1 (we have tested the classifier with depths up to 5, but found no increase in **leave-one-out** (**LOO**) validation performance, and the learning rate to $\eta = 0.1$ (default is 0.3).

When evaluating in the “native” condition, wherein the training and target data are from the same dataset, we employ LOO cross-validation. In the “cross-language adaptation” condition, wherein the model is trained on one language and tested on another (keeping the topic the same), we simply compute the evaluation metrics on the test language. Similarly, for cross-topic adaptation, we keep the language consistent and test on a target topic. In the “mixed” condition, wherein the model is trained on a set of languages and tested on one of these languages, we again employ LOO cross-validation where a domain of test language is left out at a time.

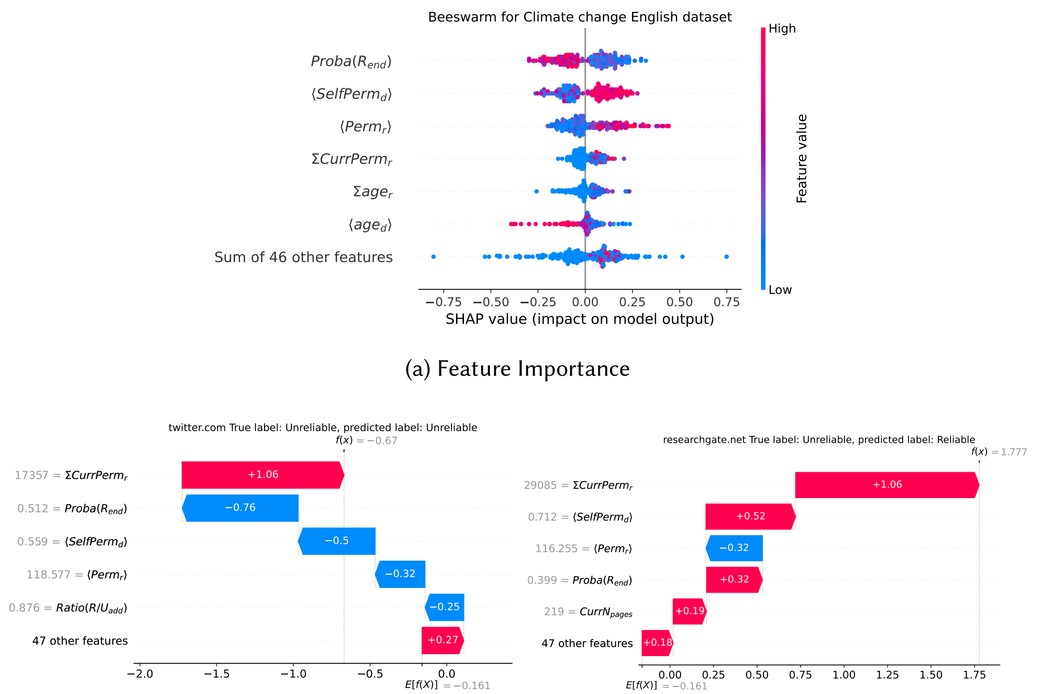
To compute a confidence interval for a metric, we employ bootstrapping wherein we resample with replacement the output of the classifier and recompute the metrics $n = 100$ times. The main metric we compute is the F1 macro score, which is the unweighted average of the F1 scores of the negative and positive classes.

4 Results

4.1 RQ1: Case Study: Climate Change in English

We begin by studying the behavior of our model for the Wikipedia articles in English around the Climate Change topic. The model performance is shown in the first row of Table 2, which lists the F1 Macro, precision and recall for each class. We find that, out of the unreliable domains, the classifier correctly identifies as unreliable 83% of them (recall), and out of the domains the classifier guessed as unreliable, 83% are truly unreliable (precision). Both precision and recall for the reliable class are slightly worse at 80%. Note that the classes are fairly balanced in this setting, with 46% of the dataset having the reliable label.

[6]We have also experimented with versions of the above features that use the aggregation over all URLs in each dataset, instead of domains. However, the resulting features were performing very similarly to those aggregated over domains, and we excluded these from the analysis.
[7]We use Python package XGBoost V1.6.2 https://pypi.org/project/xgboost/1.6.2/



(a) Feature Importance

(b) twitter.com

(c) researchgate.net

Fig. 1. Beeswarm analysis of an XGBoost model trained on Climate Change English dataset (class encoding is 1: reliable and 0: unreliable) and two example classification explanations (score shown is the log-odds ratio).

Figure 1a shows the "bee swarm" plot of **SHapley Additive exPlanations** (**SHAP**) values that illustrate the importance of each feature in the classifier [34]. The top 6 most predictive features (and an aggregation of the rest) are shown. The SHAP values for the top feature $Proba(R_{end})$ – the probability that, when a domain is removed, it is done by a registered editor – reveal that high values of this feature are associated with a negative impact on the domain reliability prediction. In other words, when a registered editor removes a domain, it is more likely to be unreliable. The opposite can be seen for the $\langle SelfPerm_d \rangle$ –average proportion of days the domain remains cited since it was first added– and $\langle Perm_r \rangle$ –the average number of revisions that the domain remains on the article–: higher values of these features are associated with more reliable sources. $\Sigma CurrPerm_r$ –the aggregated number of revisions through which the domain has existed across all articles– also correlates positively with the reliable category, as does as well the sum of the domain ages across articles, Σage_r. Finally, "younger" domains – those that have been added to an article more recently, $\langle age_d \rangle$ – are more likely to be reliable. This may be a sign of the increased focus of the platform on the quality of sources [4].

In Figures 1b and 1c we illustrate the performance of the model for two domains of special interest: *X.com*, correctly identified as unreliable, and *researchgate.net*, incorrectly identified as reliable. In the case of *twitter.com*, despite being added/present frequently on articles (high $\Sigma CurrPerm_r$), it is soon removed by registered editors ($Proba(R_{end})$ and also SelfPermanence, $\langle SelfPerm_d \rangle$ is low). For *researchgate.net*, on the other hand, model and ground truth disagree: most of the permanence features indicate that it is used as a reliable source, while the domain is considered unreliable in the perennial source list because of being a self-publishing platform lacking editorial oversight or

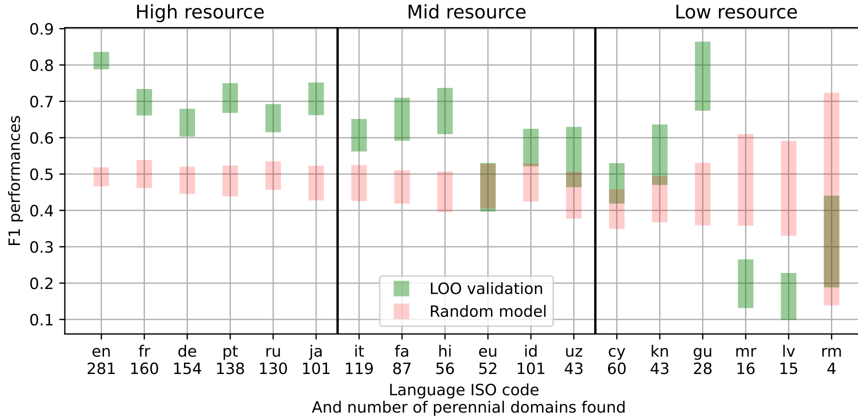


Fig. 2. Performances (mean ± stdv) of models trained and validated in languages other than English on the Climate Change datasets. The three batches of languages, separated by a vertical black line, show results on a sample of high (left), mid (center), and low (right) resource languages.

peer review. However, as our data points out, Wikipedia editors may still use it pragmatically, for example, to link to freely accessible versions of peer-reviewed articles that are otherwise behind paywalls. This example underscores the potential of our method to reveal discrepancies between Wikipedia's formal sourcing guidelines and how editors apply them in practice, highlighting edge cases that may warrant further discussion or refinement in policy.

4.2 RQ2: Topic and Language Generalizability

We continue by evaluating the performance of our language-agnostic modeling approach in other topical and linguistic settings. Table 2 shows the F1 macro and class-specific precision and recall for these topics, in the English Wikipedia. We find that, overall, the performance is consistent, with the lowest performance observed for the History dataset, with a recall of 0.69 for the unreliable class. The best performance is achieved on the COVID-19 dataset, possibly due to the ample data on the edits and the more recent nature of the corresponding articles. We find that the amount of edited data is important for the classifier's performance, as we show later in this section. Finally, we combine all English-language datasets into one and perform training and LOO validation; the performance of this model is shown in Table 2 under "all topics". Such a combined model performs on par with other topical models, with especially high precision for the unreliable class (0.89), suggesting that combining knowledge about different topics does not overall degrade the performance of the classifier.

Next, we explore the applicability of the proposed model to other languages. In these experiments, we first keep the topic constant – Climate Change – while training and evaluating models in different languages. Figure 2 shows the LOO F1 macro scores for select languages form the high-, mid-, and low-resource groups, in comparison to the performance of a random baseline model. Specifically, we show the average performance ± one standard deviation as shaded areas. To estimate the performance of a random model, we assign a random binary classification for the domains and perform bootstrapping ($n = 100$) similar to the evaluation of all other models, as discussed in Section 3.5. The performance is highest for high-resource languages and degrades for mid- and low-resource ones. The standard deviations of the scores also increase with lower resource languages (both for our model and the baselines), pointing to an increased lack of information for computing the score.

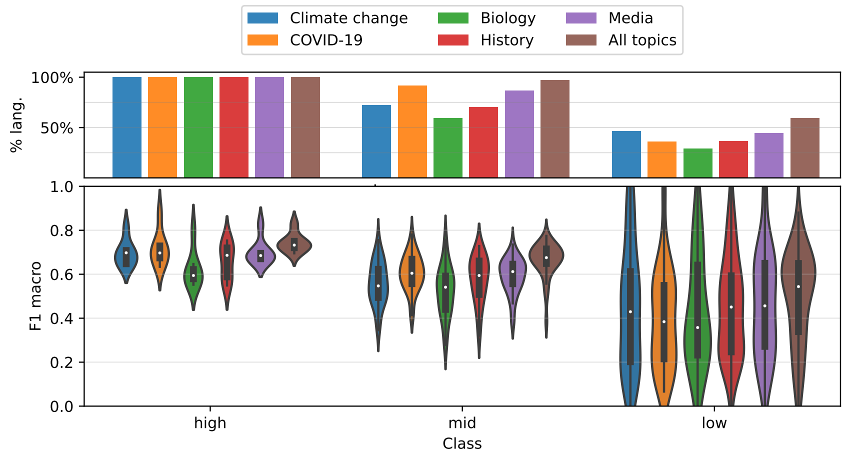


Fig. 3. Native model performance per topic for different language types. Top panel: % of models that perform better than a random classifier. Bottom panel: Violin plots of corresponding distributions of F1 macro performances.

Additionally, we show aggregate statistics of the model performance on the languages in each category in Figure 3. In particular, we are interested in whether our model outperforms the random classifier in terms of F1 macro score (top panel of the figure): we perform this comparison using Mann-Whitney scores and adjust the significance level of 0.05 using the Bonferroni correction for multiple comparisons. For instance, in the Climate Change topic, the model outperforms the random baseline in 100% of the high-resource languages, but only in 44% of low-resource languages (in an additional 12% the difference is not statistically significant). The performance on low-resource languages tends to vary widely for different topics. The brown bars show the models for which articles for all topics within a language are used for training and testing. These models perform markedly better than topic-specific ones, echoing our findings for the English language.

Effect of dataset size and introduction of perennial sources on model performance. The worse performance of our model in the lower-resource languages may be mainly (among others) due to two reasons: the datasets available for training the model are too small, or the behavior of the users of these languages toward domains is intrinsically difficult to model (their treatment of reliable and unreliable domains is difficult to distinguish). We design two experiments to address these two reasons, starting with the model's sensitivity to the amount of training data.

The red line in Figure 4 shows the performance of the model trained on different dataset sizes. In particular, we consider the English (all-topic) dataset and sample the data at log-regular intervals (starting at 10^3, then at $10^{3.25}$ and so on., until $10^{6.75}$), resulting in an increasing number of revisions available for training. We repeat this exercise 10 times (mean $\pm$ stdv. shown in the Figure). For comparison, we depict the performance of the models trained on the all-topic datasets in all other languages (shown in blue). Note that in these cases, there is no sampling, and for every language, all available data is used. We find that the performance of the English-language model is comparable to that of the other languages when there is little data available (up until about 10^5 revisions) and is about 10% higher than the other languages when almost all data is used for its training. At least in the English-language case, we find that for stable model performance around 10,000 revisions are necessary. At the higher range of data, we find that the classifier's performance has not leveled off, suggesting that adding more articles (possibly other topics) may improve the performance further.

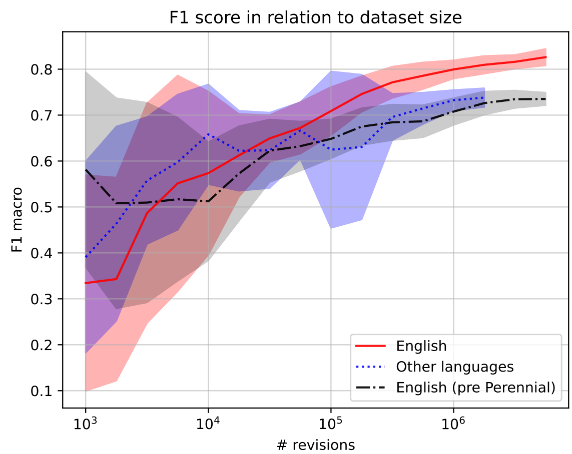


Fig. 4. Average model F1 macro (line) and standard dev. (shaded area) versus the size of the training dataset in terms of the number of revisions. In English (red), all topics are combined, and the data is sampled at regular intervals. The same experiment is repeated, considering only revisions before the implementation of the perennial sources on 2018-07-01 (grey). The performance of models in all other languages (also all-topic) is shown in blue, without sub-sampling.

Turning to editing behavior, the introduction of the perennial sources list provides a "natural experiment" to measure a possible editing behavior change in response to these additional guidelines. We only examine the impact of this change in the English-language datasets, as it is not clear whether editors from other language editions of Wikipedia used the English perennial sources list. Thus, we created a version of our English-language dataset using only revisions up to 2018-06-30 (i.e., before these guidelines were introduced). To control for the dataset size (which, as shown above, may affect model performance), we again perform sub-sampling similar to what we did with the full English-language dataset. The black dash-dotted line and the grey area in Figure 4 show the model performance based on this limited data. We observe its performance to be notably lower than with more recent data (red line and area), indicating that the editing behavior before the introduction of the perennial sources list was less predictive of reliable/unreliable domain classes than afterward. In this case, the performance is similar to that of other languages, showing that editor behavior allows extracting meaningful and predictive signals for source reliability even when the editors did not have an official guideline for source reliability. Furthermore, these findings support previous work, which shows that the quality of the Wikipedia references improved after the introduction of the perennial sources [4].

4.3 RQ3: Model Adaptation across Topics and Languages

The multilingual nature of Wikipedia presents an opportunity to supplement information available in lower-resource languages with that in the larger ones. Considering our datasets, two directions of adaptation are possible: across languages and across topics. We consider both scenarios in Figure 5, which shows the performance of the models adapted in these two ways: in a cross-language setting (trained on a language and afterward tested on a different language from the same resourcefulness class, in the same topic) and a cross-topic setting (trained on a topic, tested on a different topic, in the same language). We find that, in both cases, the performance of the adapted models decreases compared to the native performance. The decrease in performance is about the same in cross-language and cross-topic settings. This suggests that the editor's behavior around references may

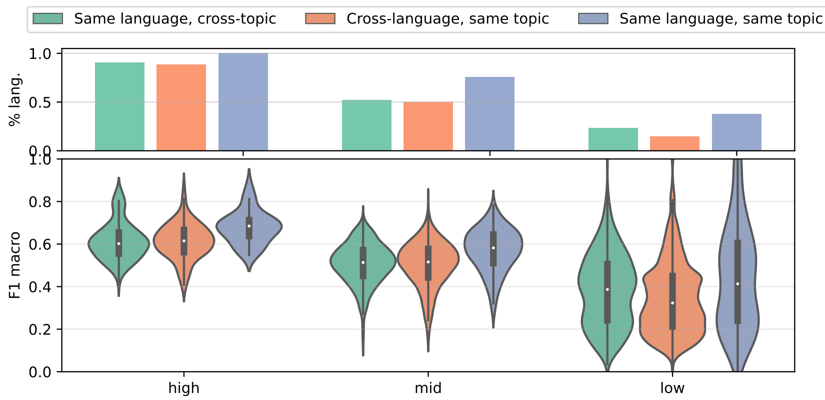


Fig. 5. Model performance in two adaptation scenarios: cross-language setting (trained on a language, tested on a different language from the same class, in the same topic, in red) and a cross-topic setting (trained on a topic, tested on a different topic, in the same language, in green). Results aggregated by language resourcefulness and compared to same-topic and language native models (in blue).

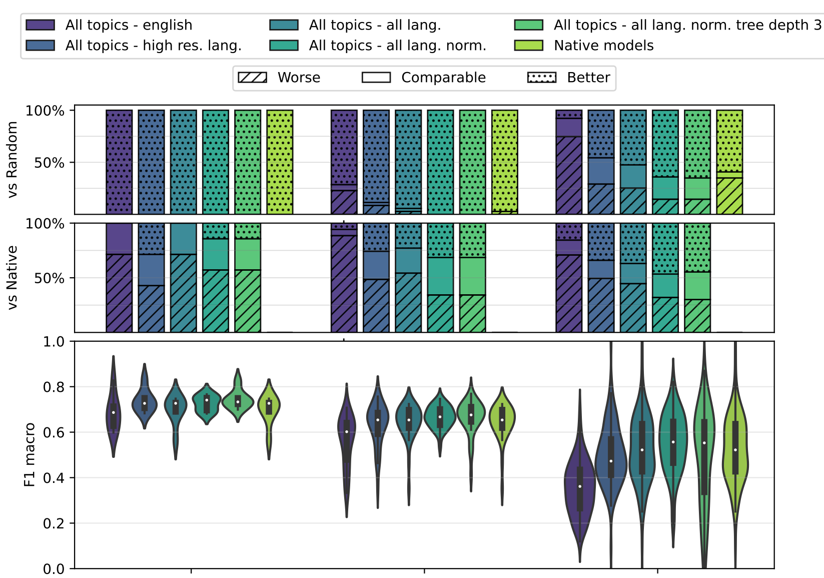


Fig. 6. Combined Model performances for different training datasets and strategies. Top panel: % of models that perform better, same as, or worse than a random classifier; Center: same, but compared to the native model; Bottom: violin plots of the corresponding distributions of the F1 macro performances.

be non-identical in different topics, as well as languages. However, again, we see the importance of resourcefulness on the performance: the models adapted across high-resource languages perform better than those across mid and low-resource ones.

Finally, we consider all topics together and analyze the performance of our model when trained with different training strategies and tested on each language. In particular, we test whether models trained on higher-resource data would be helpful to those in lower-resource settings. In Figure 6 we aggregate the corresponding results by languages of similar resourcefulness: the top panel shows how for many languages the different models perform significantly better, same as, or worse than a random model (according to Mann-Whitney tests with Bonferroni correction), the center panel shows the same, but compared to the native model, and the bottom panel shows the distribution of F1 macro scores. We find that F1 scores are generally lower when training a model on the English dataset and applying it to low-resource languages, confirming that cross-language adaptation can be problematic, even if a high-resource language such as English is used. However, when training on all high-resource languages or all languages together, the performance for low-resource languages significantly increases. This suggests that including a plurality of languages in training data may capture a variety of user behaviors that are transferable to lower-resource languages.

Nonetheless, each language-specific dataset has its own biases connected to the amount of data available, temporal peculiarities of the articles and their edits, and the volume of editing activity. To alleviate potential difficulties in adapting feature values between languages, we normalize each dataset using quantile normalization (converting feature values to ranks). This approach further improves the performance of the all-language model adaptation, even on average performing better than the native models. In more detail, the model outperforms a random classifier in more than 65% of low-resource settings, and even outperforms native models in 47% of these cases, while being on par with it in 21%. Notably, models trained solely on English fail to generalize well to low-resource languages, performing worse than random in over 70% of cases—highlighting the distinct editorial behaviors across linguistic communities.

Finally, since the size of the dataset increases with the combination of topics, we train the classifier with the maximum tree depth of 3 instead of 1, but do not find a substantial increase in performance.

4.4 In-Depth Examination of Source Usage

The above quantitative evaluation of the proposed model abstracts the rich insight that its output provides. When examining the domains that attain the highest reliability score by the classifier, we find interesting signals of the editorial treatments of different sources, in each topical sphere (here, we focus on English data). For instance, among the domains predicted (incorrectly) as reliable in the Climate Change domain is YouTube (which appeared in 644 articles in that dataset) with a positive (reliable) class probability of $P_r = 0.64$, as well as **Internet Movie Database** (**IMDB**) (appearing in 56 articles, $P_r = 0.56$). This points to the importance of popular or mainstream culture in this topic. When considering the History dataset, we find the classifier incorrectly labels as reliable domains pointing to sources that the perennial sources label as having partisan bias, such as the Jewish Virtual Library ($P_r = 0.81$), Sixth Tone ($P_r = 0.74$), and Daily Sabah ($P_r = 0.69$).

On the other hand, one can use the output of this classifier to find sources that different communities of editors consider inappropriate for their topical domain. For instance, in the Biology dataset, Science-Based Medicine, which is considered generally reliable and has "a credible editorial board" according to the perennial source list, is treated by the editors more similarly to unreliable domains ($P_r = 0.20$). Similarly, the Huffington Post, rated as "fairly reliable for factual reporting on non-political topics", scores low in the COVID dataset ($P_r = 0.07$), possibly due to the politically charged and serious nature of the rhetoric around the pandemic. As editorial policies of sources, as well as perceptions and behaviors of the Wikipedia editors, change, the metrics developed in this study may provide a guide for the improvement and updating of the perennial sources and potential adjustment of the labels in specific domains of expertise.

4.5 Alternative Indicators of Source Reliability

One limiting factor of our study was the availability of labels for the quality of sources. We consider an alternative: **Media Bias / Fact Check** (**MBFC**), a third-party website that provides ratings of "the bias, factual accuracy, and credibility of media sources",[8] which has been widely used in misinformation research and by other indexes [6, 11, 54]. We use this credibility metric and consider *very high* and *high* as the positive class, and *low* and *very low* as negative. With the same features and model as above and LOO cross-validation, we gauge its performance on MBFC. Although the class-average F1 macro is 0.61, the classifier performs especially poorly on the negative class, with a precision of 0.32 and recall of 0.55, suggesting that the features do not capture this metric. Indeed, there may be several reasons why such resources are not suitable for the analysis of Wikipedia references. First, there is little overlap between the sources noted in MBFC and in the English Wikipedia perennial sources list (out of 1156 perennial sources, 272 are in MBFC). Second, the MBFC credibility metric is weakly aligned with the perennial source reliability one. Third, MBFC itself is listed as *generally unreliable* by the English Wikipedia perennial sources list.

In conclusion, MBFC and similar resources are valuable for assessing source reliability in general research contexts. However, the above observations suggest that they are less suitable for evaluating the reliability of sources in the specific context of Wikipedia articles, especially if used as ground truth for supervised classification methods like those presented in this work.

5 Discussion

In this study, we have proposed and assessed editing behavior-based features for modeling the reliability of web domains used as references on Wikipedia. We show that these models can be a part of a language-agnostic toolkit applicable both to high- and low-resource languages. As such, this work extends previous efforts to use the edit history to measure the controversiality of Wikipedia articles [8]. Editing behavior is not the only source of useful information for language-agnostic modeling. Links between articles (already used for topical clustering of Wikipedia articles [26]) may provide semantically meaningful context for references, whereas users' reading sessions (used for entity linking [19]) may give controversiality signals from the perspective of the audience. Future work could also take a page from the social media literature and build editor-resource networks [40] to capture structural peculiarities of topics, editors, and resources associated with low-quality content. Additionally, automated feature discovery methods, such as self-supervised or representation learning on structured edit data, could help uncover latent behavioral patterns not easily captured by manually designed features. However, this would come at the cost of losing interpretability, a key value of our current approach.

Apart from finding that the proposed features capture editing behavior within a dataset spanning one topic and language, we also observe that they retain useful information during the model adaptation across languages. Although English has been the *lingua franca* in NLP, having the best-developed tools and resources, we show that adaptation from the English-language dataset to other languages is largely inferior to using a model that combines information from many languages. This improvement is especially noticeable in performance on low-resource languages, where native models often perform no better than random baselines, due to insufficient revision and citation activity for effective learning. However, when training on data pooled from all languages and applying quantile normalization to account for resource differences, the model significantly outperforms random baselines and matches or outperforms native models in most low-resource cases. This strategy highlights a practical and scalable pathway to empower under-resourced Wikipedia communities, even in the absence of extensive local data or tooling.

These findings are aligned with limitations identified by Das et al. [13], who suggest that, although language-agnostic features are used, limiting data selection to one language may limit the breadth of the captured behaviors and their applicability to new settings. As multilingual embedding models

[8] https://mediabiasfactcheck.com/about/

have gained popularity [56], it remains to be seen to what extent the latest developments in NLP will benefit the low-resource languages [57], and what the impact will be on Wikipedia specifically [27]. In this context, the need to monitor possible changes in the editorial standards and citation practices across languages becomes especially important. Community efforts could prioritize on developing robust cross-lingual transfer strategies, rather than relying solely on local data that may be too sparse to train reliable predictors

As ground truth, we have used the list of perennial sources [46] from the English Wikipedia, which is part of Wikipedia's efforts to provide clear editorial instructions to its community. We compare this resource with another popular source of reliability measure, the MBFC [55], and find significant differences between them, pointing to the necessity of Wikipedia-specific models like ours. Previous work comparing the perennial sources lists in different languages showed disagreement on whether some domains are indeed reliable [5]. Therefore, the system we propose could be used to expand lists of perennial sources and lists derived from them,[9] or even to assess cultural biases in Wikipedia sources [53, 59].

Limitations

The present study has multiple limitations. First, using language-agnostic approaches may be seen as a limitation, as it does not take advantage of the predictive power of language-dependent features. However, despite the significant advances with multilingual NLP resources such as mBERT [16], the motivation for our model was precisely to be ready for use in the 300+ language editions of Wikipedia, especially those that are less privileged and under-resourced.

Second, while the dataset contains thousands of articles and cumulatively millions of revisions, it spans only a fraction of Wikipedia. Despite choosing topical foci such that they have different reference quality and time scales, they may not capture the entire diversity of subject matter of this massive collaborative effort. For instance, differentiating topics by whether the topic is currently controversial may reveal shifts in editorial handling of sources. Additionally, future versions of our modeling approach could incorporate newly generated data by implementing re-training strategies using updated activity data, including changes in the classification of web domain reliability scores from the perennial source list.

Third, our approach may inherently favor generalist sources over topic-specific ones. Because several behavioral features (e.g., the number of articles or revisions in which a source domain appears) capture overall prevalence within Wikipedia, broadly used domains tend to accumulate stronger permanence and visibility signals. Conversely, specialized domains, despite being authoritative within their fields, may appear in fewer contexts and thus receive lower reliability estimates. This limitation suggests a potential bias toward sources covering a wider topical range. Its presence could be investigated in future research and mitigated by normalizing the corresponding features by topical diversity.

Fourth, the selection of the English perennial sources list as the gold standard disadvantaged other languages. Unfortunately, only a few other languages have a similar resource, and they are significantly less extensive and conclusive [5]. Given the global role played historically by the English edition of Wikipedia [22], our selection provided the best choice to reach the maximum potential coverage across other language editions. However, the adaptation of the domain list from English may not coincide with possibly different cultural or local interpretations of quality by the editors from other language communities. In such cases, it may be preferable to apply NLP transfer methods [14] that leverage higher-resource language editions to support lower-resource ones, particularly when they are comparable in the domain of source reliability. Our study, which

[9] https://science.feedback.org/consensus-credibility-scores-comprehensive-dataset-web-domains-credibility/

employs language-agnostic features, can serve as a foundation for identifying appropriate language pairs for transfer in future work.

Fifth, our model relies on a fixed ground truth, consistent with prior work (e.g., [4, 5]). However, perennial source lists evolve over time as editorial consensus changes and new sources are evaluated. While we treat it as static for this study, future work should incorporate time-aware labels and explore retraining strategies to adapt to such changes and maintain alignment with current community standards.

Finally, we show that the method is sensitive to the amount of data available for training (and testing), possibly resulting in insignificant results in some low-resource language settings. Further research will be needed to extend the amount of available data in these contexts.

Ethical Considerations

This study, and online content moderation in general, have important ethical aspects that should be kept in mind by both the platforms and the misinformation researchers.

First, this study has a direct impact on the access equity to encyclopedic resources in low-resource languages. Currently, the lack of resources, such as the perennial source list for these versions of Wikipedia, limits the extent to which our model is able to reflect the standards of that linguistic community. The domain scores provided by the models adapted from other (higher resource) languages that we propose here could be a starting point for bootstrapping the editorial discussion of the quality of sources used in low-resource languages.

Second, the process of determining the quality of sources bears scrutiny. In the news domain, the quality of information is often determined by professional fact-checkers. However, on Wikipedia, the editors from the community define the status of the sources via a deliberation process, guided by the community standards. Any potential biases in the judgment of these editors or failings in the deliberative process would be reflected in the quality of these labels (for instance, according to Kharazian et al. [29], some Wikipedia editions were dominated for a time by a "small group of administrators who introduced far-right bias and outright disinformation"). Generally, in 2014 Shaw and Hill [39] concluded that the Wikipedia contributors self-organize into oligarchic structures, though in 2021 Rijshouwer et al. [37] found "strong counter-tendencies" to editorial power concentration. Our proposed method may reveal community behaviors that are not necessarily explicitly stated, but which shape the platform's use of outside sources, potentially aiding the community to self-monitor the use of sources.

We believe that the code and resources we make available to the community[10] (compiled in accordance with FAIR (Findable, Accessible, Interoperable, and Reusable principles [47]) will be useful to assist future research efforts and to further evaluate and extend the features and modeling we proposed.

References

[1] Pablo Aragón and Diego Sáez-Trumper. 2021. A preliminary approach to knowledge integrity risk assessment in Wikipedia projects. arXiv:2106.15940. Retrieved from https://arxiv.org/abs/2106.15940 (Cited on 2)

[2] Akhil Arora, Robert West, and Martin Gerlach. 2024. Orphan articles: The dark matter of wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI Press, Washington, DC, USA, 100–112. (Cited on 3)

[3] Joshua Ashkinaze, Eric Gilbert, and Ceren Budak. 2024. The dynamics of (Not) unfollowing misinformation spreaders. In *Proceedings of the ACM on Web Conference 2024*. Association for Computing Machinery, New York, NY, USA, 1115–1125. (Cited on 2)

[4] Aitolkyn Baigutanova, Jaehyeon Myung, Diego Saez-Trumper, Ai-Jou Chou, Miriam Redi, Changwook Jung, and Meeyoung Cha. 2023. Longitudinal assessment of reference quality on wikipedia. In *Proceedings of the ACM Web*

[10] https://github.com/JacopoDignazi/Wiscom

Conference 2023. ACM, New York, NY, USA, 2831–2839. DOI: https://doi.org/10.1145/3543507.3583218 (Cited on 2, 3, 4, 8, 11, 16)

[5] Aitolkyn Baigutanova, Diego Saez-Trumper, Miriam Redi, Meeyoung Cha, and Pablo Aragón. 2023. A Comparative Study of Reference Reliability in Multiple Language Editions of Wikipedia. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>). ACM, New York, NY, USA, 3743–3747. DOI: https://doi.org/10.1145/3583780.3615254 (Cited on 2, 3, 4, 15, 16)

[6] Golding Barret. 2021. Iffy index of unreliable sources. (2021). Retrieved April 03, 2024 from https://iffy.news/index/ (Cited on 14)

[7] Omer Benjakob, Rona Aviram, and Jonathan Aryeh Sobel. 2022. Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. *GigaScience* 11 (2022), giab095. (Cited on 3)

[8] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal Controversies in Wikipedia Articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea). Association for Computing Machinery, New York, NY, USA, 193–196. DOI: https://doi.org/10.1145/2702123.2702436 (Cited on 2, 14)

[9] Chih-Chun Chen and Camille Roth. 2012. {{Citation needed}} the dynamics of referencing in wikipedia. In *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration.* Association for Computing Machinery, New York, NY, USA, 1–4. (Cited on 3)

[10] Ai-Jou Chou, Guilherme Gonçalves, Sam Walton, and Miriam Redi. 2020. Citation detective: a public dataset to improve and quantify wikipedia citation quality at scale. In *Proceedings of the Wiki Workshop.* Wiki Workshop. (Cited on 3)

[11] Rajdipa Chowdhury, Sriram Srinivasan, and Lise Getoor. 2020. Joint Estimation of User And Publisher Credibility for Fake News Detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event, Ireland). Association for Computing Machinery, New York, NY, USA, 1993–1996. DOI: https://doi.org/10.1145/3340531.3412066 (Cited on 14)

[12] Noam Cohen. 2021. One Woman's Mission to Rewrite Nazi History on Wikipedia. WIRED. [Online; accessed 15-Apr-2024]. (Cited on 2)

[13] Paramita Das, Isaac Johnson, Diego Saez-Trumper, and Pablo Aragón. 2024. Language-Agnostic Modeling of Wikipedia Articles for Content Quality Assessment across Languages. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 01 (2024), 1–11. (Cited on 3, 14)

[14] Paramita Das, Amartya Roy, Ritabrata Chakraborty, and Animesh Mukherjee. 2025. On the effective transfer of knowledge from English to Hindi Wikipedia. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track.* Association for Computational Linguistics, Abu Dhabi, UAE, 453–465. Retrieved from https://aclanthology.org/2025.coling-industry.39/ (Cited on 15)

[15] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. (Cited on 2)

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423 (Cited on 15)

[17] Tomás Feith, Akhil Arora, Martin Gerlach, Debjit Paul, and Robert West. 2024. Entity Insertion in Multilingual Linked Corpora: The Case of Wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Miami, Florida, USA, 22796–22819. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.1268 (Cited on 3)

[18] Mengyuan Fu, Kunhao Yang, and Yuko Fujigaki. 2023. Introducing an "invisible enemy": a case study of knowledge construction regarding microplastics in Japanese Wikipedia. *New Media and Society* 0, 0 (2023), 14614448221149747. (Cited on 3)

[19] Martin Gerlach, Marshall Miller, Rita Ho, Kosta Harlan, and Djellel Difallah. 2021. Multilingual entity linking system for Wikipedia with a machine-in-the-loop approach. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management.* ACM, New York, NY, USA, 3818–3827. (Cited on 3, 14)

[20] Hao Guo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2023. Interpretable Fake News Detection with Graph Evidence. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* Birmingham, United Kingdom. Association for Computing Machinery, New York, NY, USA, 659–668. DOI: https://doi.org/10.1145/3583780.3614936 (Cited on 2)

[21] Yuhan Guo, Qin Han, Yuke Lou, Yiming Wang, Can Liu, and Xiaoru Yuan. 2023. Edit-History Vis: An Interactive Visual Exploration and Analysis on Wikipedia Edit History. In *Proceedings of the 2023 IEEE 16th Pacific Visualization Symposium*. IEEE, New York, NY, USA, 157–166. (Cited on 3)

[22] Scott A. Hale. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM Conference on Web Science* (Bloomington, Indiana, USA). Association for Computing Machinery, New York, NY, USA, 99–108. DOI: https://doi.org/10.1145/2615569.2615684 (Cited on 15)

[23] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37. (Cited on 4)

[24] Aida Halitaj and Arkaitz Zubiaga. 2024. Providing citations to support fact-checking: Contextualizing detection of sentences needing citation on small wikipedias. *Natural Language Processing Journal* 8 (2024), 100093. (Cited on 3)

[25] Aida Halitaj and Arkaitz Zubiaga. 2025. ALPET: Active few-shot learning for citation worthiness detection in low-resource Wikipedia languages. *Expert Systems With Applications* 281 (2025), 127503. (Cited on 3)

[26] Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. 2021. Language-agnostic Topic Classification for Wikipedia. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia). Association for Computing Machinery, New York, NY, USA, 594–601. DOI: https://doi.org/10.1145/3442442.3452347 (Cited on 3, 4, 14)

[27] Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. 2024. Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, Association for Computational Linguistics, Miami, Florida, USA, 91–101. DOI: https://doi.org/10.18653/v1/2024.wikinlp-1.14 (Cited on 15)

[28] Isaac Johnson and Emily Lescak. 2022. Considerations for multilingual wikipedia research. arXiv:2204.02483. Retrieved from https://arxiv.org/abs/2204.02483 (Cited on 3)

[29] Zarine Kharazian, Kate Starbird, and Benjamin Mako Hill. 2024. Governance capture in a self-governing community: A qualitative comparison of the croatian, serbian, bosnian, and serbo-croatian wikipedias. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 61 (2024), 26 pages. DOI: https://doi.org/10.1145/3637338 (Cited on 16)

[30] Jasper W. Korte, Sabine Bartsch, Rasmus Beckmann, Roxanne El Baff, Andreas Hamm, and Tobias Hecking. 2023. From causes to consequences, from chat to crisis. The different climate changes of science and wikipedia. *Environmental Science and Policy* 148 (2023), 103553. DOI: https://doi.org/10.1016/j.envsci.2023.103553 (Cited on 4)

[31] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 591–602. DOI: https://doi.org/10.1145/2872427.2883085 (Cited on 2)

[32] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2019. Multilingual ranking of wikipedia articles with quality and popularity assessment in different topics. *Computers* 8, 3 (2019), 60. (Cited on 3)

[33] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2020. Modeling popularity and reliability of sources in multilingual Wikipedia. *Information* 11, 5 (2020), 263. (Cited on 3)

[34] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Curran Associates, Inc., Red Hook, NY, USA, 4765–4774. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (Cited on 8)

[35] Tiziano Piccardi and Robert West. 2021. Crosslingual topic modeling with WikiPDA. In *Proceedings of the Web Conference 2021*. Association for Computing Machinery, New York, NY, USA, 3032–3041. (Cited on 3)

[36] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *Proceedings of the World Wide Web Conference* (San Francisco, CA, USA). Association for Computing Machinery, New York, NY, USA, 1567–1578. DOI: https://doi.org/10.1145/3308558.3313618 (Cited on 3)

[37] Emiel Rijshouwer, Justus Uitermark, and Willem de Koster. 2023. Wikipedia: a self-organizing bureaucracy. *Information, Communication and Society* 26, 7 (2023), 1285–1302. (Cited on 16)

[38] Richard Rogers and Emina Sendijarevic. 2012. Neutral or national point of view? A comparison of srebrenica articles across Wikipedia's language versions. In *Wikipedia Academy: Research and Free Knowledge: June 29 - July 1 2012 Berlin: Accepted Submissions Wikepedia Academy*. (Cited on 3)

[39] Aaron Shaw and Benjamin M. Hill. 2014. Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication* 64, 2 (2014), 215–238. (Cited on 16)

[40] Kai Shu, H. Russell Bernard, and Huan Liu. 2019. Studying Fake News via Network Analysis: Detection and Mitigation. In *Proceedings of the Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Nitin Agarwal, Nima Dokoohaki, and Serpil Tokdemir (Eds.), Springer International Publishing, Cham, 43–65. DOI: https://doi.org/10.1007/978-3-319-94105-9_3 (Cited on 3, 14)

[41] Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. 2023. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. 51–59. (Cited on 2)
[42] Swati Swati, Adrian Mladenić Grobelnik, Dunja Mladenić, and Marko Grobelnik. 2023. A commonsense-infused language-agnostic learning framework for enhancing prediction of political bias in multilingual news headlines. *Knowledge-Based Systems* 277 (2023), 110838. (Cited on 3)
[43] Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, and Diego Saez-Trumper. 2023. Fair multilingual vandalism detection system for wikipedia. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (, Long Beach, CA, USA,). Association for Computing Machinery, New York, NY, USA, 4981–4990. DOI: https://doi.org/10.1145/3580305.3599823 (Cited on 5)
[44] Mykola Trokhymovych and Diego Saez-Trumper. 2021. WikiCheck: An end-to-end open source automatic fact-checking API based on Wikipedia. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management* (Virtual Event, Queensland, Australia). Association for Computing Machinery, New York, NY, USA, 4155–4164. DOI: https://doi.org/10.1145/3459637.3481961 (Cited on 3)
[45] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. (Cited on 2)
[46] Wikipedia contributors. 2024. English wikipedia. 2024. wikipedia:reliable sources/Perennial sources. Retrieved April 12, 2024 from https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources (Cited on 3, 15)
[47] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 1–9. (Cited on 16)
[48] KayYen Wong, Miriam Redi, and Diego Saez-Trumper. 2021. Wiki-reliability: A large scale dataset for content reliability on wikipedia. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2437–2442. (Cited on 3)
[49] Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-Align: Prompt-based social alignment for few-shot fake news detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (,Birmingham, United Kingdom,). Association for Computing Machinery, New York, NY, USA, 2726–2736. DOI: https://doi.org/10.1145/3583780.3615015 (Cited on 2)
[50] Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. MSynFD: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM on Web Conference 2024*. Association for Computing Machinery, New York, NY, USA, 4128–4137.
[51] Tianshu Xiao, Sichang Guo, Jingcheng Huang, Riccardo Spolaor, and Xiuzhen Cheng. 2023. HiPo: Detecting fake news via historical and multi-modal analyses of social media posts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom). Association for Computing Machinery, New York, NY, USA, 2805–2815. DOI: https://doi.org/10.1145/3583780.3614914
[52] Xiaofei Xu, Ke Deng, and Xiuzhen Zhang. 2022. Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1206–1214. (Cited on 2)
[53] Puyu Yang and Giovanni Colavizza. 2024. Polarization and reliability of news sources in Wikipedia. *Online Information Review* 48, 5 (2024), 908-925. (Cited on 3, 15)
[54] Junting Ye and Steven Skiena. 2019. MediaRank: Computational ranking of online news sources. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA). Association for Computing Machinery, New York, NY, USA, 2469–2477. DOI: https://doi.org/10.1145/3292500.3330709 (Cited on 14)
[55] Dave Van Zandt. 2024. Media Bias/Fact-Check. (2024). Retrieved April 23, 2024 from https://mediabiasfactcheck.com/ (Cited on 15)
[56] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *Proceedings of the Advances in Neural Information Processing Systems*. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36, Curran Associates, Inc., New York, 5484–5505. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf (Cited on 15)
[57] Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 7915–7927. (Cited on 15)
[58] Yuehua Zhao, Jingwei Da, and Jiaqi Yan. 2021. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing and Management* 58, 1 (2021), 102390. (Cited on 3)

[59] Xiang Zheng, Jiajing Chen, Erjia Yan, and Chaoqun Ni. 2023. Gender and country biases in Wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology* 74, 2 (2023), 219–233. (Cited on 15)

Appendices

A Feature List

We provide here a more detailed list of the 52 features used by our models.

A.1 Popularity Features

(1) N_{articles}: Number of articles a domain has appeared in (sensitive to the size of the dataset).
(2) $\overline{N_{articles}}$: Number of articles a domain has appeared in, normalized by the number of articles in the dataset.
(3) $CurrN_{articles}$: Number of articles a domain is used in at collection time.
(4) $\overline{CurrN_{articles}}$: Number of articles a domain is used in at collection time, normalized by the total number of articles in the dataset.

A.2 Permanence Features

(5) $\Sigma Perm_d$: Permanence, how long a domain has been used in an article (not necessarily consecutively), summed across all articles and measured in days.
(6) $\Sigma Perm_r$: Permanence, how long a domain has been used in an article (not necessarily consecutively), summed across all articles and measured by his number of revisions.
(7) $\Sigma CurrPerm_d$: How long a domain has been used in an article (not necessarily consecutively), summed across all articles and measured in days, but considering only the articles where the domain is currently used.
(8) $\Sigma CurrPerm_r$: How long a domain has been used in an article (not necessarily consecutively), summed across all articles and measured in number of revisions, but considering only the articles where the domain is currently used.
(9) $\overline{\Sigma Perm_d}$: Sum of all permanences for a domain in all articles, normalized by the sum of the ages of all articles, measured by the number of days.
(10) $\overline{\Sigma Perm_r}$: Sum of all permanences for a domain in all articles, normalized by the sum of the ages of all articles, measured by the number of revisions.
(11) $\overline{\Sigma CurrPerm_d}$: Sum of all permanences for a domain in articles where the domain was used at collection time, normalized by the sum of the ages of all articles, measured by the number of days.
(12) $\overline{\Sigma CurrPerm_r}$: Sum of all permanences for a domain in articles where the domain was used at collection time, normalized by the sum of the ages of all articles, measured by the number of revisions.
(13) $\langle Perm_d \rangle$: The average permanence of a domain (over all articles where it was used), measured by the number of days.
(14) $\langle Perm_r \rangle$: The average permanence of a domain (over all articles where it was used), measured by the number of revisions.
(15) $\langle SelfPerm_d \rangle$: Self-permanence: permanence divided by the number of days since the first time the domains were added to the article, then averaged article-wise.
(16) $\langle SelfPerm_r \rangle$: Self-permanence: permanence divided by the number of revisions since the first time the domains were added to the article, then averaged article-wise.
(17) Σage_d: The sum of ages of the domain over all the articles in a dataset, where the age of a domain in an article is the number of days since a URL from that domain has been first added to the article.

(18) Σage_r: The sum of ages of the domain over all the articles in a dataset, where the age of a domain in an article is the number of revisions since a URL from that domain has been first added to the article.
(19) $\langle age_d \rangle$: The average age of a domain over all the articles it appears at least once, measured by the number of days.
(20) $\langle age_r \rangle$: The average age of a domain over all the articles it appears at least once, measured by the number of revisions.

A.3 User-based Features

(21) U_{add} Number of unique users that have added a domain.
(22) U_{start} Number of unique users that have added a domain for the first time on a page.
(23) U_{rem} Number of unique users that have removed a domain.
(24) U_{end} Number of unique users that have removed a domain for the last time on a page.
(25) R_{add} Number of unique registered users that have added a domain.
(26) R_{start} Number of unique registered users that have added a domain for the first time on a page.
(27) R_{rem} Number of unique registered users that have removed a domain.
(28) R_{end} Number of unique registered users that have removed a domain for the last time on a page.
(29) $\overline{U_{add}}$: Number of unique users that have added a domain, divided by the total number of unique users in the dataset.
(30) $\overline{U_{start}}$: Number of unique users that have added a domain for the first time on a page, divided by the total number of unique users in the dataset.
(31) $\overline{U_{rem}}$: Number of unique users that have removed a domain, divided by the total number of unique users in the dataset.
(32) $\overline{U_{end}}$: Number of unique users that have removed a domain for the last time on a page, divided by the total number of unique users in the dataset.
(33) $\overline{R_{add}}$: Number of unique registered users that have added a domain, divided by the total number of unique users in the dataset.
(34) $\overline{R_{start}}$: Number of unique registered users that have added a domain for the first time on a page, divided by the total number of unique users in the dataset.
(35) $\overline{R_{rem}}$: Number of unique registered users that have removed a domain, divided by the total number of unique users in the dataset.
(36) $\overline{R_{end}}$: Number of unique registered users that have removed a domain for the last time on a page, divided by the total number of unique users in the dataset.
(37) $\langle U_{add} \rangle$: Average per page number of unique users that have added a domain.
(38) $\langle U_{start} \rangle$: Average per page number of unique users that have added a domain for the first time on a page.
(39) $\langle U_{rem} \rangle$: Average per page number of unique users that have removed a domain.
(40) $\langle U_{end} \rangle$: Average per page number of unique users that have removed a domain for the last time on a page.
(41) $\langle R_{add} \rangle$: Average per page number of unique registered users that have added a domain.
(42) $\langle R_{start} \rangle$: Average per page number of unique registered users that have added a domain for the first time on a page.
(43) $\langle R_{rem} \rangle$: Average per page number of unique registered users that have removed a domain
(44) $\langle R_{end} \rangle$: Average per page number of unique registered users that have removed a domain for the last time on a page.
(45) $Ratio(R_{add}/U_{add})$: The proportion of registered vs. all users that ever added a domain.

(46) $Ratio(R_{start}/U_{start})$: The proportion of registered vs. all users that ever added a domain for the first time on a page.
(47) $Ratio(R_{rem}/U_{rem})$: The proportion of registered vs. all users that ever removed a domain.
(48) $Ratio(R_{end}/U_{end})$: The proportion of registered vs. all users that ever removed a domain for the last time on a page.
(49) $Proba(R_{add})$: Probability that, when a domain is added, this is done by the registered users.
(50) $Proba(R_{start})$: Probability that, when a domain is added for the first time on a page, this is done by the registered users.
(51) $Proba(R_{rem})$: Probability that, when a domain is removed, this is done by the registered users.
(52) $Proba(R_{end})$: Probability that, when a domain is removed for the last time on a page, this is done by the registered users.

B Appendix B: Medians of resource-wise statistics

Table A1. Statistics of the Topic-Language Datasets Used in the Study Grouped by High, Mid, and Low Resource Languages

	Topic	Langs	Articles	Revs	Revs/Ar	URLs	URLs/Ar	Domains	P. Domains
High	Climate change	7	1,162	251,085	62	34,352	18	11,992	191
	COVID-19	7	595	136,470	49	29,826	21	6,031	185
	Biology	7	8,228	388,729	17	32,315	3	8,377	127
	History	7	4,839	417,275	40	30,957	5	9,324	146
	Media	7	3,542	398,138	46	43,849	8	12,441	240
Mid	Climate change	36	325	23,000	27	4,399	9	1,991	85
	COVID-19	37	182	10,026	19	4,891	9	1,326	99
	Biology	37	2,053	57,234	11	5,005	2	1,643	59
	History	37	1,687	70,988	26	5,500	3	1,901	62
	Media	37	799	30,098	20	5,347	4	1,991	123
Low	Climate change	56	100	2,403	12	633	6	373	25
	COVID-19	75	24	518	9	466	9	214	31
	Biology	41	440	4,949	7	541	3	301	14
	History	55	421	8,020	13	449	3	256	16
	Media	70	73	1,063	6	263	3	148	22

All numbers are medians (apart from the number of languages). Perennial Domains (P. Domains) refers to the number of domains for the perennial source list in English found in the different datasets. The corresponding mean values can be found in Table 1.

C APPENDIX C: Language-specific dataset sizes

Table A2. Number of Articles Per Language Editions in the Different Datasets

Language	Climate change	COVID-19	Biology	History	Media
de	479	274	463	549	472
en	975	856	931	1092	1460
es	339	258	397	461	509
fr	412	262	388	546	476
ja	217	145	230	246	382
pt	223	125	248	258	296
ru	257	162	317	435	377

(Continued)

Table A2. Continued

Language	Climate change	COVID-19	Biology	History	Media
ar	195	148	196	258	189
az	62	43	88	126	73
bg	109	44	141	175	116
bn	82	51	92	95	80
ca	176	94	217	242	148
cs	168	79	188	203	143
da	141	43	143	161	119
el	104	55	112	211	95
et	101	58	105	122	73
eu	104	36	94	144	68
fa	156	89	171	229	168
fi	184	68	184	177	178
he	184	146	230	301	225
hi	98	42	78	76	58
hr	99	33	104	148	83
hu	151	61	205	227	186
hy	56	39	70	112	83
id	136	83	129	154	122
it	259	161	289	426	427
ko	139	106	153	204	204
ms	73	63	82	91	67
nb	174	72	170	183	145
nl	238	114	270	295	233
pl	185	89	233	279	236
ro	106	45	122	171	109
simple	136	53	140	137	118
sk	105	26	118	137	91
sl	106	39	99	156	65
sr	115	59	129	230	126
sv	187	79	204	221	201
th	107	61	123	138	125
tr	157	89	160	247	182
uk	162	212	173	194	150
uz	36	39	33	84	42
vi	137	217	145	161	145
zh	205	246	212	250	326
kk	0	17	55	49	34
af	72	28	71	102	46
als	33	0	0	58	0
as	21	13	23	18	13
ast	84	20	88	130	84
ba	28	6	0	36	14
bcl	21	9	16	27	12
be	56	27	61	90	48

(Continued)

Table A2. Continued

Language	Climate change	COVID-19	Biology	History	Media
be-x-old	45	14	0	77	42
bs	74	21	71	118	52
ckb	31	16	0	44	38
cv	25	0	0	0	25
cy	63	23	556	161	55
dag	29	0	0	0	0
eo	104	26	121	137	79
ga	43	24	44	76	47
gl	108	34	127	152	81
gu	29	6	0	0	9
ha	32	23	0	20	8
ia	31	11	0	0	0
ig	17	0	7	10	9
jv	46	15	63	74	32
ka	66	24	76	113	77
kn	51	28	40	33	35
ku	30	13	0	58	31
la	69	19	89	130	54
lb	33	0	0	0	53
lt	102	26	120	129	79
lv	80	27	95	93	63
mk	72	18	73	124	59
ml	70	39	83	81	49
mn	43	0	42	0	18
mni	13	0	0	0	0
mnw	4	5	0	0	0
mr	63	20	0	0	33
my	37	18	0	29	16
ne	35	16	42	0	23
nn	87	14	83	101	59
oc	55	0	0	98	38
or	16	18	0	0	14
pa	34	24	29	36	25
pnb	28	9	38	38	16
ps	21	9	0	23	0
rm	6	0	0	0	0
rw	14	0	19	0	0
sh	67	20	62	100	52
si	51	16	32	24	16
sq	56	36	50	104	60
su	30	19	44	0	13
sw	44	14	47	74	31
ta	90	33	88	68	57
te	49	33	61	40	47

(Continued)

Table A2. Continued

Language	Climate change	COVID-19	Biology	History	Media
tg	30	9	0	0	23
tl	47	40	65	76	41
ur	57	37	58	99	42
war	34	0	0	48	17
zh-yue	58	54	62	59	49
am	0	6	0	0	0
ary	0	4	0	0	0
arz	0	23	32	79	38
awa	0	2	0	0	0
bjn	0	11	0	0	0
bo	0	4	0	0	0
br	0	13	0	90	41
bug	0	2	0	0	0
bxr	0	7	0	0	0
ceb	0	7	0	0	0
crh	0	3	0	0	0
dty	0	2	0	0	0
gn	0	6	0	0	0
gpe	0	1	0	0	0
ht	0	26	0	0	0
hyw	0	3	0	33	18
km	0	8	0	9	0
mg	0	7	0	0	0
min	0	17	0	0	0
pap	0	12	0	0	0
sco	0	13	31	31	24
shn	0	5	0	0	0
so	0	9	0	0	0
tk	0	6	0	0	0
ts	0	5	0	0	0
vo	0	3	0	0	12
wuu	0	15	0	0	0
xmf	0	10	0	0	17
yo	0	24	0	0	0
zh-min-nan	0	14	0	39	25
is	0	0	75	91	50
ky	0	0	21	20	12
pam	0	0	27	0	0
tt	0	0	26	57	14
azb	0	0	0	32	12
fo	0	0	0	26	16
fy	0	0	0	81	32
io	0	0	0	79	0
tum	0	0	0	2	0

(Continued)

Table A2. Continued

Language	Climate change	COVID-19	Biology	History	Media
an	0	0	0	0	35
frp	0	0	0	0	6
glk	0	0	0	0	3
lo	0	0	0	0	5
mai	0	0	0	0	8
nds	0	0	0	0	28
qu	0	0	0	0	15
sat	0	0	0	0	12
vec	0	0	0	0	19

Table A3. Number of revisions per language editions in the different datasets

Language	Climate change	COVID-19	Biology	History	Media
de	344,149	195,448	388,729	549,521	398,138
en	1,516,501	1,036,268	2,392,272	2,213,972	385,0137
es	251,085	142,608	412,920	417,275	447,902
fr	261,336	136,470	401,974	570,207	405,992
ja	59,235	37,293	97,684	117,733	222,585
pt	82,137	32,740	178,778	153,532	154,532
ru	90,492	56,512	230,846	348,758	257,507
ar	64,147	52,221	123,245	164,152	73,088
az	5,396	3,275	37,635	34,400	6,289
bg	14,688	4,606	50,747	77,809	24,812
bn	9,138	6,567	12,759	20,227	7,907
ca	43,861	13,878	219,014	168,055	45,265
cs	37,535	15,008	70,480	94,004	36,957
da	21,730	4,267	33,759	54,452	21,325
el	15,873	7,175	16,066	93,809	12,233
et	12,000	6,702	21,939	28,705	6,906
eu	12,238	2,086	82,623	45,030	6,865
fa	34,394	14,882	65,428	89,813	59,550
fi	41,846	10,081	76,296	70,988	58,117
he	45,760	36,839	84,560	153,974	74,220
hi	10,574	3,708	7,683	12,406	4,471
hr	11,084	2,842	21,316	40,023	9,546
hu	27,013	14,401	93,431	100,103	50,952
hy	5,149	4,177	10,735	29,745	12,491
id	23,373	14,547	111,305	55,091	32,762
it	102,487	46,926	204,053	375,945	347,881
ko	23,966	28,724	57,234	84,491	88,325
ms	6,906	7,698	1,1679	20,204	7,208
nb	36,680	10,026	81,352	78,627	40,317
nl	78,355	25,913	303,604	171,942	93,987

(Continued)

Table A3. Continued

Language	Climate change	COVID-19	Biology	History	Media
pl	45,284	22,540	160,392	173,581	116,880
ro	14,046	7,347	25,254	60,565	19,300
simple	30,367	8,317	31,239	33,583	23,031
sk	11,945	2,913	18,615	40,322	9,525
sl	12,320	2,265	12,780	39,772	5,602
sr	14,728	6,756	46,083	102,224	20,365
sv	49,949	11,079	264,477	114,709	71,292
th	12,084	11,217	22,152	36,162	30,098
tr	34,878	19,488	61,404	121,987	56,558
uk	33,040	55,024	105,302	102,155	45,483
uz	2,067	3,116	2,961	17,800	2,682
vi	22,628	66,920	407,504	45,861	30,883
zh	53,587	131,339	131,500	122,649	182,721
kk	0	517	6,564	5,187	1,523
af	6,384	2,827	12,553	20,980	3,448
als	1,320	0	0	4,863	0
as	814	466	1,025	590	270
ast	7,812	997	40,421	36,452	8,628
ba	1,316	50	0	2,004	449
bcl	1,036	182	439	1,082	158
be	3,892	1,473	10,090	23,364	2,899
be-x-old	2,385	518	0	12,592	1,722
bs	5,029	1,324	7,127	30,720	2,982
ckb	1,397	564	0	3,609	2,026
cv	860	0	0	0	746
cy	4,720	1,280	521,785	50,767	9,683
dag	790	0	0	0	0
eo	11,720	1,920	29,807	42,486	7,768
ga	2,570	1,830	4,293	8,020	2,404
gl	13,814	2,608	32,768	44,564	11,127
gu	1,191	311	0	0	157
ha	1,676	638	0	628	131
ia	1,437	265	0	0	0
ig	884	0	116	284	166
jv	2,325	492	4,949	12,716	1,271
ka	5,100	1,312	11,997	33,511	8,342
kn	2,643	1,366	2,085	1,482	1,593
ku	1,659	656	0	4,968	1,020
la	5,306	749	11,783	40,788	3,739
lb	1,925	0	0	0	2,999
lt	11,336	1,616	32,199	34,848	7,025
lv	7,180	4,234	13,517	12,402	5,800
mk	6,096	811	8,133	35,526	4,200
ml	6,461	2,373	14,983	7,912	3,545

(Continued)

Table A3. Continued

Language	Climate change	COVID-19	Biology	History	Media
mn	2,421	0	2,011	0	474
mni	385	0	0	0	0
mnw	160	217	0	0	0
mr	4,801	1,140	0	0	1494
my	1,469	1,358	0	1,317	430
ne	1,383	487	3,090	0	684
nn	8,312	1,801	10,620	20,234	4,729
oc	4,226	0	0	19,691	1,618
or	313	916	0	0	286
pa	1,609	825	1,498	2,143	832
pnb	1,076	100	4,884	4,351	290
ps	693	144	0	850	0
rm	148	0	0	0	0
rw	390	0	400	0	0
sh	4,436	619	13,554	28,448	5,352
si	2,728	421	1,249	869	469
sq	3,919	2,758	3,169	24,390	4,022
su	1,328	523	1,885	0	389
sw	3,057	416	2,273	13,290	1,027
ta	10,484	2,321	12,039	6,974	4,530
te	3,599	3,142	4,649	2,462	3,206
tg	1,414	331	0	0	637
tl	3,262	2,685	6,835	11,473	2,215
ur	4,132	2,055	3,941	18,574	2,528
war	1,436	0	0	7,500	419
zh-yue	3,844	6,256	6,081	7,762	3,438
am	0	115	0	0	0
ary	0	539	0	0	0
arz	0	1,221	12,611	18,006	10,561
awa	0	48	0	0	0
bjn	0	153	0	0	0
bo	0	57	0	0	0
br	0	530	0	22,230	2,022
bug	0	7	0	0	0
bxr	0	238	0	0	0
ceb	0	125	0	0	0
crh	0	50	0	0	0
dty	0	16	0	0	0
gn	0	89	0	0	0
gpe	0	31	0	0	0
ht	0	1,304	0	0	0
hyw	0	118	0	1,401	401
km	0	464	0	134	0

(Continued)

Table A3. Continued

Language	Climate change	COVID-19	Biology	History	Media
mg	0	138	0	0	0
min	0	518	0	0	0
pap	0	387	0	0	0
sco	0	223	1,636	1,609	689
shn	0	143	0	0	0
so	0	164	0	0	0
tk	0	95	0	0	0
ts	0	98	0	0	0
vo	0	33	0	0	412
wuu	0	341	0	0	0
xmf	0	370	0	0	342
yo	0	1,353	0	0	0
zh-min-nan	0	393	0	3,607	889
is	0	0	7,778	12,301	2,541
ky	0	0	1,317	705	297
pam	0	0	798	0	0
tt	0	0	2,568	11,215	459
azb	0	0	0	3,732	1,496
fo	0	0	0	1,252	307
fy	0	0	0	12,889	1,099
io	0	0	0	15,974	0
tum	0	0	0	16	0
an	0	0	0	0	1,325
frp	0	0	0	0	36
glk	0	0	0	0	61
lo	0	0	0	0	47
mai	0	0	0	0	104
nds	0	0	0	0	801
qu	0	0	0	0	435
sat	0	0	0	0	133
vec	0	0	0	0	569

Received 11 November 2024; revised 18 July 2025; accepted 26 October 2025