ISI Foundation

W[i]SCoM

Wikipedia Source Controversiality Metrics

Jacopo D'Ignazi, Andreas Kaltenbrunner, Yelena Mejova,
Kyriaki Kalimeri, Mariano Beiró, Michele Tizzani, Pablo Aragón

March, 2024

Motivation

- “There has been evidence that identifying non-reliable sources is an effective tool to **combat disinformation and increase the knowledge integrity of Wikipedia**”

- “This will facilitate information about source credibility to editors as well as readers and **allow the generation of perennial source lists in many other language editions**”

https://meta.wikimedia.org/wiki/Research:Wikipedia_Source_Controversiality_Metrics

Goals of the project

“The main goal is to **generate and assess actionable metrics** for source controversiality in Wikipedia.”

“To guarantee universality (i.e. applicability to all Wikipedia language editions), knowledge equity and avoid dependence on the specifics of a given language, **we will solely rely on language-agnostic approaches using mainly data from editing activity.** “

https://meta.wikimedia.org/wiki/Research:Wikipedia_Source_Controversiality_Metrics

English Wikipedia Perennial Sources List

Perennial sources

Source	Status (legend)	Discussions: List	Discussions: Last	Discussions: Summary	Uses
112 Ukraine	🚫	2019 2020 2020 1 A B	2020	112 Ukraine was deprecated following a 2019 RfC, which showed overwhelming consensus for the deprecation of a slew of sources associated with Russian disinformation in Ukraine. It was pointed out later in a 2020 RfC that 112 Ukraine had not been explicitly discussed in that first discussion prior to its blacklisting request. Further discussion established a rough consensus that the source is generally unreliable, but did not form a consensus for deprecation or blacklisting. The prior blacklisting was reversed as out of process. 112 Ukraine closed in 2021.	1 2
ABC News	✓	1 2	2021	There is consensus that ABC News, the news division of the American Broadcasting Company, is generally reliable. It is not to be confused with other publications of the same name.	1 2
Ad Fontes Media WP:ADFONTES	🚫	1 2 3 4 5	2021	There is consensus that Ad Fontes Media and their Media Bias Chart should not be used in article space in reference to sources' political leaning or reliability. Editors consider it a self-published source and have questioned its methodology.	1
Advameg (City-Data)	⊗ 🚫	2019 2019 2019 +14[c]	2019	Advameg operates content farms, including City-Data, that use scraped or improperly licensed content. These sites frequently republish content from Gale's encyclopedias; many editors can obtain access to Gale through The Wikipedia Library free of charge. Advameg's sites are on the Wikipedia spam blacklist, and links must be whitelisted before they can be used. WP:COPYLINK prohibits linking to copyright violations.	1 2 +43
The Age	✓	2021	2021	*The Age* is a newspaper based in Melbourne, Australia. There is consensus that it is generally reliable.	1
Agence France-Presse (AFP)	✓	1 2	2020	Agence France-Presse is a news agency. There is consensus that Agence France-Presse is generally reliable. Syndicated reports from Agence France-Presse that are published in other sources are also considered generally reliable.	1

- "**List of sources whose reliability and use on Wikipedia are frequently discussed**", classifying various domains as
  - generally reliable
  - generally unreliable
  - no consensus
  - blacklist and deprecated
- Since it is a well regarded and extensively reviewed list, **we will be using it as ground truth for the assessment of our metrics**.

https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

Previous work

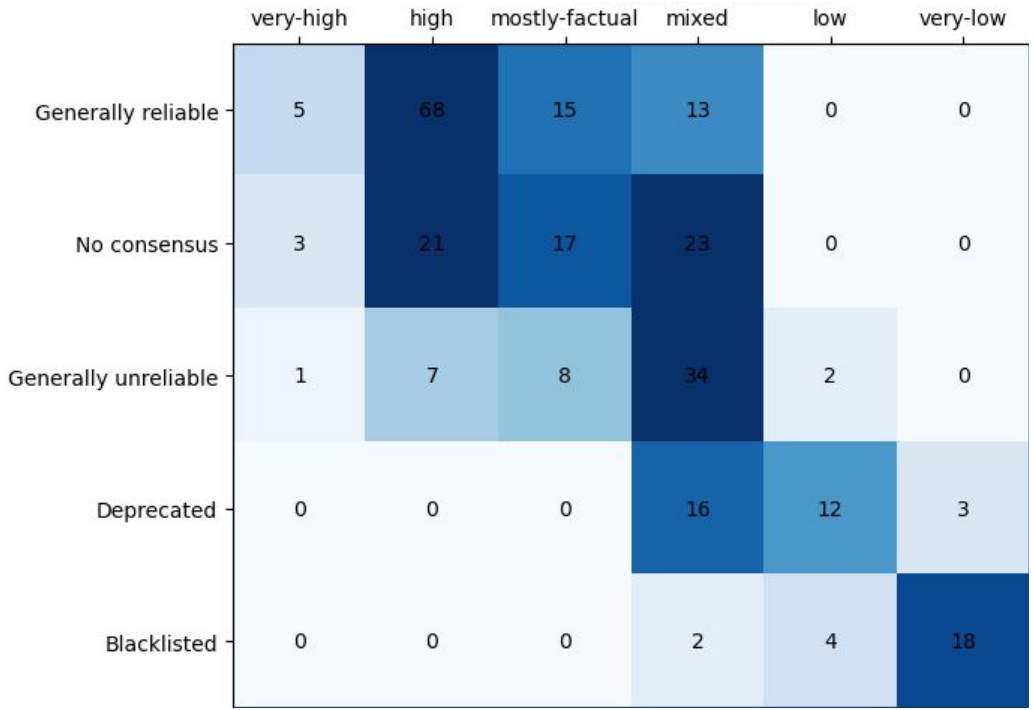
- Perennial sources lists have been found useful to maintain knowledge integrity on Wikipedia [Baigutanova et al 2023a, Baigutanova et al 2023b], but only basic metrics have been applied and only focusing on high resource languages.

- Different index have been proposed by Wikipedia community [iffy index], and outside of it, to classify domain reliability [medias bias/fact check], but none of them rely on Wikipedia editors' activity.

- Contropedia project developed a language-agnostic approach to assess controversiality of topics in Wikipedia pages [Contropedia], but does not look at sources.

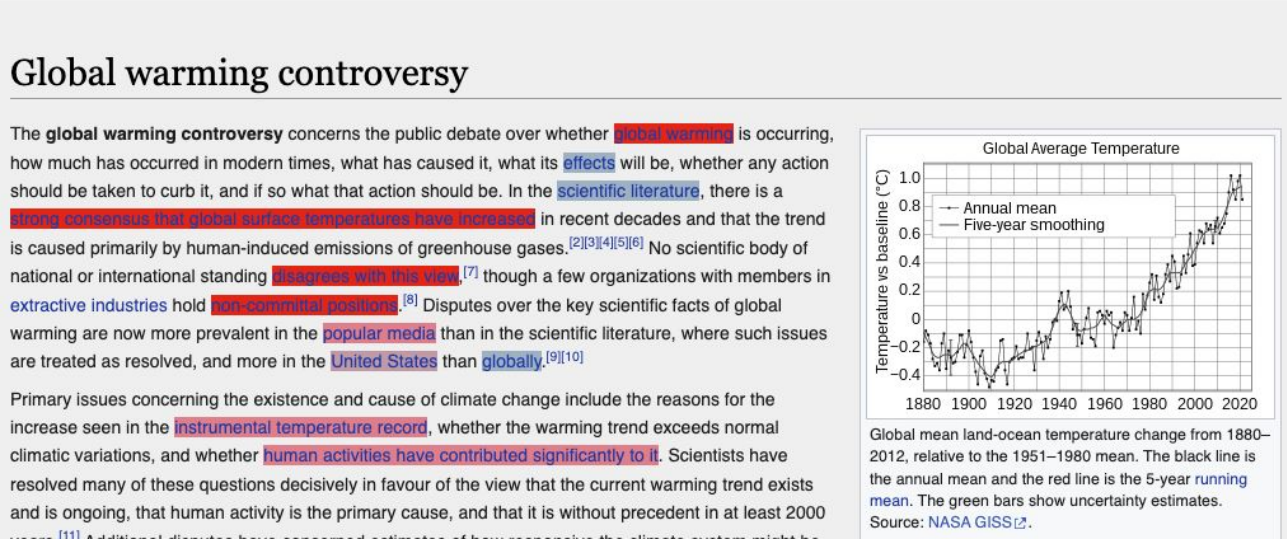
Previous work

Agreement between Perennial classification and MBFC labels



- “Generally reliable” classified mostly as high reliability by mbfc
- “Generally unreliable” have mismatched labels

Contropedia



Global warming controversy

The **global warming controversy** concerns the public debate over whether global warming is occurring, how much has occurred in modern times, what has caused it, what its effects will be, whether any action should be taken to curb it, and if so what that action should be. In the scientific literature, there is a strong consensus that global surface temperatures have increased in recent decades and that the trend is caused primarily by human-induced emissions of greenhouse gases.[2][3][4][5][6] No scientific body of national or international standing disagrees with this view,[7] though a few organizations with members in extractive industries hold non-committal positions.[8] Disputes over the key scientific facts of global warming are now more prevalent in the popular media than in the scientific literature, where such issues are treated as resolved, and more in the United States than globally.[9][10]

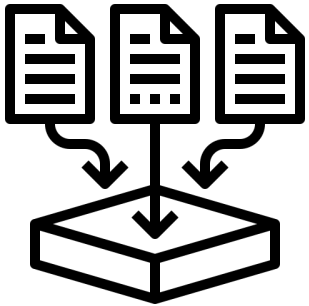
Primary issues concerning the existence and cause of climate change include the reasons for the increase seen in the instrumental temperature record, whether the warming trend exceeds normal climatic variations, and whether human activities have contributed significantly to it. Scientists have resolved many of these questions decisively in favour of the view that the current warming trend exists and is ongoing, that human activity is the primary cause, and that it is without precedent in at least 2000 years.[11] Additional disputes have concerned estimates of how responsive the climate system might be

Global mean land-ocean temperature change from 1880–2012, relative to the 1951–1980 mean. The black line is the annual mean and the red line is the 5-year running mean. The green bars show uncertainty estimates. Source: NASA GISS.

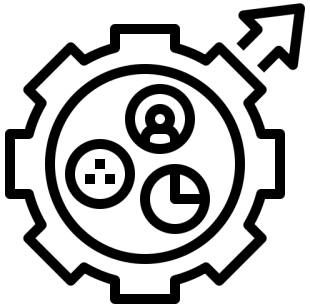
https://contropedia.net/#demo

Counting the amount of edits in the sentence containing a wikilink, it is measuring more (red) or less (blue) controversiality in relation to that topic

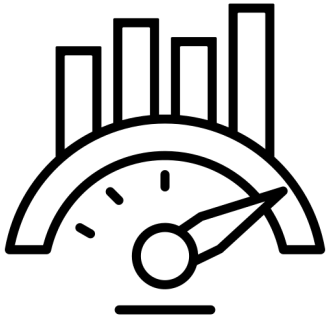
Project phases



Data Collection



Metrics development

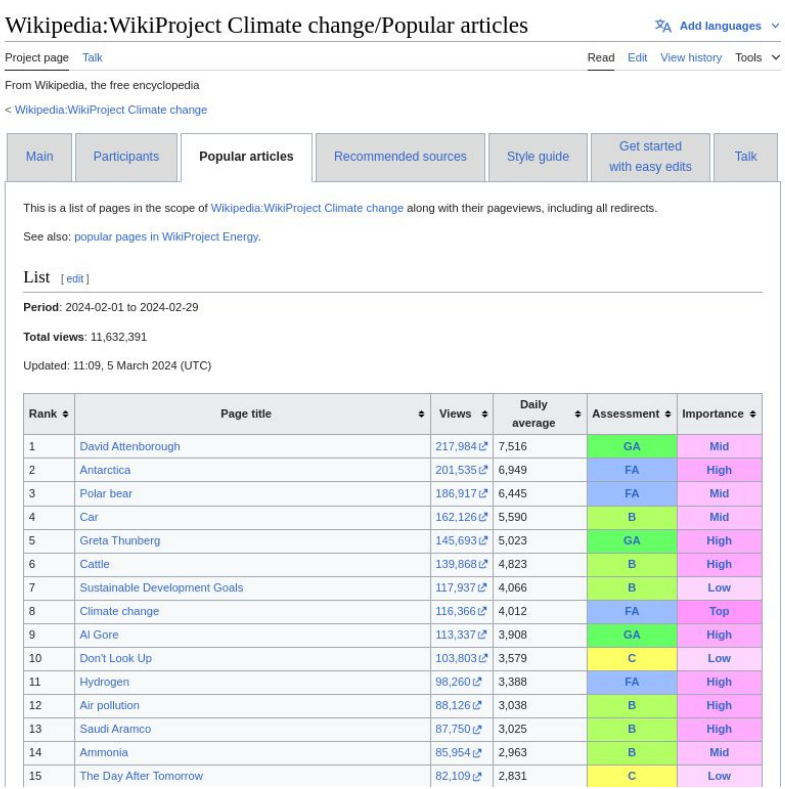


Metrics assessment

Data Collection: case study

- WikiProject related articles
  - WikiProject_Climate_change
  - WikiProject_COVID-19
- Other topics using ORES classification
  - Biology
  - History
  - Media

The main focus will be on climate change articles.



Wikipedia:WikiProject Climate change/Popular articles

Add languages

Project page Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

< Wikipedia:WikiProject Climate change

Main | Participants | **Popular articles** | Recommended sources | Style guide | Get started with easy edits | Talk

This is a list of pages in the scope of Wikipedia:WikiProject Climate change along with their pageviews, including all redirects.

See also: popular pages in WikiProject Energy.

List [edit]

Period: 2024-02-01 to 2024-02-29

Total views: 11,632,391

Updated: 11:09, 5 March 2024 (UTC)

Rank	Page title	Views	Daily average	Assessment	Importance
1	David Attenborough	217,984	7,516	GA	Mid
2	Antarctica	201,535	6,949	FA	High
3	Polar bear	186,917	6,445	FA	Mid
4	Car	162,126	5,590	B	Mid
5	Greta Thunberg	145,693	5,023	GA	High
6	Cattle	139,868	4,823	B	High
7	Sustainable Development Goals	117,937	4,066	B	Low
8	Climate change	116,366	4,012	FA	Top
9	Al Gore	113,337	3,908	GA	High
10	Don't Look Up	103,803	3,579	C	Low
11	Hydrogen	98,260	3,388	FA	High
12	Air pollution	88,126	3,038	B	High
13	Saudi Aramco	87,750	3,025	B	High
14	Ammonia	85,954	2,963	B	Mid
15	The Day After Tomorrow	82,109	2,831	C	Low

Data collection: process

23. ^ [a] [b] Lynas, Mark; Houlton, Benjamin Z.; Perry, Simon (19 October 2021). "Greater than 99% consensus on human caused climate change in the peer-reviewed scientific literature". *Environmental Research Letters*. **16** (11): 114005. Bibcode:2021ERL....16k4005L. doi:10.1088/1748-9326/ac2966. S2CID 239032360.

<ref name="EnvRschLtrs_20211019">{{cite journal |last1=Lynas |first1=Mark |last2=Houlton |first2=Benjamin Z. |last3=Perry |first3=Simon |title=Greater than 99% consensus on human caused climate change in the peer-reviewed scientific literature |journal=Environmental Research Letters |date=19 October 2021 |volume=16 |issue=11 |article=114005 |doi=10.1088/1748-9326/ac2966 |bibcode=2021ERL....16k4005L |s2cid=239032360 |url=https://iopscience.iop.org/article/10.1088/1748-9326/ac2966}}</ref>

For all the revisions of a given set of articles:

- Parse revision to **find source identifiers**
  - Inside templates (DOI, ISBN, URLS, …)
  - Unstructured URLs
- If the source identifier is a URL: extract domain

Data Collection: dataset

For each revision and for each page we extract:

- **Identifier** of any source added or removed from previous revision
- **Type of source identifier** and where it was found
  - Inside a template
  - In an unstructured form
- **Editor metadata:** username (IP if not registered).
- **Revision metadata:** id, timestamp, comment, etc.

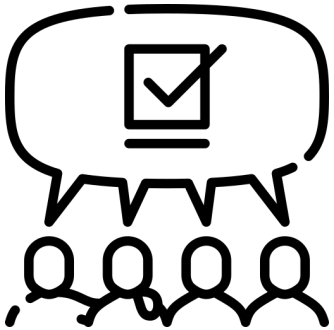
Data collection: dataset statistics

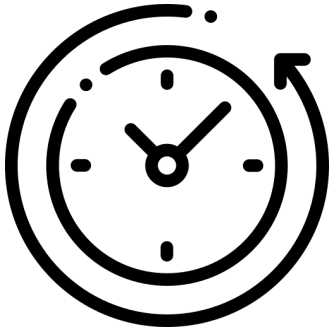
		#langs	#pages	#revisions	rev. per page	#urls	#domains	#Perennial dom.
English	Climate change	-	3,799	1,516,501	399.18	265,181	51,759	354
	COVID-19	-	2,543	1,036,268	407.50	259,246	24,966	342
	Biology	-	47,656	2,392,272	50.20	318,716	37,533	312
	History	-	12,395	2,213,972	178.62	225,630	37,442	325
	Media	-	17,634	3,850,137	218.34	478,478	80,391	410
Other languages	Climate change	265	24,123	2,385,559	98.89	323,789	75,570	359
	COVID-19	237	15,378	1,370,022	89.09	371,518	33,898	348
	Biology	266	227,027	6,118,689	26.95	503,128	60,988	305
	History	258	69,281	6,306,156	91.02	305,424	57,641	324
	Media	227	53,097	3,799,445	71.56	444,286	74,066	393

In the last column, there is the number of Perennial domain appearing in that set of pages. It is a small amount compared to all the domains -because Perennial list is indeed small- but we are using it as ground truth for our model

Feature engineering

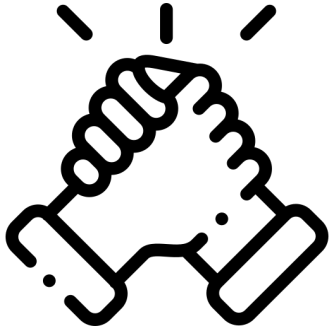
How popular is a domain?





How much time has a domain been used on an article?

How many users added or removed a domain?



Feature engineering

- Usage Statistics
  - In how many articles has that domain been used?
  - In how many articles is that domain currently used?
  - How many times has it been added and/or removed
  - …
- Permanence statistics
  - Age: how much time ago has it been added (n revisions/days)
  - Permanence: How much time has it been on a article (sum/avg/median, n revisions/days)
  - SelfPermanence: Ratio of time present since it was added: Permanence/Age (avg/median, n revisions/days)
  - …
- Editor-based
  - How many editor/registered users added/removed a domain from a article
  - Probability that the domain has been added/removed by a registered editor
  - …

Full information on features is available at: https://w.wiki/9STy

Modelling approach

- We are using an implementation of XG-Boost

- Learning to distinguish perennial sources
  generally reliable vs. generally unreliable domains (binary classification task)

- Balancing model
  weighting the model by size of positive (gen. reliable) and negative (gen. unreliable) examples

- Normalization strategies
  - metrics normalized by size of the dataset (to counteract extensive properties)
  - unnormalized metrics

Model evaluation

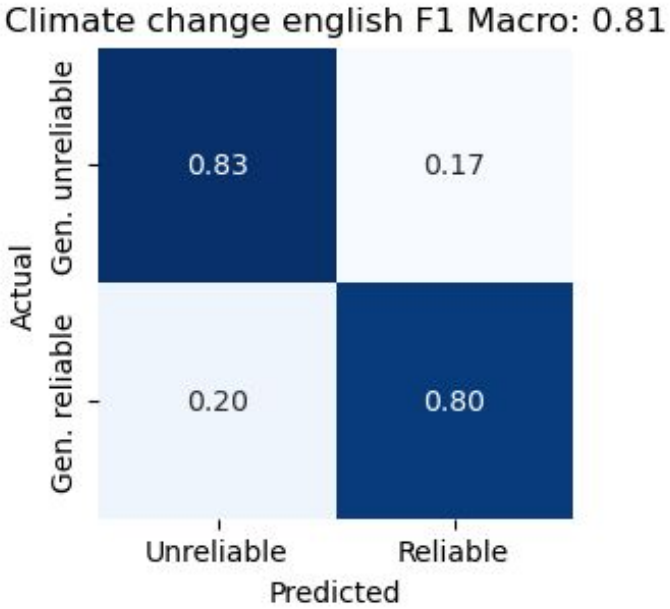
- Leave-one-out validation to assess model performances.

- F1 macro score, since we are interested in recognizing both positive (gen. reliable) and negative (gen. unreliable) classes.

- SHAP values to observe how our features “behave” in our task.

Results

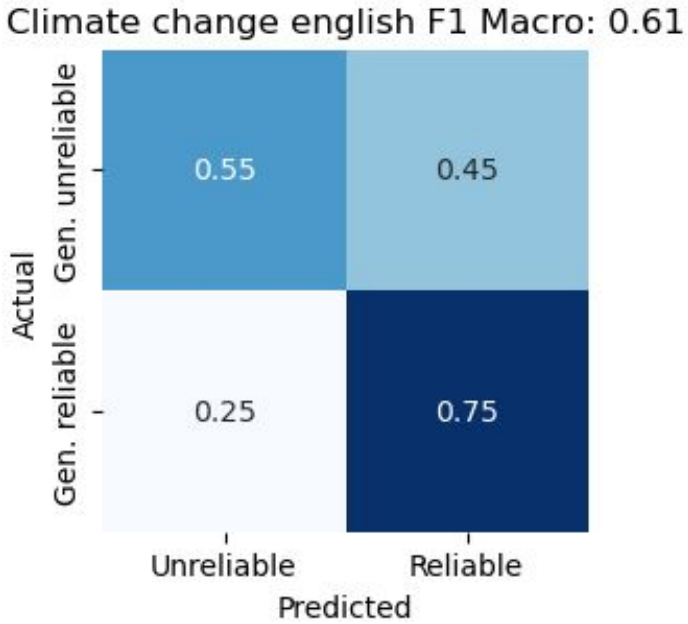
Performance metrics for climate change English model



Class	Precision	Recall	F1-Score	Support
Gen. unreliable	0.83	0.83	0.83	152
Gen. reliable	0.80	0.80	0.80	129
Accuracy			0.81	281
Macro Avg	0.81	0.81	0.81	281
Weighted Avg	0.81	0.81	0.81	281

Results

Performance metrics for climate change English model (MBFC as target)

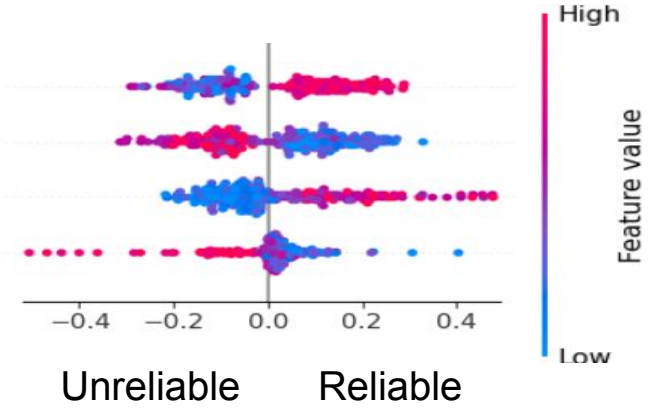


Class	Precision	Recall	F1-Score	Support
Gen. unreliable	0.32	0.55	0.40	239
Gen. reliable	0.89	0.75	0.82	1139
Accuracy			0.72	1378
Macro Avg	0.60	0.65	0.61	1378
Weighted Avg	0.79	0.72	0.74	1378

Results

Climate change articles on English Wikipedia: Top 4 most predictive features

When it is added, does it stay on the article?
Has it been removed by a registered editor?
How many revisions has it been on a article?
How much time ago has it been added?



- SelfPermanence: If a domain is added and not removed, it is a good indication that the domain is reliable.
- Prob. registered editor end: If a domain is removed by a registered editor, it is a good indication that the domain is unreliable
- …

Results

Youtube.com: an example of true negative on Climate change english pages

When added, has it stayed on the page?

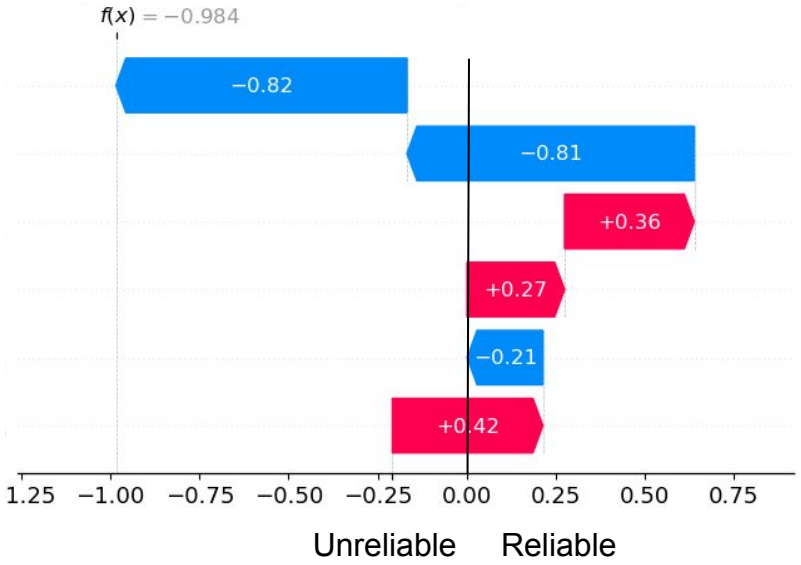
On average, how many rev. per page?

Is it currently used?

Has it been added recently?

How many days has it been visible?

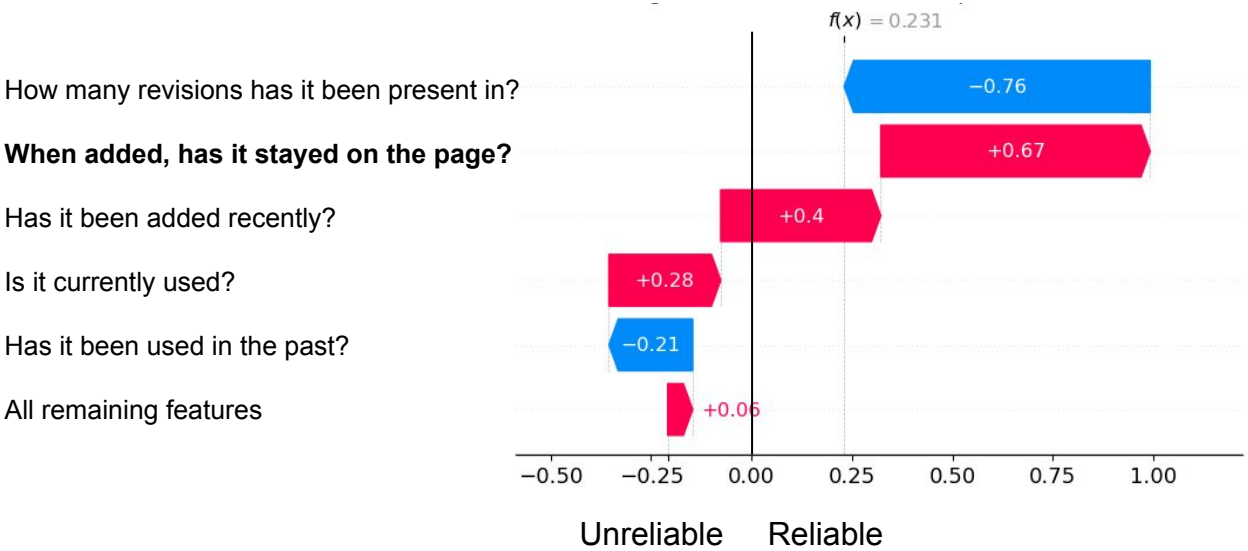
All remaining features



Despite being used on current Wikipedia pages, and being added frequently, the model correctly classify Youtube as unreliable because **when added it is generally removed soon after (low SelfPermanence and low average Permanence)**

Results

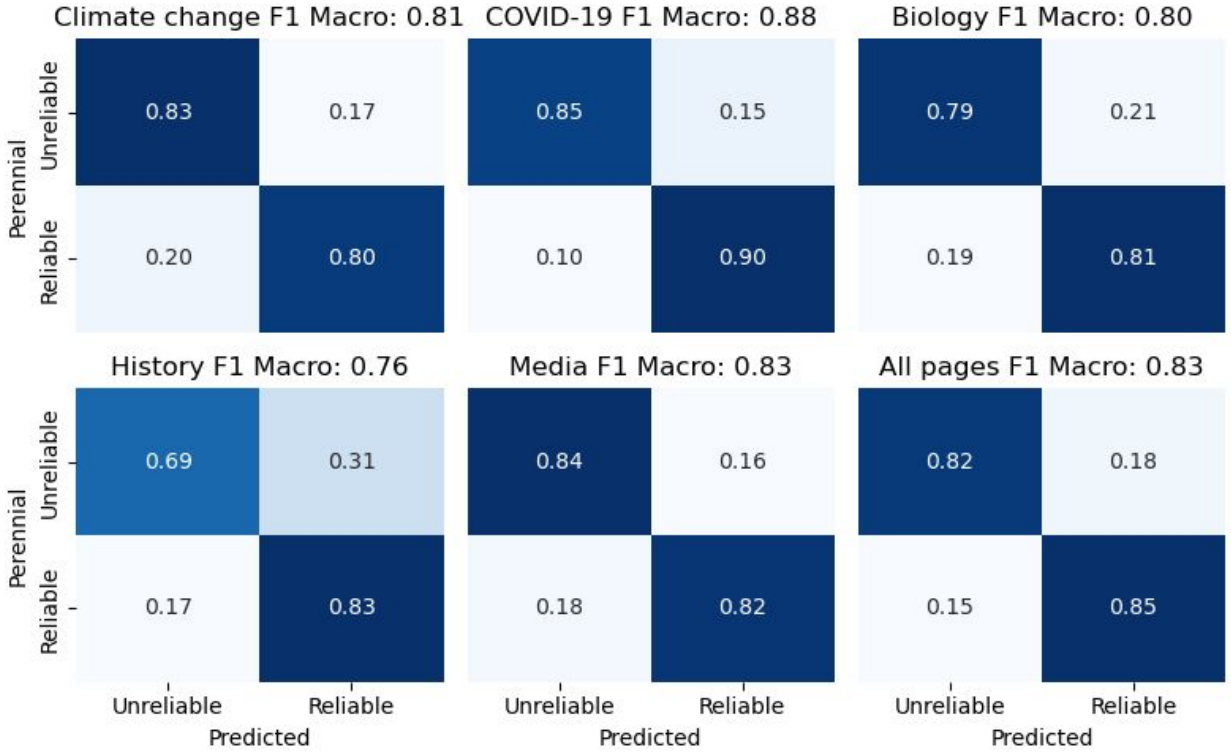
Researchgate.net: an example of false positive on Climate change english pages



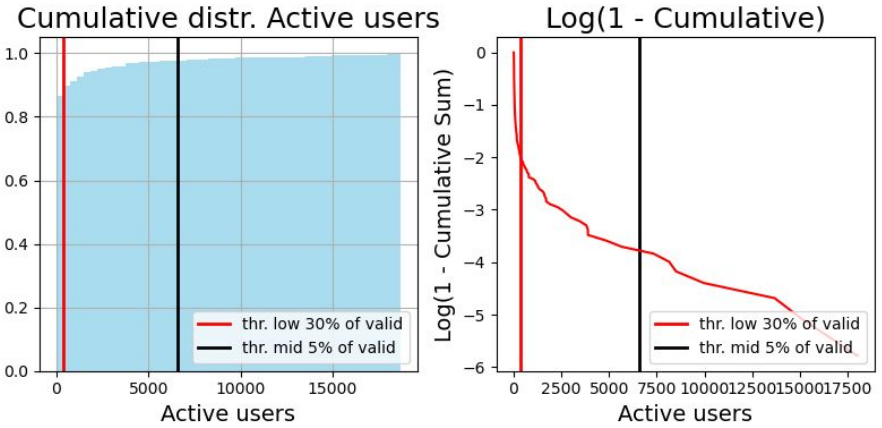
When the model fails to predict the correct label it could be an indication of domain misuse, e.g., researchgate.net (generally unreliable) is predicted as reliable because **when added it is not removed (high SelfPermanence)**.

Results

Testing this approach on other topics



Modelling other languages

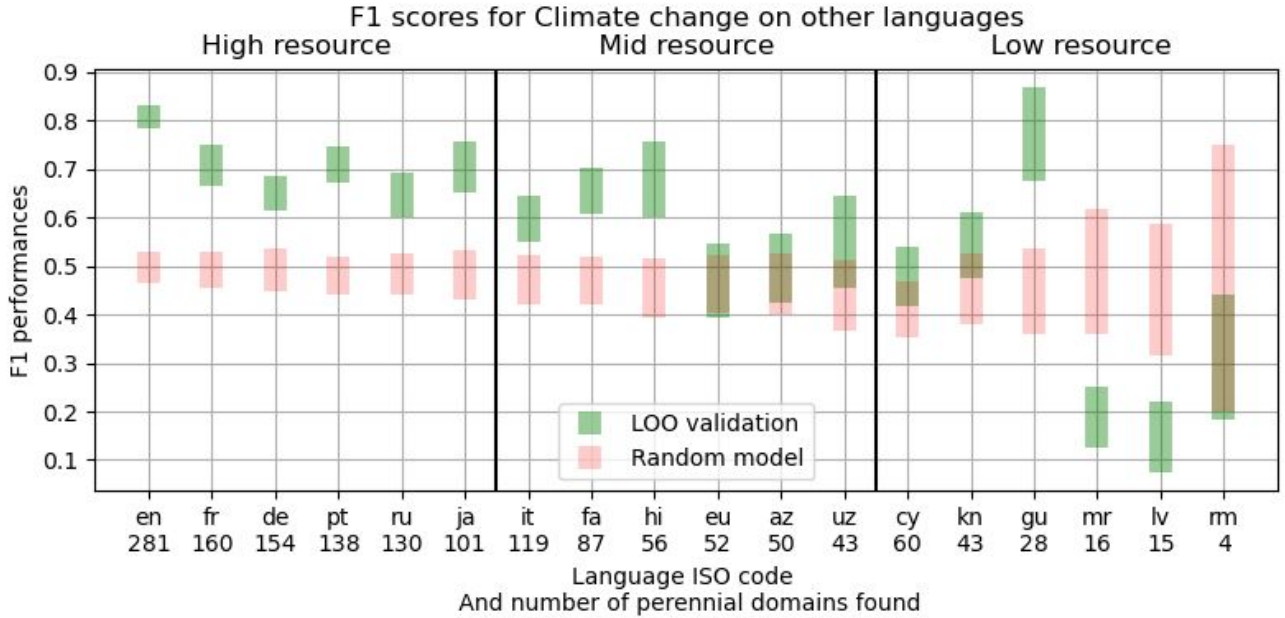


* English is omitted in these graphs

- Ranking languages by "total active users" as a measure of resource* of a language Wiki
- Remove languages with less than 2 domains from reliable and unreliable Perennial Sources
- Distinguish between
  - High resource language: Top 5% (7 lang.)
  - Mid resource language: following 25% as mid,
  - Low resource language: the remaining 70%

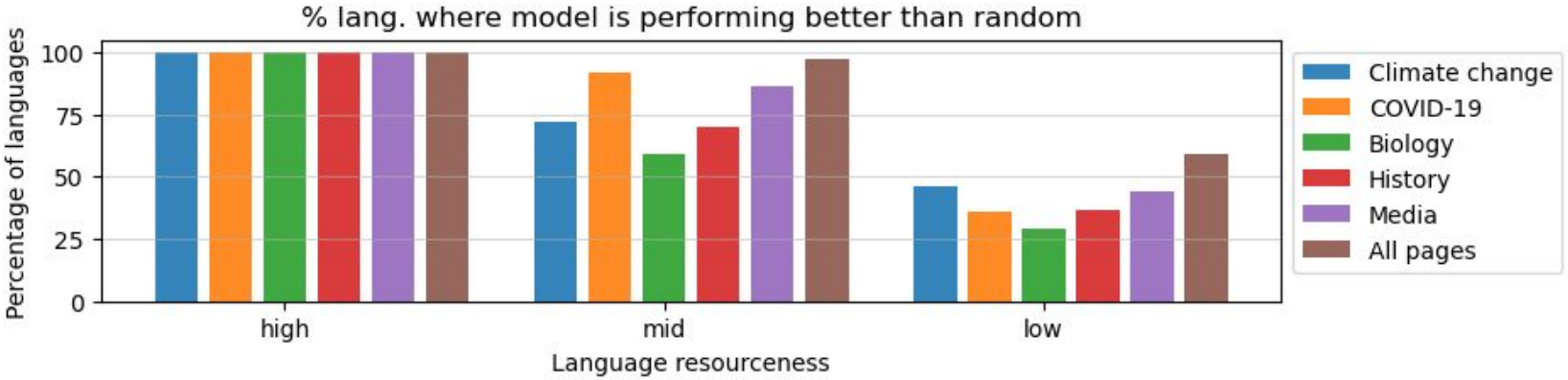
Results

Climate change articles on multiple language editions of Wikipedia



Evaluation

Articles from multiple topics on multiple language editions of Wikipedia

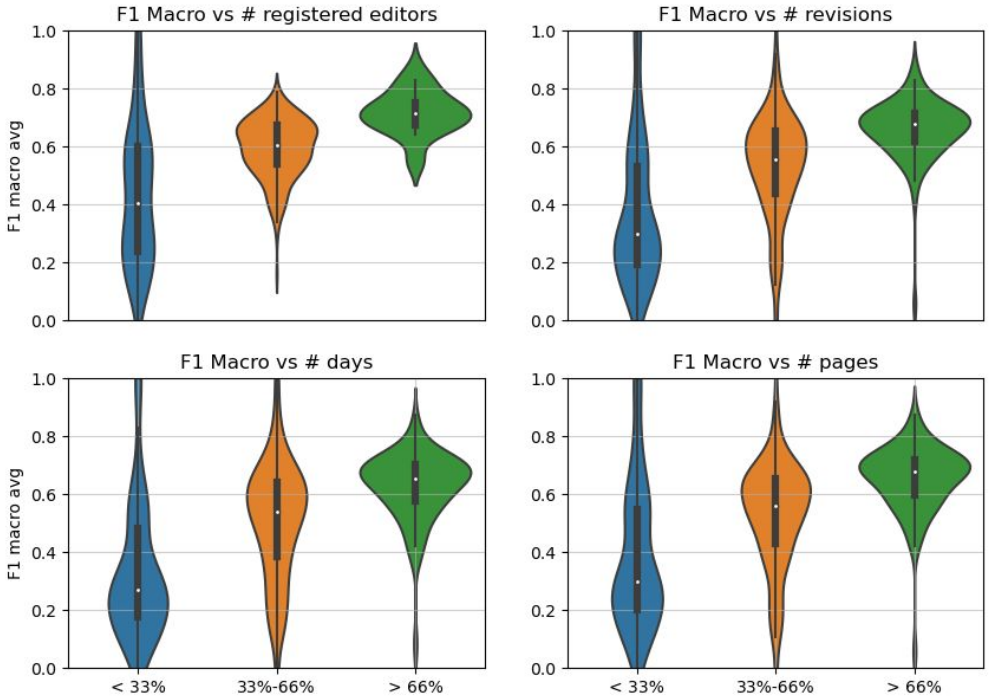


Using Mann-Whitney test for comparison against random classifier, then measuring the percentage of language where the model performs better

- Model always works on high resource languages, for every topic
- Capacity to distinguish reliable from unreliable sources decreases on lower resource languages
- Considering all pages together, the model is performing better

Evaluation

Relation between F1 macro score and dataset size

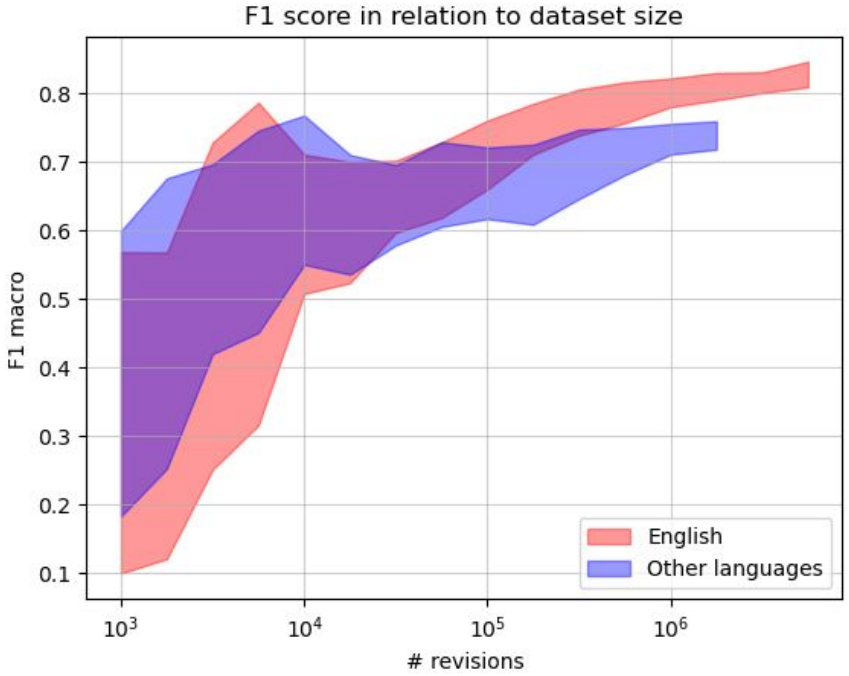


- Considering all our datasets, <u>better overall performance for high resource languages</u>.
  - This is true for every metric of extensiveness of data

- <u>There could be an effect of less attention on low resource languages to quality of sources</u>, although it is difficult to distinguish this effect from the effect of dataset size

Evaluation

Relation between F1 macro score and dataset size

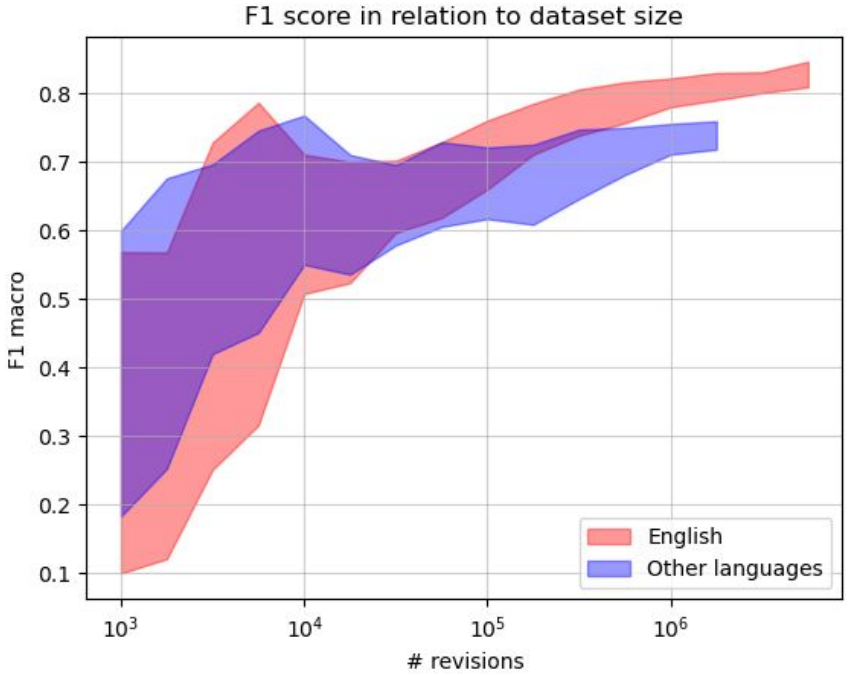


- Red area: extracting subsamples of english pages with increasing number of pages. Area is average and standard deviation of F1 macro for the model tested on that sample

- Blue area: average and standard deviation of F1 macro for languages where datasets have given amount of revisions

Evaluation

Relation between F1 macro score and dataset size

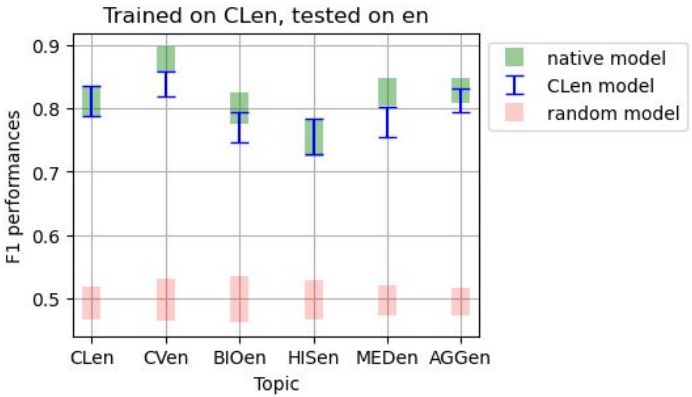


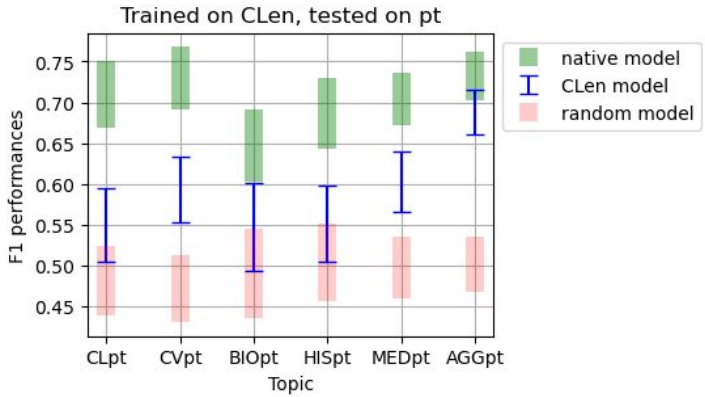
- Strong interplay between dataset size and F1 macro, either in English and other languages. An amount of around 10^5 revisions suggested to have stable results

- Once the model reaches “stable regime”, other effect come into play. Effect of attention to quality of sources (Perennial list in particular) that can be different in different wikis

Evaluation

Is it possible to train on one language-topic and apply to another?

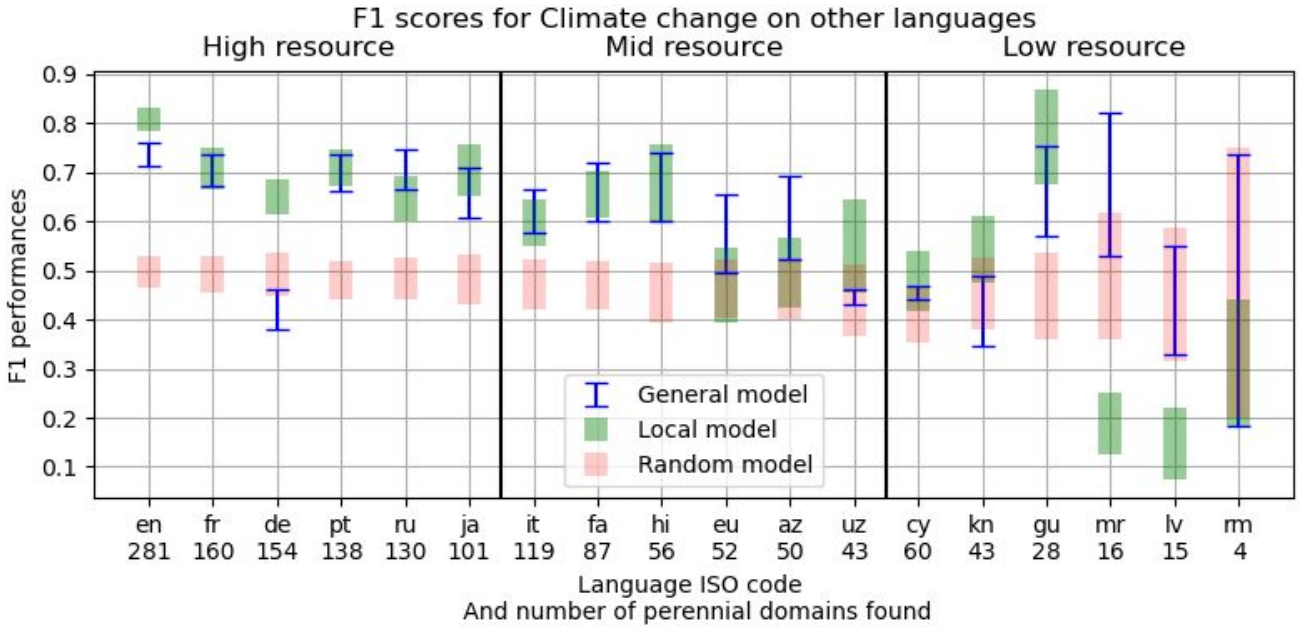




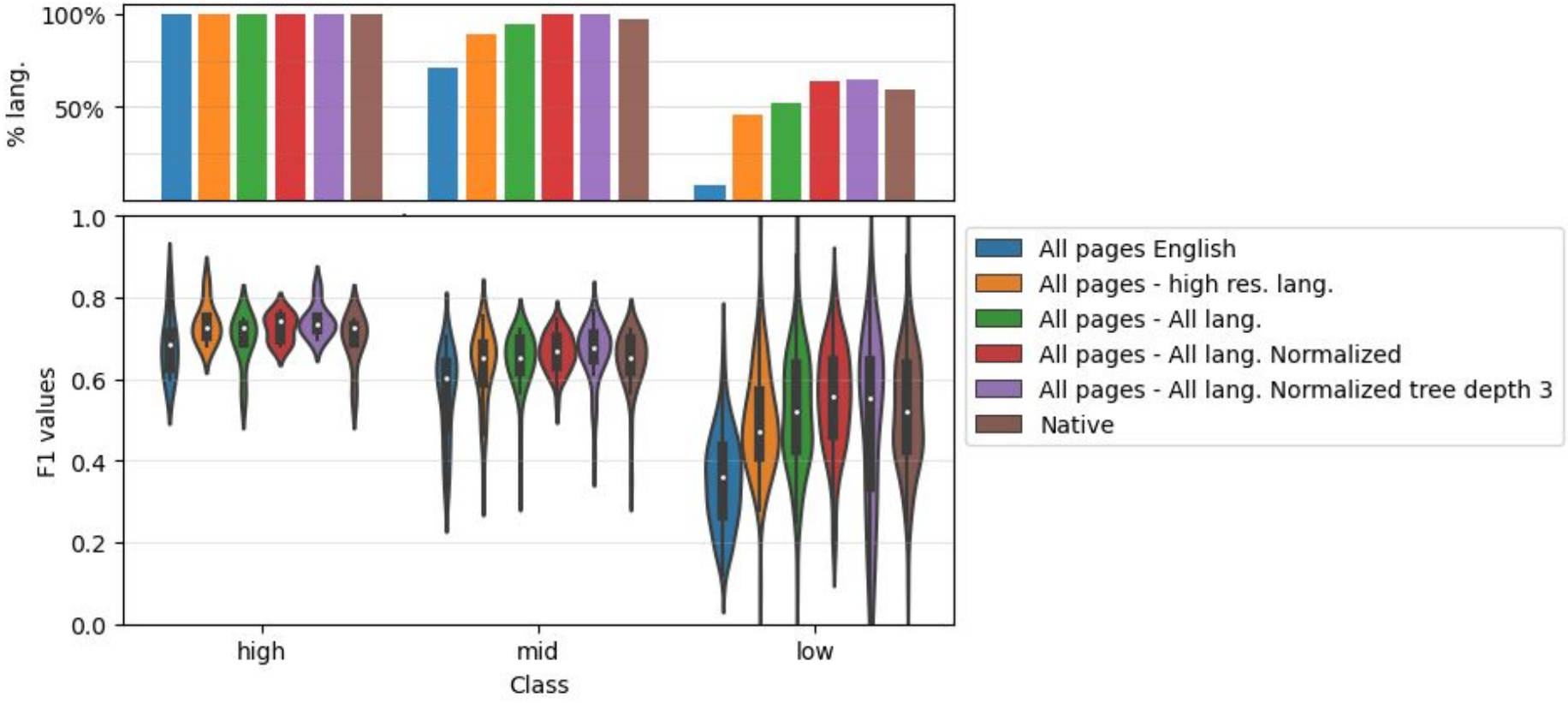
- Model is quite stable when training on a topic of English Wikipedia and testing on another topic of English Wikipedia
- Cross-Language behaviour is instead less reliable due to specificity of the editor's behaviour on different languages

Evaluation

What are the performances if we train the model on all languages?



Evaluation (general model)



Summary of findings

- The model capacity to distinguish reliable from unreliable domains is dependent on quantity of resources/activity in a context and on attention to Perennial sources in that context.

- The model works better on high resource languages, where each metric can tell insightful information about editors' behaviour with regard to perennial classified domains.

- The possibility of developing one only model to apply on all languages is compromised by low cross-language adaptability, although normalization and training on a general dataset give promising results

- The possibility of extending the model to all articles on each language Wiki is promising, since performances increase with resources.

Future work

- Can we improve model performances when training on all Wiki pages?
- **Can we use the model to assess source quality in different contexts?**
- How does model behave outside of Perennial sources?
- **Can we use the model to extend Perennial list?**
- Produce reusable code and documentation for further research

References

Baigutanova, A., Myung, J., Saez-Trumper, D., Chou, A. J., Redi, M., Jung, C., & Cha, M. (2023, April). Longitudinal assessment of reference quality on wikipedia. In Proceedings of the ACM Web Conference 2023 (pp. 2831-2839).

Baigutanova, A., Saez-Trumper, D., Redi, M., Cha, M., & Aragón, P. (2023, October). 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 (pp. 3743-3747).

Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., ... & Venturini, T. (2015, April). Societal controversies in Wikipedia articles. In Proceedings of the 33rd annual ACM conference on human factors in computing systems (pp. 193-196)

Thanks!

Contact:
jacopo.dignazi@isi.it
kaltenbrunner@gmail.com