Wiki Workshop (11th edition) – June 20, 2024

WiSCoM: Wikipedia Source Controversiality Metrics

Jacopo D'Ignazi[1], **Andreas Kaltenbrunner**[1,2,5], **Yelena Mejova**[1], **Michele Tizzani**[1], **Kyriaki Kalimeri**[1], **Mariano Beiró**[3,4], **Pablo Aragón**[5]
[1]ISI Foundation, Turin, Italy
[2]Universitat Oberta de Catalunya, Barcelona, Spain
[3]Departamento de Ingeniería, Universidad de San Andrés, Buenos Aires, Argentina
[4] CONICET, Buenos Aires, Argentina
[5]Universitat Pompeu Fabra, Barcelona, Spain

Abstract

All Wikipedia content must be verifiable through reliable sources. To assist Wikipedia editors in complying with this core content policy, we present a language-agnostic model for source controversiality using features from editorial activity. The model is trained within Wikipedia articles from different topics and evaluated against the English Wikipedia perennial source list. We achieve an F1 Macro score of around 0.8 on the English Wikipedia and in the range of 0.65 to 0.75 in other mid- and high-resource languages. We find that the permanence of a domain in an article is one of the most predictive features. Although the model deteriorates in mid- and low-resource languages, we also show that model adaptation from other higher-resource languages can boost its performance.

Keywords: knowledge integrity, knowledge equity, source reliability, source reliability, perennial sources

Introduction

The spread of disinformation is one of the main threats to knowledge integrity on Wikipedia (Aragón and Sáez-Trumper, 2021). Combating this problem has become a relevant but challenging task for volunteers. A common patrolling technique is to detect and remove statements that violate basic core content policies. In contrast to this approach, some Wikipedia editors have proven more effective in combating misinformation by first identifying unreliable sources (Cohen, 2021). Recent research revealed the positive impact of the community-curated list of perennial sources in English Wikipedia (Baigutanova et al., 2023a). However, such a list is very limited or even missing in other language editions (Baigutanova et al., 2023b).

To address this challenge, we present our research on source controversiality scoring, based on language-agnostic approaches and edit activity data from multiple Wikipedia language editions. This resource could assist Wikipedia editors, including the ones in small projects that often miss advanced tools, in identifying sources with patterns associated with low reliability and monitoring their spread across language editions.

Our research is built upon the existing work of the Contropedia project (Borra et al., 2015), which already demonstrated the potential of language-agnostic approaches to measure the controversiality of wikilinks in a given article. Here, instead, we focus on the domains of references across multiple articles and language editions.

Methods

In this study, we create five topical datasets: Climate Change, COVID-19, Biology, History, and Media. The first two correspond to articles maintained by topic-based WikiProjects and the other three are based on predicted topics from the ORES scoring system (Halfaker and Geiger, 2020). We retrieve all revisions from these articles in the English Wikipedia and the corresponding articles in other language editions. Then, we extract the references added or removed through these edits, keeping only those with an URL. Using edit metadata, we define 56 features capturing

- the permanence of domains on articles,
- how widespread domains are used,
- the number of users interacting with domains and their registration status.

We explore different ways of normalizing these statistics w.r.t. the first appearance of the domain, age of the whole dataset, in terms of time duration or number of revisions, etc. For example, the most predictive feature is *self-permanence* that reflects the length of time that a URL domain has been present on articles, divided by the time length since the domain first appeared on those articles. This is then averaged over all articles on which the domain has appeared.

Finally, for all domains in each topic/language-specific dataset, we leverage these features to train an XGBoost classifier to predict the source reliability using the English Wikipedia perennial sources list as ground truth. We limit our analysis to the 99 language editions having at least two reliable or unreliable sources in this list.

Results

We test our modelling approach using the Macro F1 score computed on leave-one-out cross-validation. First, we train our model for each topical dataset in English Wikipedia. F1 macro scores of the resulting models lie between 0.75 (History) and 0.88 (COVID-19), as shown in Table 1. Precision for the unreliable class is always above 0.80 for all the topics, and recall has a similar range, except for History.

Figure 1a shows the distribution of SHAP values of the most predictive features. We observe that domains with a higher probability $Proba(R_{end})$ to be removed by an editor are associated with the unreliable label (closer to -1), whereas a higher value of the *self-permanence* feature $\langle SelfPerm_d \rangle$ is associated with the reliable label. Figure 1b shows the contribution of the features to the single prediction for the domain `researchgate.net`, which is incorrectly classified as reliable. This is largely because of its high $\langle SelfPerm_d \rangle$ and that it has been added to many pages over an extended time period (Σage_r). Thus, this example shows an interesting mismatch between the reliability label given to a domain and the actual behaviour of editors with respect to that domain.

Next, we train the model in the datasets of other language editions, observing a progressive decay in performances in languages with fewer revisions. The blue line and the standard deviation (shaded) in Figure 2 show the performance of each of the non-English language models. For comparison, we sampled the English dataset with different sample sizes (shown in red), while for other languages all available data was used. We find that for both English and non-English models, the size of the dataset has a strong relationship with the performance of the model, with the increase in F1 macro starting to slow down at 10^5 revisions. Still, for the largest datasets, English tends to perform better, possibly due to the fact that the version of the perennial sources we use as ground truth was originally compiled for the English Wikipedia.

Finally, we consider all topics together and analyse the performance of our model when trained with different training strategies and tested on each language. In Figure 3 we aggregate the corresponding results by languages of similar resourcefulness: the top panel shows for how many languages the different models perform significantly better than a random model (according to a Mann-Whitney test), while the bottom panel shows the distribution of F1 macro for all languages. We find that the F1 scores are generally lower when training a model on the English dataset (blue bars and violins) and applying it to low-resource languages, suggesting that cross-language adaptation can be problematic. However, when training with all languages together (green), the performance for low-resource languages significantly increases. If we normalize each dataset using quantile normalization (red), the general model outperforms individual models trained specifically for every language (i.e., native models shown in purple).

Discussion/Conclusions

We have shown that the proposed controversiality features can capture source reliability for different topics and languages. Our results suggest that this language-agnostic approach could be applied to expand the list of perennial sources in English Wikipedia and other Wikipedias. As we have found that model performance is closely related to the amount of user activity data, we are interested in analyzing all articles in multiple language editions, especially in low-resource ones. Future work should also examine how Wikimedia community interventions (e.g., the introduction of perennial sources in 2018) affect model performance.

References

[Aragón and Sáez-Trumper2021] Pablo Aragón and Diego Sáez-Trumper. 2021. A preliminary approach to knowledge integrity risk assessment in Wikipedia projects. *arXiv preprint arXiv:2106.15940*.

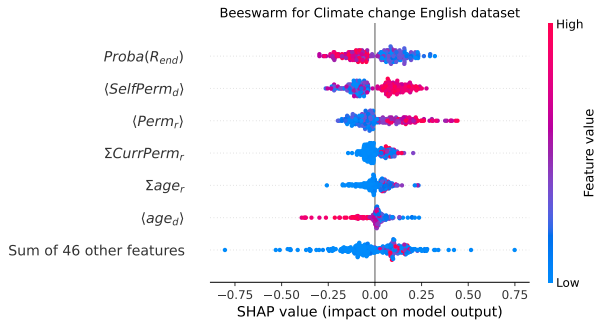
[Baigutanova et al.2023a] Aitolkyn Baigutanova, Jaehyeon Myung, Diego Saez-Trumper, Ai-Jou Chou, Miriam Redi, Changwook Jung, and Meeyoung Cha. 2023a. Longitudinal Assessment of Reference Quality on Wikipedia. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

[Baigutanova et al.2023b] Aitolkyn Baigutanova, Diego Saez-Trumper, Miriam Redi, Meeyoung Cha, and Pablo Aragón. 2023b. 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*, CIKM '23, page 3743–3747. ACM.

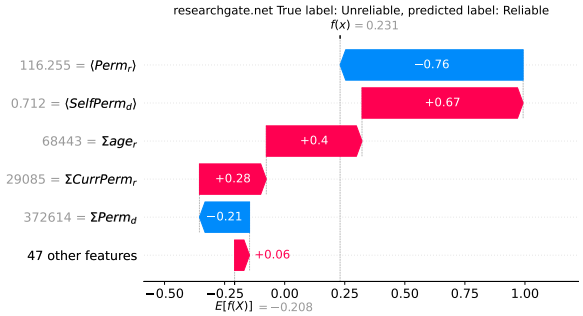
[Borra et al.2015] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal Controversies in Wikipedia Articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 193–196. ACM.

[Cohen2021] Noam Cohen. 2021. One Woman's Mission to Rewrite Nazi History on Wikipedia. WIRED. [Online; accessed 15-Apr-2024].

[Halfaker and Geiger2020] 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):1–37.



(a) Beeswarm plot of the SHAP values distribution for the most predictive features.



(b) Waterfall plot of the SHAP values for a single prediction of the domain `researchgate.net` .

Figure 1: SHAP values distribution (left) and single prediction example (right) for the English climate change dataset.

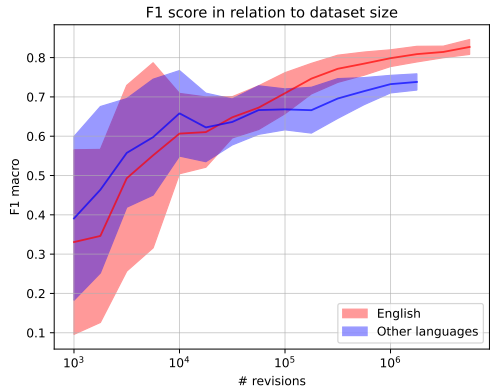


Figure 2: Macro F1 model performance score of models trained (all topics combined) in English (red) versus those of other language editions (blue) for different sizes of collected data. For English, data was sampled with different sizes; for other languages, all available data was used.

Table 1: Leave-one-out validation scores for the English model trained on each topic individually, or for all topics combined (last row). Precision and recall scores are calculated for the unreliable domains, according to their perennial classification.

Topic	F1 Macro	Precision	Recall
Climate change	0.81	0.83	0.83
COVID-19	0.88	0.89	0.85
Biology	0.80	0.80	0.80
History	0.75	0.82	0.69
Media	0.83	0.85	0.84
All topics	0.83	0.89	0.82

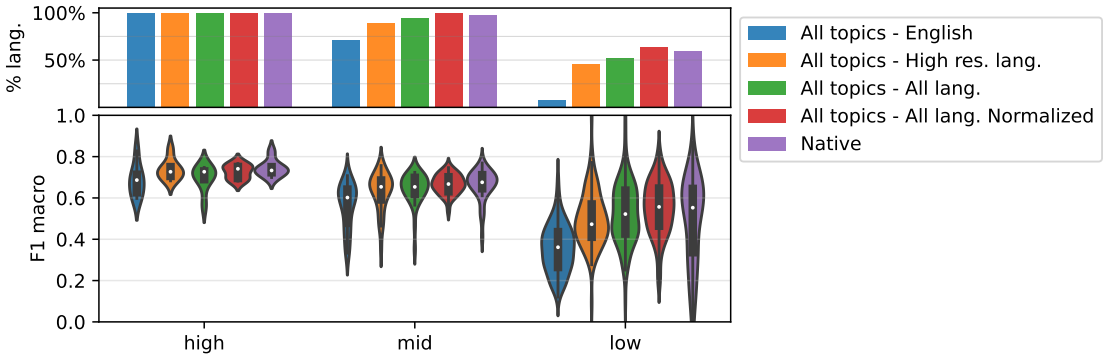


Figure 3: Model performances in terms of % of languages in each group (high-, mid-, and low-resource languages) for which our models perform statistically better than a random baseline (top panel), and the distributions of F1 macro scores in each language (bottom panel). Colours represent different training strategies of a "general" model.