Language-agnostic Topic Classification for Wikipedia

Isaac Johnson
isaac@wikimedia.org
Wikimedia Foundation
United States

Martin Gerlach
mgerlach@wikimedia.org
Wikimedia Foundation
United States

Diego Sáez-Trumper
diego@wikimedia.org
Wikimedia Foundation
United States

ABSTRACT

A major challenge for many analyses of Wikipedia dynamics—e.g., imbalances in content quality, geographic differences in what content is popular, what types of articles attract more editor discussion—is grouping the very diverse range of Wikipedia articles into coherent, consistent topics. This problem has been addressed using various approaches based on Wikipedia’s category network, WikiProjects, and external taxonomies. However, these approaches have always been limited in their coverage: typically, only a small subset of articles can be classified, or the method cannot be applied across (the more than 300) languages on Wikipedia. In this paper, we propose a language-agnostic approach based on the links in an article for classifying articles into a taxonomy of topics that can be easily applied to (almost) any language and article on Wikipedia. We show that it matches the performance of a language-dependent approach while being simpler and having much greater coverage.

CCS CONCEPTS
• Human-centered computing → Empirical studies in collaborative and social computing.

KEYWORDS
Wikipedia, language-agnostic, topic classification

ACM Reference Format:
Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. 2021. Language-agnostic Topic Classification for Wikipedia. In Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442442.3452347

1 INTRODUCTION

As of January 2021, Wikipedia has over 300 language editions with 55.7 million articles\(^1\) about 20.4 million distinct entities\(^2\) and an additional 250 thousand articles created every month.\(^3\) These articles cover a very wide range of content and it can be difficult to track and understand these dynamics—e.g., what types of content are attracting the interest of editors or readers?—especially across the different language editions.

Wikipedia itself has a number of editor-curated annotation systems that bring some order to all of this content. Most directly perhaps is the category network, but content can also be categorized based on properties stored in Wikidata, tagging by WikiProjects (groups of editors who focus on a specific topic), or inclusion of templates such as infoboxes. While these annotation systems are quite powerful and extensive, they ultimately are human-generated and semi-structured and thus have many edge cases and under-coverage in languages or communities that do not have editors who can maintain these annotations (see [10]). Researchers have developed many approaches to improve these annotation systems by using Wikipedia’s category network [17, 22], WikiProjects [2, 30], depending on DBPedia’s manually-curated taxonomy that then assigns topics based on infobox templates [17], or throwing out the editor-based annotation systems completely and learning a set number of topics through unsupervised techniques such as topic modeling [18, 20, 23, 27].

However, these approaches generally suffer from two limitations related to coverage in terms of language and article. First, not all language communities have the editor base or need to maintain these annotations to the same degree. For example, Arabic Wikipedia has the most categories per article at 25.5, and English Wikipedia with its approximately 40,000 monthly active editors,\(^4\) has 1.5M categories that are collectively applied 66M times across its 6.2 million articles.\(^5\) Wu Chinese Wikipedia, however, with approximately 20 active editors and 41,231 articles only has 6,990 categories that are applied 26,396 times (leaving many articles with no categories at all). Similar variation is seen with other systems as well: linkage to Wikidata is higher but still as low as 87.5% in Cebuano Wikipedia with its 5.5M articles,\(^6\) only 92 Wikipedia languages have a page describing WikiProjects,\(^7\) and many articles lack infoboxes.\(^8\) Second, approaches that seek to expand article coverage by also predicting topics for non-annotated articles depend on hand-labeling of topics (which requires language expertise) [20, 23, 27] or language modeling that does not easily scale to all languages on Wikipedia [2].

In this paper, we make the following contributions:

- We present an approach to automatically labeling (almost) all Wikipedia articles across every language of Wikipedia with a consistent set of topics. Specifically, we build on work

---

\(^{1}\)https://wikistats.wikimedia.org/display.php?f=wp

\(^{2}\)Personal calculation based on 4 January 2021 Wikidata JSON dump: https://dumps.wikimedia.org/wikidatawiki/entities/20210104/

\(^{3}\)http://stats.wikimedia.org/#/all-wikipedia-projects/contributing/new-pages/normal[bar][2-year][total][monthly]

\(^{4}\)https://stats.wikimedia.org/#/en.wikipedia.org/contributing/active-editors/normal[2-year][page_type=content]non-content[monthly]

\(^{5}\)Personal calculations using the December 2020 dumps and categorylinks table (https://www.mediawiki.org/wiki/Manual:Categorylinks_table) filtered by page table to articles in namespace 0 (https://www.mediawiki.org/wiki/Manual/Page_Table)

\(^{6}\)https://wikidata-analytics.wikimedia.org/app/WD_percentUsageDashboard

\(^{7}\)https://www.wikidata.org/wiki/Q4234303

\(^{8}\)Only about one-third of English Wikipedia articles have infoboxes per DBPedia’s statistics: https://wikidata.org/services-resources/ontology
Three general approaches have been taken to classifying Wikipedia articles into a consistent and coherent set of topics: 1) directly applying existing editor-generated annotations on Wikipedia, 2) linking Wikipedia articles to an external taxonomy, and, 3) learning unsupervised topics and manually labeling them. This work pulls most directly from Section 2.1 with additional modeling similar to Section 2.3.

2.1 Annotations on Wikipedia
The most common and simplest strategy for classifying Wikipedia articles by topic is using existing annotations that editors have added to articles. For instance, Wikipedia has a category network that roughly forms a tree with approximately 40 root topics.9 Not all language communities of Wikipedia, however, use the same set of high-level categories or label articles with categories to the same extent [17]. The category network also is messy, requiring careful rules and pruning to avoid loops when linking a given category to its high-level topic [1, 17, 22].

WikiProjects represent another annotation system where groups of editors interested in a specific topic area—e.g., Medicine [12]—tag articles that are relevant so that they may be evaluated and improved. These labels have been used directly [30] but also can be aggregated into higher-level topics based on an editor-curated mapping of WikiProjects to topics10. Asthana and Halfaker [2] use these aggregated topics as well as building a machine-learning model to predict these topics based on article text for filling in gaps in the annotations and applying them articles that have not yet been labeled. As described in Section 3, this taxonomy of topics and approach of using machine learning to expand existing editor-based annotations is deemed the most appropriate for this task.

2.2 External Ontologies
Researchers have also turned to external ontologies—most notably DBPedia [15]—for classifying Wikipedia articles. DBPedia has built a manually-curated ontology of around 800 topics that are linked to Wikipedia articles based on which infobox templates are present on a given Wikipedia article. While this has the benefit of being manually-curated and DBPedia also provides linkages to many other ontologies, coverage is limited by how many infobox templates are present and mapped to DBPedia’s ontology.

Wikidata offers another ontology that is much more closely-linked to Wikipedia. Wikidata items often either have an instance-of property (P31) or subclass-of property (P279), the network of which can be used to categorize Wikidata items (and their corresponding Wikipedia articles) into a set of high-level topics [24]. Wikidata’s ontology contains loops, dead-ends, and other inconsistencies that limit its usage, however, in applying coherent topics to articles [4, 24]. It also is a step removed from Wikipedia articles, which removes the direct connection and feedback loop between the topics applied to an article and what content is included in the article.

2.3 Unsupervised Approaches
Some researchers have also avoided these existing ontologies in favor of unsupervised learning of topics and post-hoc labeling. These generally are learned via topic models, most notably Latent Dirichlet Allocation (LDA), with article text as input [18, 20, 27]. These unsupervised approaches have the benefit of generating continuous topic vectors that can be valuable for modeling and having high coverage because they do not rely on annotations. However, there are limitations of these approaches for topic labeling of articles systematically. First, the identified latent topics cannot always easily be interpreted in terms of its content. Second, text-based approaches usually require custom adaptations when being applied across all languages due to issues arising from parsing and pre-processing different scripts.

The approach by Piccardi and West [23], which learns a topic model over articles as represented by their links mapped to the language-agnostic Wikidata vocabulary, is very similar to the approach taken in this paper but unsupervised and with additional preprocessing that imputes links that are not there. They also consider the downstream task of supervised topic classification but do so with a balanced dataset, making their performance not comparable to our evaluations with a representative and highly imbalanced dataset.

3 METHODS
We describe our criteria and chosen approach for developing a language-agnostic approach to labeling Wikipedia articles with a consistent set of topics.

3.1 Guiding Principles
While high performance is important, it is not the only (or even the most important) criterion that guided the development of the approach described in this paper.

3.1.1 Coverage. The main motivation for development of this model was to expand beyond the limited coverage of past approaches in terms of languages and articles within these languages. Knowledge equity, a core component of the Wikimedia Foundation’s strategy,11 explicitly calls out the importance of fairly allocating resources, and that includes technologies that can be used to support research and tooling for Wikipedia communities. While

---

9The English Wikipedia system: https://en.wikipedia.org/wiki/Category:Main_topic_classifications
10On English Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory
11https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction#Knowledge_equity:_Knowledge_and_communities_that_have BEEN_LEFT_OUT_BY_structures_of_power_and_privilege
this may feel less relevant for e.g., understanding pageview trends in very small wikis (where simply examining the most popular articles may be sufficient), the importance of equity in models such as topic classification is especially salient when considering experiences such as Newcomer Tasks that use topics to provide a personalized, supportive editing experience for new Wikipedia editors. Limiting a model to only languages where pre-trained word embeddings are available or that already have high coverage of annotations would mean that the communities that could benefit the most from growth are cut-off from these technical supports.

To achieve maximal coverage, we use a language-agnostic approach for article features — i.e. the vocabulary for our model is shared across all languages and does not require any language-specific parsing or modeling such as pre-trained word embeddings. Our approach does not consider the actual words in an article at all (just a “bag-of-links”). As such, it can make topic predictions for (almost) any Wikipedia article, as the only requirement is that it is linked to other Wikipedia articles.

3.1.2 Size and Simplicity. We aimed for the model to be fast to train, simple to maintain, low memory requirements, and an overall minimal footprint. These requirements come out of practical concerns—models require maintenance and hardware that must be resourced by the Wikimedia Foundation, which is a non-profit that does not have the money and large staff available to many technology companies. They also come out of a recognition of the environmental costs of training and deploying large machine learning models, a growing trend among technology companies that puts an emphasis on complexity and size instead of seeking simpler, more sustainable solutions. These concerns are exacerbated by growing evidence that, at least for certain tasks, simpler model architectures combined with appropriate tuning can outperform much more complex models.

This criterion is reflected in our model architecture —i.e. the vocabulary for our model is shared across all languages and does not require any language-specific parsing or modeling such as pre-trained word embeddings. By choosing a 50-dimensional embeddings (a relatively small dimensionality) and the fastText model architecture, a very simple and low-footprint model architecture, we take approximately 10 minutes to train across 16 CPU threads and takes up only 863MB on disk. Furthermore, the training data for any given article is based solely on the existing links in that article. The model can therefore be trained based on (relatively small) database tables while not processing the much larger and more complex article text dumps.

3.1.3 Feedback. Given the long history of community-driven decision-making on Wikipedia, including the management of bots and other automated technologies, machine-learning technologies that are developed to support these editor communities should also be developed with clear pathways for feedback and iteration so as to best fulfill the values held by the editor communities they are intended to support.

We design our model per this criterion in two ways. First, the features — i.e. the vocabulary for our model is shared across all languages and does not require any language-specific parsing or modeling such as pre-trained word embeddings. Our approach does not consider the actual words in an article at all (just a “bag-of-links”). As such, it can make topic predictions for (almost) any Wikipedia article, as the only requirement is that it is linked to other Wikipedia articles.

3.1.4 Performance. The resulting model needs to have high performance so that it is useful for research and supporting editors. The machine-in-the-loop nature of many Wikipedia tools means that models do not need to be perfect, they just need to be useful for supporting editors in their work. This positioning of the model as support for editors removes the need for an authoritative, state-of-the-art model and instead centers the importance of interpretability and simple feedback loops for adapting and improving the model. We demonstrate in the evaluation section for comparison.

3.2 Our Approach

Described below are the four components of the chosen model: ground-truth topic labels, input features for representing articles, model architecture, and evaluation technique.

3.2.1 Topic labels. We generate topic labels following the approach by Asthana and Halfaker [2], which is based on WikiProject tagging. Specifically, WikiProjects, which are groups of editors who organize to focus on specific topic areas, tag Wikipedia articles that they deem relevant to their topic area. Almost all articles on English Wikipedia have been tagged by at least one WikiProject and many articles are relevant to multiple WikiProjects. To generate our topic labels, for each article, we identify all its WikiProject tags which can be efficiently extracted via the PageAssessments extension. Each of the tags are then mapped to one of 64 topics (such as “Biography”, “History and Society”, or “Mathematics”). We chose

14 The mapping of WikiProject to topic is maintained here: https://github.com/wikimedia/wiki-tax
15 See [5] for a delineation of “machine-in-the-loop” vs. “human-in-the-loop”.
16 e.g., WikiProject Medicine is an exemplar.
17 https://meta.wikimedia.org/wiki/Data_dumps/FAQ#How_big_are_the_en_wikipedia_dumps_uncompressed?
18 https://meta.wikimedia.org/wiki/Extension:PageAssessments#Database_tables
19 https://www.mediawiki.org/wiki/ORES/ArticleTopicTaxonomy
this approach because it was simple, transparent, and Asthana and Halfaker’s topic taxonomy was already incorporated into various language-specific tools and research. It has the drawback that it draws its labels solely from English Wikipedia, which we address in the Future Work section. The category network likely represents the best alternative for extending community-driven topics, but lack of hierarchy, inconsistency across languages, and constant evolution pose additional challenges.

For the December 2020 snapshot, this WikiProject-based approach yielded labels for 5,970,598 articles (96% of all articles on English Wikipedia) and 15,180,479 groundtruth topics (2.5 topics per article). Through the language mappings stored in Wikidata, these labels were also directly associated with articles in any of the approximately 300 other languages on Wikipedia. Thus, the labels for 5,970,598 English Wikipedia articles actually becomes labels for 30,581,076 Wikipedia articles across all languages (55% of all articles on Wikipedia). A small number of articles were then filtered out if they lacked links that could be mapped to Wikidata IDs and a train/validation/test split of 90%/2%/8% was applied. This lead to 5,088,621 items (27,225,747 articles)\(^{19}\) in the training data, 113,241 items (603,582 articles) in the validation data, 451,909 items (2,417,700 articles) in the test data, and 25,047,098 articles with data but no labels.

The labels are highly imbalanced ranging from Libraries and Information with 10,043 articles on English Wikipedia to Biographies with 1,862,363 articles on English Wikipedia with the average topic having 235,748 English Wikipedia articles and median topic having 100,240 English Wikipedia articles.

### 3.2.2 Features

We represent each Wikipedia article as an unordered set of links in a language-agnostic way using Wikidata items as the underlying vocabulary. While the central feature of Wikipedia articles is the actual textual content, an incredibly important and widespread feature of articles is their links. Even stub articles—i.e. those with just a few sentences—often contain several links within the text (or associated templates\(^{19}\)) to other articles that are mentioned or relevant. Almost every article on Wikipedia also is mapped to an associated Wikidata item, which, among other things, stores an editor-curated listing of all the other articles in other Wikipedia language editions that are about the same concept.

Consider the English Wikipedia article for City Bureau,\(^{20}\) a civic journalism non-profit based in Chicago, Illinois, United States. As of January 2021, this article links to 19 other articles in English Wikipedia. These articles, which include other journalistic institutions and Chicago-area places, provide insight into what the City Bureau article itself is about. Furthermore, all of these articles have associated Wikidata items,\(^{21}\) so if there was a City Bureau article in another language that also linked to the same Chicago-area and journalistic concepts, it could be represented in the exact same way.

For every article in every language edition of Wikipedia, we gather all of the links to other articles\(^{22}\). Redirects\(^{23}\) are resolved, e.g., replacing links to “Chicago, Illinois” with the canonical article title “Chicago”. Each article is mapped to its corresponding Wikidata item,\(^{24}\) e.g., the article on “Chicago” in, both, the English and the German Wikipedia is represented by “Q1297”\(^{25}\). As a result, the representation of each article as a set of links is language-agnostic.

For the December 2020 snapshot, this yielded 55,334,958 articles that aggregated to 20,119,210 distinct concepts (Wikidata items). These articles were associated with 3,122,037,115 links that were mapped to a vocabulary of 18,037,708 distinct Wikidata items (though only 4,145,064 of these items were retained in the final model’s vocabulary). As described in the next section, about half of these articles have associated groundtruth labels and can therefore be used for training and testing. The distribution of links per article is shown in Figure 1. As can be seen, the vast majority of articles (99.71%) have at least one link and thus can be covered by the model.

### 3.2.3 Model Architecture

We model the problem of topic classification as a multi-class supervised learning task. We specifically chose the fastText supervised classification architecture\(^{14}\), which has been demonstrated to be efficient and highly effective at (multi-class) classification tasks. The bag-of-links approach we take here is functionally equivalent to the bag-of-words approach the fastText model was designed to use as input.

The fastText model learns embeddings for each word in its vocabulary, averages the embeddings together for all the words in a document, and then trains a simple, one-vs-all logistic regression over this average embedding to make its predictions. It requires the following hyperparameters: embedding dimensionality, threshold for how many times a link must appear to be included in the model.

---

\(^{19}\)Item here meaning just English Wikipedia articles while the articles number includes all the articles in other languages about the same concept as well.

\(^{20}\)https://en.wikipedia.org/wiki/City_Bureau

\(^{21}\)https://en.wikipedia.org/w/index.php?action=query&generator-links&titles=City_Bureau&prop=pageprops&ppprop=wikibase_item&gplimit=100&gplnamespace=0&redirects&format=json&formatversion=2

\(^{22}\)https://www.mediawiki.org/wiki/Manual:Links_table

\(^{23}\)https://www.mediawiki.org/wiki/Manual:Redirect_table

\(^{24}\)https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Content/Wikidata_item_page_link

\(^{25}\)https://www.wikidata.org/wiki/Q1297
vocabulary, learning rate, and number of epochs. We fix the learning rate at 0.1 and conduct grid search over dimensionality (50, 100), vocabulary threshold (5, 10, 20), and epochs (2, 3).26 We saw minimal variation (~1%) in F1 scores and so use the lowest-footprint values: dimensionality of 50, epochs of 2, and vocabulary threshold of 20. The model also has a window-size hyperparameter used for training, but given that there is no order to the bag-of-links we use as input features,27 we randomize the link order in the training data and set the window size to be relatively large (20) to reduce the chance of learning any spurious signals.

3.2.4 Evaluation. We evaluate the resulting model in two different ways: a quantitative evaluation using held-out test data and qualitative evaluation of predictions by the model in several languages by expert Wikipedians. Details about how to access the code, data, and model for further evaluation or usage are included in Appendix as well.

For the quantitative evaluation, we use a standard held-out test set of data. Given that articles can appear in multiple languages, we randomly sort articles into the train-validation-test splits based on their associated Wikidata item—i.e. if the English article for Curt Flood28 is randomly placed into the test set, then any other language versions of that article29 will also be placed in the test set. Even though the other language versions of the article will very likely have different links associated with them [11], they share labels and we choose to not consider them to be independent data points. Given the heavy imbalance in data, we report both micro and macro precision, recall, and F1 statistics. We also report average precision, which summarizes the precision-recall curve and is therefore threshold-agnostic (the other approaches uses the default threshold of 0.5 for mapping probabilities to binary predictions).

The results for our language-agnostic link-based model are also compared to two other models. The first is the Wikidata-based model described in Johnson et al. [13], which predicts topics for a Wikipedia article based on the properties and values contained in its associated Wikidata item. It is trained with the same fastText architecture as the language-agnostic links-based model, a learning rate of 0.1, 25 epochs, 50 dimensions, and a vocabulary threshold of 3. The second comparison model is the language-dependent, text-based model described in Asthana and Halfaker [2], which predicts topics for a Wikipedia article based on a gradient-boosted classifier learned over top a document embedding formed by averaging together word embeddings associated with each word in the article. This is very similar to the fastText architecture and an API for the model is available through the Wikimedia Foundation.30

The different training paradigms and choice of input features make it difficult to directly compare these different approaches under the exact same conditions.31 We instead choose to compare these models as they would be deployed so as to directly compare their potential value to the Wikimedia community.

For the qualitative evaluation, we focus on five languages chosen for their geographic coverage and the willingness of their communities to experiment with new tools: Arabic (ar), Czech (cs), English (en), French (fr), and Vietnamese (vi). For each language, we generate ten random examples of articles that are confidently32 labeled with each topic. These examples are presented to Wikipedians who are familiar with that language and they evaluate whether the topic is relevant to the article or not (a binary yes/no with the opportunity for explanation). Across the 64 topics, 5 languages, and 10 articles per topic, this results in 3,200 evaluations. The same evaluation is done (by the same Wikipedians) for the language-dependent, text-based model in each language. The Wikidata-based model was left out of this evaluation because it did not clearly outperform the language-agnostic link-based model in the quantitative evaluation.

4 RESULTS

We first look at the quantitative results for just English-language articles for all three models as the most-directly comparable set of results (Rows 1-3 of Table 1). The models perform very similarly (within 2-3% of each other) though the text-based model performs the best in English and the Wikidata model has lower recall (and therefore F1 and average precision). Looking at the fourth row in Table 1, we see that the language-agnostic link-based model does not suffer in performance when making predictions for articles in other languages. Together, these results provide evidence that this model is able to provide both high performance and high coverage.

The quantitative results are limited, however, by the fact that articles in languages outside of English only will have groundtruth labels if they have an English-language equivalent, which is not true for almost half of the articles. The results from the qualitative evaluation of English and four additional languages (Arabic, Czech, French, Vietnamese) provide some insight into whether the model generally seems to work for non-English languages.

We see that the link-based model consistently outperforms the language-specific, text-based model in all five languages tested (Table 2). The increase in performance varies by language but ranges from 4% (absolute) better in Arabic and English to almost 12% better in Vietnamese. There are certain topics where the language-specific, text-based model performs better (though the low sample size per topic makes it difficult to read too far into this) such as the women biographies topic. This suggests that for certain topics, especially given the sensitive nature of gender, relying on Wikidata’s information may be a far better solution. Finally, the feedback from raters highlighted the limitations of some of these topics—e.g., in one instance, while all of the articles in the topic for Northern Africa were correct, they all came from the same country (Morocco), which may reflect bias in the wiki or bias in the recall of the model. Either way, some topics are clearly too high-level for certain use-cases and complementary models will be required to address these. The results

26The number of epochs may seem low but remember that an article may exist in many languages, so the model practically speaking sees these training examples more than once per epoch.

27Links do have an order in an article, but our approach is unaware of this order and would require an additional and much lengthier preprocessing step to extract that information from an article’s text.

28https://en.wikipedia.org/wiki/Curt_Flood

29In this case, French, Arabic, Egyptian Arabic, Japanese, and Polish as can be seen under sitelinks on Wikidata: https://www.wikidata.org/wiki/Q5195372

30https://www.mediawiki.org/wiki/ORES#Topic_routing

31While all three models share the same set of labels and groundtruth data, the choice of input features—article text, article links, Wikidata statements—means that each model has a different number of possible training examples.

32We use the threshold of 0.5 though it’s likely that better results would be achieved with topic-specific thresholds.
We developed a language-agnostic topic classification model that can automatically label (almost) all Wikipedia articles across all languages. The results demonstrate that the advantages of this approach by meeting the criteria laid out in Section 3.1 around coverage, simplicity, and feedback, while at the same time performing with high precision and recall. In fact, the evaluation showed that the model performs equally well or better than alternative approaches.

More generally, we believe language-agnostic approaches to building machine-learning models in the service of Wikimedia projects can be applied to many other classification tasks. For instance, labeling the quality of articles is a common maintenance task on Wikipedia but one that is difficult to constantly keep up-to-date [29]. While not all language editions use the same set of quality classes [17], Lewoniewski et al. [16] built a simple language-agnostic quality model that can be applied to any language. While certain tasks such as vandalism detection may prove too nuanced for language-agnostic models, future work should explore how to apply language-agnostic methods to other tasks such as identifying more specific topics like the countries associated with an article or the intention behind edits [31].

5 DISCUSSION

While we demonstrate that language-agnostic topic classification model presented here is performant and adheres to the principles laid out for its development, there is still plenty of room for improvement. The clearest downside to the approach presented is its reliance on groundtruth labels and a taxonomy derived solely from English Wikipedia. A few other language editions (Arabic, Hungarian, Turkish, French) use the PageAssessments extension that allows for easy extraction of WikiProject tags and many other languages have active WikiProjects whose data could be mined through other methods. This would directly provide some labels for articles that do not exist in English Wikipedia but have been tagged by WikiProjects that have corollaries on English Wikipedia (as determined by Wikidata-maintained sitelinks). Even better would be working with these communities to map their other WikiProjects whose data could be mined.

While almost all articles have links, many have only a few, making it difficult for the model to predict topics. There are language-agnostic approaches to inferring what links might be missing from a given Wikipedia article—e.g., the approach presented by Piccardi and West [23]—but more work is needed to ensure that that process is efficient. Their work applied it across 28 languages and still required language-specific parsing of the article text, but it is a promising approach. In the meantime, tools such as link recommendation for new editors can be used to help add links directly to articles where they are most needed.

We considered only one model architecture for a language-agnostic model in this paper. There are many other potential language-agnostic architectures though such as graph-based approaches [21].

### Table 1: Overall model results. Model evaluation results for the three models. The first three rows show results just for test data from English Wikipedia. The two entries for the link-based language-agnostic model are for the same model but different test sets: held-out English-language articles only (row 2) and held-out articles from all languages (row 4). Average-precision summarizes the precision-recall curve for all possible thresholds at which a prediction probability may be turned into a binary yes. See https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html for more information.

| Model                  | Pre. (micro) | Pre. (macro) | Rec. (micro) | Rec. (macro) | F1 (micro) | F1 (macro) | Avg. Pre. (micro) | Avg. Pre. (macro) |
|------------------------|--------------|--------------|--------------|--------------|------------|------------|-------------------|-------------------|
| Wikidata               | 0.891        | 0.854        | 0.796        | 0.679        | 0.838      | 0.752      | 0.895             | 0.797             |
| Link-Based (en)        | 0.866        | 0.826        | 0.808        | 0.719        | 0.835      | 0.766      | 0.892             | 0.814             |
| Text-Based (en)        | 0.883        | 0.846        | 0.817        | 0.723        | 0.847      | 0.776      | 0.908             | 0.841             |
| Link-Based (all)       | 0.877        | 0.836        | 0.793        | 0.678        | 0.83       | 0.744      | 0.891             | 0.795             |

Table 2: Evaluation of quality of predictions by topic for language-agnostic and language-dependent models. This table summarizes the qualitative evaluation of the language-agnostic (LA) model and language-dependent (LD) model. It reports the overall precision across the 64 topics (10 articles per topic) for each model as well as how many of the 64 topics did not have at least 70% precision (7 out of 10 correct). The language-agnostic model outperforms the language-dependent model for all languages in terms of both metrics. Keep in mind that while the language-agnostic model is a single model, there are in fact five language-dependent models, one for each language, with their own word embeddings and model parameters.

| Model                  | Overall Precision | Topics < 7/10 |
|------------------------|-------------------|---------------|
| Arabic (LD)            | 90.9%             | 4             |
| Arabic (LA)            | 94.7%             | 2             |
| Czech (LD)             | 74.5%             | 17            |
| Czech (LA)             | 81.4%             | 11            |
| English (LD)           | 85.5%             | 7             |
| English (LA)           | 89.5%             | 6             |
| French (LD)            | 79.7%             | 11            |
| French (LA)            | 88.0%             | 5             |
| Vietnamese (LD)        | 79.7%             | 14            |
| Vietnamese (LA)        | 91.3%             | 2             |

6 FUTURE WORK

While we demonstrate that language-agnostic topic classification model presented here is performant and adheres to the principles laid out for its development, there is still plenty of room for improvement. The clearest downside to the approach presented is its reliance on groundtruth labels and a taxonomy derived solely from English Wikipedia. A few other language editions (Arabic, Hungarian, Turkish, French) use the PageAssessments extension that allows for easy extraction of WikiProject tags and many other languages have active WikiProjects whose data could be mined through other methods. This would directly provide some labels for articles that do not exist in English Wikipedia but have been tagged by WikiProjects that have corollaries on English Wikipedia (as determined by Wikidata-maintained sitelinks). Even better would be working with these communities to map their other WikiProjects to the topic taxonomy and propose adjustments to the taxonomy, which may itself reflect English-language and cultural biases. Given the evaluations presented here, we think this taxonomy has utility in languages outside of English, but it is open to iteration.

While almost all articles have links, many have only a few, making it difficult for the model to predict topics. There are language-agnostic approaches to inferring what links might be missing from a given Wikipedia article—e.g., the approach presented by Piccardi and West [23]—but more work is needed to ensure that that process is efficient. Their work applied it across 28 languages and still required language-specific parsing of the article text, but it is a promising approach. In the meantime, tools such as link recommendation for new editors can be used to help add links directly to articles where they are most needed.

We considered only one model architecture for a language-agnostic model in this paper. There are many other potential language-agnostic architectures though such as graph-based approaches [21].
or applying sequence models to reader navigation [26]. While not all of these language-agnostic models are simple and adhere closely to the guiding principles for this model, these other approaches may complement the simple bag-of-links approach described here and be necessary for achieving sufficient performance and coverage on other tasks.

ACKNOWLEDGMENTS

We would like to thank the many supporters of this language-agnostic topic classification project and the Wikipedians who have built out such extensive annotation systems on Wikipedia. In particular, Aaron Halfaker for his constant support and guidance in building on his initial work and Marshall Miller, Habib Mhenni, Benoit Evellin, Martin Urbanec, Bluppt, and HAKSOAT for their support in evaluating these models.

REFERENCES

[1] Ahmad Aghaрабrahim, Andy Stauder, and Michael Utsaszewski. 2020. Testing the validity of Wikipedia categories for subject matter labeling of open-domain corpus data. Journal of Information Science and Engineering (Dec. 2020), 0165551520977438. https://doi.org/10.1177/0165551520977438

[2] Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–18.

[3] EM Bender, T Gehu, A McMillan-Major, et al. 2021. On the dangers of stochastic parrots: Can language models be too big. Proceedings of PLoCIT (2021).

[4] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2021. Why are we really making much progress? A worrying analysis of recent neural recommenda-
tion approaches. In Proceedings of the 18th ACM Conference on Recommender Systems. 91–109.

[5] Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03995 (2020).

[6] Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–18.

[7] Volodymyr Miz, Joelle Hanna, Nicolas Aspert, Benjamin Ricas, and Pierre Vanderheyden. 2020. What is trending on Wikipedia? capturing trends and language biases across Wikipedia editions. In Companion Proceedings of the Web Conference 2020. 794–801.

[8] Alessandro Piscopo and Elena Simperl. 2018. Who models the world? Collaborative ontology creation and user roles in Wikidata. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–18.

[9] Paul Middleman, Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking spatial visualization through semantic enhancement. In Proceedings of the 22nd international conference on intelligent user interfaces. 179–190.

[10] Andrew Hall, Sarah McRoberts, Jochen Theubt-Spieker, Yilun Lin, Shalad Sen, Brent Hecht, and Loren Terveen. 2017. Freedom versus standardization: structured data generation in a peer production community. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 6552–6562.

[11] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In Proceedings of the 26th international conference on world wide web. 1591–1600.

[12] Philipp Singer, Florian Lemmerich, Brent Hecht, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we need Wikipedia. In Proceedings of the 26th international conference on world wide web. 1591–1600.

[13] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2021. Why are we really making much progress? A worrying analysis of recent neural recommenda-
tion approaches. In Proceedings of the 18th ACM Conference on Recommender Systems. 91–109.

A AVAILABILITY

We make as much of the code, data, and predictions open for further usage and critique. All of the topics predicted for every single Wikipedia article are available for download and use. The code for the model is available under open license, though the data pipeline is written for a Hadoop cluster operated by the Wikimedia Foundation and thus likely would require adjustments to use the public dump files. For interested users, an API to compare the different models is made available via a simple interface.

B COMPLETE RESULTS

https://figshare.com/articles/dataset/Wikipedia_Article_Topics_for_All_Languages_based_on_article_outlinks_/12619766

https://github.com/geoeh/wikipedia-language-agnostic-topic-classification

https://wiki-topic.toolforge.org/comparison
Table 3: Topic-specific results for the language-agnostic model. For each topic, the number of test articles (n), true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), precision, recall, F1, and average precision (Ave. Prec.) are given. Average-precision summarizes the precision-recall curve for all possible thresholds at which a prediction probability may be turned into a binary yes. See https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html for more information. The model performs quite well for almost all topics (precision over 0.7) with the exception of Education. Recall suffers for several topics as well, mostly those with fewer examples.