Automatically Labeling Low Quality Content on Wikipedia by Leveraging Patterns in Editing Behavior

SUMIT ASTHANA, University of Michigan, USA
SABRINA TOBAR THOMMEL, University of Michigan, USA
AARON LEE HALFAKER, Microsoft, USA
NIKOLA BANOVIC, University of Michigan, USA

Wikipedia articles aim to be definitive sources of encyclopedic content. Yet, only 0.6% of Wikipedia articles have high quality according to its quality scale due to insufficient number of Wikipedia editors and enormous number of articles. Supervised Machine Learning (ML) quality improvement approaches that can automatically identify and fix content issues rely on manual labels of individual Wikipedia sentence quality. However, current labeling approaches are tedious and produce noisy labels. Here, we propose an automated labeling approach that identifies the semantic category (e.g., adding citations, clarifications) of historic Wikipedia edits and uses the modified sentences prior to the edit as examples that require that semantic improvement. Highest-rated article sentences are examples that no longer need semantic improvements. We show that training existing sentence quality classification algorithms on our labels improves their performance compared to training them on existing labels. Our work shows that editing behaviors of Wikipedia editors provide better labels than labels generated by crowdworkers who lack the context to make judgments that the editors would agree with.

CCS Concepts: • **Human-centered computing → Social recommendation**; **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**; *Wikis*; *Social tagging systems.*

Additional Key Words and Phrases: Wikipedia, Content Labeling, Machine Learning.

ACM Reference Format:
Sumit Asthana, Sabrina Tobar Thommel, Aaron Lee Halfaker, and Nikola Banovic. 2021. Automatically Labeling Low Quality Content on Wikipedia by Leveraging Patterns in Editing Behavior. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 359 (October 2021), 23 pages. https://doi.org/10.1145/3479503

1 INTRODUCTION

Wikipedia [46], an online encyclopedia, aims to be the ultimate source of encyclopedic knowledge by achieving a high quality for all its articles. High quality articles are definitive source of knowledge on the topic and serve the purpose of providing information to Wikipedia readers in a concise manner, without causing confusion and wasting time [44]. Thus, Wikipedia editors have defined a comprehensive content assessment criteria, called the WP1.0 Article Quality Assessment scale [47] to grade article quality on a scale from the most basic "stub" (articles with basic information about the topic, without proper citations and Wikipedia-defined structure) to the exemplary "Featured Articles" (well-written and well-structured, comprehensive and properly cited articles).

Authors' addresses: Sumit Asthana, asumit@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Sabrina Tobar Thommel, sabtt@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Aaron Lee Halfaker, ahalfaker@microsoft.com, Microsoft, Seattle, Washington, USA; Nikola Banovic, nbanovic@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

2573-0142/2021/10-ART359 $15.00
https://doi.org/10.1145/3479503

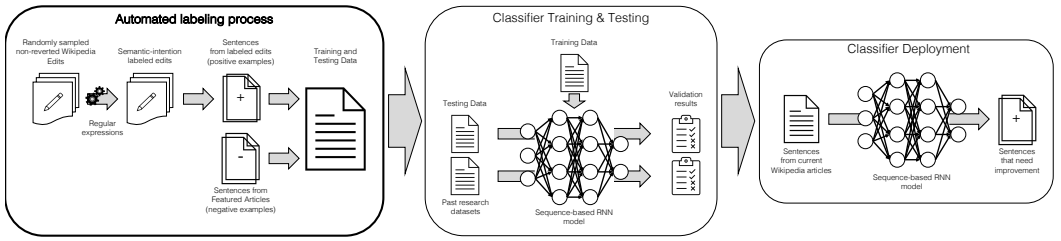


Fig. 1. Our pipeline for labeling low-quality sentences on Wikipedia. We start with our automated labeling approach (left), where we obtain a large corpus of historic Wikipedia sentence edits, and label their semantic intent using programmatic rules. We extract positive sentences from relevant semantic edits and negative sentences from Featured Articles. We then use our labels to train existing Machine Learning models, and test them by comparing with labeling approaches from past research (middle). Existing models trained on our labels can then be deployed to automatically detect Wikipedia sentences that require improvement (right).

Article maintenance, as opposed to creating new articles and content, has become a significant portion of what Wikipedia editors do [27]. Currently, editors rate article quality and identify and make required improvements manually, which is taxing and time-consuming. Being a collaborative editing platform, articles are in a constant state of churn and current assessments are quickly outdated because articles will have been modified by others. For the limited number of experienced editors on Wikipedia, performing such assessments across a set of 6.5 million Wikipedia articles is a huge bottleneck [42]; currently only about 7,000 of articles have "Featured Article" status and only about 33,000 have the second best "Good Article" status [47].

With continuously declining number of editors on Wikipedia [43], automating quality assessment tasks could reduce the workload of remaining editors. Supervised Machine Learning (ML) has already automated tasks like vandalism detection [19] and overall article quality prediction [45]. Such ML approaches require labeled sets of examples of Wikipedia content that requires improvement (positive examples) and content that does not (negative examples). One of the main reasons for the success of those existing ML approaches [19, 45] (both have been deployed to Wikipedia) is the relative ease of obtaining labels either because they are visually salient (e.g., in case of vandalism) or already part of existing practices (e.g., editors manually record article quality on talk pages of Wikipedia articles as part of existing article assessment).

However, automating other quality assessment tasks (e.g., identifying sentences that require citation, sentences with non-neutral point of view, sentences that require clarification) requires labels at the Wikipedia sentence level which makes automating such tasks difficult. Wikipedia editors rarely manually flag outstanding Wikipedia sentence quality issues as part of their editing process [1]. Even existing crowdsourcing-based labeling method [22, 39, 55] could produce noisy Wikipedia sentence quality labels, especially when crowdworkers, who are not domain experts, lack knowledge about Wikipedia policies on content quality [13, 18, 24, 26].

Here, we present a method for automatically labeling Wikipedia sentence quality across improvement categories directly from past Wikipedia editors' editing behavior to enable automated detection of sentences that need quality improvements (Figure 1). To label positive examples (sentences that need improvements), we implemented Wikipedia core content principles guidelines [48] as syntax-based rules to capture the meaning or intent of a historic Wikipedia edit (i.e., the smallest recorded unit of change in a Wikipedia article, paragraph, or sentence, such as added citations, removed bias, clarification) for each quality category we want to classify (e.g., needs citation, needs bias-removal, or needs clarification). Each historic edit then indicates that the edited sentence

needed that particular improvement resulting in a positive example. We follow Redi et al's. [39] approach and label all sentences in featured articles as negative examples (sentences that do not need improvements).

To illustrate our approach, we built three Wikipedia sentence quality detection pipelines (including corresponding rules) for three Wikipedia quality improvement categories: 1) citations (adding or modifying references and citations for verifiability), 2) Neutral Point of View (NPOV) edits (rewriting using encyclopedic, neutral tone; removing bias), and 3) clarifications (specifying or explaining an existing fact or meaning by example or discussion without adding new information) [55]. We first evaluated our rules in a user study with nine Wikipedia editors, in which they manually labeled improvement category of 434 historic Wikipedia edits. We then compared the outputs of our rules with the ground truth and showed that our rules could effectively extract positive examples. Our results also revealed high ambiguity amongst participants' manual labels, which further underline the importance of our automated rule-based approach.

We then validated the usefulness of our automated labeling approach by comparing the performance of *existing* deep learning models [2] trained using existing, baseline labeling approaches (e.g., implicit labeling [39], crowdsourcing [22]) and our automatically extracted labels. Our results showed that existing models trained using our automatic labeling method achieved 29% and 22% improvement in F1-score for citations and NPOV respectively than the same models trained on data labeled using existing approaches.

Our work provides further evidence that the edits produced by Wikipedians working in their context provide better signal for supporting their work than labels generated by crowdworkers who lack the context to make judgments about sentence quality that Wikipedians would agree with. Learning from implicit editing behavior of Wikipedia editors allowed us to produce labels that capture the nuances of Wikipedia quality policies [25]. Our work has implications for the growth of collaborative content spaces where different people come together to curate content adhering to the standards and purpose of the space [32]. With Wikipedia behavior policies becoming more decentralized [14], our strategy of learning from implicit behaviors has additional relevance of enforcing norms that may only be enacted through reading and observing policies up until now.

2 COLLABORATIVE PLATFORM CONTENT QUALITY LABELING CHALLENGES

Research has given considerable attention to improving and maintaining good quality of user generated content on collaborative platforms. Such research has explored both assisted and automated editing tools [17] for content creation & recommendation [54], vandalism detection [15, 19], and content regulation (e.g., content that violates platform policy) using both programmatic rules [11, 16, 23, 38] and Machine Learning-based methods [4]. A subset of such research focuses specifically on automatically detecting content quality (e.g., [7, 45]) and bringing the issues to the attention of the community.

Such automated efforts have been possible in part because of the availability of quality labels for such tasks. For example, a small subset of visually-salient, hand-labeled examples are sufficient for simple ML models to identify vandalism with high accuracy [15]. Also, training existing article quality models [45] involves using existing quality labels for over 6.5 million articles that Wikipedians have generated when manually rating article quality as part of existing processes.

Unfortunately, that kind of automated assistance to editors does not easily extend to other collaborative editing tasks because labels for other quality assessment and improvement tasks are not immediately available. For example, although Wikipedia encourages its editors to manually flag outstanding content issues with cleanup templates markup (e.g., marking a sentence with *{citation needed}* template [53]) or label their Wikipedia edits with a free-form edit intent summary

(e.g., point-of-view), their usage is not standardised and only few Wikipedia edits or Wikipedia sentences that need improvement actually have them [1].

Existing attempts to supplement such labels *via* crowdsourcing [22, 39, 55] produce too few labels when using Wikipedia editors as labelers or produce noisy labels when using crowdworkers who are not Wikipedia editors. Such non-editor crowdworkers do not always provide reliable judgments on what content needs improvement [8], often due to their lack of knowledge about the nuances of Wikipedia policies [13, 26].

Although crowdsourcing has been used in the past [37] to successfully label examples of vandalism, it is important to note that annotating vandalism is simpler than examples related to concepts like the need for citations, neutrality of point-of-view, and clarifications, since the concept of vandalism could be commonly shared between lay Web users and Wikipedians. In the absence of a widely accepted clear standard of categorization of Wikipedia sentence quality, most of the other tasks that editors perform are hard to label.

To get around explicitly asking editors or crowdworkers to label the quality of Wikipedia sentences, existing research [39] has attempted to obtain labels implicitly. Redi et al. [39] showed that citation labels are easy to obtain because presence/absence of citations in "Featured Articles" acts as an implicit label that sentences with citations needed them and those without did not. While such implicit labeling strategy can be used to label negative examples across semantic improvement categories, they cannot extract positive examples of needed improvements for categories, such as neutrality of point-of-view or clarification.

Recently, Yang et al. [55] created a taxonomy of Wikipedia edits based on the semantic intention behind the edit to build a classifier to automatically categorize the semantic intent of Wikipedia edits. This taxonomy comprehensively covers the tasks Wikipedians do, ranging from fighting vandalism and copy-editing to making content clarifications and simplifications. Existing research [40, 55] has used this taxonomy to automatically classify article quality using ML-based approaches, but not individual Wikipedia sentence quality. Thus, such existing approaches do not pinpoint which specific parts of a Wikipedia article need improvement. Unfortunately, also, it is not immediately obvious how to adapt such existing article quality labeling methods to the problem of automatically labeling Wikipedia sentence quality using the taxonomy above.

3 METHOD FOR AUTOMATICALLY LABELING LOW QUALITY CONTENT

To build a pipeline for automatic detection of Wikipedia sentences that need quality improvements (Figure 1), we need examples of sentences that require improvement (i.e., positive examples) and examples of sentences that do not need any further improvement (i.e., negative examples). We first focus on extracting positive examples, which is the main contribution of our work. We then use an existing method [39] for extracting negative examples, which assumes sentences from Featured Articles do not need further improvements (we only briefly summarize it in this section). We can then train different ML models for each semantic category on such labeled sentences, and use the classifier to classify if previously unseen Wikipedia sentences (e.g., newly added or edited Wikipedia sentences that we have not trained on) need a particular kind of semantic improvement.

3.1 Identifying Semantic Intents in Wikipedia Edits to Extract Positive Examples

Here, we leverage traces of Wikipedia editors' collaborative editing behaviors to learn which Wikipedia edits are attempts by the editors to improve quality of Wikipedia sentences. Our approach has three benefits: 1) collaborative editing behavior data is readily available because Wikipedia automatically logs all edits as part of its editing process, 2) editing is common across all Wikipedia languages, hence provides for a common approach for detecting quality issues for Wikipedia in all

languages, and 3) Wikipedia already provides a definition for categorization of edits based on their semantic intent [48] (e.g., adding citations, removing point of view, clarifying Wikipedia sentences).

Our insight is that it is easier to identify semantic intent of edits given their syntax compared to identifying quality issues directly in free form natural language sentences. Unlike existing work [55] that attempted to identify the semantic intent across multiple Wikipedia paragraphs (each composed of multiple Wikipedia sentences), we identify semantic improvements at Wikipedia sentence level to pinpoint which sentences need improvement. This also minimizes the noise associated with propagating weak, paragraph-level labels to other sentences in the paragraph that do not need such specific improvement.

Thus, we start with automatically identifying the semantic intent of different Wikipedia edits, which we later use to indicate that the Wikipedia sentence prior to being edited required that semantic improvement (e.g., that it required adding citations, removing point of view, or clarifying). Each time an editor edits an article, Wikipedia represents all of the changes the editor made at that time as an *edit diff* (Figure 2), which consists of Wikipedia content (e.g., section headings, paragraphs, sentences) before editor changes, the same content after editor changes, and an optional comment from the editor that contains their explanation for the edit. Note that an *edit diff* can span multiple Wikipedia article sections and paragraphs. Wikipedia splits each *edit diff* into one or more *lines*, which indicate exact content that the editor changed and the surrounding content (i.e., *context*). Each changed *line* contains one or more *segments*, each representing a continuous unit of change (i.e., there is no content within the segment that is unchanged).

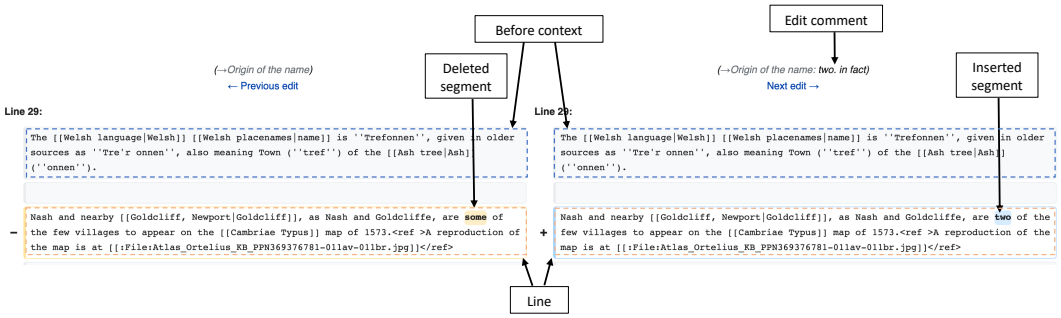


Fig. 2. An example of an edit diff showing two segments - inserted and deleted. Before context is shown with blue dashed line. Orange dashed line outlines one full line.

Since we want to label individual Wikipedia sentences within an *edit diff*, we only extract sentences from the *edit diff* that contain at least one *segment*. To extract positive examples, we test each changed sentence against a set of rules, which translate natural language description of semantic edits categorization [48] to computer code. Here, we illustrate our method on three semantic categories: 1) *Citations* ("add or modify references and citations; remove unverified text"), 2) *Point-of-View (POV)* ("rewrite using encyclopedic, neutral tone; remove bias; apply due weight"), 3) *Clarifications* ("specify or explain an existing fact or meaning by example or discussion without adding new information"). Table 1 specifies the rules for the three semantic categories. A specific semantic category (e.g., point-of-view) is applied only when all the rules under the given category satisfy. Table 2 shows regular expressions that we implemented for each rule in Table 1.

Category	Rule
Citations	is_citation_inserted
Point-of-view	(para_changes == 1) AND (comment_matches "POV"|"pointy") AND NOT (is_citation_inserted) or deleted AND NOT (is_template_inserted or deleted) AND NOT (is_wikilink_inserted or deleted) AND NOT (is_infobox_inserted or deleted) AND NOT (is_multiline_inserted or deleted)
Clarification	(inserted_length_words b/w [0,10]) AND (deleted_length_words b/w [0,5]) AND NOT (is_citation_inserted_or_deleted) AND NOT (is_template_inserted_or_deleted) AND NOT (is_wikilink_inserted_or_deleted) AND NOT (is_infobox_inserted or deleted) AND NOT (is_multiline_inserted or deleted)

Table 1. Rules for citations, point-of-view and clarification edits

Rule	Regex	Match on
is_citation_inserted	<ref>|{{Cite}}	segment_inserted
is_citation_inserted_or_deleted	<ref>|{{Cite}}	both
is_template_inserted_or_deleted	{{[^{]+}}	both
is_wikilink_inserted_or_deleted	[[^[+]]	both
is_infobox_inserted_or_deleted	^|[a-zA-Z0-9]+=	both
is_multiline_inserted_or_deleted	\n	both
comment_matches	"pov|pointy"	edit_comment
para_changes == 1	len(paragraphs) == 1	both
inserted_length_words b/w [0,10]	len(segment_inserted.split())	segment_inserted
deleted_length_words b/w [0,5]	len(segment_deleted.split())	segment_deleted

Table 2. Regular expressions for the rules used in citations, point-of-view and clarification edits. "Match on" specifies the portion of the edit diff on which the regular expression is applied. "Both" specifies segment_inserted and segment_deleted. See Figure 2 for reference.

We started with a set of rules that implemented our own subjective interpretation of the semantic intent categories. We then iterated on our rules by evaluating them on single-line *edit diffs* from the *edittypes* dataset [55] and comparing the outputs of our rules with the *edit diffs* labels from that dataset to tune the parameters of our rules (e.g., to determine the value of inserted_length_words parameter in Table 1 which represents the number of inserted words in a *segment*).

Note that the existing dataset [55] contains crowdsourced labels, which could be noisy. We therefore used the dataset as a reference, and not as ground truth. We excluded any *edit diffs* we used in this stage from our future evaluations. We now describe the rationale for each of our individual rule categories below.

3.1.1 Citations. Citations on Wikipedia are added using the *<ref>* or *{{Cite}}* tags. Note that here we focus on cases when a citation is needed, so we do not consider modifications that editors make to

existing citations (e.g., changing reference URL) or when they remove existing citations. Thus, we implement our *is_citation_inserted* rule in Table 1 as a regular expression that looks for complete additions of the two tags in each sentence (Table 2).

3.1.2 Point-of-View. Wikipedia's Neutral Point of View (NPOV) policy [51] is a very broad policy covering a variety of cases of bias in text. Wikipedia editors may choose to remove bias in the content inline[1], if a few words in the paragraph are violating the neutral point-of-view (NPOV) policy, by removing or rephrasing the violating words. They may also remove an entire paragraph if the paragraph is written in a manner violating the NPOV policy[2]. To illustrate our method, we focus on inline point-of-view edits, where words in a single line violating the NPOV policy are removed or rephrased.

We implemented our rules as a regular expression that matches words "POV" and "pointy" in the edit comment. We only consider *edit diffs* that contain only a single changed *line* to reduce the uncertainty of which changed *line* the editor referred to in their comment. We skip sentences in the changed *line* where all *segments* contain only Wikipedia markup (e.g., citations, templates, links).

3.1.3 Clarifications. For the clarification category, we write regular expression rules to only consider segments with insertions between 0 to 10 words and deletions 0 to 5 words. We skip sentences in the changed *line* where all *segments* contain only Wikipedia markup (e.g., citations, templates, links). We also skip addition of new sentences because this often indicates adding new information (which is part of elaboration category).

3.1.4 Using Semantic Intent Rules to Assign Positive Example Labels. For each *edit diff* that our rules detected as a specific semantic improvement category (e.g., added citation, point-of-view improvement, clarifications), we weakly label the *original* Wikipedia sentence in the *edit diff* with the corresponding label (e.g., needs a citation, needs point-of-view improvement, needs clarifying). For example, Figure 3 shows an *edit diff*, whose semantic intent is clarification, along with the original sentence that the editor clarified, resulting in the new clarified sentence. The original sentence is *"While the exact cause is unknown, it is believed to involve a combination of [[Genetics|genetic]] and environmental factors."*. This sentence is a positive example for "needs clarification" because it was clarified to *"While the exact cause is unknown, Tourette's is believed to involve a combination of [[Genetics|genetic]] and environmental factors."*

3.2 Leveraging Article Quality Labels to Extract Negative Examples

To extract sentences that do not need improvement (i.e., negative examples) for each semantic category, we extract sentences from highest quality Wikipedia articles (i.e., Featured Articles). Sentences in Featured Articles have sparse issues as they have gone through an extensive review process. Our insight is that it is unlikely for sentences needing improvements by Wikipedia standards to make their way into "Featured Articles". This approach of using Featured Article sentences as negative examples has been explored for detecting citations [39], where every sentence without a citation is a negative example (i.e., a sentence that does not need a citation). We generalize this in our work to extract negative examples corresponding to point-of-view and clarifications and assume that Wikipedia sentences in such articles do not need further changes to point-of-view or further clarifications.

[1] An example of inline point-of-view edit -https://en.wikipedia.org/wiki/?diff=745183020
[2] An example of full paragraph deletion - https://en.wikipedia.org/wiki/?diff=745912558

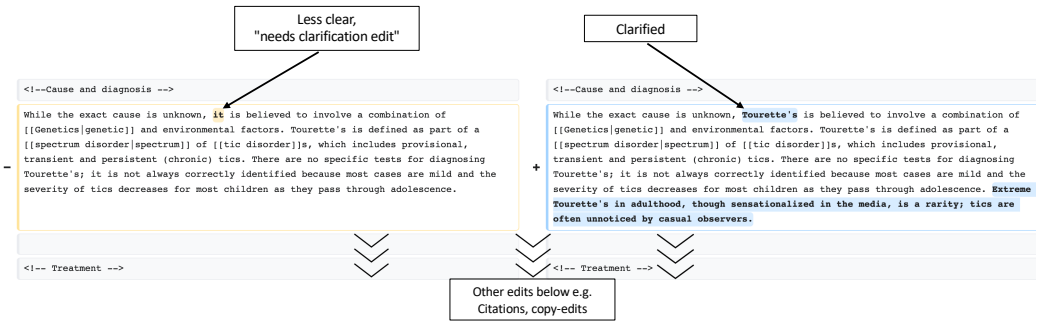


Fig. 3. An edit having one of the intents as clarification and the clarified sentence. Note that the edit also contains other changes but for the purpose of "clarification", we discard the rest of the changes, that are not caught by the rules, as irrelevant.

4 EVALUATION OF SEMANTIC INTENT RULES

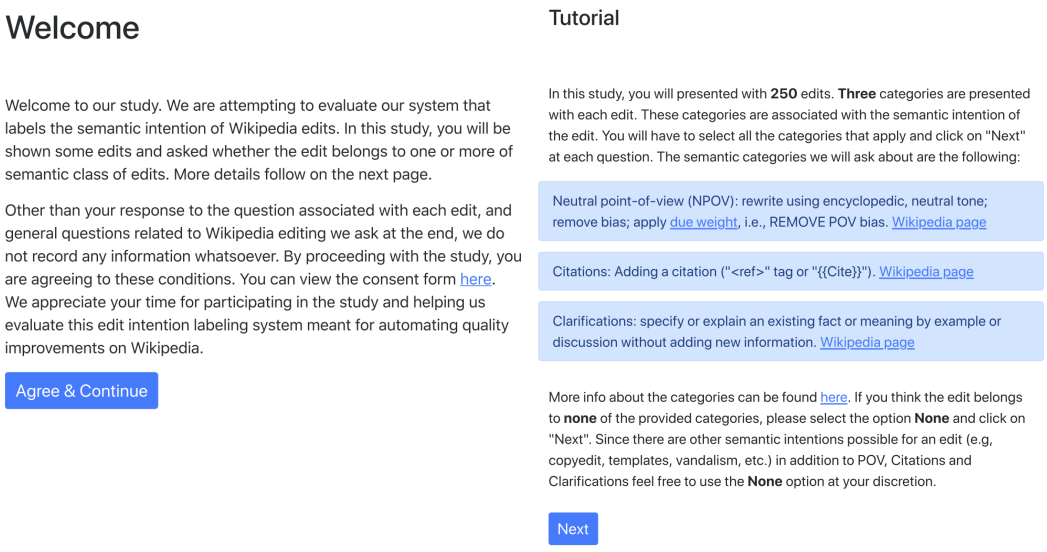
Here, we obtain ground truth semantic intent labels (citations, point-of-view, and clarifications) for a random sample of Wikipedia edits and compare them with the output of our semantic intent rules to evaluate our automatic Wikipedia labeling approach. We obtain the ground truth labels from Wikipedia editors. The purpose of this study is two-fold: 1) to assess the effectiveness of our rule-based method compared to ground truth, and 2) to understand to what extent Wikipedia editors agree on such labels among themselves.

The effectiveness of our labeling approach relies on two key factors: 1) there are more than a billion Wikipedia edits to extract positive examples from, and 2) there are enough curated Featured Articles to extract negative examples from. Even if our rules have a low recall (i.e., they miss to extract sentences that need improvement), the enormous amount of edits ensures that we are still left with a large amount of data to train deep learning models on, as long as our rules have high precision (i.e., they do not wrongly mark sentences that do not need improvement as positive examples). Note that a low recall of our rules that we use to extract positive examples would not necessarily impact the quality of our labels because our negative examples (i.e., sentences that do not need improvement) come from Featured Articles.

4.1 Study Software

To conduct our study, we built an online labeling interface for labeling semantic intention of different Wikipedia edits (Figure 4 and Figure 5). The interface first shows the welcome page which contains a study description, a link to the consent form, and a button to consent and continue (Figure 4a). After the study participant clicks on the button, the interface shows the tutorial page (Figure 4b), which contains the instructions for labeling the edits using our interface along with the exact Wikipedia definitions of semantic edit intentions [49] for the three types of edits we used in our study: 1) citations, 2) point-of-view (POV), and 3) clarifications.

To label edits, our labeling interface shows one Wikipedia *edit diff* at a time, similar to the Wikipedia interface (Figure 5). The participant can label each edit by selecting checkboxes next to the three semantic intention labels (*citations, point-of-view, clarifications*), or selecting *None* (if none of those three labels apply), adding optional comment, and clicking on the submit button. We removed metadata information (edit comment, author, and the date of edit) to remove any source of bias or leaking of labels. The interface always displays one practice edit to get familiar with our labeling interface.



(a) Welcome page (b) Definitions page

Fig. 4. User study software welcome and the definitions pages.

4.2 Edit Diff Sampling Method

Just like it is costly and time consuming to get training labels manually, it is similarly costly and time consuming to get ground truth labels to validate our rules. However, we can still evaluate the effectiveness of our rules on much smaller samples than it would take to train a large model. Therefore, we only sample a small random set of *edit diffs* that Wikipedia experts can manually examine and label in a reasonable amount of time. Through pilot studies, we determined that 250 *edit diffs* is a reasonable upper bound that a Wikipedia editor can label in an hour.

We wanted to ensure a balanced representation of *edit diffs* with different semantic intents, but also have enough representative samples to properly estimate the false positive rate (i.e., the percentage of *edit diffs* that our rules wrongly label as positive examples). We hypothesized that the percentage of edits with some semantic intent categories would be small (e.g., point-of-view) compared to more common ones (e.g., copy editing, wikification), and that we would risk not including them in our sample if we simply randomly selected a small subset of edits from over one billion Wikipedia edits. Thus, we started with a random sample of about 100,000 *edit diffs* and pre-labeled them using our rules. We then created a stratified sample of 1,000 *edit diffs* by randomly selecting (without replacement) 100 pre-labeled point-of-view and clarification *edit diffs* each, and then randomly selecting another 800 *edit diffs* from the remaining *edit diffs* without replacement.

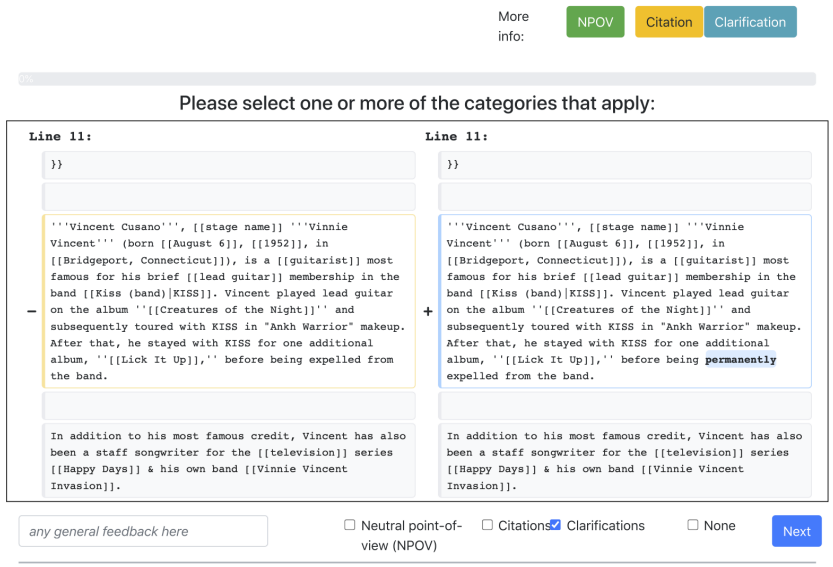


Fig. 5. User study interface for labeling semantics of a Wikipedia *edit diff*.

4.3 Participants

We recruited nine English Wikipedia editors as participants in our study. We recruited the participants by sending emails to Wikipedia editors (using the "mail" feature on their Wikipedia user pages) who frequently discuss articles on the "Featured Article Criteria" (FAC) discussion board [50] on the English Wikipedia. This discussion board is used for promoting articles to the "Featured Article" status; editors engaging on this forum are expected to have a good knowledge of Wikipedia editing and Wikipedia quality criteria. The average experience of the participants was approximately 10 years. The topic interests of the editors were varied including but not limited to: military history, medicine, weather, gaming, politics. Two participants were coordinators on individual Wikiprojects, one participant had an administrator role, and another participant was a new page patroller. The remaining five indicated they did not have any specific roles on Wikipedia. We compensated participants 30$ per hour of their time they spent participating in our study.

4.4 Tasks and Procedures

One of the authors acted as the investigator and conducted the study with each participant separately using video conferencing. The investigator briefed each participant about the study and then provided the participants with a link to our study labeling interface (Figure 4). Only participants that read the consent form, which was approved by our Institutional Review Board (IRB), and consented were allowed to proceed with the study.

The participants then read the tutorial (Figure 4b) and labeled one practice edit to get familiar with our labeling interface (Figure 5). This was followed by an *edit diff* labeling session where the interface asked participants to label the semantic intent category of 250 Wikipedia *edit diffs* one at a time. The investigator asked the participants to stop labeling when the hour was up or when they labeled 250 *edit diffs*, whichever came first. The participants each on average labeled 130 *edit diffs* (min=45, max=250). At the end of the study, the investigator asked the participants about their Wikipedia editing experience.

4.5 Results

The participants labeled a total of 434 out of 1,000 *edit diffs* in our sample. We ensured that each of the 434 *edit diffs* had labels from at least three different participants. We had multiple participants label the same *edit diffs* because point-of-view and clarifications are highly subjective categories, and a single participant may easily miss them. Thus, we assign a ground truth category to an *edit diff* if at least one participant labeled the *edit diff* with that category.

4.5.1 Participant Agreement. We first investigated the quality and ambiguities of the ground truth labels. The proportion of ground truth citations, point-of-view, and clarifications were: 24%, 41%, and 75% respectively. Note that they do not sum up to 100% because an *edit diff* can be in multiple categories. We then computed the agreement between different participants and their labels for the three categories using the Krippendorf alpha [28], which measures the inter-annotator agreement with more than two raters. The Krippendorf alpha [28] values for citations, point-of-view, and clarifications were: 0.59, 0.02, 0.17, indicating medium agreement for citations, but very low or no agreement for the other two categories.

We interpret that the high agreement of citation labels was likely due participants' common grounding about what constitutes adding a citation and what does not. However, some of the disagreement could have been due to a few participants who labeled some *edit diffs* as "citations added" even when an existing citation was modified, but no new citations were added.

Point-of-view and clarification ground truth labels had very high disagreement which could stem from the inherit ambiguity of the neutral point-of-view (NPOV) policy [51] and the clarification guidelines [52], and the differences in how different editors understand and interpret them. Three participants also mentioned that some *edit diffs* required specific knowledge on the topic of the article to determine if the changes were addressing point-of-view.

Previous work showed such disagreements in part-of-speech tagging [36], disagreements because of lack of labelers' expertise in the area [29, 41], and disagreements on controversial matters [21] (e.g., use of "annexation" *vs* "liberation" when describing a conflict based on the political and ideological stand of the editor). Disagreements on the question of neutrality is a reflection of the pluralist voices of a society [9]. Thus, we expect that such disagreement will also impact the effectiveness of our automated rules. In Section 6, we discuss how learning from spaces with such disagreements could provide a way to probe existing policies of a socio-technical system.

4.5.2 Effectiveness of Automated Rules. We then evaluated the effectiveness of our rules. The proportion of citations, point-of-view, and clarifications assigned by our rules were: 12%, 10%, and 18%. We computed the precision (i.e., how many of the *edit diffs* that our rules extracted were actually positive examples) and recall (i.e., how many of the *edit diffs* positive examples did our rules extract) of our automated labels compared to the ground truth (Figure 6). The precision was 94%, 70%, and 73% for citations, point-of-view, and clarification labels respectively. The recall was 49%, 17%, and 17% for citations, point-of-view, and clarifications respectively.

Our results show that our citation labels had high precision. This means that our rules are able to extract positive examples that are free from false positives (i.e., sentences that our rules indicate need improvement, but that do not actually need that specific improvement category). Although it is possible that some participants confused modifying citations for adding citations (as we noted above), note that this only affected the recall (made it lower), which does not have an impact on the ability of our rules to extract positive examples.

The precision of our point-of-view and clarification labels was not as high as with citations. However, considering that our point-of-view rules use comments from editors that explicitly indicate the semantic intent behind their *edit diff* (i.e., they were positive examples based on

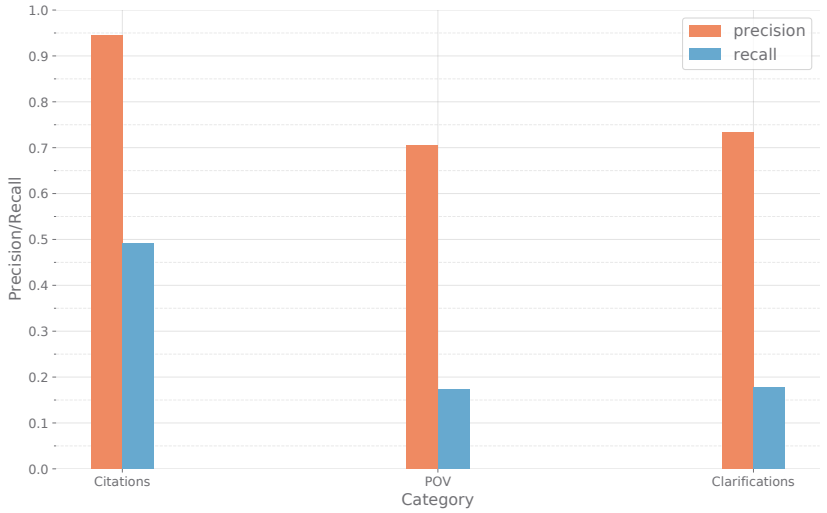


Fig. 6. Precision and Recall of our rules.

judgement from at least one editor), it is likely that our participants simply missed to properly label them or that they disagreed due to inherit ambiguity of point-of-view *edit diffs*.

Similarly, clarifications can also be ambiguous. However, since our clarification rules do not use explicit labels like point-of-view rules do, we had to perform additional manual error analysis to understand the disagreement between our rules and the ground truth. Thus, we performed visual examination of false positive *edit diffs* (i.e., *edit diffs* that our rules extracted as positive examples, but that our participants did not label as such). We found that most false positive clarification *edit diffs* could easily be confused for copy editing (i.e., adding small pieces of information), which often does add clarity to the edited sentence. In a few cases, the original editor indicated that they attempted to clarify content in the *edit diff* comment.

Nevertheless, the precision for all three categories is still encouraging. The low recall presents no concerns as long as we are able to extract enough positive examples using our rules-based method. Thus, we next evaluate our ability to extract labels using our semantic intent rules.

5 EVALUATION OF SENTENCE QUALITY LABELS

We evaluate the effectiveness of our automatically generated sentence quality labels, by training Machine Learning models with our labeled sentences to perform the task of classifying whether a Wikipedia sentence needs a specific semantic improvement or not. We will refer to sentence quality labels generated using our approach as *Edit-labels* henceforth.

We compare the effectiveness of our labels on classification of two semantic improvement tasks that existing research has attempted previously: 1) citations–given a Wikipedia sentence, identify whether it needs citations or not [39], and 2) point-of-view–given a Wikipedia sentence, identify whether it is biased or not according to Wikipedia's NPOV policy [22]. For the semantic category clarification, we do not have any prior work to compare against, but we include it to showcase a category that would have been challenging (or even impossible) to detect using the existing automated labeling methods. We provide our own analysis for this category and the interpretations

of the model output. Our code for labeling quality of Wikipedia sentences and associated sentence quality labeled data can be found here [3].

5.1 Wikipedia Sentence Sampling Method

We extracted 6.5 million Wikipedia *edit diffs* from a random sample of 100,000 Wikipedia articles with quality ranging from the most basic "stub" articles to "Featured Articles". We filtered out *edit diffs* that were reverts [10], because reverted changes are not part of any of our semantic intent categories. We used the *mwreverts* library[4] to label edits that were reverted and exclude them from the labeling step. We used a revert window of 15 edits and a maximum revert time of 2 days to identify *edit diffs* that were reverted [12].

We then labeled our set of non-reverted *edit diffs* with their semantic intent using our rules to get positive examples as we describe it in Section 3.1.4 . We extracted the negative examples from "Featured Articles" in our sample as described in Section 3.2. Table 3 shows the record counts in our final dataset for each of the three semantic categories: 1) citations, 2) point-of-view, and 3) clarifications. We further split our dataset into training (70%), validation (10%), and test (20%) sets.

Dataset	Train	Validation	Test	Total
Citations	68,000	9,780	19,561	97,341
Point-of-view	129,500	18,500	37,000	185,000
Clarifications	139,608	19,944	39,888	199,440

Table 3. Dataset splits for citations, point-of-view, and clarification categories created from edit-labels.

5.2 Model Training and Wikipedia Sentence Representation

In our evaluation, we used the same Recurrent Neural Network (RNN) models from previous work [22, 39] and only varied the labeled datasets they were trained on. We used GRU-based RNNs with global attention [2, 5], and implemented the models using tensorflow[5] with keras[6]. The input to these models is the sentence represented as a sequence of numerical features, one for each word. When training the models, we strip the input sentences of all the wiki-markup, remove special characters and convert all text to lowercase. We used the following input features:

5.2.1 Word Representation. First part of sentence representation consists of word embeddings $(w_1, w_2 ... w_n)$. We used GloVe word embeddings [35] to represent each word. The dimensions for each word embedding were $W_{emb} \in \mathbb{R}^{100}$.

5.2.2 Part of Speech Representation. In addition to word embeddings described above, we used the sequence of Part-of-speech (POS) tags (one for each word) $(p_1, p_2, ...p_n)$ as input features. Like the word feature, each POS tag is represented in the 100-dimensional embedding space $W_{pos} \in \mathbb{R}^{100}$. POS tags [3] have found extensive use in the NLP community because of their usefulness in capturing additional information about relations between words. Unlike GloVe word embeddings which were pre-trained, we trained the pos-tag embeddings with the classification task. We used POS tag representation for training only *point-of-view* and *clarification* detection models because baseline work on citations [39] did not use POS tags as input.

[3]https://github.com/comp-hci-lab/wikipedia_automatically_label_lowquality_content
[4]https://pythonhosted.org/mwreverts/
[5]https://www.tensorflow.org/
[6]https://keras.io/

5.2.3 Section representation. Previous work on detecting sentences that need citations [39] has shown that adding Wikipedia section representation along with word representations improved the model performance. This is because on English Wikipedia, different sections have different guidelines for citations[7]. For example, sections like *History* describing historical events need more citations than a section on the plot of a movie.

We used the section inputs only for citations detection. We first made Wikipedia section titles consistent with Wikipedia database format by converting spaces to underscores and the first character in the title to upper-case using *mediawiki-utilities*[8]. We then trained the section embedding matrix $W_S \in \mathbb{R}^{100}$ as part of the classification objective and combined it with the word embeddings.

5.3 Results

Here, we report the results of evaluation of individual semantic intent categories. Note that, unlike our label extraction that favors only extracting as few false negatives (i.e., high precision), the goal for sentence quality improvement detection task is to detect as many examples that need improvements as possible (i.e., high recall) without selecting too many false negatives (causing low precision). Thus, in our evaluation we focus on the F1-score (the harmonic mean of precision and recall) and favor models with high F1-score. However, when comparing models with similar F1-scores, having a higher recall should be preferred to detect as many sentences that need improvement.

5.3.1 Evaluation of labels for citations. Table 5 shows a comparison between the models trained on the baseline citations labeling method and our method. We tested both the model trained from the Featured Article dataset and the model trained on two datasets: 1) LQN-full, and 2) Edit-labels. We extracted the "LQN-full" dataset from the same articles as "LQN" in Redi et al.[39]. We used both citations and "citation-needed" tags as positive signals. Wikipedia editors add "citation-needed" tags on Wikipedia pages where they think a sentence needs a citation. Thus, in addition to actual citations, this is also a positive signal for "needing citations" because at least one Wikipedia editor thought so. We take sentences with "citation-needed" tags or actual citations on these articles as positive examples of needing citations. We take sentences in paragraphs with no citations or citation-needed tags as negative examples. We call our dataset "LQN-full" because we extract all the sentences from these articles giving us a large set of test sentences to evaluate citations (~300,000) compared to (~20,000) in the "LQN" dataset of the previous work. "Edit-labels test" dataset is the test split of the citation dataset created from "Edit-labels" described in Table 3.

		Training		Testing	
Training	Testing	Positive	Negative	Positive	Negative
Featured Articles	LQN-full	10,000	10,000	149,000	151,000
Featured Articles	Edit-labels	10,000	10,000	9,782	9,782
Edit-labels	LQN-full	39,122	39,122	149,000	151,000
Edit-labels	Edit-labels	39,122	39,122	9,782	9,782

Table 4. Statistics for the citation datasets used for training and testing.

For both the testing datasets, we can see that the model trained on edit-labels outperforms the model trained on limited examples from Featured Articles. We also see that when we test on a large set of sentences from low quality articles, both the approaches do not perform very well in terms of identifying citations. One of the reasons why the model trained on sentences from Featured articles

[7]https://en.wikipedia.org/wiki/Wikipedia:Citing_sources
[8]https://pythonhosted.org/mediawiki-utilities/lib/title.html

Training	Testing					
	LQN-full (300,000 samples)			Edit-labels test		
	P	R	F-1	P	R	F-1
Featured Articles	**0.65**	0.38	0.48	0.48	0.34	0.40
Edit-labels	0.54	**0.67**	**0.60**	**0.65**	**0.73**	**0.69**

Table 5. Testing results for citations when trained on labels from Featured Articles vs Edit-labels

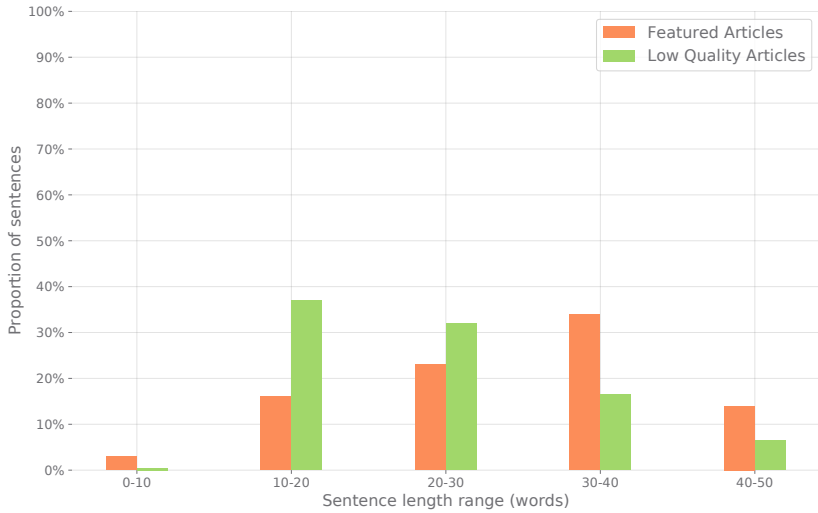


Fig. 7. Distribution of sentence length (words) in Featured and low-quality articles

does not perform very well is that the RNN model also learns other, potentially irrelevant patterns in the Featured articles (e.g., the length of the sentence). Figure 7 shows the proportion of sentences of varying lengths (in words) in Featured Articles and low-quality articles from our datasets.

The majority of Featured Article sentences lie in the range of 20-40 words whereas low-quality article sentences have lengths in the range 10-30 words. This is expected because Wikipedia editors perform an extensive review of Featured Article sentences and ensure that they are well formed[9]. Sentences in low-quality articles may just state a fact, not necessarily conveying the information in an elegant manner. Training on edit-labels captures this variation because we are training on sentences across the entire encyclopedia. We see that models trained on a large set of sentences extracted from edits across articles of all quality are better at generalizing to the task of identifying the need for citations.

5.3.2 Evaluation of labels for point-of-view. Table 6 shows the statistics of the datasets that we use for evaluating the labels for sentences with point-of-view issues. The "crowdsourced" set is taken from the baseline [22] work for point-of-view and consists of two datasets: "featured" and "cw-hard". Hube et al. [22] first randomly sampled 5,000 sentences from Wikipedia edit diffs that

[9]https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

		Training		Testing	
Training	Testing	Positive	Negative	Positive	Negative
Crowdsourced	Crowdsourced	1,290	1,290	370	370
Edit-labels	Crowdsourced	74,000	74,000	1,843	1,843
Edit-labels	Edit-labels	74,000	74,000	18,500	18,500

Table 6. Statistics for the POV datasets used for training and testing. Crowdsourced is from previous work

Testing	Featured (Crowdsourced)			cw-hard (Crowdsourced)			Edit-labels			
Training	P	R	F-1	P	R	F-1	P	R	F-1	ROC-AUC
Crowdsource trained	**0.89**	0.71	0.79	**0.71**	0.64	**0.67**	0.78	0.67	0.72	0.70
Edit-labels trained	0.79	**0.85**	**0.82**	0.42	**0.85**	0.56	**0.81**	**0.86**	**0.83**	**0.92**

Table 7. Testing results of Point-of-view with crowdsourcing and edit-labels training

	Sentence	Comment
1.	After the Battle of Nicopolis in 1396 and the fall of the Vidin Tsardom three years later, the Ottomans conquered all Bulgarian lands south of the Danube, with sporadic resistance ending when the Ottomans gained a firm hold on the Balkans by conquering Constantinople in 1453.	**OK**: This looks alright to me assuming it comes with a citation that supports the firm (long term) control of the Balkans. If it's not supported by the citation, it would be POV at best.
2.	It is widely accepted that the band is joke and is collectively thought to be the biggest sale out in rock history.	**POV**. *joke* is problematic
3.	Also known to have lots of friends, and it is possibly due to the fact they are known to be charming as well as loyal to friends and family members.	**POV**. *charming* should be removed
4.	The family was a branch of the FitzGerald dynasty, or Geraldines, related to the Earls of Desmond (extinct), who were questionably granted extensive lands in County Limerick by the Duke of Normandy by way of conquest.	**POV**. "Questionably" needs to be cited and it isn't phrased in a neutral tone. E.g. I would prefer "<someone> questioned the granted lands".
5.	In fact, this participation may be a reaction to the Catholic church's active political involvement.	**POV**. May be? Is? {{who}} said that?

Table 8. Manual assessment of a small sample of crowd-labeled neutral sentences

contained the comment "POV" and were single line changes. They labeled these 5,000 sentences as "biased" or "neutral" through crowdsourcing. They use the 1,843 crowdsourced labeled biased sentences from this set as positive examples in both the datasets: "featured" and "cw-hard". They

used the crowdsource labeled "neutral" sentences (3,157) as negative examples in the "cw-hard" dataset. They use sentences from "Featured Articles" as negative examples for the "featured" dataset, similar to our work.

Table 7 shows testing statistics of the models when trained on the two crowdsourced datasets from previous work and our test split of the edit labels dataset. We directly report the results of the model evaluation (GRU-based RNN with global attention and word and POS tags as input) from the previous work [22]. The model trained on our Edit-labels dataset outperforms the model trained on "featured" dataset when tested on the "featured" and Edit-labels dataset. The model did not perform at par with the baseline for the "cw-hard" dataset. The "cw-hard" dataset consists of negative examples which the crowdworkers labeled as "neutral". However, these sentences were sourced from Wikipedia edits with comment as "POV" (meaning at least one Wikipedian thought there was a bias in the sentence).

To get a better sense of the crowd-labeled neutral examples from this dataset, we manually assessed some "neutral" labeled examples from the "cw-hard" dataset for clarity. One of the authors, who has researched Wikipedia editors for about a decade and is also a Wikipedia editor, identified some issues with the crowd-labeled neutral sentences in the "cw-hard" dataset. Table 8 shows a small random sample of the negative (neutral labeled) examples from the "cw-hard" dataset with our own assessments along with the reasons. Four out of five sentences have point-of-view issues but they are not obvious without knowledge of the Wikipedia NPOV policies. One of the common reasons for a sentence having a point-of-view issue is not having a citation. This is because a citation pushes the opinion on the content, taking it away from the content writer which is acceptable[10]. Crowdworkers cannot be expected to be aware of such nuances unless explicitly explained.

Testing	Edit-labels			
Training	P	R	F-1	ROC-AUC
Edit-labels	0.75	0.75	0.75	0.83

Table 9. Testing results for the clarification category

5.3.3 Evaluation of labels for clarification . No prior research attempted to automatically detect whether a sentence needs clarification, hence we report and interpret our evaluation on the test set for clarifications. Table 9 shows the results for clarification evaluations. We are able to achieve an F1-score of 0.75 for the clarification category and ROC-AUC of 0.83. In table 10 we show some examples where clarifications were made, what clarifications were made and whether the model correctly flagged the need for the same. TP stands for true positives and FP stands for false positives. Edit-labels is used for both training and testing. Refer to Table 3 for the statistics of the training and testing splits of the Edit-labels dataset for the clarification category.

5.4 Attention

Attention over RNN models [2] allows RNN models to place different weights on different words in the input sentence while making the predictions. We use these weights to visualize the focus that the classifier places on the different words for the given sentence. This is useful for analyzing the model outputs by the users in actual deployments. Editors can look at the most important words for a given task and take quick decisions instead of reading the full sentence. Figure 8 shows the attention weights for some sentences from each of citations, point-of-view and clarifications. Blue

[10]https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view#Explanation_of_the_neutral_point_of_view

revision-id	Wikipedia sentence	Prediction	Clarification
21577990	In sum, NLP promotes methods which are verifiable and have so far been found to be largely false, inaccurate or ineffective.	TP	..which are *largely* verifiable..
459001268	It debuted at #1 on the "New York Times" Bestseller List ("Fallen" came in at # 2), remaining at that position through the week of October 17.	TP	..came in *that week* at..
669670844	During this period, Sonnenblick made the two great contributions that would define his career.	TP	..two great *scientific* contributions
N/A	In his 2013 autobiography, Jackson stated that there was, and that Martin and some white Yankees would tell racist jokes.	FP	N/A
810160688	with his leg injured he is barely able to get away but is rescued by a kingdom soldier that is still alive	TP	*..rescued by Alavaro*, a kingdom..

Table 10. Manual assessment of clarification examples

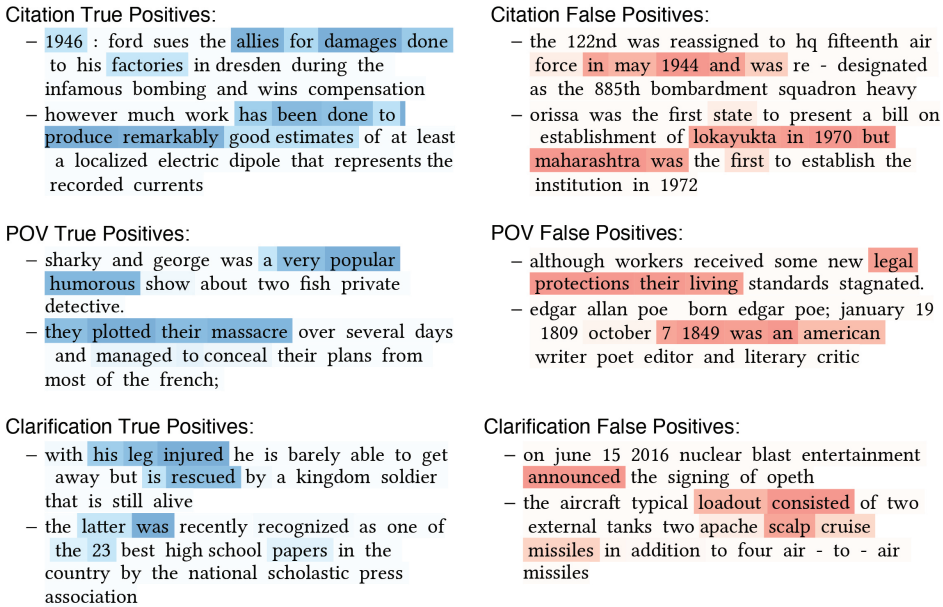


Fig. 8. Attention visualization of examples for the three categories

is used for true positives and red is used for false positives. The darker the word, the more the attention on that word for prediction for the specific sentence.

For citations, attention is given to reporting verbs like *damages done* or *produce remarkably good estimates* which require proof of opinion. This is in line with previous work[39] which reports that such verbs or presence of facts lead to high likelihood of needing citations. For point-of-view,

attention seems to be particularly helpful as focus is given on words such as *very popular*, *formidable*, *plotted their massacre* which are the typical deletions made in inline point-of-view edits as per the guideline - *avoid stating opinion as facts* [11]. This is because, if such words are uncited, they cause a point-of-view issue as the opinion becomes an opinion of the content writer. For clarifications, attention is placed near the words where additions were made in the sentence. For example, consider sentence 1 in clarification true positives, *rescued by a Kingdom soldier* is clarified to *rescued by Alavaro, a kingdom soldier* and a high attention is placed on rescued (see table 10 - last example).

6 DISCUSSION

Our Wikipedia sentence quality labeling method and detection pipeline has several implications for both Wikipedia and collaborative systems more broadly. By learning implicitly from expert behaviors, we introduce a flexible and fast labeling method. However, our labeling method is not free from inherent ambiguities of assessing content quality, which has implications not only for how we pick our labels, but also how we train our models to detect low quality content on collaborative platforms. Finally, the models that we train from implicit behaviors enable a new family of technical probes that can provide insight into how Wikipedia editors enact existing Wikipedia policies and guidelines in practice.

6.1 Flexible, Fast, and Effective Labeling

Our method can extract more data, more quickly compared to existing methods. By demonstrating the application of our sentence quality detection pipeline on three categories of semantic edit intents (citations, point-of-view, and clarifications), we showed that our method can be used to extract order of magnitude more data than what crowd-sourcing can get. Further, we have shown that models trained from Wikipedian behavior traces outperformed those trained using existing labeling methods.

One of the benefits of our method is that it can continue to grow the size of training datasets as new Wikipedia edits become available. This means that our method could quickly adapt labels to reflect changes in editor behaviors over time, something that would render models trained on existing labels ineffective. Even if there are updates to the semantic edit categorization, our automated method only requires updates to the corresponding rules, unlike existing manual labeling methods that would require us to hire crowdworkers or Wikipedia editors to relabel the data.

We use precision and recall metrics to evaluate our labeling approach against prior work [22, 39]. However, for practical use, sentence quality detection models trained using our labeling approach can be used to flag problematic sentences (e.g., needing citations, needing point-of-view edits) on Wikipedia articles based on a manually chosen threshold. If more Wikipedia editors are available to assess the predictions, a low confidence threshold of the classifier can be used to flag more examples with issues (high recall). If the Wikipedia community wants only high quality predictions to avoid false positives, predictions with a high confidence threshold can be made (high precision).

6.2 Inherit Ambiguity of Semantic Intent Category Descriptions and Labels

Our findings point to the inherent ambiguity of content quality labels and different semantic edit intents that we base our rules on. This was evident in the high disagreements between editors in our study. The diverse background of Wikipedia editors brings in a pluralist mindset [31], which introduces different interpretations of what requires improvement and in turn introduces ambiguity into content quality labels. Also, some edit type categories, such as point-of-view (POV), are highly

[11] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view#Explanation_of_the_neutral_point_of_view

topic dependent and some editors expressed their inability to correctly judge the category of an edit because of being non-experts in that area.

However, some of the disagreements could also come from ambiguities in the Wikipedia's description of different edit type categories [49]. For example, our participants' comments during the study implied that the existing definition of clarifications is fairly ambiguous, which could account for some of our participants' disagreement on the clarification labels. However, their comments also indicated that the ambiguity could come from overlaps in edit type definitions. For example, clarifications are also difficult to distinguish from elaborations when information is added in the same line. One potential way to reduce some of this ambiguity could be to introduce more concrete examples for each semantic category (e.g., what needs a clarification: a missing date, missing location, or missing profession).

Our findings also indicate that point-of-view issues have more nuance beyond "peacock terms /weasel words" that reflect opinionated language. A major cause of point-of-view issues arises due to lack of citations as lack of credibility indicates that the content is likely an opinion of the author of the content. Many point-of-view issues also arise from different ideological stands of the editors. For example, a contention could be regarding the use of the word "annexation" vs "liberation" when describing a conflict on Wikipedia, based on which ideological side the editor belongs to.

6.3 Models Learned from Implicit Behaviors as Technical Probes for Wikipedia Norms

Our method operationalizes Wikipedia written policy genres (e.g., policies, guidelines, and "essays" [33]) to automatically extract quality labels from Wikipedia editor behaviors that capture how editors interpret and enact current policy genres. Wikipedia written policy genres are examples of Wikipedia online community injunctive norms (i.e., shared beliefs within a group, community or culture that "ought" to be followed), while Wikipedia editor behaviors constitute a set of descriptive norms (i.e., typical behaviors of individuals within a group, community or culture) [6].

The most common role of Machine Learning (ML) algorithms that we can train using our labels is supporting and enforcing Wikipedia's norms. For example, an ML algorithm that automatically flags low quality Wikipedia sentences is a technical representation of policy. Recent work in this area of algorithmic governance [30, 34] discussed the rising role of algorithms as norm enforcement mechanisms in Wikipedia. However, that work strictly explores enforcing Wikipedia injunctive norms. Instead, our approach of modeling Wikipedian behavior—capturing not only the explicit rules, but rather describing how the rules are applied in practice—extends this notion of governance from originating from formal consensus (top down) to a description of actual practice (bottom up).

We further envision the role of models trained on our labels as supporting resources to reflect on Wikipedia norms themselves. Much like Wikipedia "essays" that document editor practices and provide a "soft regulatory mechanism" where a formal policy is insufficiently specific to apply [33], predictive models that describe Wikipedian behavior when referencing a formal policy afford the ability to explore probable applications of that policy in new contexts. Like "essays", this mixing of context with formal norm interpretation has the potential to also fill a niche in the broader normative ecology of Wikipedia and to extend the interpretability and consistent application of more abstract, formal policies in the socio-technical practice of editing Wikipedia articles.

7 CONCLUSION AND FUTURE WORK

Our work showed how content policies of Wikipedia can be coded into programmatic rules to identify the semantic intents of Wikipedia edits, which can then be used to detect sentences with quality issues at scale. We showed that deep learning models can be trained using these sentences labeled with quality issues. These models can then identify the same type of quality issues in new unseen Wikipedia sentences.

Future work can directly leverage our proposed method to build new sentence quality identification pipelines to flag issues in sentences along other semantic intent categories. For example, experienced editors in our study mentioned performing a lot of copy edits, which involve fixing the grammar in the articles as a lot of Wikipedia editors are not native speakers of English. Rephrasing the tone of the article to a more encyclopedic one is another commonly performed action on the articles. One editor mentioned that "articles are written in a victorian style storytelling manner and that has to be rephrased to a more encyclopedic tone".

Our sentence quality labeling approach could be used to assess whether Wikipedia editors improve content quality according to existing injunctive norms. This is because policies [14] mediate behavior in socio-technical systems like Wikipedia and the models trained on edits of Wikipedians capture such policy guided behaviors. By providing a way to bridge the gap between descriptive and injunctive norms [6] future work should explore how to make normative behavior more consistent on Wikipedia. Moreover, future work should explore how our approach could give more structure to the traditional way of guiding newcomers and passing knowledge by distilling policy guided human behaviors of the community into machine learning models and show a way to use the output of the models as an aid to understand and reason about the policies.

Our entire pipeline (labeling semantic intention of Wikipedia *edit diffs* and extracting sentences from Wikipedia articles that need improvements) is language agnostic. Future work should explore how to develop effective rules for other language Wikipedias. For example, for low resource Wikipedia language editions, effective rules can be built by exploiting semantic relatedness [20] of concepts from low resource to high resource Wikipedia language editions. Point-of-view edits in a target language Wikipedia could be detected by starting with point-of-view edits in English Wikipedia, and finding edits in the target language Wikipedia that are semantically "close".

ACKNOWLEDGMENTS

This work was funded by Toyota Research Institute (TRI). We thank members of the Machine Assisted Cognition group at TRI for valuable discussions about this work. We are grateful to Jonathan Morgan from Wikimedia Foundation for insightful discussions on injunctive and disjunctive norms in online spaces, and Diyi Yang from Georgia Tech for providing us with the classifier for the semantic edit intentions dataset. We also thank members of the CompHCI lab at the University of Michigan, Nel Escher and Divya Ramesh, for their input on ambiguity of labels in collaborative spaces, and Anindya Das Antar and Snehal Prabhudesai, for their help in testing an early prototype of the study software.

REFERENCES

[1] Anderka, M., Stein, B., and Lipka, N. Predicting quality flaws in user-generated content: the case of wikipedia. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012* (2012), W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, Eds., ACM, pp. 981–990.

[2] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Y. Bengio and Y. LeCun, Eds.

[3] Biber, D. *Variation across speech and writing*. Cambridge University Press, 1991.

[4] Chandrasekharan, E., Gandhi, C., Mustelier, M. W., and Gilbert, E. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction 3*, CSCW (2019), 1–30.

[5] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP* (2014), ACL, pp. 1724–1734.

[6] Cialdini, R. B., and Goldstein, N. J. Social influence: Compliance and conformity. *Annu. Rev. Psychol. 55* (2004), 591–621.

[7] Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. Automatic quality assessment of content created

collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, Austin, TX, USA, June 15-19, 2009* (2009), F. Heath, M. L. Rice-Lively, and R. Furuta, Eds., ACM, pp. 295–304.

[8] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR) 51*, 1 (2018), 1–40.

[9] Daston, L. Objectivity and the escape from perspective. *Social Studies of Science 22*, 4 (1992), 597–618.

[10] Ekstrand, M. D., and Riedl, J. rv you're dumb: identifying discarded work in wiki article history. In *Proceedings of the 2009 International Symposium on Wikis, 2009, Orlando, Florida, USA, October 25-27, 2009* (2009), D. Riehle and A. Bruckman, Eds., ACM.

[11] Fiesler, C., Jiang, J., McCann, J., Frye, K., and Brubaker, J. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media* (2018), vol. 12.

[12] Flöck, F., Vrandecic, D., and Simperl, E. Revisiting reverts: accurate revert detection in wikipedia. In *23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25-28, 2012* (2012), E. V. Munson and M. Strohmaier, Eds., ACM, pp. 3–12.

[13] Forte, A., Andalibi, N., Gorichanaz, T., Kim, M. C., Park, T., and Halfaker, A. Information fortification: An online citation behavior. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork* (New York, NY, USA, 2018), GROUP '18, Association for Computing Machinery, p. 83–92.

[14] Forte, A., Larco, V., and Bruckman, A. Decentralization in wikipedia governance. *Journal of Management Information Systems 26*, 1 (2009), 49–72.

[15] Geiger, R. S., and Halfaker, A. When the levee breaks: without bots, what happens to wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, August 05 - 07, 2013* (2013), A. Aguiar and D. Riehle, Eds., ACM, pp. 6:1–6:6.

[16] Geiger, R. S., and Halfaker, A. Operationalizing conflict and cooperation between automated software agents in wikipedia: A replication and expansion of 'even good bots fight'. *Proc. ACM Hum. Comput. Interact. 1*, CSCW (2017), 49:1–49:33.

[17] Geiger, R. S., and Ribes, D. The work of sustaining order in wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (2010), pp. 117–126.

[18] Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from? In *FAT** (2020), ACM, pp. 325–336.

[19] Halfaker, A. Interpolating quality dynamics in wikipedia and demonstrating the keilana effect. In *OpenSym* (2017), ACM, pp. 19:1–19:9.

[20] Hassan, S., and Mihalcea, R. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (2009), pp. 1192–1201.

[21] Hickman, M. G., Pasad, V., Sanghavi, H., Thebault-Spieker, J., and Lee, S. W. Wiki hues: Understanding wikipedia practices through hindi, urdu, and english takes on evolving regional conflict. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development* (2020), pp. 1–5.

[22] Hube, C., and Fetahu, B. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019* (2019), J. S. Culpepper, A. Moffat, P. N. Bennett, and K. Lerman, Eds., ACM, pp. 195–203.

[23] Jhaver, S., Birman, I., Gilbert, E., and Bruckman, A. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI) 26*, 5 (2019), 1–35.

[24] Kairam, S., and Heer, J. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), pp. 1637–1648.

[25] Keegan, B., and Fiesler, C. The evolution and consequences of peer producing wikipedia's rules. In *Proceedings of the International AAAI Conference on Web and Social Media* (2017), vol. 11.

[26] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), CHI '08, Association for Computing Machinery, p. 453–456.

[27] Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. He says, she says: conflict and coordination in wikipedia. In *CHI* (2007), ACM, pp. 453–462.

[28] Klaus, K. Content analysis: An introduction to its methodology, 1980.

[29] Lee, S. W., Krosnick, R., Park, S. Y., Keelean, B., Vaidya, S., O'Keefe, S. D., and Lasecki, W. S. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–23.

[30] Lovink, G., Tkacz, N., Reagle, J. M., O'Sullivan, D., Liang, L., Salah, A., Gao, C., Suchecki, K., Scharnhorst, A., Geiger, R., et al. Critical point of view: A wikipedia reader.
[31] Matei, S. A., and Dobrescu, C. Wikipedia's "neutral point of view": Settling conflict through ambiguity. *The Information Society 27*, 1 (2011), 40–51.
[32] Morgan, J. T., and Filippova, A. 'welcome'changes? descriptive and injunctive norms in a wikipedia sub-community. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–26.
[33] Morgan, J. T., and Zachry, M. Negotiating with angry mastodons: the wikipedia policy environment as genre ecology. In *Proceedings of the 16th ACM international conference on Supporting group work* (2010), pp. 165–168.
[34] Müller-Birn, C., Dobusch, L., and Herbsleb, J. D. Work-to-rule: the emergence of algorithmic governance in wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies* (2013), pp. 80–89.
[35] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (2014), A. Moschitti, B. Pang, and W. Daelemans, Eds., ACL, pp. 1532–1543.
[36] Plank, B., Hovy, D., and Søgaard, A. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2014), pp. 507–511.
[37] Potthast, M. Crowdsourcing a wikipedia vandalism corpus. In *SIGIR* (2010), ACM, pp. 789–790.
[38] Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K. A., Terveen, L. G., and Riedl, J. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2007, Sanibel Island, Florida, USA, November 4-7, 2007* (2007), T. Gross and K. Inkpen, Eds., ACM, pp. 259–268.
[39] Redi, M., Fetahu, B., Morgan, J. T., and Taraborelli, D. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (2019), L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds., ACM, pp. 1567–1578.
[40] Ruprechter, T., Santos, T., and Helic, D. Relating wikipedia article quality to edit behavior and link structure. *Applied Network Science 5*, 1 (2020), 1–20.
[41] Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., Russell, J., Wang, Z., and Hecht, B. Turkers, scholars," arafat" and" peace" cultural communities and algorithmic gold standards. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (2015), pp. 826–838.
[42] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. Information quality work organization in wikipedia. *Journal of the American society for information science and technology 59*, 6 (2008), 983–1001.
[43] Suh, B., Convertino, G., Chi, E. H., and Pirolli, P. The singularity is not near: slowing growth of wikipedia. In *Proceedings of the 2009 International Symposium on Wikis, 2009, Orlando, Florida, USA, October 25-27, 2009* (2009), D. Riehle and A. Bruckman, Eds., ACM.
[44] Wang, R. Y., and Strong, D. M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst. 12*, 4 (1996), 5–33.
[45] Warncke-Wang, M., Cosley, D., and Riedl, J. Tell me more: an actionable quality model for wikipedia. In *OpenSym* (2013), ACM, pp. 8:1–8:10.
[46] Wikipedia. Wikipedia, sep 2020.
[47] Wikipedia. Wikipedia:content assessment, sep 2020.
[48] Wikipedia. Wikipedia:core content policies, sep 2020.
[49] Wikipedia. Wikipedia:edittypes taxnonmy, sep 2020.
[50] Wikipedia. Wikipedia:featured articles candidates, sep 2020.
[51] Wikipedia. Wikipedia:neutral point of view, sep 2020.
[52] Wikipedia. Wikipedia:please clarify, sep 2020.
[53] Wikipedia. Wikipedia:template cleanup, sep 2020.
[54] Wulczyn, E., West, R., Zia, L., and Leskovec, J. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web* (2016), pp. 975–985.
[55] Yang, D., Halfaker, A., Kraut, R. E., and Hovy, E. H. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017* (2017), M. Palmer, R. Hwa, and S. Riedel, Eds., Association for Computational Linguistics, pp. 2000–2010.

Received October 2020; revised April 2021; accepted July 2021