Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability

Miriam Redi
Wikimedia Foundation
London, UK

Besnik Fetahu
L3S Research Center
Leibniz University of Hannover

Jonathan Morgan
Wikimedia Foundation
Seattle, WA

Dario Taraborelli
Wikimedia Foundation
San Francisco, CA

ABSTRACT

Wikipedia is playing an increasingly central role on the web, and the policies its contributors follow when sourcing and fact-checking content affect million of readers. Among these core guiding principles, *verifiability* policies have a particularly important role. Verifiability requires that information included in a Wikipedia article be corroborated against *reliable secondary sources*. Because of the manual labor needed to curate Wikipedia at scale, however, its contents do not always evenly comply with these policies. Citations (i.e. reference to external sources) may not conform to verifiability requirements or may be missing altogether, potentially weakening the reliability of specific topic areas of the free encyclopedia. In this paper, we aim to provide an empirical characterization of the reasons *why* and *how* Wikipedia cites external sources to comply with its own verifiability guidelines. First, we construct a *taxonomy of reasons* why inline citations are required, by collecting labeled data from editors of multiple Wikipedia language editions. We then crowdsource a large-scale dataset of Wikipedia sentences annotated with categories derived from this taxonomy. Finally, we design algorithmic models to determine if a statement *requires a citation*, and to predict the *citation reason*. We evaluate the accuracy of such models across different classes of Wikipedia articles of varying quality, and on external datasets of claims annotated for fact-checking purposes.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; *Natural language processing*; • **Information systems** → *Crowdsourcing*; • **Human-centered computing** → *Wikis*.

KEYWORDS

Citations; Wikipedia; Crowdsourcing; Neural Networks;

ACM Reference Format:
Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313618

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313618

1 INTRODUCTION

Wikipedia is playing an increasingly important role as a "neutral" arbiter of the factual accuracy of information published in the web. Search engines like Google systematically pull content from Wikipedia and display it alongside search results [38], while large social platforms have started experimenting with links to Wikipedia articles, in an effort to tackle the spread of disinformation [37].

Research on the accuracy of information available on Wikipedia suggests that despite its radical openness—anyone can edit most articles, often without having an account —the confidence that other platforms place in the factual accuracy of Wikipedia is largely justified. Multiple studies have shown that Wikipedia's content across topics is of a generally high quality[21, 34], that the vast majority of vandalism contributions are quickly corrected [20, 33, 42], and that Wikipedia's decentralizedprocess for vetting information works effectively even under conditions where reliable information is hard to come by,such as in breaking news events [27].

Wikipedia's editor communities govern themselves through a set of collaboratively-created policies and guidelines [6, 19]. Among those, the Verifiability policy[1] is a key mechanism that allows Wikipedia to maintain its quality. Verifiability mandates that, in principle, "all material in Wikipedia... articles must be verifiable" and attributed to reliable secondary sources, ideally through inline citations, and that unsourced material should be removed or challenged with a *{citation needed}* flag.

While the role citations serve to meet this requirement is straightforward, the process by which editors determine which claims require citations, and why those claims need citations, are less well understood. In reality, almost all Wikipedia articles contain at least some unverified claims, and while high quality articles may cite hundreds of sources, recent estimates suggest that the proportion of articles with few or no references can be substantial [35]. While as of February 2019 there exists more than 350,000 articles with one or more *{citation needed}* flag, we might be missing many more.

Furthermore, previous research suggests that editor citation practices are not systematic, but often contextual and ad hoc. Forte et al. [17] demonstrated that Wikipedia editors add citations primarily for the purposes of "information fortification": adding citations to protect information that they believe may be removed by other editors. Chen et al. [10] found evidence that editors often add citations to existing statements relatively late in an article's lifecycle. We submit that by understanding the reasons why editors prioritize

[1] https://en.wikipedia.org/wiki/Wikipedia:Verifiability

adding citations to some statements over others we can support the development of systems to scale volunteer-driven verification and fact-checking, potentially increasing Wikipedia's long-term reliability and making it more robust against information quality degradation and coordinated disinformation campaigns.

Through a combination of qualitative and quantitative methods, we conduct a systematic assessment of the application of Wikipedia's verifiability policies at scale. We explore this problem throughout this paper by focusing on two tasks:

(1) Citation Need: identifying *which* statements need a citation.
(2) Citation Reason: identifying *why* a citation is needed.

By characterizing qualitatively and algorithmically these two tasks, this paper makes the following contributions:

- We develop a Citation Reason Taxonomy[2] describing reasons why individual sentences in Wikipedia articles require citations, based on verifiability policies as well as labels collected from editors of the English, French, and Italian Wikipedia (See Sec. 3).
- We assess the validity of this taxonomy and the corresponding labels through a crowdsourcing experiment, as shown in Sec. 4. We find that sentences needing citations in Wikipedia are more likely to be historical facts, statistics or direct/reported speech. We publicly release this data as a Citation Reason corpus.
- We train a deep learning model to perform the two tasks, as shown in Secc. 5 and 6. We demonstrate the high accuracy (F1=0.9) and generalizability of the Citation Need model, explaining its predictions by inspecting the network's attention weights.

These contributions open a number of further directions, both theoretical and practical, that go beyond Wikipedia and that we discuss in Section 7.

2 RELATED WORK

The contributions described in this paper build on three distinct bodies of work: crowdsourcing studies comparing the judgments of domain experts and non-experts, machine-assisted citation recommendations on Wikipedia, and automated detection and verification of factual claims in political debates.

Crowdsourcing Judgments from Non-Experts. Training machine learning models to perform the citation need and citation reason tasks requires large-scale data annotations. While generating data for the first task necessarily requires expert knowledge (based on understanding of policies), we posit that defining the *reasons* why a citation that has already been deemed appropriate is needed can be effectively performed by people without domain expertise, such as crowdworkers.

Obtaining consistent and accurate judgments from untrained crowdworkers can be a challenge, particularly for tasks that require contextual information or domain knowledge. However, a study led by Kittur [31] found that crowdworkers were able to provide article quality assessments that mirrored assessments made by Wikipedians by providing clear definitions and instructions, and by focusing the crowdworkers attention on the aspects of the article that provided relevant evaluation criteria. Similarly, Sen et al. [46] demonstrated that crowdworkers are able to provide semantic relatedness judgments as scholars when presented with keywords related to general knowledge categories.

Our labeling approach aims to assess whether crowdworkers and experts (Wikipedians) agree in their understanding of verifiability policies—specifically, whether non-experts can provide reliable judgments on the reasons why individual statements need citations.

Recommending Sources. Our work is related to a body of bibliometrics works on citation analysis in academic texts. These include unsupervised methods for citation recommendation in articles [24], and supervised models to identify the purpose of citations in academic manuscripts[1]. Our work explores similar problems in the different domain of Wikipedia articles: while scholarly literature cites work for different purposes[1] to support original research, the aim of Wikipedia's citations is to verify existing knowledge.

Previous work on the task of source recommendation in Wikipedia has focused on cases where statements are marked with a *citation needed* tag [14–16, 44]. Sauper et al. [14, 44] focused on adding missing information in Wikipedia articles from external sources like news, where the corresponding Wikipedia entity is a salient concept. In another study [16], Fetahu et al. used existing statements that have either an outdated citation or *citation needed* tag to query for relevant citations in a news corpus. Finally, the authors in [15], attempted to determine the *citation span*—that is, which parts of the paragraph are covered by the citation—for any given existing citation in a Wikipedia article and the corresponding paragraph in which it is cited.

None of these studies provides methods to determine whether a given (untagged) statement *should* have a citation and *why* based on the citation guidelines of Wikipedia.

Fact Checking and Verification. Automated verification and fact-checking efforts are also relevant to our task of computationally understanding verifiability on Wikipedia. Fact checking is the process of assessing the veracity of factual claims [45]. Long et al. [36] propose TruthTeller computes annotation types for all verbs, nouns, and adjectives, which are later used to predict the truth of a clause or a predicate. Stanovsky et al. [47] build upon the output rules from TruthTeller and use those as features in a supervised model to predict the factuality label of a predicate. Chung and Kim [13] assess source credibility through a questionnaire and a set of measures (e.g. informativeness, diversity of opinions, etc.). The largest fact extraction and verification dataset FEVER [49] constructs pairs of factual snippets and paragraphs from Wikipedia articles which serve as evidence for those factual snippets. However,these approaches cannot be applied in our case because they make the assumption that any provided statement is of factual nature.

Research on the automated fact detectors in political discourse [23, 32, 39] is the work in this domain that is most closely related to ours. While these efforts have demonstrated the ability to effectively detect the presence of facts to be checked, they focus on the political discourse only, and they do not provide explanation for the models' prediction. In our work, we consider a wide variety of topics—any topic covered in Wikipedia—and design models able to not only detect claims, but also explain the reasons why those claims require citations.

[2]We use here the term "taxonomy" in this context as a synonym of *coding scheme.*

Table 1: A taxonomy of Wikipedia verifiability: set of reasons for adding and not adding a citation. This taxonomy is the result of a qualitative analysis of various sources of information regarding Wikipedia editors' referencing behavior.

Reasons why citations are needed	
Quotation	The statement appears to be a direct quotation or close paraphrase of a source
Statistics	The statement contains statistics or data
Controversial	The statement contains surprising or potentially controversial claims - e.g. a conspiracy theory
Opinion	The statement contains claims about a person's subjective opinion or idea about something
Private Life	The statement contains claims about a person's private life - e.g. date of birth, relationship status.
Scientific	The statement contains technical or scientific claims
Historical	The statement contains claims about general or historical facts that are not common knowledge
Other	The statement requires a citation for reasons not listed above (please describe your reason in a sentence or two)
Reasons why citations are not needed	
Common Knowledge	The statement only contains common knowledge - e.g. established historical or observable facts
Main Section	The statement is in the lead section and its content is referenced elsewhere in the article
Plot	The statement is about a plot or character of a book/movie that is the main subject of the article
Already Cited	The statement only contains claims that have been referenced elsewhere in the paragraph or article
Other	The statement does not require a citation for reasons not listed above (please describe your reason in a sentence or two)

3 A TAXONOMY OF CITATION REASONS

To train models for the Citation Need and Citation Reason tasks, we need to develop a systematic way to operationalize the notion of verifiability in the context of Wikipedia. There is currently no single, definitive taxonomy of reasons why a particular statement in Wikipedia should, or should not, have a supporting inline citation. We drew on several data sources to develop such a taxonomy using an inductive, mixed-methods approach.

Analyzing Citation Needed Templates. We first analyzed reasons Wikipedia editors provide when requesting an inline citation. Whenever an editor adds a *citation needed* tag to a claim that they believe should be attributed to an external source, they have the option to specify a reason via a free-form text field. We extracted the text of this field from more than 200,000 *citation needed* tags added by English Wikipedia editors and converted it into a numerical feature by averaging the vector representations of each sentence word, using Fasttext [8]. We then used k-means to cluster the resulting features into 10 clusters (choosing the number of clusters with the elbow method [28]). Each cluster contains groups of consistent reasons why editors requested a citation. By analyzing these clusters, we see that the usage of the "reason" field associated with the *citation needed* tag does not consistently specify the reason why these tags are added. Instead, it is often used to provide other types of contextual information—for example, to flag broken links or unreliable sources, to specify the date when the tag was added, or to provide very general explanations for the edit. Therefore, we did not use this data to develop our taxonomy.

Analyzing Wikipedia Citation Policies. As a next step, we analyzed documentation developed by the editor community to describe rules and norms to be followed when adding citations. We examined documentation pages in the English, French, and Italian language editions. Since each Wikipedia language edition has its own citation policies, we narrowed down the set of documents to analyze by identifying all subsidiary rules, style guides, and lists of best practices linked from the main Verifiability policy page, which exists across all three languages. Although these documents slightly differ across languages, they can be summarized into 28 distinct rules [3]. Rules that we identified across these pages include a variety of types of claims that should usually or always be referenced to a source, such as claims of scientific facts, or any claim that is likely to be unexpected or counter-intuitive. These documentation pages also contain important guidance on circumstances under which it is appropriate to *not* include an inline citation. For example, when the same claim is made in the lead section as well as in the main body of the article, it is standard practice to leave first instance of the claim unsourced.

Asking Expert Wikipedians. To expand our Citation Reason Taxonomy, we asked a group of 36 Wikipedia editors from all three language communities (18 from English Wikipedia, 7 from French Wikipedia, and 11 from Italian Wikipedia) to annotate citations with reasons. Our experiment was as follows: we extracted sentences with and without citations from a set of Featured Articles and removed the citation metadata from each sentence. Using WikiLabels[4], an open-source tool designed to collect labeled data from Wikipedia contributors, we showed our annotators the original article with all citation markers removed and with a random selection of sentences highlighted. Editors were then asked to decide whether the sentence needed a citation or not (Citation Need task), and to specify a reason for their choices (Citation Reason task) in a free-text form. We clustered the resulting answers using the same methodology as above, and used these clusters to identify additional reasons for citing claims.

[3]The full guideline summary and the cluster analysis can be found here: https://figshare.com/articles/Summaries_of_Policies_and_Rules_for_Adding_Citations_to_Wikipedia/7751027

[4]https://meta.wikimedia.org/wiki/Wiki_labels

Our final set of 13 discrete reasons (8 for adding and 5 for not adding) is presented in Table 1. In Section 4, we evaluate the accuracy of this taxonomy and use it to label a large number of sentences with citation-needed reasons.

4 DATASETS

In this Section, we show how we collected data to train models able to perform the Citation Need task, for which we need sentences with binary citation/no-citation labels, and the Citation Reason task, for which we need sentences labeled with one of the reason category from our taxonomy.

4.1 Citation Need Dataset

Previous research [17] suggests that the decision of whether or not to add a citation, or a *citation needed* tag, to a claim in a Wikipedia article can be highly contextual, and that doing so reliably requires a background in editing Wikipedia and potentially domain knowledge as well. Therefore, to collect data for the Citation Need task we resort to expert judgments by Wikipedia editors.

Wikipedia articles are rated and ranked into ordinal quality classes, from "stub" (very short articles) to "Featured". Featured Articles[5] are those articles that are deemed as the highest quality by Wikipedia editors based on a multidimensional quality assessment scale [6]. One of the criteria used in assessing Featured Articles is that the information in the article is *well-researched*.[7] This criterion suggests that Featured Articles are more likely to consistently reflect best practices for when and why to add citations than lower-quality articles. The presence of *citation needed* tags is an additional signal we can use, as it indicates that at least one editor believed that a sentence requires further verification.

We created three distinct datasets to train models predicting if a statement requires a citation or not[8]. Each dataset consists of: (i) *positive instances* and (ii) *negative instances*. Statements with an *inline citation* are considered as *positives*, and statements *without an inline citation* and that appear in a *paragraph with no citation* are considered as *negatives*.

Featured – FA. From the set of 5,260 *Featured* Wikipedia articles we randomly sampled 10,000 positive instances, and equal number of negative instances.

Low Quality (citation needed) – LQN. In this dataset, we sample for statements from the 26,140 articles where at least one of the statements contains a *citation needed* tag. The *positive instances* consist solely of statements with *citation needed* tags.

Random – RND. In the random dataset, we sample for a total of 20,0000 positive and negative instances from all Wikipedia articles. This provides an overview of how editors cite across articles of varying quality and topics.

[5]https://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[6]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment#Quality_scale
[7]"[the article provides] a thorough and representative survey of the relevant literature; claims are verifiable against high-quality reliable sources and are supported by inline citations where appropriate." https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria
[8]Unless otherwise specified, all data in the paper is from English Wikipedia

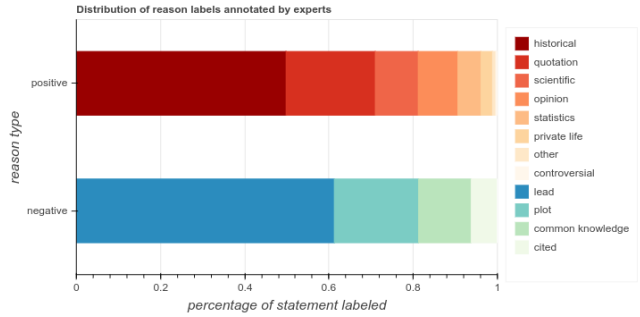

Figure 1: Distribution of labels assigned by Wikipedia Editors through the Wikilabels platform to characterize the reason why statements need citations.

4.2 Citation Reason Dataset

To train a model for the Citation Reason task,we designed a labeling task for Wikipedia editors in which they are asked to annotate Wikipedia sentences with both a binary judgment (citation needed/not needed) and the reason for that judgment using our Citation Reason Taxonomy. We used these annotations as ground truth for a larger-scale crowdsourcing experiment, where we

asked micro-workers to select reasons for why *positive* sentences require citations. We compared how well crowdworkers' judgments matched the Wikipedia editor judgments. Finally, we collected enough annotations to train machine learning algorithms.

4.2.1 Round 1: Collecting Data from Wikipedia Editors. To collect "expert" annotations from Wikipedia editors on why sentences need citations, we proceeded as follows.

Interface Design. We created a modified version of the free-text WikiLabels labeling task described in Section 3. We selected random sentences from Featured Articles, and removed citation markers when present. We presented the participants with the unsourced sentence highlighted in an article and asked them to label the sentence as needing an inline citation or not, and to specify a reason for their choice using a drop-down menu pre-filled with categories from our taxonomy. We recruited participants through mailing lists, social media and the English Wikipedia's Village pump (the general discussion forum of the English Wikipedia volunteer community).

Results. We collected a total of 502 labels from 35 English Wikipedia editors. Of the valid[9] annotated sentences, 255 were labeled as needing a citation (*positive*), and 80 as not needing a citation. Fig. 1 shows the breakdown of results by selected reason.

We found that the reason given for roughly 80% of the *positive* sentences is that they are "historical facts", "direct quotations", or "scientific facts". Furthermore, we observed that only a small percentage of participants selected the "Other" option, which suggests that our Citation Reason Taxonomy is robust and makes sense to editors, even when they are asked to provide these reasons outside of their familiar editing context.

[9]Due to a bug in the system, not all responses were correctly recorded.

Table 2: Example of sentences annotated with different categories by Wikipedia experts and Mechanical Turk contributors.

Non-Expert judgment	Expert judgment	Sentence extracted from Wikipedia Featured Article
historical	*quotation*	He argued that a small number of Frenchmen could successfully invade New Spain by allying themselves with some of the more than 15,000 Native Americans who were angry over Spanish enslavement
life	*historical*	Actor Hugh Jackman is also a fan of the club, having been taken to Carrow Road as a child by his English mother, though he turned down an opportunity to become an investor in the club in 2010
statistics	*historical*	The act, authored by Ohio senator and former Treasury secretary John Sherman, forced the Treasury to increase the amount of silver purchased to 4,500,000 troy ounces (140,000 kg) each month
quotation	*historical*	"This stroke", said Clark, "will nearly put an end to the Indian War." Clark prepared for a Detroit campaign in 1779 and again in 1780, but each time called off the expedition because of insufficient men and supplies

4.2.2 Round 2: Collecting Data from non-Experts. We adapted the task in Round 1 to collect data from crowdworkers to train a Citation Reason model.

Task adaptation. Adapting classification tasks that assume a degree of domain expertise to a crowdwork setting, where such expertise cannot be relied upon, can create challenges for both reliability and quality control. Crowdworkers and domain experts may disagree on classification tasks that require special knowledge [46]. However, Zhang et al.[51] found that non-expert judgments about the characteristics of statements in news articles, such as whether a claim was well supported by the evidence provided, showed high inter-annotator agreement and high correlation with expert judgments. In the context of our study, this suggests that crowdworkers may find it relatively easier to provide reasons for citations than to decide which sentences require them in the first place. Therefore, we simplified the annotation task for crowdworkersto increase the likelihood of eliciting high-quality judgments from non-experts. While Wikipedia editors were asked to both identify whether a sentence required citation and provide a reason, crowdworkers were only asked to provide a reason why a citation was needed.

Experimental Setup. We used Amazon Mechanical Turk for this annotation task. For each task, workers were shown one of 166 sentences that had been assigned citation reason categories by editors in round 1. Workers were informed that the sentence came from a Wikipedia article and that in the original article it contained a citation to an external source. Like editors in the first experiment, crowdworkers were instructed to select the most appropriate category from the eight citation reasons 1. Each sentence was classified by 3 workers, for a total of 498 judgments. For quality control purposes, only crowdworkers who had a history of reliable annotation behavior were allowed to perform the task. Average agreement between workers was 0.63% (vs random 1/8 =0.125).

4.2.3 Comparing Expert and Non-Expert annotations. The distribution of citation reasons provided by crowdworkers is shown in Fig. 2. The overall proportions are similar to that provided by Wikipedia editors in Round 1 (See Fig. 1). Furthermore, the confusion matrix in Fig. 3 indicates that crowdworkers and Wikipedia editors had high agreement on four of the five most prevalent reasons: *historical, quotation, scientific* and *statistics*. Among these five categories, non-experts and experts disagreed the most on *opinion*. One potential reason for this disagreement is that identifying whether a statement

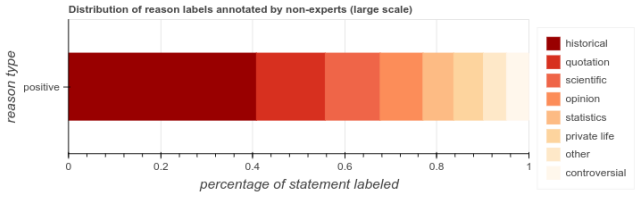

Figure 2: Citation reason distribution from the small-scale (166 sentences) crowdsourcing experiment.

is an opinion may require additional context (i.e. the contents of the preceding sentences, which crowdworkers were not shown).

The confusion matrix in Fig. 3) shows the percentage of different kinds of disagreement—for example, that crowdworkers frequently disagreed with editors over the categorization of statements that contain *"claims about general or historical facts."* To further investigate these results, we manually inspected a set of individual sentences with higher disagreement between the two groups. We found that in these cases the reason for the disagreement was due to a sentence containing multiple types of claims, e.g. a historical claim and a direct quote (see Table 2). This suggests that in many cases these disagreements were not due to lower quality judgments on the part of the crowdworkers, but instead due to ambiguities in the task instructions and labeling interface.

4.2.4 The Citation Reason Corpus: Collecting Large-scale Data. Having verified the agreement between Wikipedia editors and crowdworkers, we can now reliably collect larger scale data to train a Citation Reason model. To this end, we sampled 4,000 sentences that contain citations from Featured articles, and asked crowdworkers to annotate them with the same setup described above (see Sec 4.2.2). The distribution of the resulting judgments is similar to Fig. 2: as in Round 1, we found that the top categories are the *scientific*,*quotation* and *historical* reasons.[10]

[10]Our *Citation Reason corpus* is publicly available here: https://figshare.com/articles/Citation_Reason_Dataset/7756226.

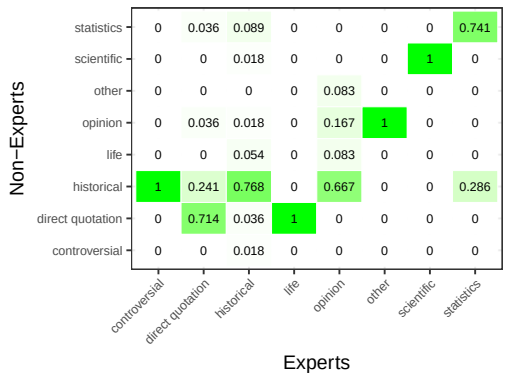

Figure 3: Confusion matrix indicating the agreement between Mechanical Turk workers ("non-experts") and Wikipedia editors ("experts"). The darker the square, the higher the percent agreement between the two groups

5 A CITATION NEED MODEL

We design a classifier to detect when a statement needs a citation. We proposea neural based Recurrent Neural Network (RNN) approach with varying representations of a statement, and compare it with a baseline feature-based model.

5.1 Neural Based Citation Need Approach

We propose a neural based model, which uses a recurrent neural network (RNN) with GRU cells[11] to encode statements for classification. We distinguish between two main modes of statement encoding: (i) vanilla RNN, fed with 2 different representations of a sentence (words and section information, indicated with RNN^{w} and RNN^{+S}), and (ii) RNN with global attention RNN_a (with similar representation).

5.1.1 Statement Representation. For a given Wikipedia sentence, for which we want to determine its citation need, we consider the words in the statement and the section the statement occurs in. To feed the network with this information, we transform sentence words and section information into features, or representations. Through the word representation we aim at capturing cue words or phrases that are indicators of a statement requiring a citation. Section representation, on the other hand, allows us to encode information that will play a crucial role in determining the Citation Reason later on.

Word Representation. We represent a statement as a sequence of words $s = (w_1, \ldots, w_n)$. We use GloVe pre-trained word embeddings [40] to represent the words in s. Unknown words are randomly initialized in the word embedding matrix $W_{glove} \in \mathbb{R}^{k \times 100}$, where k is the number of words in the embedding matrix.

Section Representation. The section in which the statement occurs in a Wikipedia article is highly important. The guidelines for *inline citations* suggest that when a statement is in the *lead section*, and that is referenced elsewhere in the article, editors should avoid multiple references [11]. Additionally, since sections can be seen as a topically coherent group of information, the reasons for citation will vary across sections (e.g. *"Early Life"*). We train the *section embedding* matrix $W_S \in \mathbb{R}^{l \times 100}$, and use it in combination with W_{glove}, where l is the number of sections in our dataset.

5.1.2 Statement Classification. We use 2 types of Recurrent Neural Networks to classify the sentence representations.

Vanilla RNNs. RNNs encode the individual words into a hidden state $h_t = f(w_t, h_{t-1})$, where f represents GRU cells [11]. The encoding of an input sequence from s is dependent on the previous hidden state. This dependency based on f determines how much information from the previous hidden state is passed onto h_t. For instance, in case of GRUs, h_t is encoded as following:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (1)$$

where, the function z_t and $\tilde{h}_t$ are computed as following:

$$z_t = \sigma(W_z w_t + U_z h_{t-1} + b_z) \quad (2)$$

$$\tilde{h}_t = \tanh\left(W_h w_t + r_t \odot (U_h h_{t-1} + b_h)\right) \quad (3)$$

$$r_t = \sigma(W_r w_t + U_r h_{t-1} + b_r) \quad (4)$$

The RNN encoding allows us to capture the presence of words or phrases that incur the need of a citation. Additionally, words that do not contribute in improving the classification accuracy are captured through the model parameters in function r_t, allowing the model to ignore information coming from them.

RNN with Global Attention – RNN_a. As we will see later in the evaluation results, the disadvantage of vanilla RNNs is that when used for classification tasks, the classification is done solely based on the last hidden state h_N. For long statements this can be problematic as the hidden states, respectively the weights are highly compressed across all states and thus cannot capture the importance of the individual words in a statement.

Attention mechanisms [4] on the other hand have proven to be successful in circumventing this problem. The main difference with standard training of RNN models is that all the hidden states are taken into account to derive a *context vector*, where different states contribute with varying weights, or known with *attention weights* in generating such a vector.

Fig. 4 shows the RNN_a^{+S} model we use to classify a statement. We encode the statement through a bidirectional RNN based on its word representation, while concurrently a separate RNN encodes the section representation. Since not all words are equally important in determining if a statement requires a citation, we compute the *attention weights*, which allow us to compute a *weighted representation* of the statement based on the hidden states (as computed by the GRU cells) and the *attention weights*. Finally, we *concatenate* the weighted representation of the statement based on its words and section, and push it through a dense layer for classification.

The vanilla RNN, and the varying representations can easily be understood by referring to Fig. 4, by simply omitting either the section representation or the attention layer.

5.1.3 Experimental Setup. We use Keras [12] with Tensorflow as backend for training our RNN models. We train for 10 epochs (since the loss value converges), and we set the batch size to 100. We use Adam [29] for optimization, and optimize for *accuracy*. We set the number of dimensions to 100 for hidden states h, which represent the words or the section information.

[11] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

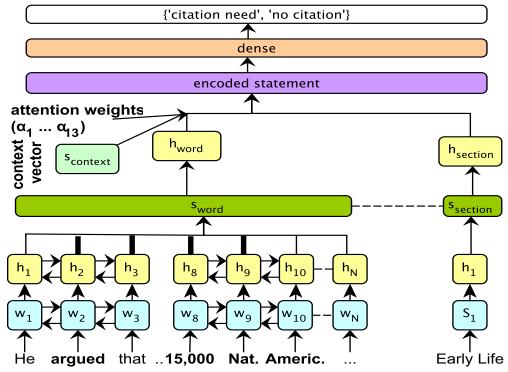

Figure 4: Citation Need model with RNN and global attention, using both word and section representations.

Table 3: Point-Biserial Correlation Coefficient between citation need labels and individual feature values

FA		LQN		RND	
Section	-0.621	*underline*	0.054	*say*	0.084
say	0.107	*say*	0.0546	*underline*	0.0842
underline	0.107	*believe*	0.042	*Section*	-0.072
realize	0.068	*disagree*	0.040	*report*	0.062
suggest	0.068	*claim*	0.039	*tell*	0.062

We train the models with 50% of the data and evaluate on the remaining portion of statements.

5.2 Feature-based Baselines

As we show in Table 1, where we extract the reasons why statements need a citation based on expert annotators, the most common reasons (e.g. *statistics*, *historical*) can be tracked in terms of specific *language frames* and *vocabulary use* (in the case of *scientific* claims). Thus, we propose two baselines, which capture this intuition of language frames and vocabulary. From the proposed feature set, we train standard supervised models and show their performance in determining if a statement requires a citation.

5.2.1 Dictionary-Based Baseline – Dict. In the first baseline, we consider two main groups of features. First, we rely on a set of lexical dictionaries that aim in capturing words or phrases that indicate an activity, which when present in a statement would imply the necessity of a citation in such cases. We represent each statement as a feature vector where each element correspond to the frequency of a dictionary term in the statement.

Factive Verbs. The presence of *factive verbs* [30] in a statement presumes the truthfulness of information therein.

Assertive Verbs. In this case, assertive verbs [25] operate in two dimensions. First, they indicate an assertion, and second, depending on the verb, the credibility or certainty of a proposition will vary (e.g. *"suggest"* vs. *"insist"*). Intuitively, *opinions* in Wikipedia fall in this definition, and thus, the presence of such verbs will be an indicator of opinions needing a citation.

Entailment Verbs. As the name suggests, different verbs entail each other, e.g. *"refrain"* vs. *"hesitate"* [5, 26]. They are particularly interesting as the context in which they are used may indicate cases of *controversy*, where depending on the choice of verbs, the framing of a statement will vary significantly as shown above. In such cases, Wikipedia guidelines strongly suggest the use of citations.

Stylistic Features. Finally, we use the frequency of the different POS tags in a statement. POS tags have been successfully used to capture linguistic styles in different genres [41]. For the different citation reasons (e.g. *historical*, *scientific*), we expect to see a variation in the distribution of the POS tags.

5.2.2 Word Vector-Based Baseline – WV. Word representations have shown great ability to capture word contextual information, and their use in text classification tasks has proven to be highly effective [22]. In this baseline, our intuition is that we represent each statement by averaging the individual word representations from a pre-trained word embeddings [40]. Through this baseline we aim at addressing the cases, where the *vocabulary use* is a high indicator of statements needing a citation, e.g. *scientific* statements.

5.2.3 Feature Classifier. We use a Random Forest Classifier [9] to learn Citation Need models based on these features. To tune the parameters (depth and number of trees), similar to the main deep learning models, we split the data into train, test and validation (respectively 50%,30% and 20% of the corpus). We perform cross-validation on the training and test set, and report accuracy results in terms of F1 on the validation set.

5.3 Citation Need Indicators

We analyze here how algorithms associate specific sentence features with the sentence's need for citations.

5.3.1 Most Correlated Features. To understand which sentence features are more related to the need for citation, we compute the Point Biserial Correlation coefficient [48] between the binary citation/no-citation labels and the frequency of each word in the baseline dictionary of each sentence, as well as the *Section* feature.

We report in Table 3 the top-5 most correlated features for each dataset. In featured articles, the most useful features to detect statements needing citation is the position of the sentence in the article, i.e. whether the sentence lies in the lead section of the article. This might be due to the fact that FA are the result of a rigorous formal process of iterative improvement and assessment according to established rubrics [50], and tend to follow the best practices to write the lead section, i.e. including general overview statements, and claims that are referenced and further verified in the article body. In the *LQN* dataset we consider as "positives" those sentences tagged as *Citation Needed*. Depending on the article, these tags can appear in the lead section too, thus explaining why the *Section* feature is not discriminative at all for this group of sentences. Overall, we see that *report verbs*, such as *say, underline, claim* are high indicators of the sentence's need for citations.

5.3.2 Results from Attention Mechanisms in Deep Learning. Fig. 5 shows a sample of positive statements from Featured Articles

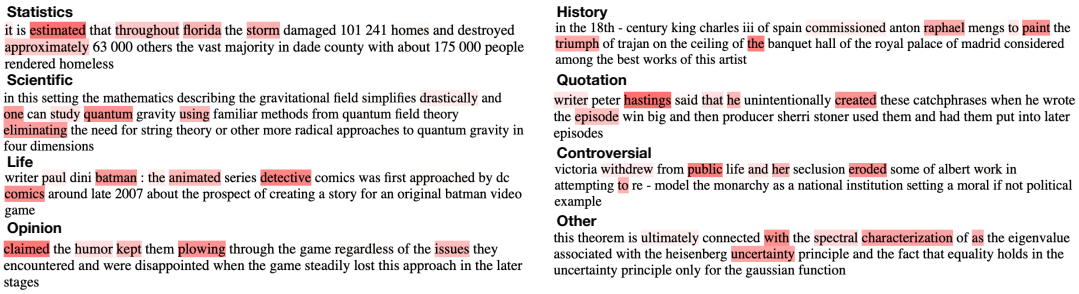

Figure 5: Attention mechanism for RNN_a^{+S} visualizing the focus on specific words for the different citation reasons. It is evident that the model is able to capture patterns similar to those of human annotators (e.g. *"claimed"* in the case of *opinion*.)

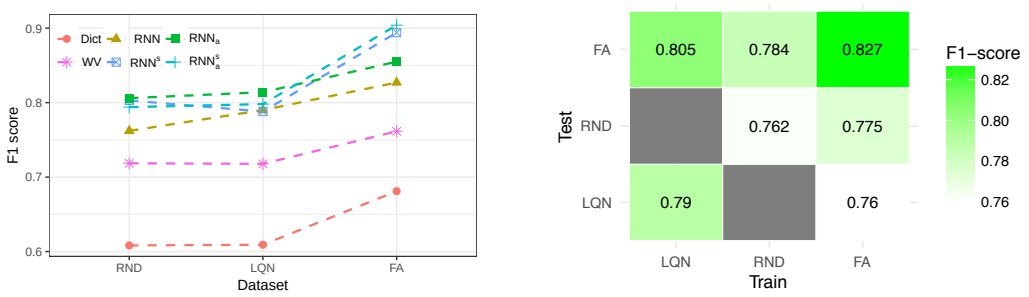

Figure 6: (a) F1 score for the different Citation Need detection models across the different dataset. (b) Confusion Matrix visualizing the accuracy (F1 score) of a Citation Need model trained on Featured Articles and tested on other datasets, showing the generalizability of a model trained on Featured Articles only.

Table 4: Accuracy (F1 score) of Citation Need classification models on Featured Article vs individual expert editor annotations on the same set of Featured Articles.

	no citation	citation	average
individual editor	0.608	**0.978**	0.766
RNN_a^{+S}	**0.902**	0.905	**0.904**

grouped by citation reason. The words are highlighted based on their *attention weight* from the RNN_a^{+S} model. The highlighted words show very promising directions. It is evident that the RNN_a^{+S} model attends with high weights words that are highly intuitive even for human annotators. For instance, if we consider the *opinion* citation reason, the highest weight is assigned to the word *"claimed"*. This is case is particularly interesting as it capture the *reporting verbs* [43] (e.g. *"claim"*) which are common in opinions. In the other citation reasons, we note the *statistics* reason, where similarly, here too, the most important words are again verbs that are often used in reporting numbers. For statements that are *controversial*, the highest attention is assigned to words that are often used in a negative context, e.g. *"erode"*. However, here it is interesting, that the word *"erode"* is followed by context words such as *"public"* and *"withdrew"*. From the other cases, we see that the attention mechanism focuses on domain-specific words, e.g *scientific* citation reason.

5.4 Evaluating the Citation Need model

In this section, we focus on assessing the performance of our model at performing the Citation Need task, its generalizability, and how its output compares with the accuracy of human judgments.

5.4.1 Can an Algorithm Detect Statements in Need of a Citation? We report the classification performance of models and baselines on different datasets in Fig. 6.

Given that they are highly curated, sentences from Featured Articles are much easier to classify than sentences from random articles: the most accurate version of each model is indeed the one trained on the Featured Article dataset.

The proposed RNN models outperform the featured-based baselines by a large margin. We observe that adding attention information to a traditional RNN with GRU cells boosts performances by 3-5%. As expected from the correlation results, the position of the sentence in an article, i.e. whether the sentence is in the lead section, helps classifying Citation Need in Featured Articles only.

5.4.2 Does the Algorithm Generalize? To test the generalizability of one the most accurate models, the RNN Citation Need detection model trained on Featured Articles, we use it to classify statements from the *LQN* and the *RND* datasets, and compute the F1 score over such cross-dataset prediction. The cross-dataset prediction reaches a reasonable accuracy, in line with the performances models trained and tested on the other two noisier datasets. Furthermore, we test the performances of our RNN_a model on 2 external datasets: the

Table 5: Citation reason prediction based on a pre-trained RNN_a^{+S} model on the *FA* dataset, and a RNN_a^{+S} which we train only on the Citation Reason dataset.

	pre-trained			*no pre-training*		
	P	R	F1	P	R	F1
direct quotation	**0.44**	**0.65**	**0.52**	0.43	0.46	0.45
statistics	0.20	**0.20**	**0.20**	**0.28**	0.15	0.19
controversial	**0.12**	**0.02**	**0.04**	0.04	0.01	0.02
opinion	**0.20**	0.12	0.15	0.19	0.12	0.15
life	0.13	0.06	0.09	**0.30**	0.06	**0.10**
scientific	**0.62**	0.56	**0.59**	0.54	**0.58**	**0.56**
historical	**0.56**	0.67	0.61	0.54	**0.74**	**0.62**
other	0.13	0.05	0.07	**0.14**	**0.08**	**0.10**
avg.	0.30	**0.29**	**0.28**	**0.31**	0.28	0.27

claim dataset from Konstantinovskiy et al. [32], and the CLEF2018 Check-Worthiness task dataset [39]. Both datasets are made of sentences extracted from political debates in UK and US TV-shows, labeled as positives if they contain facts that need to be verified by fact-checkers, or as negative otherwise. Wikipedia's literary form is completely different from the political debate genre. Therefore, our model trained on Wikipedia sentences, cannot reliably detect claims in the fact-checking datasets above: most of the sentences from these datasets are outside our training data, and therefore the model tends to label all those as negatives.

5.4.3 Can the Algorithm Match Individual Human Accuracy? Our Citation Need model performs better than individual Wikipedia editors under some conditions. Specifically, in our first round of expert citation labeling (Section 3 above), we observed that when presented with sentences from Featured Articles in the WikiLabels interface, editors were able to identify claims that already had a citation in Wikipedia with a high degree of accuracy (see Table 4), but they tended to *over-label*, leading to a high false positive rate and lower accuracy overall compared to our model. There are several potential reasons for this. First, the editorial decision about whether to source a particular claim is, especially in the case of Featured Articles, an iterative, deliberate, and consensus-based process involving multiple editors. No single editor vets all the claims in the article, or decides which external sources to cite for those claims. Furthermore, the decisions to add citations are often discussed at length during the FA promotion process, and the editors involved in writing and maintaining featured articles often have subject matter expertise or abiding interest in the article topic, and knowledge of topic-specific citation norms and guidelines [18]. By training on the entire corpus of Featured Articles, our model has the benefit of the aggregate of hundreds or thousands of editors' judgments of when (not) to cite across a range of topics, and therefore may be better than any individual editor at rapidly identifying general lexical cues associated with "common knowledge" and other statement characteristics that indicate citations are not necessary.

6 A CITATION REASON MODEL

In this Section, we analyze the *Citation Reason Corpus* collected in Sec. 4, and fine-tune the Citation Need model to detect reasons why statements need citations.

6.1 Distribution of Citation Reasons by Topic

Understanding if Wikipedia topics or article sections have different sourcing requirements may help contributors better focus their efforts. To start answering this question, we analyze citation reasons as a function of the article topic and the section in which the sentence occurs. We rely on DBpedia [3] to associate articles to topics and we show in Table 6 the most topics and article sections associated with each citation reason. We note that the distribution of citation reasons is quite intuitive, both across types and sections. For instance, *"direct quotation"* is most prominent in section *Reception* (the leading section), which is intuitive, where the statements mostly reflect how certain *"Athlete"*, *"OfficeHolders"* have expressed themselves about a certain event. Similarly, we see for *"historical"* and *"controversial"* the most prominent section is *History*, whereas in terms of most prominent article types, we see that *"MilitaryConflict"* types have the highest proportion of statements.

While the distribution of citation reasons is quite intuitive across types and sections, we find this as an important aspect that can be leveraged to perform targeted sampling of statements (from specific sections or types) which may fall into the respective citation reasons s.t we can have even distribution statements across these categories.

6.2 Evaluating the Citation Reason model

To perform the Citation Reason task, we build upon the pre-trained model RNN_a^{+S} in Fig. 4. We modify the RNN_a^{+S} model by replacing the dense layer such that we can accommodate all the eight citation reason classes, and use a *softmax* function for classification.

The rationale behind the use of the pre-trained RNN_a^{+S} model is that by using the much larger training statements from the binary datasets, we are able to adjust the model's weights to provide a better generalization for the more fine-grained citation reason classification. An additional advantage of using the model with the pre-trained weights is that in this way we can retain a large portion of the contextual information from the statement representation, that is, the context in which the words appear for statement requiring a citation.

The last precaution we take in adjusting the RNN_a^{+S} for Citation Reason classification is that we ensure that the model learns a balanced representation for the different citation reason classes.

Table 5 shows the accuracy of the pre-trained RNN_a^{+S} model trained on 50% of the Citation Reason dataset, and evaluate on the remaining statements. The pre-trained model has a better performance for nearly all citation reasons. It is important to note that due to the small number of statements in the Citation Reason dataset and additionally the number of classes, the prediction outcomes are not optimal. Our goal here is to show that the citation reason can be detected and we leave for future work a large scale evaluation.

7 DISCUSSION AND CONCLUSIONS

In this paper, we presented an end-to-end system to characterize, categorize, and algorithmically assess the verifiability of Wikipedia contents. In this Section we discuss the theoretical and practical implications of this work, as well as limitations and future directions.

7.1 Theoretical Implications

A Standardization of Citation Reasons. We used mixed methods to create and validate a Citation Reason Taxonomy. We then used this taxonomy to label around 4,000 sentences with reasons why they need to be referenced, and found that, in English Wikipedia, they are most often *historical facts, statistics or data about a subject, or direct or reported quotations*. Based on these annotations, we produced a Citation Reason corpus that we are

Table 6: Most common article topics and article sections for the different citation reasons.

Article Section						
quotation	**statistics**	**controversial**	**opinion**	**life**	**scientific**	**historical**
reception	history	history	reception	biography	description	history
history	reception	background	history	history	history	background
legacy	legacy	reception	development	early life	taxonomy	abstract
production	abstract	legacy	production	career	habitat	aftermath
biography	description	aftermath	background	background	characteristics	life and career
Article Topics						
quotation	***statistics***	***controversial***	***opinion***	***life***	***scientific***	***historical***
videogame	athlete	military conflict	videogame	athlete	animal	conflict
athlete	settlement	videogame	athlete	office holder	fungus	military person
book	videogame	settlement	album	royalty	plant	royalty
officeholder	infrastructure	athlete	single	military	military unit	office holder
album	country	royalty	book	artist	band	settlement

making available to other researchers as open data[12]. While this taxonomy and corpus were produced in the context of a collaborative encyclopedia, given that they are not topic- or domain-specific, we believe they represent a resource and a methodological foundation for further research on online credibility assessments, in particular seminal efforts aiming to design controlled vocabularies for credibility indicators[51].

Expert and Non-expert Agreement on Citation Reasons. To create the verifiability corpus, we extended to crowdworkers a labeling task originally designed to elicit judgments from Wikipedia editors. We found that *(non-expert) crowdworkers and (expert) editors agree about why sentences need citations in the majority of cases*. This result aligns with previous research [31], demonstrating that while some kinds of curation work may require substantial expertise and access to contextual information (such as norms and policies), certain curation subtasks can be entrusted to non-experts, as long as appropriate guidance is provided. This has implications for the design of crowd-based annotation workflows for use in complex tasks where the number of available experts or fact-checkers doesn't scale, either because of the size of the corpus to be annotated or its growth rate.

Algorithmic Solutions to the Citation Need Task. We used Recurrent Neural Networks to classify sentences in English Wikipedia as to whether they need a citation or not. We found that algorithms can effectively perform this task in English Wikipedia's Featured Articles, and generalize with good accuracy to articles that are not featured. We also found that, contrary to most NLP classification tasks, our Citation Need model outperforms expert editors when they make judgments out of context. We speculate that this is because when editors are asked to make judgments as to what statements need citations in an unfamiliar article without the benefit of contextual information, and when using a specialized microtask interface that encourages quick decision-making, they may produces more conservative judgments and default to Wikipedia's general approach to verifiability—dictating that all information that's likely to be challenged should be verifiable, ideally by means of an inline citation. Our model, on the other hand, is trained on the complete Featured Article corpus, and therefore learns from the wisdom of the whole editor community how to identify sentences that need to be cited.

Algorithmic Solutions to the Citation Reason Task We made substantial efforts towards designing an interpretable Citation Need model. In Figure 5 we show that our model can capture words and phrases that describe citation reasons. To provide full explanations, we designed a model that can classify statements needing citations with a reason. To determine the citation reason, we modified the binary classification model RNN_a^{+S} to predict the eight reasons in our taxonomy. We found that using the pre-trained model in the binary setting, we could re-adjust the model's weights to provide reasonable accuracy in predicting citation reasons. For citation reason classes with sufficient training data, we reached precision up to $P = 0.62$. We also provided insights on how to further sample Wikipedia articles to obtain more useful data for this task.

7.2 Limitations and Future Work

Labeling sentences with reasons why they need a citation is a non-trivial task. Community guidelines for inline citations evolve over time, and are subject to continuous discussion: see for example the discussion about why in Wikipedia "you need to cite that the sky is blue" and at the same time "you don't need to cite that the sky is blue" [13]. For simplicity, our Citation Reason classifier treats citation reason classes as mutually exclusive. However, in our crowdsourcing experiment, we found that, for some sentences, citation reasons are indeed not mutually exclusive. In the future, we plan to add substantially more data to the verifiability corpus, and build multi-label classifiers as well as annotation interfaces that can account for fuzzy boundaries around citation reason classes.

In Sec. 5 we found that, while very effective on Wikipedia-specific data, our Citation Need model is not able to generalize to fact-checking corpora. Given the difference in genre between the political discourse in these corpora, and the Wikipedia corpus, this limitation is to be expected. We explored, however, two other generalizability dimensions: domain expertise and language. We demonstrated that, for this task, annotation can be effectively performed by non-experts, facilitating the solution of this task at scale and distributing it beyond expert communities. Moreover, we built a general multilingual taxonomy by evaluating policies from different Wikipedia language communities, and by testing its effectiveness with expert contributors from English, Italian, and French Wikipedia.

More broadly, this work is designed for multilingual generalizability. In the future, we aim to replicate the large annotation efforts across languages. This should be fairly straight-forward, since Featured Articles exist in 163 Wikipedia language editions[14]. Moreover, the RNN model can be fed with word vectors such as fasttext[8], which now exist in more than 80 languages [7] and that one can re-train with any language from a Wikipedia project.

Finally, in this study we consider the application of verifiability policies to a static snapshot of Wikipedia articles, not taking into account their revision

[12]URL hidden for double blind submission

[13]https://en.wikipedia.org/wiki/Wikipedia:You_do_need_to_cite_that_the_sky_is_blue

[14]https://www.wikidata.org/wiki/Q16465

history. We also used general text features, and limited the encyclopedic-specific features to the *main section* feature. We expect that the distribution of citation reasons may vary over the course of an article's development, as a function of how controversial or disputed particular sections are. Performing this analysis is beyond the scope of the current study, but it might surface important exceptions to our findings or uncover interesting editing dynamics.

7.3 Practical Implications

Our study also has practical implications for the design of collaborative systems to support information verifiability and fact-checking. Our results of a robust agreement between non-experts and experts around citation reasons suggests that this type of task can be effectively crowdsourced, potentially allowing systems to recruit non-experts to triage or filter unverified statements based on model predictions, and allowing subject-matter experts to focus their attention on identifying reliable sources for these statements, or improving articles or topics with the highest rate of unreferenced factual claims. On Wikipedia, existing microtasking tools such as CitationHunt[15] could be modified to surface unverified claims that have not yet been flagged *citation needed* by humans and provide users with model predictions to guide their research, while also allowing users to provide feedback on those predictions to refine and improve the model. The model could also be used to surface article or statement-level information quality indicators to Wikipedia readers, using scores or lightweight visualizations, as suggested by Forte and others [2, 18], to support digital literacy and to allow readers to make more informed credibility assessments. Finally, downstream re-users of Wikipedia content, such as search engines and social media platforms, could also draw on model outputs to assign trust values to the information they extract or link to.

Beyond Wikipedia, our work complements existing attempts to assess how experts and non-experts assess the credibility of digital information, [51] and suggests that it is possible to develop robust *verifiability taxonomies* and automated systems for identifying unverified claims in complex information spaces even without substantial domain knowledge. Such capabilities could support large-scale, distributed fact checking of online content, making the internet more robust against the spread of misinformation and increasing the overall information literacy of internet users.

ACKNOWLEDGMENTS

We would like to thank the community members of the English, French and Italian Wikipedia for helping with data labeling and for their precious suggestions, and Bahodir Mansurov and Aaron Halfaker from the Wikimedia Foundation, for their help building the WikiLabels task. This work is partly funded by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), and BMBF Simple-ML project (grant no. 01IS18054A).

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 596–606.

[2] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning Trust to Wikipedia Content. In *Proceedings of the 4th International Symposium on Wikis (WikiSym '08)*. ACM, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/1822258.1822293

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] Jonathan Berant, Ido Dagan, Meni Adler, and Jacob Goldberger. 2012. Efficient tree-based approximation for entailment graph learning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 117–125.

[6] Ivan Beschastnikh, Travis Kriplean, and David W McDonald. 2008. Wikipedian Self-Governance in Action: Motivating the Policy Lens.. In *ICWSM*.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] Chih-Chun Chen and Camille Roth. 2012. Citation Needed: The Dynamics of Referencing in Wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym '12)*. ACM, New York, NY, USA, Article 8, 4 pages. https://doi.org/10.1145/2462932.2462943

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[12] François Chollet et al. 2015. Keras.

[13] Chung Joo Chung, Hyunjung Kim, and Jang-Hyun Kim. 2010. An anatomy of the credibility of online newspapers. *Online Information Review* 34, 5 (2010), 669–685. https://doi.org/10.1108/14684521011084564

[14] Besnik Fetahu, Katja Markert, and Avishek Anand. 2015. Automated News Suggestions for Populating Wikipedia Entity Pages. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*. 323–332. https://doi.org/10.1145/2806416.2806531

[15] Besnik Fetahu, Katja Markert, and Avishek Anand. 2017. Fine Grained Citation Span for References in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 1990–1999. https://aclanthology.info/papers/D17-1212/d17-1212

[16] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding News Citations for Wikipedia. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. 337–346. https://doi.org/10.1145/2983323.2983808

[17] Andrea Forte, Nazanin Andalibi, Tim Gorichanaz, Meen Chul Kim, Thomas Park, and Aaron Halfaker. 2018. Information Fortification: An Online Citation Behavior. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork (GROUP '18)*. ACM, New York, NY, USA, 83–92. https://doi.org/10.1145/3148330.3148347

[18] Andrea Forte, Nazanin Andalibi, Thomas Park, and Heather Willever-Farr. 2014. Designing Information Savvy Societies: An Introduction to Assessability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2471–2480. https://doi.org/10.1145/2556288.2557072

[19] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia Governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72. https://doi.org/10.2753/MIS0742-1222260103 arXiv:https://doi.org/10.2753/MIS0742-1222260103

[20] R. Stuart Geiger and Aaron Halfaker. 2013. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*. ACM, New York, NY, USA, Article 6, 6 pages. https://doi.org/10.1145/2491055.2491061

[21] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438, 7070 (Dec. 2005), 900–901. http://dx.doi.org/10.1038/438900a

[22] Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. 427–431. https://aclanthology.info/papers/E17-2068/e17-2068

[23] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1803–1812.

[24] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 755–764.

[25] Joan B Hooper. 1974. *On assertive predicates*. Indiana University Linguistics Club.

[26] Lauri Karttunen. 1971. Implicative verbs. *Language* (1971), 340–358.

[27] Brian Keegan, Darren Gergle, and Noshir Contractor. 2013. Hot Off the Wiki: Structures and Dynamics of WikipediaâĂŹs Coverage of Breaking News Events. *American Behavioral Scientist* 57, 5 (2013), 595–622. https://doi.org/10.1177/0002764212469367 arXiv:https://doi.org/10.1177/0002764212469367

[28] David J Ketchen and Christopher L Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal* 17, 6 (1996), 441–458.

[29] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] https://tools.wmflabs.org/citationhunt/

[30] Paul Kiparsky and Carol Kiparsky. 1968. *Fact.* Linguistics Club, Indiana University.

[31] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[32] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *arXiv preprint arXiv:1809.08193* (2018).

[33] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 591–602. https://doi.org/10.1145/2872427.2883085

[34] Stacey M Lavsa, Shelby L Corman, Colleen M Culley, and Tara L Pummer. 2011. Reliability of Wikipedia as a medication information source for pharmacy students. *Currents in Pharmacy Teaching and Learning* 3, 2 (2011), 154–158.

[35] WĂĆodzimierz Lewoniewski, Krzysztof WĂŹcel, and Witold Abramowicz. 2017. Analysis of References Across Wikipedia Languages. *Communications in Computer and Information Science Information and Software Technologies* (2017), 561âĂŞ573. https://doi.org/10.1007/978-3-319-67642-5_47

[36] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake News Detection Through Multi-Perspective Speaker Profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*. 252–256. https://aclanthology.info/papers/I17-2043/i17-2043

[37] Louise Matsakis. 2018. Facebook and Google must do more to support Wikipedia. https://www.wired.co.uk/article/wikipedia-google-youtube-facebook-support

[38] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *11th International Conference on Web and Social Media, ICWSM 2017*. AAAI Press.

[39] Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Wajdi Zaghouani, Tamer Elsayed, Reem Suwaileh, and Pepa Gencheva. 2018. CLEF-2018 Lab on Automatic Identification and Verification of Claims in Political Debates. In *Proceedings of the CLEF-2018*.

[40] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[41] Philipp Petrenz and Bonnie L. Webber. 2011. Stable Classification of Text Genres. *Computational Linguistics* 37, 2 (2011), 385–393. https://doi.org/10.1162/COLI_a_00052

[42] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, Destroying, and Restoring Value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/1316624.1316663

[43] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1650–1659.

[44] Christina Sauper and Regina Barzilay. 2009. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*. 208–216. http://www.aclweb.org/anthology/P09-1024

[45] Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation* 43, 3 (2009), 227–268. https://doi.org/10.1007/s10579-009-9089-9

[46] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. 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 (CSCW '15)*. ACM, New York, NY, USA, 826–838. https://doi.org/10.1145/2675133.2675285

[47] Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. 352–357. https://doi.org/10.18653/v1/P17-2056

[48] Robert F Tate. 1954. Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of mathematical statistics* 25, 3 (1954), 603–607.

[49] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. 809–819. https://aclanthology.info/papers/N18-1074/n18-1074

[50] Fernanda B. Viégas, Martin Wattenberg, and Matthew M. McKeon. 2007. The Hidden Order of Wikipedia. In *Online Communities and Social Computing*, Douglas Schuler (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 445–454.

[51] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 603–612. https://doi.org/10.1145/3184558.3188731