

Providing Citations to Support Fact-Checking: Contextualizing Detection of Sentences Needing Citation on Small Wikipedias

Aida Halitaj*, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, 327 Mile End Rd, London, E1 4NS, United Kingdom



ARTICLE INFO

Keywords:

Citation-worthiness detection
Low-resource languages
Fact-checking
Information integrity
Wikipedia

ABSTRACT

Authoritative citations are critical to ensure information integrity, especially in encyclopedias like Wikipedia. To date, research on automating citation worthiness detection has largely focused on the most resourceful language, English Wikipedia, neglecting the applicability to smaller Wikipedias. In addition, previous research proposed models that analyze the content inherent to a sentence to determine its citation worthiness, overlooking the potential of additional context to improve the prediction. Addressing these gaps, our study proposes a transformer-based contextualized approach for smaller Wikipedias, presenting a novel method to compile high-quality datasets for the Albanian, Basque, and Catalan editions. We develop the Contextualized Citation Worthiness (CCW) model, employing sentence representations enriched with adjacent sentences and topic categories for enhanced contextual insight. Empirical experiments on three newly created datasets demonstrate significant performance improvements of our contextualized CCW model, with 6%, 3% and 6% absolute improvements over the baseline for Albanian, Basque and Catalan datasets, respectively. We conduct an in-depth analysis to understand the influence and extent to which preceding and succeeding context as well as topic categories contribute to the accuracy of citation-worthiness predictions. Our findings suggest that incorporating such contextual information aids in the automatic identification of sentences in need of citations, not least when both the preceding and succeeding context are incorporated. This has implications for supporting Wikipedia projects across low-resource languages, promoting better article validation and fact-checking.

1. Introduction

Research suggests that humans have a tendency to accept statements as true unless they are explicitly prompted to question them, hence debunking of information requiring more effort and motivation (Li et al., 2022; Newman et al., 2022; Lewandowsky et al., 2012; Gilbert, 1991), a phenomenon that exacerbates the problem of misinformation in society (Lutzke et al., 2019). This has significant negative implications in modern society, where individuals are increasingly exposed to inaccurate or incomplete information, with online sources playing a prominent role in their spread (Xu et al., 2023; Olan et al., 2022; Wang et al., 2019; Shin et al., 2018; Del Vicario et al., 2016; Dordevic et al., 2016). To help readers in assessing the validity of information, as well as further research in automated fact-checking (Zeng et al., 2021), a way forward is to enable the provision of citations linking information with relevant pieces of evidence.

One established way of backing up a claim is by providing a citation in the form of external evidence, i.e. linking it to reputable sources

that help verify the validity of the claim. The use of citations serves as a mechanism to validate claims and consequently to help prevent the spread of misinformation (Przybyla et al., 2022). Despite the importance of providing citations with relevant sentences in articles, as is the case for example with the requirements set out by Wikipedia's verifiability policy,¹ there are cases where citations are missing and it is important to support with the detection of these incomplete sentences. This calls for the need of studying methods for automated citation worthiness detection (Redi et al., 2019).

The task of citation worthiness detection, consisting of automatically identifying sentences needing a citation but currently lacking one, has attracted some attention in the scientific community in recent years to alleviate the otherwise burdensome manual task. Research in this area has explored several features to aid in detection, such as word embeddings, topic models, sentence length, keywords, or the position of the sentence within the article (Zeng and Acuna, 2020; Redi et al., 2019; Bonab et al., 2018). Much of this research has primarily studied citation worthiness detection in scientific articles (Gosangi et al., 2021;

* Corresponding author.

E-mail addresses: a.halitaj@qmul.ac.uk (A. Halitaj), a.zubiaga@qmul.ac.uk (A. Zubiaga).

URL: <https://www.zubiaga.org/> (A. Zubiaga).

¹ <https://en.wikipedia.org/wiki/Wikipedia:Verifiability>.

Bonab et al., 2018; Sugiyama et al., 2010), with much less research on Wikipedia articles which are more diverse in terms of domains and languages, and more heterogeneous in writing style given its collaborative nature. Citation worthiness detection in Wikipedia therefore poses an extra challenge that has been understudied to date.

Our work tackles two key limitations of previous research in citation worthiness detection for Wikipedia articles. First, where prior research has focused on the more active English Wikipedia supported by a larger community of editors, we study citation worthiness detection for low-resource languages with a smaller community of editors behind it, i.e. Albanian, Basque, and Catalan. This introduces additional challenges, both in collecting high-quality datasets, which becomes more difficult, and in conducting experiments, due to the limited resources available for Natural Language Processing (NLP) in these languages. And second, while previous work has modeled each input sentence independently, in this work we propose to contextualize the sentences for improved citation worthiness detection. We study how, and the extent to which, two kinds of context can help with the task, namely (i) the adjacent sentences including previous and next sentences, and (ii) the topic categories associated with the sentence.

In addition, our work aims to improve the generalizability of automated citation prediction approaches by preserving the original class distribution of a dataset. Unlike previous studies that used dataset balancing techniques, in our approach, we maintained the original distribution of classes, i.e. citation worthiness and no citation worthiness. This allows the model to learn from the full range of examples found in the real world and it enhances the practicality and applicability of our research to the reality of Wikipedia.

2. Research objectives

This study aims to achieve two primary objectives:

1. **O1:** Identify effective methods to prepare high-quality datasets from smaller Wikipedias for the task of detecting sentences needing a citation. Specifically, utilize the quality scores of articles to build a more reliable and scientifically valid dataset.
2. **O2:** Explore how the contextualization of sentences affects the ability to predict their need for citations in low-resource languages.

To achieve these objectives, we devise a novel methodology to gather data on citation worthiness for smaller versions of Wikipedia. We applied this methodology to the Albanian, Basque, and Catalan Wikipedias, creating three new datasets: SQ-citation-needed, EU-citation-needed, and CA-citation-needed. Using these datasets, we conducted experiments with the Contextualized Citation Worthiness (CCW) model, leveraging Transformer models to test two key hypotheses about the importance of contextualizing sentences:

1. **H1:** Given a sentence, its adjacent sentences, including the previous and next sentence, will benefit the citation detection model.
2. **H2:** The topic categories associated with the given sentence will help the model determine if it needs a citation.

Through the evaluation of our hypothesis, we make the following novel contributions:

- We developed a new data collection and labeling methodology suitable for smaller Wikipedias. Using this approach, we created the CA-citation-needed, EU-citation-needed, and SQ-citation-needed datasets for the Catalan, Basque, and Albanian languages, respectively. This expands the scope of the study to include low-resource languages, addressing a gap left by previous research that primarily focused on English.

- By experimenting with CCW, we are the first to study the usefulness of contextualized modeling for citation worthiness detection in Wikipedia articles. Our contextualization experiments include incorporating adjacent sentences as well as topic categories associated with a given sentence.
- We introduced the CCW (Contextualized Citation Worthiness) model, which uses mBERT (Devlin et al., 2019) contextualized embeddings to better preserve the meaning of sentences compared to previous studies (Redi et al., 2019; Bonab et al., 2018) that relied on embeddings like GloVe or fastText. Our approach captures contextual meaning more effectively, providing variable-length representations and handling out-of-vocabulary words more robustly.
- We contribute to the field of NLP by focusing on low-resource, understudied languages such as Albanian, Basque and Catalan, which have seen limited research to date.

3. Related work

The need to automate the identification of sentences that require support from external citations is commonplace and has been studied in different contexts. These include primarily (i) detecting citeworthy sentences in academic writing, (ii) detecting citeworthy sentences in Wikipedia articles, and (iii) in the context of fact-checking, detecting sentences that need to be verified, also known as claim detection. While (ii) is more relevant to our research, (i) represents the earlier development in this field. In what follows we discuss work in these three directions, following a discussion on Wikipedia's policy regarding citations.

3.1. Wikipedia's verifiability policy

In line with the standards expected with a high-quality encyclopedia, Wikipedia also requires that its articles provide sufficient links to evidence to help validate its integrity. This is documented in Wikipedia's verifiability policy,² which requires that Wikipedia content be accompanied with a relevant citation where appropriate. Conversely, cases of claims lacking a citation should either be removed or flagged as such by using the `{{citation needed}}` tag³ for future attention. In addition to the policy, it has also been found that the use of citations positively correlates with higher quality of articles (Hu et al., 2023; Chou et al., 2020).

Wikipedia's verifiability policy is a rigorous editorial procedure which has strengthened encyclopedia's reputation as a trustworthy source of information, not only for fact-checkers and journalists but also for major platforms like Google, YouTube, and Facebook, which rely on it in their efforts to combat misinformation (Saez-Trumper, 2019; McGrew et al., 2017; McMahon et al., 2017). Where Wikipedia is edited and maintained by a community of volunteer editors, providing them with the appropriate mechanisms to make good use of citations along with their content is crucial, for example by suggesting to them when a sentence requires a citation.

Despite its proven usefulness, the task of identifying citeworthy Wikipedia sentences can be challenging for inexperienced editors (Logan et al., 2010) and time-consuming for more experienced editors (Kaffee and Elshahar, 2021). While the task of identifying citeworthy sentences can be to some extent manageable for Wikipedias of languages with large communities behind them (e.g. English, German or French), it becomes even more challenging for smaller Wikipedias with smaller communities of active editors (Hara et al., 2010), as is the case in our study with the Albanian, Basque or Catalan language. This calls for the development of automated mechanisms to support Wikipedia editors by detecting sentences needing a citation.

² <https://en.wikipedia.org/wiki/Wikipedia:Verifiability>.

³ https://en.wikipedia.org/wiki/Wikipedia:Citation_needed.

3.2. Citation worthiness detection in scientific articles

Research conducted with the aim of automating the identification of sentences needing citation in scientific articles emerged in the early 2010s as the increase in the volume of published research demanded it. In most cases, the issue at hand is a classification task, where the goal is to determine if a sentence from a research article requires a citation. Researchers developed models that could take in a single sentence as input, as well as models that analyzed other text features such as the presence of proper nouns, unigrams, bigrams, citation keywords, sentence length, placement within the document (e.g. introduction, literature review, methodology), and labels of adjacent sentences (Sugiyama et al., 2010; Bonab et al., 2018).

Different models were used, such as linear classifiers and neural networks. The linear classifiers consisted mostly of Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM), while the neural network models consisted of Bidirectional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Networks (CNN) based classifiers.

According to a study by Zeng and Acuna (2020), exploiting the contextual mechanism proved to be useful in detecting meaningful citation in scientific articles. In addition to surrounding sentences, other contextual features like the title, paragraphs, and nearby citations of a research article were found to be useful in predicting cite-worthiness (Gosangi et al., 2021). This improvement was attributed to both the contextual features and the usage of advanced machine learning techniques, particularly the attention mechanism with a BiLSTM model. The importance of contextualization has been strengthened in CiteWorth study which reached the best performance with a model based on Longformer that took as an input paragraph-level contextualized sentences (Wright and Augenstein, 2021). Another study (Roostaei, 2022) reinforced the significance of identifying sentence-level citation worthiness as a crucial step in citation recommendation systems. They analyzed the ACL-ARC dataset and proposed a citation-worthiness identification model, which uses syntactic embedding and ConvNets classifier architecture to identify citation contexts. They performed a down-sampling analysis to address the imbalanced nature of the dataset.

More recent studies show that research has expanded to new domains, like legal texts. Specifically, CiteCaseLAW is a recent approach (Khatri et al., 2023) aiming to identify cite-worthy sentences in the text of the legal domain. Several models were developed and evaluated, including Logistic Regression, CRNN, Transformers, Longformer, and BERT. These models were trained and tested on a novel dataset of 178M sentences extracted from the Caselaw Access Project (CAP) and were compared to established baselines for other legal text classification tasks. In the context of academic writing, surrounding information proved to be helpful in predicting citation. This has however not been studied in the case of Wikipedia articles, where contextualization can be significantly different, not least because in a collaboratively edited Wikipedia, a sentence may be written by an editor whereas surrounding sentences may be written by other editors. Hence, our research investigates how contextualization can support citation worthiness detection in Wikipedia articles, not only by using surrounding sentences but also topic categories.

3.3. Citation worthiness detection in Wikipedia

Early work on citing sources in Wikipedia initially focused on analyzing and clarifying the editing process which encompasses the use of citations (Viégas et al., 2004; Korfiatis et al., 2006; Blumenstock, 2008; Laniado and Tasso, 2011; Chen and Roth, 2012). This work did not however study the ability to detect the need for citations, but rather studied citation patterns. This type of research in Wikipedia has been the subject of extensive study by the academic community (Schmidt et al., 2023).

Research on identifying sentences in Wikipedia articles that require citations has emerged more recently compared to similar studies in academic writing. Citation detection differs between academic and Wikipedia settings. Academic papers have structured citation practices, making citation identification and extraction easier. Wikipedia adopts a more flexible approach that prioritizes reliability by encouraging the use of credible and reputable references to ensure the accuracy and trustworthiness of its content. Validating citation data in academic settings relies on expert-placed citations, while Wikipedia data usually requires additional means of data validation such as manual efforts of Wikipedia editors or other means of crowd-sourcing.

Previous research (Fetahu et al., 2017) has addressed the issue of identifying the citation span in Wikipedia articles, aiming to predict which textual fragments in an article are covered by citation. However, this approach did not address the broader question of whether a given sentence within the article should have a citation. This question was later investigated in Redi et al. (2019) and it aimed to explore how and why Wikipedia uses citations to ensure the trustworthiness of information in its articles. In addition to predicting if a sentence needs a citation, they also provided a taxonomy of reasons why the citation is needed.

A study used Redi's framework to address the issue of citation needed in Wikipedia (Wright and Augenstein, 2020). It aimed to determine the credibility of statements in political speeches, spot rumors on Twitter, and recognize when citations are necessary. Positive Unlabeled learning was utilized, and the study revealed successful transfer of citation needed detection to rumor detection, but not to political speeches and debates.

Redi's research is particularly relevant to our work. Therefore, it serves as one of the foundational pillars, as we seek to build upon and expand our understanding of citation practices specifically in languages with limited resources. Wikipedia helps readers to assess the quality and reliability of the content by assigning some quality classes⁴ to the articles. The highest quality articles in Wikipedia are categorized as Featured Articles.⁵ To be marked as a featured article, it must pass a rigorous peer-review process⁶ and meet the criteria established by the Wikipedia community.⁷ In the study led by Redi et al. (2019), they relied heavily on the content of featured articles in the English language. This reliance prevents the approach from being generalized for languages with low resources and smaller communities of Wikipedia editors. In contexts where such communities are underdeveloped, the capacity for peer-reviewing articles with the aim of verifying their accuracy may be limited resulting in a smaller number of featured articles. While Redi's approach is not readily applicable to languages with limited resources, it demonstrated that the quality of articles significantly impacts the prediction of citation needed. Thus, in our study, we show an alternative, more flexible approach⁸ that generates quality scores for all types of articles regardless of the language. In addition, our approach is completely automated and does not rely on the manual work of Wikipedia editors.

Redi's proposed citation need model just recently has been expanded into an end-to-end inference pipeline that examines trends in a decade's worth of data to assess the reliability of information on Wikipedia by analyzing the quality of its Refs. Baigutanova et al. (2023). This involves the metric which calculates the percentage of sentences that require a citation but do not have one and the metric which determines the proportion of non-authoritative sources. While this updated model is an improvement from the previous version, it still has limitations that particularly affect low-resource languages. It continues to remain efficient for English.

⁴ https://en.wikipedia.org/wiki/Wikipedia:Content_assessment.

⁵ https://en.wikipedia.org/wiki/Wikipedia:Featured_articles.

⁶ https://en.wikipedia.org/wiki/Wikipedia:Featured_article_review.

⁷ https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

⁸ https://meta.wikimedia.org/wiki/Research:Prioritization_of_Wikipedia_Articles/Language-Agnostic_Quality#cite_note-wikirank-1.

Others have looked at related problems of content integrity in Wikipedia, however not focused on citation worthiness detection. For example, researchers studied detection of self-contradiction in Wikipedia articles (Hsu et al., 2021), labeling of low-quality Wikipedia content (Asthana et al., 2021) or sentence quality estimation (Ando et al., 2024). There has also been limited effort in building citation datasets out of Wikipedia (Ando et al., 2024; Singh et al., 2021), however, these have been limited to English only and have not studied models for citation detection.

All in all, citation worthiness detection in Wikipedia has been understudied to date, limited to English only and neglecting the role of sentence contextualization. Our study addresses this gap by studying citation worthiness detection for smaller Wikipedias. We specifically study the importance of contextualization in two ways: by incorporating surrounding sentences and by integrating topic categories.

3.4. Claim checkworthiness detection

A related task of citation needed is claim detection, which is the first step of the fact-checking process (Zeng et al., 2021). It aims to identify statements requiring verification (Thorne and Vlachos, 2018; Panchendrarajan and Zubiaga, 2024). This specific step is similar to identifying sentences that require citation, as both need some sort of support that was not provided by the person making the assertion.

Various studies conducted by academics and fact-checking organizations proposed effective approaches to automate this step of the process (Hassan et al., 2015, 2017; Arslan et al., 2020; Jaradat et al., 2018; Konstantinovskiy et al., 2021; Abumansour and Zubiaga, 2023; Bai et al., 2023; Sheikhi et al., 2023). Early work in claim detection included traditional machine learning methods, like ClaimBuster (Hassan et al., 2015, 2017; Arslan et al., 2020) which was first end-to-end automated fact-checking system; ClaimRank that is the first multilingual automated system designed to detect check-worthy claims in a given text of political domain (Jaradat et al., 2018). A successful collaboration between academics and fact-checking organization aimed to develop an annotation schema and a benchmark for automated claim detection (Konstantinovskiy et al., 2021). The authors introduced a novel approach that utilized universal sentence representations for classification, successfully achieving improvements over the ClaimBuster and ClaimRank methods. Apart from these efforts, research in claim checkworthiness detection has been supported by a number of shared tasks under the umbrella of CheckThat! (Nakov et al., 2022).

Due to the necessity of domain expertise, fact-checking organizations are usually specialized on the political domain, and their research usually is based on datasets generated from political debates. In such cases their models perform well in data coming from a similar domain but not others. In our work, we do not impose any restrictions concerning the topic under examination. Instead, we consider any topic covered in Wikipedia that satisfies the methodology of our framework. Notably, the aforementioned studies primarily focused on the English language. However, our study seeks to broaden the scope of detecting sentences needing some type of support, to encompass under-resourced languages such as the Albanian, Basque and Catalan languages.

4. Collection and labeling of citation needed datasets

Existing datasets for citation worthy sentences in Wikipedia have mainly been collected for English. They depend on the assumption that high-quality articles – featured articles, which undergo rigorous peer-review and adhere to the highest encyclopedic standards – already contain all necessary citations (Redi et al., 2019). This allows the automatic labeling of sentences thus the final dataset is created. This approach is however not extensible to smaller Wikipedias with fewer active editors, as is the case with the Albanian, Basque, or Catalan Wikipedias.

The main challenge with smaller Wikipedias lies in the scarcity of an active community capable of conducting the rigorous peer-review process required for designating featured articles. For example, while the English Wikipedia thrives with around 38,000 active editors each month, the Albanian, Basque, and Catalan Wikipedias have far fewer — approximately 60, 219, and 434 monthly active editors, respectively, over two years.⁹ This limited editorial capacity affects their ability to generate and approve featured content, resulting in a much smaller pool of featured articles, as evidenced by the mere 33 featured articles on the Albanian Wikipedia.¹⁰ Thus, applying the same data collection strategies used in previous studies to these smaller Wikipedias would lead to very small datasets, insufficient for research purposes.

To overcome this limitation, we devise an alternative data collection and quality assessment methodology that adapts to the realities of smaller Wikipedias. This new approach allows us to collect SQ-citation-needed, EU-citation-needed, and CA-citation-needed, citation worthiness detection datasets for the Albanian, Basque, and Catalan Wikipedias. In what follows, we describe in more detail this methodology.

4.1. Data collection

We use a combination of publicly available Wikipedia dumps and the Wikipedia API to collect all the data for our datasets. Initially, we collected a total of 93,442 articles from the Albanian Wikipedia,¹¹ 417,739 articles from the Basque Wikipedia,¹² and 723,899 articles from the Catalan Wikipedia.¹³ Having all these articles, we defined a set of regular expressions to parse the texts and break them down into sentences. We then follow a set of steps to complement the dataset with additional information. The data processing flow is presented in Fig. 1 and further details explaining this workflow are described next.

Data labeling. The main goal of Wikipedia is to be a source of concise but detailed information across all knowledge areas.¹⁴ One of the main policies of Wikipedia’s content is verifiability.¹⁵ This principle allows users of the encyclopedia to check the credibility of information through sources provided in the article’s text, typically as inline citations. In our study, the presence of these citations serves as a benchmark for assessing the trustworthiness of a statement. We treat each sentence in a Wikipedia article as a claim; if the sentence has an inline citation we say that the claim is check-worthy, and if it does not have an inline citation we mark it as not check-worthy.

Therefore, in the process of breaking down articles into individual sentences, the classification becomes automated: sentences with inline citations are labeled as “citation” (1), indicating a claim that can be checked for verifiability, whereas those without inline citations are labeled as “no-citation” (0), signaling a claim that may not be check-worthy. Given that this assumption can be risky if done for all articles of an entire Wikipedia, next we further elaborate on how we choose high-quality articles where we can rely on existing citations.

Determining the quality of articles. Where previous research tackling English Wikipedia has determined article quality based on whether the article is highlighted as a featured article, this is impractical for small Wikipedias as the number of featured articles is very small. Alternatively, we use Wikipedia’s own open source language-agnostic quality framework.¹⁶ This approach helps us generate quality scores for

⁹ <https://stats.wikimedia.org/#/all-projects>.

¹⁰ https://sq.wikipedia.org/wiki/Wikipedia:Artikuj_t%C3%AB_p%C3%ABrkrer.

¹¹ <https://sq.wikipedia.org/>.

¹² <https://eu.wikipedia.org/>.

¹³ <https://ca.wikipedia.org/>.

¹⁴ <https://en.wikipedia.org/wiki/Wikipedia:Purpose>.

¹⁵ <https://en.wikipedia.org/wiki/Wikipedia:Verifiability>.

¹⁶ https://meta.wikimedia.org/wiki/Research:Prioritization_of_Wikipedia_Articles/Language-Agnostic_Quality.

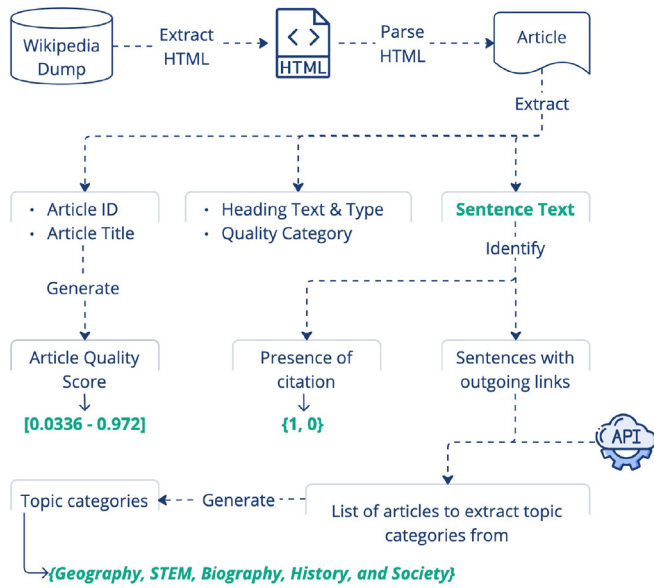


Fig. 1. Overview of the dataset processing workflow. This flowchart illustrates the procedure applied to Wikipedia dumps in languages sq, eu, and ca. The process involves extracting HTML content, parsing the articles, and identifying key elements such as article IDs, titles, headings, quality scores, citation presence, and topic categories. The final output includes sentences classified by citation needs and categorized by topic categories (Geography, STEM, Biography, History, and Society). Each step is automated and integrated, as indicated by the dashed lines, with an API used for topic category extraction. The green text in the figure highlights the final elements of the dataset.

any given article from any Wikipedia, with scores between zero and one. We explain how we use them to filter high-quality articles in the next section.

Collecting topic categories for articles. We are interested in investigating how the topics associated with a particular sentence help predicting if it needs a citation or not. To retrieve the list of topics relevant to a sentence, we follow a two-step process. First, we extract the outgoing links from a sentence to other Wikipedia articles, which gives us a list of Wikipedia articles linked from that particular sentence. Second, we use the Wikipedia language-agnostic topic classification framework (Johnson et al., 2021), which allows us to extract the topic categories associated with each of the linked Wikipedia article. We then aggregate the topic categories across all of the outgoing links to come up with a consolidated list of topic categories associated with the sentence in question. The topic categories can include any of the 64 categories provided by the system; these categories belong to one of the five high-level categories (Geography, STEM, Biography, History, and Society), however in our work we use the 64 lower-level categories. Note that on occasions this method can return an empty set of categories when the sentence does not have any outgoing links.

Resulting data structure. Once we complete the steps above, each of the articles in our dataset contains the following information:

- **Sentences:** The main component of the dataset is the textual content of sentences of a Wikipedia article, which is a list of sentences within the article. For the purposes of experimentation, these are the sentences after stripping the citation, where there was one.
- **Citation labels:** Binary labels indicating the presence (1) or absence (0) of inline citations, aligned with each sentence in the article.
- **Article’s quality score:** This is the score, between zero and one, indicating the overall quality of the article.

- **Topic categories:** For each sentence, the aggregated set of categories associated with the outgoing links, which we use as indicative of the topics associated with the topic itself.

4.2. Data cleaning and preparation

After completing the data collection with the additional topic categories, quality scores and labels, as indicated above, we perform two filtering steps to increase and ensure the quality of the resulting dataset.

Filtering articles with fewer than 5 sentences. To get rid of very short articles lacking depth, we removed articles with fewer than 5 sentences. The threshold of five sentences was chosen based on the general guidelines of paragraph construction, which typically consists of three to five sentences. This approach ensured that the articles under consideration contained more complete ideas even if they were only one paragraph in length.

Filtering article by quality scores. To filter out low-quality articles using the quality scores retrieved as described above, we opted to perform a quantile-based discretization of the continuous variable. We split all the quality scores into five quintiles, such that each of the splits has (nearly) the same number of sentences pertaining to different score ranges. This led to five quality-based groups, ranging from 1 to 5, 1 indicating the lowest quality and 5 the best quality. We conducted preliminary experiments with train-test splits within each of the quality-based groups, observing an obvious increase in performance as the quality of articles increased. This, along with previous research suggesting that article quality and use of citations correlate (Chou et al., 2020), led us to rely on these quality scores to filter the data for high-quality articles. After this process, we only keep articles pertaining to category 5, the highest-quality articles whose quality score ranges from 0.745 to 0.966 for SQ-citation-needed, 0.716 to 0.957 for EU-citation-needed, and from 0.712 to 0.972 for CA-citation-needed dataset. Further details, like the number of sentences per quality group and topics with citation ratio, are shown in Table 1. Whereas, a visualization showing the distribution of quality scores across articles and sentences is shown in Fig. 2.

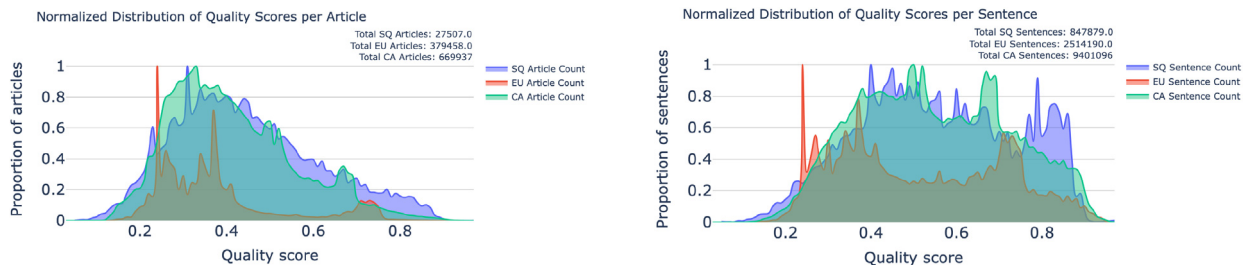
Imbalanced nature of the dataset. Previous research in citation worthiness detection has balanced the classes to facilitate the experimentation, which however also hinders its realistic applicability. With our three datasets, we keep the original class imbalance for the experimentation, which for the final version of the Albanian dataset is 3:1 (no citation to citation); for the Basque dataset it is 3:1 (no citation to citation), and for the Catalan dataset it is 2:1 (no citation to citation).

4.3. Dataset statistics

In this section, we provide basic statistics of the datasets created from the Albanian (sq), Basque (eu), and Catalan (ca) editions of Wikipedia.

Initially, our collection of data yielded a total of 93,442 Albanian articles, 417,739 Basque articles, and 723,899 Catalan articles. Upon applying the inclusion criterion – requiring a minimum of five sentences per article – our dataset was refined to 27,507 Albanian articles, 379,458 Basque articles, and 669,937 Catalan articles. Subsequently, we classified these articles into quality groups. Each group was composed to represent 20% of the total sentences across articles per language. Table 1 provides a summary of the datasets derived from the sq, eu, and ca Wikipedia editions. It showcases the relationship between article quality scores, the number of articles and sentences, and the distribution of topic categories and citation ratios within each dataset.

Fig. 2, with Figs. 2(a) and 2(b), present normalized distributions of quality scores of articles in Wikipedia differentiated by language Albanian (sq), Basque (eu), and Catalan (ca).



(a) Distribution of quality scores per article: This graph shows frequencies of articles within specific quality score ranges for Albanian (sq), Basque (eu), and Catalan (ca) Wikipedia editions, normalized to the highest article count for each language.

(b) Distribution of quality scores per sentence: This graph shows the normalized distribution of sentences from articles across different quality scores for the three language editions of Wikipedia.

Fig. 2. Distribution of quality scores.

Table 1

Quality scores in relation to the number of sentences, topic categories, and inline citation presence: A summary of article counts, sentence counts, topic distributions, and citation ratios across different quality intervals for the SQ, EU, and CA citation-needed datasets.

Dataset	Quality label	Quality score interval	# of articles	# of sentences	Topic ratio	Citation ratio	Topic and citation ratio
SQ-citation-needed	1	(0.0461, 0.391]	11,607	169,583	15.33%	2.79%	0.37%
	2	(0.391, 0.49]	6136	169,834	15.21%	4.69%	1.01%
	3	(0.49, 0.604]	4724	169,382	17.13%	7.76%	1.79%
	4	(0.604, 0.745]	3365	169,514	20.20%	14.03%	3.09%
	5	(0.745, 0.966]	1675	169,566	25.54%	28.51%	6.79%
EU-citation-needed	1	(0.066, 0.321]	160,940	502,838	66.24%	2.74%	1.53%
	2	(0.321, 0.411]	128,226	502,843	68.66%	9.26%	7.03%
	3	(0.411, 0.571]	41,383	502,842	47.75%	10.71%	6.39%
	4	(0.571, 0.716]	26,980	502,835	40.69%	15.16%	7.00%
	5	(0.716, 0.957]	21,929	502,832	44.55%	23.82%	11.36%
CA-citation-needed	1	(0.0336, 0.38]	318,576	1,880,222	53.93%	18.23%	10.17%
	2	(0.38, 0.486]	162,797	1,880,220	48.04%	23.11%	11.93%
	3	(0.486, 0.596]	102,695	1,880,221	42.78%	24.98%	12.12%
	4	(0.596, 0.712]	63,726	1,880,244	34.28%	26.36%	9.89%
	5	(0.712, 0.972]	22,143	1,880,189	41.13%	29.81%	13.38%

Fig. 2(a) shows the proportion of articles across different quality scores, normalized such that the highest number of articles for each language corresponds to a value of 1. This figure indicates an inverse relationship between the number of articles and quality scores, with the number of articles gradually decreasing, as the quality score increases. This also shows the selective nature of choosing the top-quality articles, which only a small number of articles satisfy.

Fig. 2(b), presents the distribution of quality scores of articles but split into sentences. It plots the proportion of sentences, rather than articles, across the same range of quality scores. The area plots shown in this figure do not show a clear inverse relationship like Fig. 2(a), instead, the distributions are more varied. This is affected by the normalization of the data — although the number of high-quality articles is smaller, as Table 1 indicates, these articles disproportionately contribute more sentences to the dataset than their lower-quality counterparts. This can be attributed to the fact that higher quality articles are often more detailed and longer, requiring more effort and time to research and write. Moreover, these articles are more likely to include inline citations and links to other sources or related Wikipedia articles, a trend that Table 1 supports by showing an increase in sentences with citations and links alongside rising quality scores.

Together, these findings imply that fewer articles achieve overall high quality and that the quality of an article, as measured by its score, is positively associated with the article’s length in terms of sentence count.

5. The Contextualized Citation Worthiness (CCW) model

In this study, we address the problem of identifying sentences that require citations within articles in low-resource languages on

Wikipedia. We refer to this as the citation worthiness detection task and we conduct it separately for three languages: Albanian (sq), Basque (eu), and Catalan (ca).

In tackling this task, our main objective is to build a model that demonstrates our hypothesis revolving around the effectiveness of leveraging contextualized representations for citation worthiness detection. Our hypotheses are twofold: (1) *H1*, sentences adjacent to the sentence in question, including the previous and next sentences, will benefit the citation detection model, and (2) *H2*, the topic categories associated with the sentence will help determine if it needs a citation.

To study this, we build our citation worthiness detection model, the Contextualized Citation Worthiness (CCW) model, on top of a transformer-based multilingual BERT (Devlin et al., 2019) model. This was a reasonable choice given the limited availability of pre-trained language models for low-resource languages such as the Albanian, Basque, and Catalan language. Our CCW model incorporates different inputs to test our hypotheses regarding contextualized modeling. More specifically, we test with variants of the model that leverage different portions of adjacent sentences as well as the use or not of information from topic categories. In addition, we compare our models with an additional baseline model of our own, as well as a replicated version of a competitive baseline from the literature.

5.1. Problem statement

Let D_{lang} represent the dataset for a specific language $lang$, where $lang \in \{sq, eu, ca\}$ denotes Albanian, Basque, and Catalan, respectively. Each dataset D_{lang} consists of multiple Wikipedia articles, such that:

$$D_{lang} = \{A_{lang}^1, A_{lang}^2, \dots, A_{lang}^n\} \quad (1)$$

where A_{lang}^i denotes the i th article in the dataset D_{lang} , and n is the total number of articles in the dataset. Each Wikipedia article A_{lang}^i is composed of a sequence of sentences:

$$A_{lang}^i = \{s_{lang}^{i1}, s_{lang}^{i2}, \dots, s_{lang}^{im_i}\} \quad (2)$$

where s_{lang}^{ij} denotes the j th sentence in the i th article of the $lang$ dataset and m_i is the total number of sentences in the i th article. Each sentence s_{lang}^{ij} may be associated with a set of topic categories:

$$T_{lang}^{ij} = \{t_{lang}^{ij1}, t_{lang}^{ij2}, \dots, t_{lang}^{ijk_{ij}}\} \quad (\text{case when a sentence has associated topic categories}) \quad (3a)$$

or

$$T_{lang}^{ij} = \emptyset \quad (\text{case when a sentence does not have associated topic categories}) \quad (3b)$$

where T_{lang}^{ij} denotes the set of topics associated with the j th sentence in the i th article; t_{lang}^{ijk} denotes the k -th topic in the set T_{lang}^{ij} , and k_{ij} is the total number of topics associated with the j -th sentence in the i -th article.

The problem of citation worthiness involves assigning each sentence s_{lang}^{ij} a label from one of two categories: $y_{lang}^{ij} \in \{y_{c,lang}^{ij}, y_{n,lang}^{ij}\}$. Specifically, $y_{c,lang}^{ij}$ indicates that the sentence is considered check-worthy and it might need a citation, while $y_{n,lang}^{ij}$ signifies that the sentence is not considered check-worthy.

To address this problem, we investigate different input representations for each sentence s_{lang}^{ij} considering its context and associated topic categories. We elaborate on the inputs' representation design in the section below.

5.2. Input representation

Our CCW model utilizes various input representations, combining components such as the current sentence (required), the previous sentence, the next sentence, and topic categories. We design three main types of input representations: single sentence input, sentence pair input, and sentence triad input. Each type includes variants with and without topic categories. Specifically, the variants are:

1. Without Topic Categories:

- Single sentence input (SSI)
- Sentence pair input with the previous sentence (SPIPS)
- Sentence pair input with the next sentence (SPINS)
- Sentence triad input (STI)

2. With Topic Categories:

- Single sentence input with topics (SSIT)
- Sentence pair input with the previous sentence and topics (SPIPST)
- Sentence pair input with the next sentence and topics (SPINST)
- Sentence triad input with topics (STIT)

These different input variants are fed into the CCW model, with the elements separated by the special token $[SEP]$. The CCW model can process up to 512 tokens as input, which accommodates the average sentence length of 21 words in our dataset without issue. The length constraint is not a problem for building different input representations. The resulting contextual embeddings are then fed into a multilingual BERT model. The advantage of using BERT compared to RNN or other traditional networks is that the model learns the bidirectional representation (i.e. the inner representation of the language). This then will allow the model to make predictions for citation worthiness while preserving the semantics of the low-resource languages.

Single sentence input (SSI). The single sentence input representation is the one used in previous research and therefore it represents the input of the main baselines (Redi et al., 2019; Gosangi et al., 2021). To build this representation for an article A_{lang}^i , the input to the model for each underlying sentence s_{lang}^{ij} is:

$$ssi_{lang}^{ij} = s_{lang}^{ij} \quad (4)$$

Single sentence input with topic (SSIT). The single sentence input with topics builds upon the single sentence input by incorporating associated topic categories when they exist. This input representation provides additional context to the model, potentially improving performance by leveraging the semantic information of the topics. To build this representation for an article A_{lang}^i , the input to the model for each underlying sentence s_{lang}^{ij} is:

$$ssit_{lang}^{ij} = s_{lang}^{ij} + [SEP] + T_{lang}^{ij} \quad (5)$$

In cases where there are no associated topic categories for a given sentence, T_{lang}^{ij} is an empty set and the input defaults back to the ssi_{lang}^{ij} .

Example:

For s_{sq}^{11} , if $T_{sq}^{11} \neq \emptyset$:

$$ssit_{sq}^{11} = s_{sq}^{11} + [SEP] + \{t_{sq}^{111}, t_{sq}^{112}, t_{sq}^{113}\}$$

showing that s_{sq}^{11} is linked to three topic categories. The CCW architecture of models handling single sentence inputs is shown in Fig. 3(b).

Sentence pair input (SPI). Another input variant we propose is the sentence pair classification, which integrates the surrounding sentence and topic categories. This approach comes in two forms: one that incorporates the previous sentence in addition to the current, and another one that incorporates the next sentence along with the current sentence. Each form has a version, which incorporates topic categories, and a basic version without them. By evaluating both, we aim to determine whether context from the previous or the subsequent sentence is more effective in enriching the representation and thus, enhancing the model's performance.

For a given sentence s_{lang}^{ij} the sentence pair input with previous sentence (SPIPS) takes the form:

$$sps_{lang}^{ij} = s_{lang}^{i(j-1)} + [SEP] + s_{lang}^{ij} \quad (6)$$

if $j = 1$, indicating the first sentence in a Wikipedia article, then the previous sentence is undefined and can be denoted as $s_{lang}^{i(j-1)} = \emptyset$. The sentence pair input representation with the next sentence (SPINS) takes the form:

$$spins_{lang}^{ij} = s_{lang}^{ij} + [SEP] + s_{lang}^{i(j+1)} \quad (7)$$

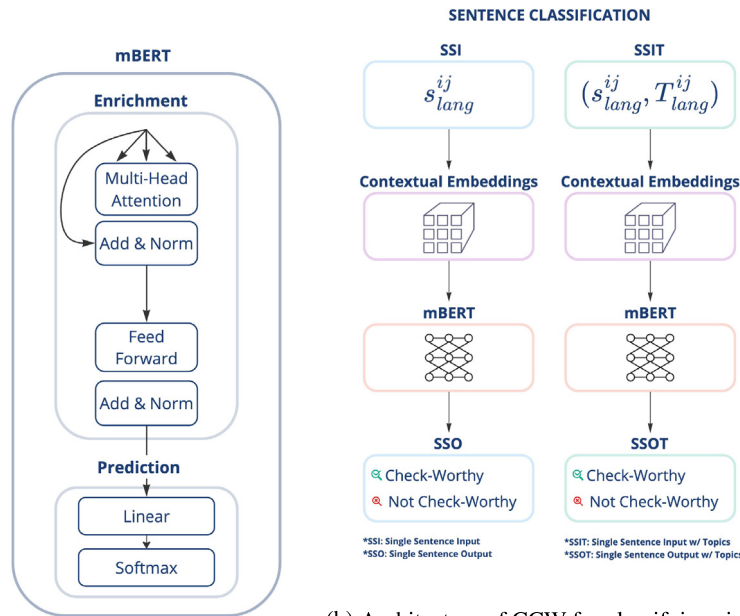
if $j = m$, indicating the last sentence in a Wikipedia article, then the next sentence is undefined and can be denoted as $s_{lang}^{i(j+1)} = \emptyset$. The CCW architecture of models handling sentence pair inputs is shown in Fig. 4.

Sentence pair input with topics (SPIPST). The sentence pair input representation with topics builds upon sentence pair input by incorporating topic categories when they exist. In cases when a given sentence s_{lang}^{ij} has associated topic categories and $T_{lang}^{ij} \neq \emptyset$ the sentence pair input with the previous sentence and topics (SPIPST) and the input with the next sentence and topics (SPINST) take the form:

$$sps_{lang}^{ij} = s_{lang}^{i(j-1)} + [SEP] + s_{lang}^{ij} + [SEP] + T_{lang}^{ij} \quad (8)$$

and:

$$spinst_{lang}^{ij} = s_{lang}^{ij} + [SEP] + T_{lang}^{ij} + [SEP] + s_{lang}^{i(j+1)} \quad (9)$$



(a) Architecture of mBERT model (b) Architecture of CCW for classifying single sentence inputs without (left) and with (right) topic categories

Fig. 3. Overview of mBERT architecture and CCW architecture for single sentence input classification. Fig. 3(a) illustrates the architecture of mBERT which is incorporated in all of the eight variants of CCW Model. Fig. 3(b) illustrates the CCW classification of single sentence input (SSI) and single sentence input with topics (SSIT) which incorporates additional topic categories information they exist. From each representation, we generate contextual embeddings which then pass through mBERT and results in classification for Check-Worthy or Not Check-Worthy sentences.

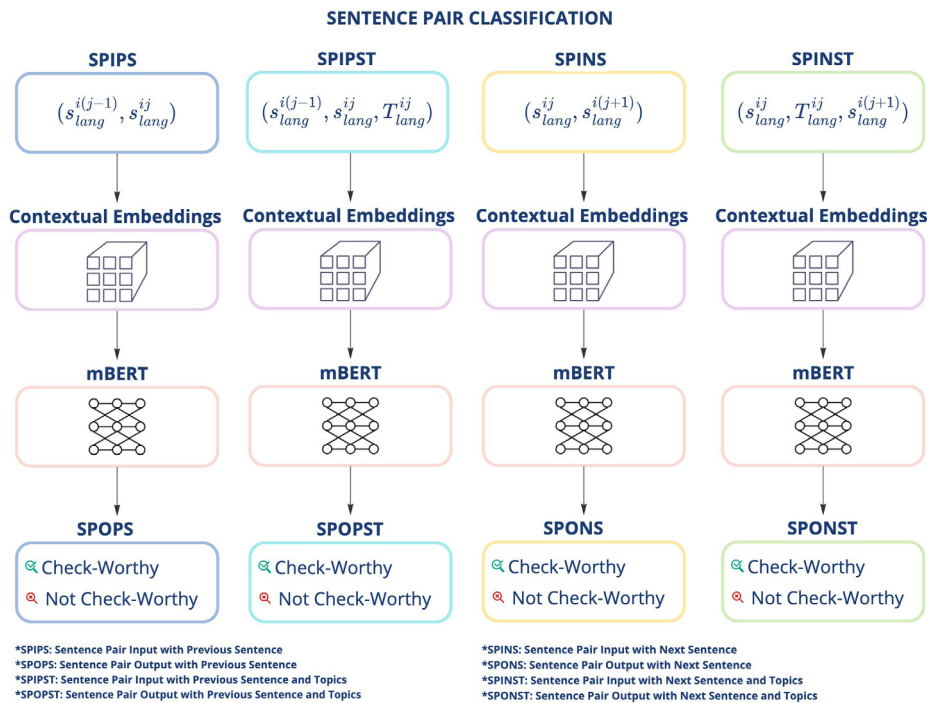


Fig. 4. This diagram illustrates the architecture of the CCW model for sentence pair input representations, with and without topic categories. The representations include Sentence Pair Input with Previous Sentence (SPIPS), Sentence Pair Input with Previous Sentence and Topics (SPIPST), Sentence Pair Input with Next Sentence (SPINS), and Sentence Pair Input with Next Sentence and Topics (SPINST). From each representation we generate contextual embeddings which then are passed through mBERT, and results in classifications for check-worthy or not check-worthy sentences.

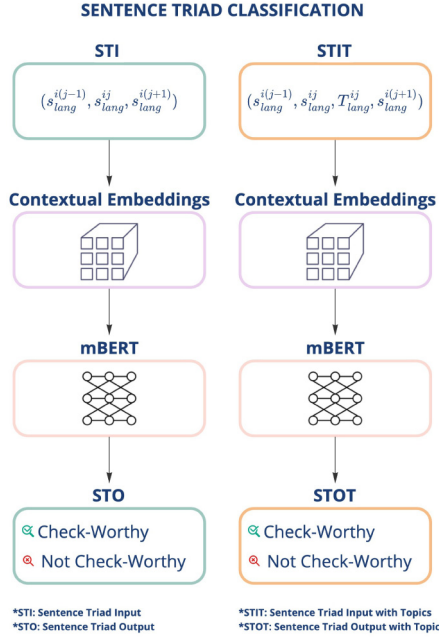


Fig. 5. This diagram illustrates the architecture of the CCW model for sentence triad input representations, with and without topic categories. The representations include Sentence Triad Input (STI), and Sentence Triad Input With Topics (STIT). The STIT incorporate additional topic categories information when they exist. From each representation we generate contextual embeddings which then are passed through mBERT, and results in classifications for check-worthy or not check-worthy sentences.

Sentence triad input (STI). The last type of input representation is a sentence triad, where the current sentence is accompanied by both adjacent sentences (its previous and next sentences). For a given sentence s_{lang}^{ij} , the sentence triad input representation is:

$$sti_{lang}^{ij} = s_{lang}^{i(j-1)} + [\text{SEP}] + s_{lang}^{ij} + [\text{SEP}] + s_{lang}^{i(j+1)} \quad (10)$$

Sentence triad input with topics (STIT). Sentence triad input with topics builds upon sentence triad input by incorporating topic categories when they exist. In cases when a given sentence s_{lang}^{ij} has associated topic categories and $T_{lang}^{ij} \neq \emptyset$ the STIT is:

$$stit_{lang}^{ij} = s_{lang}^{i(j-1)} + [\text{SEP}] + s_{lang}^{ij} + [\text{SEP}] + T_{lang}^{ij} + [\text{SEP}] + s_{lang}^{i(j+1)} \quad (11)$$

For both variants STI and STIT if $j = 1$, indicating the first sentence in a Wikipedia article, then the previous sentence is undefined and can be denoted as $s_{lang}^{i(j-1)} = \emptyset$. Likewise, if $j = m$, indicating the last sentence in a Wikipedia article, then the next sentence is undefined and can be denoted as $s_{lang}^{i(j+1)} = \emptyset$. The CCW architecture of models handling sentence triad inputs is shown in Fig. 5.

5.3. Experiment settings

We set up the CCW model with the following hyperparameters and settings:

Model settings: We implemented our method using Hugging Face¹⁷ and PyTorch¹⁸ For all of our experiments, we used multilingual BERT (mBERT) as the input text representation generator. The sequence classification model is a fully-connected layer that takes the input text representation generated by the BERT multilingual and outputs citation

needed labels through the Softmax function. The architecture of mBERT model is shown in Fig. 3(a).

Hyperparameter tuning: We performed hyperparameter tuning to optimize the model’s performance. Various learning rates were tested, including 0.00001, 0.00002, 0.00003, 0.00005, 0.0002, 0.001, 0.01, and 0.1. A learning rate of 0.0002 yielded the best results. Additionally, we experimented with different weight decay parameters (0.00001, 0.0001, 0.001, 0.01, and 0.1) and found that a weight decay of 0.001 performed optimally. The Adam optimizer was used with the best-performing learning rate of 0.0002 and weight decay of 0.001.

Experimental reproducibility: We set the batch size to 32 and ran all experiments with 10 different seeds. The final reported results are the average of these 10 runs. Each experiment ran for a maximum of 50 epochs, employing early stopping if the F1 Score did not improve in the last three epochs. The best-performing epoch on the validation dataset was then selected and applied to the test dataset.

Train-test splits: For each language dataset – Albanian, Basque, and Catalan – we partitioned the data into three subsets, with 60% of the original dataset allocated to the training set, 20% to the validation set, and 20% to the test dataset. To maintain the integrity of the evaluation process, the splitting was performed at the article level, ensuring that the model is tested on completely unseen articles, preventing any overlap where sentences from the training articles might appear in the test data. Consequently, this approach guarantees that sentences adjacent to those in the test set – either previous or subsequent – have not been exposed to the model during its training phase. Table 2 shows the details of train, validation, and test splits for the dataset of each language.

Class imbalance: In our experimental design, we chose to maintain the original class imbalance to accurately reflect real-world scenarios. As stated in the “Imbalanced Nature of the Dataset” Section 4.2, the class distributions in our datasets vary from 2:1 to 3:1, no citation to citation across datasets. This approach aims to demonstrate the model’s performance under conditions that mirror the real-world distribution of citation needs in Wikipedia articles, thus enhancing the practical relevance of our findings. In citation worthiness detection, it is crucial to minimize false negatives (i.e., sentences that need citations but are not flagged). Maintaining class imbalance allows us to prioritize this, ensuring that more sentences requiring citations are correctly identified. Although we expect the model to be more proficient at identifying the larger class (no citation needed), it provides insights into how the model performs across different classes without artificially altering the dataset. Our experiments show that by maintaining class imbalance, we have successfully reduced the false negative rate by 25% for the Catalan dataset, 13% for the Basque dataset, and 15% for the Albanian dataset. This is critical in ensuring that sentences needing citations are appropriately flagged, thus improving the reliability of the content. While we expect the overall performance in predicting true positive cases not be as high as in a balanced dataset, the focus on realistic class distributions and reducing false negatives we aim to ensure that our model remains practically useful and relevant.

Evaluation measure: For all experiments, our primary focus will be on reporting the F1 score. In all tables, we present precision and recall, as these are the fundamental metrics upon which the F1 score is based. Each table will display the precision, recall, and F1 score of the positive class ($y_{c,lang}^{ij}$), alongside the weighted precision, recall, and F1 score for the entire test dataset. Additionally, we will visualize confusion matrix metrics to provide insights into the model’s predictive performance and identify areas where errors occur. The emphasis on the weighted F1 score is due to the imbalanced nature of our datasets, where the number of sentences without citations significantly exceeds those with citations.

5.4. Baseline models

Our study aims to explore contextualization through our CCW model. As a benchmark, we consider the simplest of our CCW models which only takes a single sentence as an input and throughout experiments will be denoted as CCW-SSI model.

¹⁷ <https://huggingface.co/>.

¹⁸ <https://pytorch.org/>.

Table 2
Statistics of the final datasets.

Dataset	Data	Articles	# of sentences w/ Inline citation	# of sentences w/o Inline citation	# of all sentences
SQ-citation-needed	Train	1005	29,928	77,105	107,033
	Dev	335	9147	22,190	31,337
	Test	335	9263	21,933	31,196
	Total	1675	48,338	121,228	169,566
EU-citation-needed	Train	13,157	73,086	232,372	305,458
	Dev	4386	24,770	79,530	104,300
	Test	4386	21,935	71,139	93,074
	Total	21,929	119,791	383,041	502,832
CA-citation-needed	Train	13,285	335,538	802,052	1,137,590
	Dev	4429	109,590	258,428	368,018
	Test	4429	115,365	259,216	374,581
	Total	22,143	560,493	1,319,696	1,880,189

For evaluation, we compare our model (CCW-SSI) against two adapted versions of a baseline model from prior studies. Specifically, we adapted the model from Redi et al. (2019) for Albanian, Basque, and Catalan languages by integrating BERT and fastText embeddings due to the lack of pre-trained GloVe models for these languages. We refer to these adaptations as the **Redi-MB** and **Redi-FT** models.

5.4.1. Architecture of Redi-MB and Redi-FT

Both models are based on Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) cells, following the architecture proposed by Redi et al. (2019). Redi-MB uses BERT embeddings, and Redi-FT model employs fastText Continuous Bag of Words (CBOW) embeddings (Grave et al., 2018) for word representation. Redi-FT uses an RNN with GRU cells to encode word sequences. The final hidden state is then used for classification. Whereas, Redi-MB enhances the GRU-based RNN with an attention mechanism. Attention weights are applied to each hidden state, creating a weighted representation of the sentence. This context vector is then passed through a fully connected layer for the final output predictions.

5.5. Statistical analysis

To ensure that the observed results of our study are credible and not a random chance we have performed statistical significance tests. We compare our proposed models against three baseline models – Redi-FT, Redi-MB, and CCW-SSI – across three datasets: CA-citation-needed, EU-citation-needed, and SQ-citation-needed. For each dataset, we conduct separate statistical significance tests to compare the performance of each proposed models against single sentence input models. We utilize the paired t-test (ttest_rel from SciPy library) to compute t-statistics and p-values for all model comparisons. For all combinations, we use a degrees of freedom (df) value of 9 for weighted F1 Score performance metric. The paired t-tests evaluated our null hypothesis:

- **H0:** The enhancements made to the CCW models do not result in a statistically significant improvement in performance compared to single sentence input models. In other words, the weighted F1 scores of the contextually enriched models are equal to those of the baseline models.

Based on the p-values from the tests, we determine whether to reject the null hypothesis. If the p -value is less than the chosen significance level of 0.05, we reject the null hypothesis.

6. Experimental results

In this section, we present the evaluation of our Contextualized Citation Worthiness (CCW) model alongside the baseline models, Redi-FT and Redi-MB, on the SQ-citation-needed, EU-citation-needed, and CA-citation-needed datasets. The results are organized into two main

Table 3

Comparison of single-sentence baseline approaches. P, R, and F1 are precision, recall, and F1 scores for the positive class $y_{c,lang}^{ij}$. W-P, W-R, W-F1 ARE the weighted metrics on the entire test dataset.

Language	Input	Model	P	R	F1	W-P	W-R	W-F1
sq	s_{sq}^{ij}	REDI-FT	0.46	0.25	0.32	0.67	0.71	0.68
		REDI-MB	0.51	0.31	0.39	0.70	0.73	0.70
		CCW-SSI	0.57	0.36	0.44	0.73	0.75	0.73
eu	s_{eu}^{ij}	REDI-FT	0.57	0.25	0.34	0.78	0.81	0.77
		REDI-MB	0.71	0.20	0.31	0.80	0.82	0.77
		CCW-SSI	0.61	0.28	0.39	0.79	0.82	0.79
ca	s_{ca}^{ij}	REDI-FT	0.51	0.17	0.25	0.71	0.76	0.71
		REDI-MB	0.62	0.12	0.20	0.74	0.77	0.70
		CCW-SSI	0.57	0.27	0.37	0.75	0.77	0.74

parts. First, we report the performance using standard single-sentence inputs. Second, we present the outcomes for our proposed enriched representations, which incorporate contextualization including adjacent sentences. For both sets of results, we analyze the impact of incorporating additional context derived from the associated topic categories.

6.1. Results for single-sentence models

We first compare the performance of all single-sentence models, including both, the Redi-FT and Redi-MB baseline models, as well as our CCW-SSI benchmark model for each language. To evaluate the overall performance of the classifiers, especially considering the impact of class imbalance, we have chosen positive F1 score and the weighted F1 score as presented in Table 3. Given the imbalanced nature of the data and the fact that we are dealing with datasets in multiple languages where class distribution varies, we emphasize the F1 score as our primary metric for assessing classifiers performance. This metric, representing the harmonic mean of precision and recall, is selected for its robustness where one class is overrepresented (like the ‘no-citation’ class in our datasets).

The CCW-SSI model obtains an F1 score of 0.73 for Albanian (sq), 0.79 for Basque (eu), and 0.74 for Catalan (ca) dataset, which is superior to the REDI-FT and REDI-MB models across three languages. Despite the overall low positive F1 scores across all languages, CCW-SSI consistently demonstrates an improvement over the baseline methods (REDI-FT and REDI-MB). This indicates that our proposed method is more effective in identifying the positive class, even though there is still room for further enhancement. The improvements, albeit modest, suggest that CCW-SSI could be a promising direction for refining our models and achieving better performance in the future.

Table 4 with confusion matrices offers a detailed breakdown of each classifier performance and the specific types of errors each model makes. The confusion matrices reinforce the superiority of the CCW-SSI model against REDI-FT and REDI-MB as previously evidenced by

Table 4
Confusion matrices for single sentence classification models on SQ-citation-needed, EU-citation-needed, and CA-citation-needed datasets.

Language	REDI-FT	REDI-MB	CCW-SSI	
sq				
	eu			
		ca		

F1 scores. Across all three languages, the CCW-SSI matrices reveal a consistent pattern of higher true positive rates compared to other baseline models.

Examining the error patterns presented in Table 4, it is evident that all models exhibit a higher rate of false negatives as opposed to false positives. In practice, this trend suggests that models are more likely to overlook sentences that should be flagged for citations rather than mistakenly flagging those that do not require them. Notably, the CCW-SSI model shows the lowest rate of false negatives for the Albanian (sq) dataset, failing to correctly identify 64% of sentences needing citations. This is an improvement of 11% as compared to REDI-FT. Conversely, the highest frequency of false negatives is recorded by the REDI-MB model in the Catalan (ca) dataset, where it misses 88% of sentences that should be cited. This observation emphasized the challenge of accurately detecting citation-needed instances and reflects the potential of the CCW model to contribute to the advancement of this task.

In what follows we further experiment with the CCW model by enhancing it with contextualized representations that incorporate surrounding sentences and topic categories. This progression is directly informed by our current findings, which highlight the critical role of context in model performance. By enriching the model with a broader linguistic context, we aim to address the challenges in detecting sentences that need citation. These enhancements are expected to ease the high rates of false negatives observed in the baseline models.

6.2. Results for contextualized models

Having evaluated the performance of single-sentence models and demonstrated the competitive nature of the CCW model, we next move onto experimenting with different input representations that help us assess our hypotheses revolving around contextualization of sentences with surrounding sentences and topic categories. We present the results for this set of experiments with one question at a time. As described in the ‘Input Representation’ section above, we experiment with seven different contextualized representations in addition to the baseline, single-sentence input (SSI).

Table 5
Comparison of single-sentence baseline approaches for our CCW model showing the impact of topic categories. P, R, and F1 are precision, recall, and F1 scores for the positive class $y_{c,lang}^{ij}$. W-P, W-R, W-F1 ARE the weighted metrics on the entire test dataset.

Language	Model	P	R	F1	W-P	W-R	W-F1
sq	CCW-SSI	0.57	0.36	0.44	0.73	0.75	0.73
	CCW-SSIT	0.54	0.40	0.46	0.72	0.74	0.73
eu	CCW-SSI	0.61	0.28	0.39	0.79	0.82	0.79
	CCW-SSIT	0.60	0.29	0.39	0.79	0.81	0.79
ca	CCW-SSI	0.57	0.27	0.37	0.75	0.77	0.74
	CCW-SSIT	0.57	0.27	0.37	0.74	0.77	0.74

How effective are topic categories on the single-sentence model? As a first attempt at contextualizing the CCW model, we experiment with the model incorporating topic categories (denoted as CCW-SSIT) on top of the base single-sentence model (denoted as CCW-SSI). Table 5 shows precision, recall, and F1-Score of the positive class as well as the weighted precision, recall, and F1-Score for all test datasets. For Albanian (sq) dataset, including topic categories improves recall and F1 score, with a slight drop in precision. The true positive rate, as presented in Table 6, also improves, indicating better identification of positive cases. For Basque (eu) dataset, including topic categories has a minimal impact, slightly improving recall and the true positive rate with no significant change in other metrics. Lastly, for Catalan (ca) dataset including topic categories does not affect any of the metrics, suggesting that topic categories do not provide additional value for this language in this context. Overall, the effectiveness of including topic categories in single sentence input models varies by language, showing the most benefit for Albanian and minor impact for Catalan language. Further investigation could be conducted to understand why topic categories are more effective for certain languages.

Is the surrounding context helpful for determining whether a sentence needs citation? In the pursuit of contextualizing the CCW model, we expand our approach by including the adjacent sentences on top of the

Table 6
Confusion matrices for single sentence classification models on SQ-citation-needed, EU-citation-needed, and CA-citation-needed datasets.

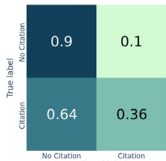
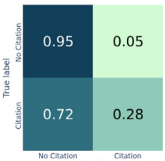
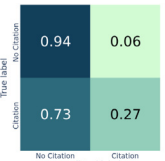
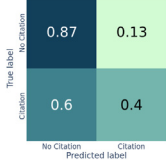
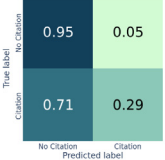
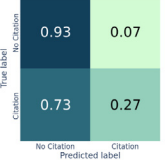
	sq	eu	ca
CCW-SSI			
CCW-SSIT			

Table 7
Comparison of context-based approaches for our CCW model showing the impact of adjacent sentences. P, R, and F1 are precision, recall, and F1 scores for the positive class $y_{c,lang}^{ij}$. W-P, W-R, W-F1 are the weighted metrics on the entire test dataset.

Language	Model	P	R	F1	W-P	W-R	W-F1
sq	CCW-SSI	0.57	0.36	0.44	0.73	0.75	0.73
	CCW-SPIPS	0.57	0.34	0.43	0.72	0.75	0.72
	CCW-SPINS	0.60	0.36	0.45	0.74	0.76	0.74
	CCW-STI	0.59	0.39	0.47	0.74	0.76	0.74
eu	CCW-SSI	0.61	0.28	0.39	0.79	0.82	0.79
	CCW-SPIPS	0.61	0.30	0.40	0.79	0.82	0.79
	CCW-SPINS	0.65	0.32	0.43	0.80	0.82	0.80
	CCW-STI	0.66	0.32	0.43	0.80	0.82	0.80
ca	CCW-SSI	0.57	0.27	0.37	0.75	0.77	0.74
	CCW-SPIPS	0.57	0.31	0.40	0.75	0.77	0.75
	CCW-SPINS	0.60	0.35	0.45	0.76	0.79	0.76
	CCW-STI	0.61	0.36	0.45	0.77	0.79	0.77

base single-sentence model. For these experiments, we compare three additional models: CCW-SPIPS, CCW-SPINS, CCW-STI. Table 7 shows the results obtained from the classification report of three context-based models and our baseline model CCW-SSI. Whereas, Table 8 visualizes their confusion matrices.

For the Albanian dataset, including adjacent sentences, especially with the CCW-STI model, improves recall and F1 score. The true positive rate also increases, indicating better identification of positive cases. For the Basque dataset, the CCW-STI model shows improvements in precision, recall, and F1 score. The true positive rate increases, and the false negative rate decreases, reflecting better model performance when surrounding context added. For Catalan dataset, the CCW-STI model significantly improves recall and F1 score. The true positive rate increases, and the false negative rate decreases, showing enhanced model performance. Overall, including adjacent sentences in the context-based models (especially CCW-STI) demonstrates clear improvements in recall and F1 scores across all languages, with better identification of positive cases reflected in the confusion matrices. The consistency across languages reinforces the conclusion that the surrounding context plays a crucial role in enhancing the model's ability to recognize the need for citations. This indicates that adjacent sentences are not just helpful, but perhaps essential for improving the performance of models designed for citation necessity detection in text.

Should we rely on the preceding, succeeding or both sentences as context? Based on the analysis of Tables 7 and 8 we conclude that preceding sentence through CCW-SPIPS model shows some improvements but is generally less effective than using the succeeding sentence

or both sentences; succeeding sentence through CCW-SPINS model provides significant improvements in precision, recall, and F1 scores across all languages, indicating a more substantial benefit from considering the succeeding context; both sentences through CCW-STI model consistently shows the highest improvements in recall and F1 scores across all languages. The inclusion of both preceding and succeeding sentences provides the best performance enhancements, suggesting that a comprehensive context is most beneficial.

In summary, our findings indicate that the incorporation of both adjacent sentences to the current sentence enhances the overall performance of the model. Specifically, including the succeeding sentence, which is likely to offer additional details or evidence that supports the statement in the current sentence, aids in more accurately determining if a citation is required as opposed to the previous sentence. The decision to use preceding, succeeding, or both sentences for context may depend on the characteristics of the language being modeled. However, our experiments consistently show that the triad approach is advantageous, with the pairing of the current sentence with the next as the second-best option.

Does the inclusion of both preceding and succeeding sentences, along with topic categories in the three-sentence model improve the prediction accuracy compared to the two-sentence models? The CCW-STI models in Table 7 show that contextual information enhances the performance of the base single-sentence models (CCW-SSI) specifically in the positive class across languages. The inclusion of both adjacent sentences, along with the current sentence and topic categories, provide a more comprehensive set of contextual information for the model to analyze. The information in Table 9 show that the inclusion of both preceding and succeeding sentences, along with topic categories in the three-sentence model (CCW-STIT), results in highest recall and matched highest F1 score compared to two-sentence models in the Albanian dataset; slight improvement in recall with similar F1 scores compared to the best two-sentence model (CCW-SPINS) in the Basque dataset; and finally the results show an improved recall and F1 scores to the best two-sentence model (CCW-SPINS) in the Catalan dataset.

Overall, CCW-STIT demonstrates the best performance in recall across all languages and maintains the highest F1 scores, indicating that using both preceding and succeeding sentences along with topic categories improves prediction accuracy over the two-sentence models.

6.3. Statistical analysis results

In this section, we present the outcomes of our statistical significance tests, emphasizing the t-statistics and their corresponding p-values. The heatmap shown in Fig. 6 illustrates the t-statistic scores,

Table 8
Confusion matrices of CCW context-based approaches showing the impact of adjacent sentences.

Language	CCW-SSI	CCW-SPIPS	CCW-SPINS	CCW-STI
sq				
eu				
ca				

Table 9

This table presents the precision (P), recall (R), and F1 scores for the positive class $y_{c,lang}^{jj}$, as well as weighted precision (W-P), weighted recall (W-R), and weighted F1 (W-F1) for the entire test dataset, comparing models using preceding sentence (CCW-SPINS), succeeding sentence (CCW-SPINS), both sentences (CCW-STI), and both sentences with topic categories (CCW-STIT) across Albanian (sq), Basque (eu), and Catalan (ca) datasets.

Language	Model	P	R	F1	W-P	W-R	W-F1
sq	CCW-SPINS	0.57	0.34	0.43	0.72	0.75	0.72
	CCW-SPINS	0.60	0.36	0.45	0.74	0.76	0.74
	CCW-STIT	0.58	0.40	0.47	0.73	0.75	0.74
eu	CCW-SPINS	0.61	0.30	0.40	0.79	0.82	0.79
	CCW-SPINS	0.65	0.32	0.43	0.80	0.82	0.80
	CCW-STIT	0.64	0.33	0.43	0.80	0.82	0.80
ca	CCW-SPINS	0.57	0.31	0.40	0.75	0.77	0.75
	CCW-SPINS	0.60	0.35	0.45	0.76	0.79	0.76
	CCW-STIT	0.60	0.36	0.45	0.76	0.79	0.77

where lighter shades represent more extreme negative values (e.g., -80), and darker shades represent less extreme negative values (e.g., -10). This color gradient facilitates the identification of the most substantial differences. We did not visualize p-values because most were extremely small, making visualization less informative. Instead, we reference the p-values in terms of a set threshold: tests are considered significant if the p -value is less than 0.05; otherwise, they are not significant.

When comparing our proposed CCW models to the REDI-FT models, we observe highly significant t-statistics, with the most extreme value being -82.16 , and p-values less than 0.05 across all comparisons.

In the comparison between our proposed models and the REDI-MB models, we also obtain significant t-statistics, with the highest being -21.84 , and p-values less than 0.05 in all cases.

However, when comparing our proposed models against the CCW-SSI model, we observe a mix of significant and non-significant results. The non-significant results are observed for CCW-SSIT (t-stat = 1.01, p -value = 0.33), CCW-SPIPS (t-stat = -1.86 , p -value = 0.09), and CCW-SPIPS (t-stat = -2.22 , p -value = 0.05). Conversely, the remaining models (CCW-SPINS, CCW-SPINST, CCW-STI, CCW-STIT) show statistically significant results with p-values less than 0.05.

The next section moves on to discuss additional insights we have gained regarding the effectiveness of utilizing different contextual information sources.

7. Analysis of results

Next, we further delve into the results of our experiments, particularly looking at additional dimensions and details. We look at how topic categories are distributed across citation sentences, followed by analysis of correctly classified instances thanks to context, and concluding with an error analysis looking at near-mistakes.

7.1. Topic categories that are more likely to contain citation

By incorporating topic categories into the methodology, we aimed to find out if sentences belonging in articles of particular categories are more prone to requiring citation than others. Our supplementary analysis in Table 10 shows the most likely and least likely topic categories to contain citation in the Albanian (sq), Basque (eu), and Catalan (ca) datasets. A commonality across all three languages is the presence of four topic categories that frequently require citations: articles on the topic of Europe, as well as those in the Biographical, Historical, and Science, Technology, Engineering, and Mathematics (STEM) categories. The Europe topic category is the most prevalent in terms of citations, which is understandable given that all three languages are spoken on the European continent. Biographical articles can require extensive research from a variety of sources including primary sources belonging to the main character of the article such as speeches, letters, or secondary biographies written by other authors. STEM topics typically build upon prior research and findings, citations of those sources are necessary. History relies heavily on evidence-based research, as well as date and time information; features that often have proven to be important factors that a sentence needs citation. Sports is similar to history in that it often relies on evidence-based research or reports to support claims.

Controversially, topics least likely to have citations include Performing Arts, Comics and Anime, Fashion, Radio, and Video Games. These topics may have fewer citations due to their dynamic and contemporary nature. These fields are often influenced by current trends and personal experiences rather than historical records or research studies that require formal citation. For example, the subjective interpretation as seen in Performing Arts, the evolving and speculative content in Comics and Anime, and the trend-driven nature of Fashion are less likely to require citations like academic papers or highly factual content demands. Similarly, Radio content is typically time-sensitive and focused on news or entertainment that does not always reference a source. Video

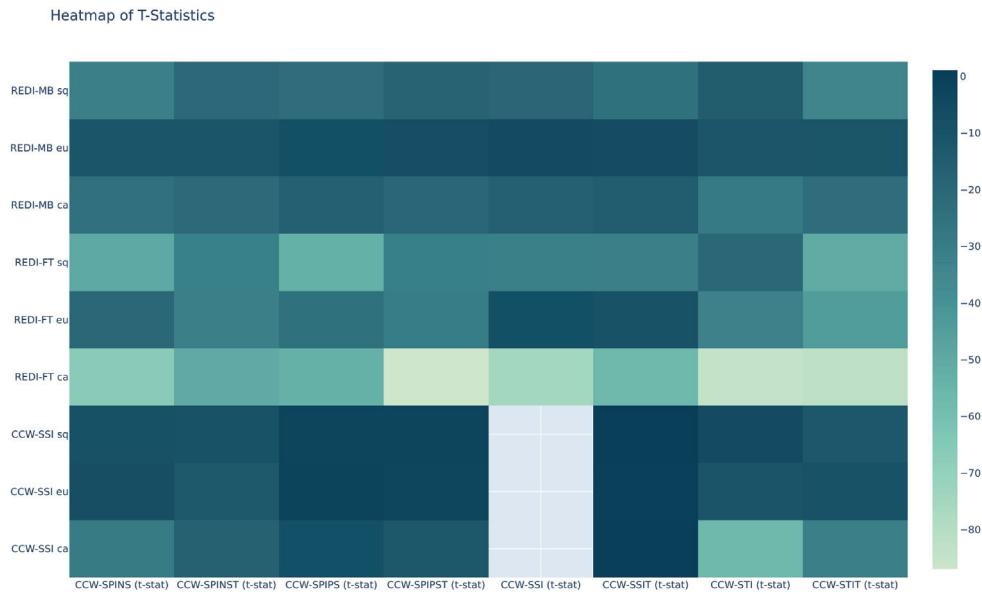


Fig. 6. Heatmap of t-statistics for weighted F1 scores comparing baseline models (REDI-FT, REDI-MB, CCW-SSI) with proposed variants across three datasets (sq, eu, ca). The t-statistics were computed using paired t-tests (ttest_rel) to assess the statistical significance of the F1 scores between the baseline models and their respective variants. All statistical significance tests are done with a 9 degree of freedom $df = 9$.

Table 10
Analysis of topic categories for citation presence in Albanian, Basque, and Catalan datasets.



Games, while a technical field, are often discussed in terms of player experience and industry trends, which do not always lend themselves to citation. These fields are characterized by rapid innovation and personal expression, which often rely on primary sources or creative works that are not conventionally cited in the same way as scientific or historical information.

7.2. Correctly classified examples when context added

In Table 11 we have given three constructive examples of the model input before adding contextual information and after adding contextual information. When the model was fed with only one sentence the prediction was wrong, whereas when adding context the prediction improved and the data point was correctly classified.

In the first example the single model input, “On 23 November 2013, Lampard scored his 206th and 207th Chelsea goals against his former club West Ham in a 3–0 win”. was predicted wrongly as not needing a citation because the model may not have had enough contextual information to recognize the significance of the achievement. The

model may have only focused on the factual information provided in the sentence, such as the date, the score, and the teams involved, without understanding the significance of Lampard’s goals and their impact on his career or the history of the club. In contrast, the second input provides additional contextual information, such as Lampard becoming the fourth highest goalscorer in Premier League history, which helps the model better understand the importance of the event being described. This additional information provides a clearer picture of the significance of Lampard’s achievement, allowing the model to correctly classify the sentence as requiring a citation.

7.3. Error radius analysis

To further inspect the performance of our model, we look at an alternative evaluation method by looking at near-misses. Our motivation for this approach is the complexity of the task, which often involves uncertainty in deciding whether a sentence requires a citation or not. In some cases, the span of a citation placed in one sentence may extend to the surrounding sentences, making it difficult to determine the exact

Table 11

Correctly classified examples when context added (Note: For clarity, the following example is a translated excerpt from Albanian to English using Google Translate.).

Misclassified before adding context		Correctly classified after adding context	
sq	en	sq	en
Më 23 nëntor 2013, Lampard shënoi golin e 206të dhe 207të me Çelsin kundër ish klubit të tij West Ham në një fitore 3-0.	On 23 November 2013, Lampard scored his 206th and 207th Chelsea goals against his former club West Ham in a 3-0 win.	Më 23 nëntor 2013, Lampard shënoi golin e 206të dhe 207të me Çelsin kundër ish klubit të tij West Ham në një fitore 3-0. Ai u bë golashënuesi i katërt më i mirë në historinë e Premier Ligës më 2 dhjetor duke kaluar Robbie Fowler i cili kishte shënuar 164 gola.	On 23 November 2013, Lampard scored his 206th and 207th goals for Chelsea against his former club West Ham in a 3-0 win. He became the fourth highest goalscorer in Premier League history on 2 December, overtaking Robbie Fowler who had scored 164 goals.
Infeksionet akute kanë sëmundshmëri dhe tropizëm të lartë për gjëndrat e pështymës, lachrymal dhe harderian.	Acute infections have high morbidity and tropism for the salivary, lachrymal and harderian glands.	Infeksionet akute kanë sëmundshmëri dhe tropizëm të lartë për gjëndrat e pështymës, lachrymal dhe harderian. Një koronavirus i shklopit të lidhur me HKU2 i quajtur koronavirus i sindromës diarre akute të derrit (SADS-CoV) shkakton diarre te derrat.	Acute infections have high morbidity and tropism for the salivary, lachrymal and harderian glands. An HKU2-related stick coronavirus called porcine acute diarrhoea syndrome coronavirus (SADS-CoV) causes diarrhoea in pigs.
Ajo priti festivalin për të dytin vit radhazi në vitin 2019, me Miley Cyrus që ishte pjesë e formacionit të interpretuesve.	She hosted the festival for the second consecutive year in 2019, with Miley Cyrus as part of the line-up of performers.	Ish-kryetari i Komunës së Prishtinës, Shpend Ahmeti i dha asaj Çelësin e Prishtinës, hera e parë që ishte dhënë. Ajo priti festivalin për të dytin vit radhazi në vitin 2019, me Miley Cyrus që ishte pjesë e formacionit të interpretuesve. Lipa ka 15 tatuazhe, duke përfshirë një kushtuar Sunny Hill. <i>Biography* Films Media* Music North America Women</i>	The former mayor of the Municipality of Pristina, Shpend Ahmeti, gave her the Key of Pristina, the first time it had been given. She hosted the festival for the second consecutive year in 2019, with Miley Cyrus as part of the lineup of performers. Lipa has 15 tattoos, including one dedicated to Sunny Hill. <i>Biography* Films Media* Music North America Women</i>

placement. That is, when two sentences discuss the same event or fact, one could arguably place the citation in either of those sentences. To approximate this, we measure the proximity between the predicted label and the true label, measured as the number of sentences. It is a measure of how far off the model’s prediction is from the actual citation.

Table 12 shows the error radius for SQ-citation-needed, EU-citation-needed, and CA-citation-needed datasets, evaluated in two different scenarios based on the legitimate span of error, which may extend up to two sentences before or after the target sentence. For this analysis, used the CCW-STIT model which incorporates all contextual information (the previous sentence, the current sentence, the next sentence, and the topic categories of the current sentence).

The row, labeled “Radius Zero”, reflects the model’s citation prediction performance without error radius. We are showing it for ease of comparison with other radius.

The row, labeled “Radius One”, reflects the model’s citation prediction performance with an error radius of one. This implies that the true citation is located either one sentence before or after the predicted citation for the languages sq, eu, and ca.

The row, labeled “Radius Two”, presents the model’s performance when the error radius is expanded to encompass up to two sentences either before or after the predicted citation. This can be analogous to predicting the need for citation within a paragraph that contains up to five sentences. In this case, the prediction of the model significantly improves compared to predicting only for one sentence.

Our error radius analysis provides several key insights into the model’s performance. There is a clear trend across languages showing that as the error radius increases, the model becomes better at predicting citations, indicated by the improved true positives rate. This suggests that citations tend to be relevant over a span of multiple sentences rather than a single one.

A false negative occurs when the model does not predict the need for citation where there actually should be one. When increasing the error radius, the number of false negatives decreases. So, where it might have previously overlooked a sentence that needed a citation, with the increased radius, it now catches it and makes a correct prediction more often. This can reduce the effort that Wikipedia editors have to put in. They can focus their attention on the flagged paragraphs, knowing that these highlighted sections are where they should check

for citations. This is much more efficient than going through every sentence individually, especially in articles that may be very long.

The false positive rate just slightly decreases when considering the two-sentence error span, suggesting that the model does not increase incorrect citation predictions even as the radius expands.

With the expanded error radius, the true negative rate shows a slight improvement. In the context of our research, the negative class forms the majority. Even without an error span, the model already demonstrated strong performance in predicting true negatives. Nonetheless, the sentence span error analysis reveals that incorporating paragraph citation prediction further enhances the true negative prediction rate.

In summary, by introducing and then increasing the error radius, the model shows substantial gains in its ability to predict paragraphs where citations are needed. The error radius approach reflects the real-world use of citations given that the citation span can often extend beyond the current sentence, and the results demonstrate that considering this wider context is beneficial for the model’s performance.

8. Discussion

Theoretical and practical implications of the research. Where the limited body of previous research in citation worthiness detection in Wikipedia had only focused in the English Wikipedia, in this work we have studied the more challenging scenario of tackling small Wikipedias. Smaller Wikipedias, such as those in Albanian, Basque, and Catalan, present the challenge of being curated by a smaller community of editors and therefore one needs to be more careful when relying on the existing citations to build a dataset. To address this issue, we have devised a methodology that relies on article quality scores to choose the articles making up the top quintile of sentences in the Albanian, Basque, and Catalan Wikipedias, hence sampling high-quality content. By following this methodology, we have collected and labeled the SQ-citation-needed, EU-citation-needed, and CA-citation-needed datasets. This is particularly crucial for languages considered low-resource in the digital media, which often lack the extensive corpus of reviewed content available to their larger counterparts. Our research not only fills a gap in the existing literature by focusing on these underrepresented languages but also sets a precedent for future studies in citation detection by demonstrating consistent, replicable results across multiple languages.

Table 12
Citation prediction error radius with zero, one, and two sentence span for Albanian (sq), Basque (eu), and Catalan (ca) Datasets. The model used is CCW-STIT.

	sq	eu	ca																											
Radius Zero	<table border="1"> <tr> <td>True label No Citation</td> <td>0.89</td> <td>0.11</td> </tr> <tr> <td>True label Citation</td> <td>0.6</td> <td>0.4</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.89	0.11	True label Citation	0.6	0.4		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.95</td> <td>0.05</td> </tr> <tr> <td>True label Citation</td> <td>0.71</td> <td>0.29</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.95	0.05	True label Citation	0.71	0.29		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.92</td> <td>0.08</td> </tr> <tr> <td>True label Citation</td> <td>0.64</td> <td>0.36</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.92	0.08	True label Citation	0.64	0.36		No Citation Predicted label	Citation Predicted label
True label No Citation	0.89	0.11																												
True label Citation	0.6	0.4																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.95	0.05																												
True label Citation	0.71	0.29																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.92	0.08																												
True label Citation	0.64	0.36																												
	No Citation Predicted label	Citation Predicted label																												
Radius One	<table border="1"> <tr> <td>True label No Citation</td> <td>0.95</td> <td>0.05</td> </tr> <tr> <td>True label Citation</td> <td>0.19</td> <td>0.81</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.95	0.05	True label Citation	0.19	0.81		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.98</td> <td>0.02</td> </tr> <tr> <td>True label Citation</td> <td>0.17</td> <td>0.83</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.98	0.02	True label Citation	0.17	0.83		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.96</td> <td>0.04</td> </tr> <tr> <td>True label Citation</td> <td>0.18</td> <td>0.82</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.96	0.04	True label Citation	0.18	0.82		No Citation Predicted label	Citation Predicted label
True label No Citation	0.95	0.05																												
True label Citation	0.19	0.81																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.98	0.02																												
True label Citation	0.17	0.83																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.96	0.04																												
True label Citation	0.18	0.82																												
	No Citation Predicted label	Citation Predicted label																												
Radius Two	<table border="1"> <tr> <td>True label No Citation</td> <td>0.96</td> <td>0.04</td> </tr> <tr> <td>True label Citation</td> <td>0.08</td> <td>0.92</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.96	0.04	True label Citation	0.08	0.92		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.98</td> <td>0.02</td> </tr> <tr> <td>True label Citation</td> <td>0.05</td> <td>0.95</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.98	0.02	True label Citation	0.05	0.95		No Citation Predicted label	Citation Predicted label	<table border="1"> <tr> <td>True label No Citation</td> <td>0.97</td> <td>0.03</td> </tr> <tr> <td>True label Citation</td> <td>0.06</td> <td>0.94</td> </tr> <tr> <td></td> <td>No Citation Predicted label</td> <td>Citation Predicted label</td> </tr> </table>	True label No Citation	0.97	0.03	True label Citation	0.06	0.94		No Citation Predicted label	Citation Predicted label
True label No Citation	0.96	0.04																												
True label Citation	0.08	0.92																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.98	0.02																												
True label Citation	0.05	0.95																												
	No Citation Predicted label	Citation Predicted label																												
True label No Citation	0.97	0.03																												
True label Citation	0.06	0.94																												
	No Citation Predicted label	Citation Predicted label																												

First, we gathered and pre-processed three raw dataset, incorporating articles' quality scores. Prior studies (Redi et al., 2019; Wright and Augenstein, 2020) relied on featured articles that are marked as such by English Wikipedia community through manual efforts, a method challenging to replicate for languages with fewer resources due to smaller, less active communities. To address this limitation, we utilized a language-agnostic quality framework,¹⁹ generating quality scores based on page length, references, sections, wikilinks, categories, and media. These scores were crucial for selecting top-quality articles. Compared to the previous method that relied on featured articles, which are limited in smaller Wikipedias (e.g., only 33 featured articles²⁰ in Albanian Wikipedia), our approach offers a more versatile method for identifying quality articles in small Wikipedias. This is evidenced by its successful application to the Albanian, Basque and Catalan editions in our research. The relevance of this methodology is particularly important knowing that Albanian, Basque, and Catalan come from distinct linguistic origins and families, which is why they do not share many common linguistic features. Despite these differences, the automated approach we have proposed for detecting citation needs has demonstrated effectiveness across all three languages.

Second, using our newly created datasets SQ-citation-needed, EU-citation-needed, and CA-citation-needed we experimented with our proposed CCW model which leverages contextualized representations of sentences for citation worthiness prediction. We have studied two ways of contextualizing sentences to incorporate adjacent sentences and topic categories. While previous research had balanced the two classes in the dataset (citation and no citation) for the experimentation, in this work we are the first to tackle the more realistic and challenging setting of keeping the original class imbalance. Through our experiments, we observe that surrounding sentences can indeed lead to substantial improvements, whereas the improvement achieved through the use of topic categories is more modest with regard to the

¹⁹ https://meta.wikimedia.org/wiki/Research:Prioritization_of_Wikipedia_Articles/Language-Agnostic_Quality#Basic_Approach.

²⁰ https://en.wikipedia.org/wiki/Wikipedia:Featured_articles_in_other_languages/Albanian.

overall model performance, but significant with regard to improving the prediction of positive class. While our methodology has been tested and validated using data collected from the Albanian, Basque, and Catalan Wikipedia, we believe that the methodology and model have the potential to be extended to other languages with Wikipedia projects of similar size. Further studies are needed to confirm its effectiveness in these additional contexts.

Our findings indicate the feasibility of building automated models that detect the need for citation in a given sentence, even in Wikipedias with limited resources. Our approach is distinct from previous methods that relied on both automated and manual efforts. We have developed a data collection and labeling method, along with a transformers-based model, that operate entirely autonomously, eliminating the need of any input from human labor.

Implications for researchers. First, this study enriches and expands the theory and methodology of the scientific citation needed tasks in smaller Wikipedias by incorporating contextual features in the prediction model. To the best of our knowledge, we are the first to tackle the task of citation needed for small Wikipedias through the use of adjacent sentences and topic categories as contextual features. The outcome of our research encourages future research to extend cite-worthy models with the rich surrounding information contained in Wikipedia articles.

Second, our research emphasize the potential of machine learning and NLP techniques in enhancing the quality of information available on digital platforms like Wikipedia. By developing a model that can identify areas in an article that require citations, we are not only improving the reliability of the information but also the overall user experience. This has broader implications for the development of automated fact-checking and information verification tools, which are increasingly important in the digital age where misinformation can spread rapidly.

Third, we recognize the importance of supporting low-resource and understudied languages like Albanian, Basque or Catalan which currently have limited research in the field of NLP. We have contributed to these languages and the NLP field with three new datasets that can be used for citation needed task or related task like claim detection

for fact-checking. The method used to put together this dataset can be utilized to create other datasets, especially in low-resource languages.

By investing low-resource languages, we aim to improve accessibility and inclusivity in the field of NLP. In addition, we believe it is crucial to provide people who speak these languages with reliable tools that support responsible and credible information dissemination.

Limitations and future work. While our work advances research in citation worthiness detection for small Wikipedias, this is not without limitations. Our data labeling strategy is supported by a consolidated method to ensure that the articles we consider are of high quality. Through empirical experimentation, we have also demonstrated the validity of this approach to ensure a better quality dataset compared to the use of the entire Wikipedia without filtering. However, given the automated approach to the large-scale labeling, we cannot ensure that all the labels are correct and therefore the labeling strategy is inevitably bound to some inaccurate labels. While we validated our approach and CCW model in the Albanian, Basque, and Catalan Wikipedia, we can ascertain the extensibility of the method to other languages without needing significant changes.

Another limitation of our study is the scalability of automated annotations for large datasets. While our methodology has been tested on datasets of varying sizes from low-resource languages (e.g., 1675 articles for Albanian and 22,143 articles for Catalan), further empirical validation is needed for larger datasets. Potential strategies to mitigate annotation errors at scale include incorporating human-in-the-loop approaches. Our focus on low-resource and small Wikipedia projects means that large-scale annotation was beyond the scope of this study. Future research could explore the applicability and effectiveness of our methodology on larger datasets to address this scalability issue.

Our dataset construction method is specifically designed for Wikipedia. This methodology is tailored to utilize the unique structure and metadata available in Wikipedia articles, which may not be present in other types of text corpora. As a result, the direct applicability of our approach to other text datasets is limited. Future research could investigate how our methodology can be adapted to different text corpora to improve its generalizability beyond Wikipedia.

CRedit authorship contribution statement

Aida Halitaj: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Arkaitz Zubiaga:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.²¹

References

- Abumansour, A.S., Zubiaga, A., 2023. Check-worthy claim detection across topics for automated fact-checking. *PeerJ Comput. Sci.* 9, e1365.
- Ando, K., Sekine, S., Komachi, M., 2024. WikiSQE: A large-scale dataset for sentence quality estimation in wikipedia. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 17656–17663.
- Arslan, F., Hassan, N., Li, C., Tremayne, M., 2020. A benchmark dataset of check-worthy factual claims. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14.
- Asthana, S., Tobar Thommel, S., Halfaker, A.L., Banovic, N., 2021. Automatically labeling low quality content on wikipedia by leveraging patterns in editing behaviors. *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW2), 1–23.
- Bai, Y., Colas, A., Wang, D.Z., 2023. MythQA: Query-based large-scale check-worthy claim detection through multi-answer open-domain question answering. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 3017–3026.
- Baigutanova, A., Myung, J., Saez-Trumper, D., Chou, A.-J., Redi, M., Jung, C., Cha, M., 2023. Longitudinal assessment of reference quality on wikipedia. In: *Proceedings of the ACM Web Conference 2023*. pp. 2831–2839.
- Blumenstock, J.E., 2008. Size matters: word count as a measure of quality on wikipedia. In: *Proceedings of the 17th International Conference on World Wide Web*. pp. 1095–1096.
- Bonab, H., Zamani, H., Learned-Miller, E., Allan, J., 2018. Citation worthiness of sentences in scientific reports. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. pp. 1061–1064.
- Chen, C.-C., Roth, C., 2012. {{Citation needed}} the dynamics of referencing in wikipedia. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. pp. 1–4.
- Chou, A.-J., Goncalves, G., Walton, S., Redi, M., 2020. Citation detective: a public dataset to improve and quantify wikipedia citation quality at scale. In: *Proceedings of the Wiki Workshop*.
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W., 2016. The spreading of misinformation online. *Proc. Natl. Acad. Sci.* 113 (3), 554–559.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 4171–4186.
- Dordevic, M., Safieddine, F., Masri, W., Pourghomi, P., 2016. Combating misinformation online: identification of variables and proof-of-concept study. In: *Social Media: The Good, the Bad, and the Ugly: 15th IFIP WG 6.11 Conference on E-Business, E-Services, and E-Society, I3E 2016, Swansea, UK, September 13–15, 2016, Proceedings 15*. Springer, pp. 442–454.
- Fetahu, B., Markert, K., Anand, A., 2017. Fine grained citation span for references in wikipedia. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 1990–1999.
- Gilbert, D.T., 1991. How mental systems believe.. *Am. Psychol.* 46 (2), 107.
- Gosangi, R., Arora, R., Gheisarieha, M., Mahata, D., Zhang, H., 2021. On the use of context for predicting citation worthiness of sentences in scholarly articles. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 4539–4545.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T., 2018. Learning word vectors for 157 languages. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Hara, N., Shachaf, P., Hew, K.F., 2010. Cross-cultural analysis of the wikipedia community. *J. Am. Soc. Inf. Sci. Technol.* 61 (10), 2097–2108.
- Hassan, N., Arslan, F., Li, C., Tremayne, M., 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*. pp. 1803–1812.
- Hassan, N., Li, C., Tremayne, M., 2015. Detecting check-worthy factual claims in presidential debates. In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. pp. 1835–1838.
- Hsu, C., Li, C.-T., Saez-Trumper, D., Hsu, Y.-Z., 2021. WikiContradiction: Detecting self-contradiction articles on wikipedia. In: *2021 IEEE International Conference on Big Data, Big Data, IEEE*, pp. 427–436.
- Hu, Z., Cui, J., Lin, A., 2023. Identifying potentially excellent publications using a citation-based machine learning approach. *Inf. Process. Manage.* 60 (3), 103323.
- Jaradat, I., Gencheva, P., Barrón-Cedeño, A., Márquez, L., Nakov, P., 2018. ClaimRank: Detecting check-worthy claims in arabic and english. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. pp. 26–30.
- Johnson, I., Gerlach, M., Sáez-Trumper, D., 2021. Language-agnostic topic classification for wikipedia. In: *Companion Proceedings of the Web Conference 2021*. pp. 594–601.
- Kaffee, L.-A., Elshahr, H., 2021. References in wikipedia: The editors' perspective. In: *Companion Proceedings of the Web Conference 2021*. pp. 535–538.

²¹ <http://doi.org/10.5281/zenodo.438045>.

- Khatri, M., Wadhwa, P., Satija, G., Sheik, R., Kumar, Y., Shah, R.R., Kumaraguru, P., 2023. CiteCaseLAW: Citation worthiness detection in caselaw for legal assistive writing. arXiv preprint arXiv:2305.03508.
- Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A., 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digit. Threats: Res. Pract.* 2 (2), 1–16.
- Korfiatis, N.T., Poulos, M., Bokos, G., 2006. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Inf. Rev.*
- Laniado, D., Tasso, R., 2011. Co-authorship 2.0: Patterns of collaboration in wikipedia. In: *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*. pp. 201–210.
- Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J., 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest* 13 (3), 106–131.
- Li, M.-H., Chen, Z., Rao, L.-L., 2022. Emotion, analytic thinking and susceptibility to misinformation during the COVID-19 outbreak. *Comput. Hum. Behav.* 133, 107295.
- Logan, D.W., Sandal, M., Gardner, P.P., Manske, M., Bateman, A., 2010. Ten simple rules for editing wikipedia.
- Lutzke, L., Drummond, C., Slovic, P., Árvai, J., 2019. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. *Glob. Environ. Change* 58, 101964.
- McGrew, S., Ortega, T., Breakstone, J., Wineburg, S., 2017. The challenge that's bigger than fake news: Teaching students to engage in civic online reasoning. *Am. Educ.* 41 (3), 4.
- McMahon, C., Johnson, I., Hecht, B., 2017. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. In: *Proceedings of the International AAAI Conference on Web and Social Media*. pp. 142–151.
- Nakov, P., Barrón-Cedeño, A., da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouni, W., et al., 2022. Overview of the clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*. Springer, pp. 495–520.
- Newman, D., Lewandowsky, S., Mayo, R., 2022. Believing in nothing and believing in everything: The underlying cognitive paradox of anti-COVID-19 vaccine attitudes. *Pers. Individ. Differ.* 189, 111522.
- Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J., Liu, S., 2022. Fake news on social media: the impact on society. *Inf. Syst. Front.* 1–16.
- Panchendrarajan, R., Zubiaga, A., 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Nat. Lang. Process. J.* 7, 100066.
- Przybyla, P., Borkowski, P., Kaczyński, K., 2022. Countering disinformation by finding reliable sources: a citation-based approach. In: *2022 International Joint Conference on Neural Networks. IJCNN, IEEE*, pp. 1–8.
- Redi, M., Fetahu, B., Morgan, J., Taraborelli, D., 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In: *The World Wide Web Conference*. pp. 1567–1578.
- Roostae, M., 2022. Citation worthiness identification for fine-grained citation recommendation systems. *Iran. J. Sci. Technol. Trans. Electr. Eng.* 46 (2), 353–365.
- Saez-Trumper, D., 2019. Online disinformation and the role of wikipedia. arXiv preprint arXiv:1910.12596.
- Schmidt, M., Kircheis, W., Simons, A., Potthast, M., Stein, B., 2023. A diachronic perspective on citation latency in wikipedia articles on CRISPR/Cas-9: an exploratory case study. *Scientometrics* 1–25.
- Sheikhi, G., Touileb, S., Khan, S.A., 2023. Automated claim detection for fact-checking: A case study using norwegian pre-trained language models. In: *The 24rd Nordic Conference on Computational Linguistics*.
- Shin, J., Jian, L., Driscoll, K., Bar, F., 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Comput. Hum. Behav.* 83, 278–287.
- Singh, H., West, R., Colavizza, G., 2021. Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quant. Sci. Stud.* 2 (1), 1–19.
- Sugiyama, K., Kumar, T., Kan, M.-Y., Tripathi, R.C., 2010. Identifying citing sentences in research papers using supervised learning. In: *2010 International Conference on Information Retrieval & Knowledge Management. CAMP, IEEE*, pp. 67–72.
- Thorne, J., Vlachos, A., 2018. Automated fact checking: Task formulations, methods and future directions. In: *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 3346–3359.
- Viégas, F.B., Wattenberg, M., Dave, K., 2004. Studying cooperation and conflict between authors with history flow visualizations. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 575–582.
- Wang, Y., McKee, M., Torbica, A., Stuckler, D., 2019. Systematic literature review on the spread of health-related misinformation on social media. *Soc. Sci. Med.* 240, 112552.
- Wright, D., Augenstein, I., 2020. Claim check-worthiness detection as positive unlabelled learning. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. pp. 476–488.
- Wright, D., Augenstein, I., 2021. CiteWorth: Cite-worthiness detection for improved scientific document understanding. In: *Findings of the Association for Computational Linguistics: ACL-LJCNLP 2021*. pp. 1796–1807.
- Xu, Y., Zhou, D., Wang, W., 2023. Being my own gatekeeper, how I tell the fake and the real-fake news perception between typologies and sources. *Inf. Process. Manage.* 60 (2), 103228.
- Zeng, X., Abumansour, A.S., Zubiaga, A., 2021. Automated fact-checking: A survey. *Lang. Linguist. Compass* 15 (10), e12438.
- Zeng, T., Acuna, D.E., 2020. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. *Scientometrics* 124 (1), 399–428.