# Countering Disinformation by Finding Reliable Sources: a Citation-Based Approach

Piotr Przybyła
*Institute of Computer Science,*
*Polish Academy of Sciences*
Warsaw, Poland
piotr.przybyla@ipipan.waw.pl

Piotr Borkowski
*Institute of Computer Science,*
*Polish Academy of Sciences*
Warsaw, Poland
piotr.borkowski@ipipan.waw.pl

Konrad Kaczyński
*Institute of Computer Science,*
*Polish Academy of Sciences*
and *University of Warsaw*
Warsaw, Poland
konrad.kaczynski@ipipan.waw.pl

***Abstract*—We propose a new task aimed at countering dis- and misinformation, called Finding Reliable Sources. Given a one-sentence claim, the challenge is to automatically find a knowledge source (e.g. a book, a research article, a web page) that could support or refute the claim. We show that this capability could be learnt by observing associations between sentences in English Wikipedia and citations provided for them. Thus, we collect a corpus of over 50 million references to 24 million identified sources with the citation context from Wikipedia, and build search indices using several meaning representation methods. For evaluation, apart from the Wikipedia corpus, we prepare another test set based on the FEVER fact-checking dataset.**

***Index Terms*—finding reliable sources, misinformation, fact-checking, fake news**

## I. Introduction

One of the solutions most commonly applied to the challenge of misinformation is fact-checking: a careful inspection of a given text, e.g. a fake news piece, identification of the core claims and analysis of their veracity using reliable sources. The method is widely used by fact-checking organisations (e.g. *FactCheck.org*, *PolitiFact* or *Snopes*) but the amount of human effort involved makes it hard to keep up with all the erroneous claims appearing in the media [1].

Automating this effort might appear as a promising direction, but it remains a challenging problem. One of the difficulties lies in the low availability of machine-readable reference sources. In the attempts made so far, e.g. FEVER shared tasks [2], a corpus of Wikipedia articles often plays this role. But the online encyclopaedia has its disadvantages: its content may not include all the information pertinent to a developing news story; can be modified anonymously; and is not commonly accepted as a perfectly reliable source, e.g. in a scholarly context.

Nevertheless, Wikipedia can be useful for information verification, as it is considered *a road map to reliable sources* by professional journalists [3]. This means consulting the primary sources cited there (e.g. textbooks or research articles) rather than the textual content of Wikipedia.

Through this work, we show that the provision of relevant sources can be realised by an automatic solution based on natural language processing (NLP) and machine learning (ML) and we propose a new task, namely *Finding Reliable Sources* (FRS). The input for FRS is a single sentence expressing a certain claim or a fact and the expected output is a ranked list of identifiers of external sources that could be used to verify it. An example with a claim and recommended sources (returned by the solution described in this study) is shown in Table I.

TABLE I
Example of data involved in the FRS task: a claim and a list of reliable sources that could be used to verify it.

| **Input: Claim** |
|---|
| *Smoking tobacco is good for your health.* |

| **Output: Sources** |
|---|
| DOI:10.3390/ijerph110606459<br>The Case in Favor of E-Cigarettes for Tobacco Harm Reduction |
| http://www.cancer.gov/cancertopics/.../light-cigarettes<br>"Light" Cigarettes and Cancer Risk |
| http://www.drkoop.com/ency/93/002032.html<br>Smoking and Smokeless Tobacco – DrKoop |
| DOI:10.1093/jnci/djh144<br>Association Between Exclusive Pipe Smoking and Mortality From Cancer and Other Diseases |
| DOI:10.1523/JNEUROSCI.2139-05.2005<br>Monoamine Oxidase Inhibition Dramatically Increases the Motivation to Self-Administer Nicotine in Rats |
| . . . |

Note that the FRS task is closely related to the research on *local citation recommendation*, which seeks to facilitate the process of writing a scholarly article by recommending references from the literature that fit a specific sentence [4]. However, there are important differences, making the current task more difficult:

- Challenge A: the FRS input lacks the wider context of a scholarly article and consists of a single sentence, e.g. extracted from a social network post.
- Challenge B: FRS involves a great variety of sources, both in terms of domain and genre, while the citation recommendation solutions are evaluated using corpora of academic papers, often from a single domain (e.g. computer science).
- Challenge C: a reliable source is relevant to a claim even

This work was supported by the *Polish National Agency for Academic Exchange* through a *Polish Returns* grant number PPN/PPO/2018/1/00006.

when contradicting it, while scholarly citations support the sentences they accompany.

Despite these difficulties, here we show that it is possible to build a solution for the problem by applying the literature recommendation methods to a non-scholarly training data based on Wikipedia. We provide the following contribution:

- a corpus based on English Wikipedia, containing over 50 million citations to 24 million sources with the associated context (larger than any existing local citation corpora),
- a collection of baseline solutions for the FRS task, relying on a search index to find sources linked to sentences that are similar to the query in terms of vector-based meaning representation,
- two FRS evaluation datasets, one using the Wikipedia corpus and FEVER-FRS created based on the FEVER fact-checking shared task data,
- experiments illustrating the impact of the three challenges, differentiating our task from scholarly citation recommendation.

We hope the presented study will encourage more work into countering disinformation by promoting reliable information, rather than only detecting unreliable content. The resources created within this work are available for download[1][2][3].

## II. Related work

### A. Fact-checking

The task of automatic fact-checking denotes the assessment of the truthfulness of claims made in text [5]. While fully automated solutions with human-level performance remain a distant goal [6], some progress has been made in the subtasks leading in this direction, such as extracting claims worthy of checking [7], [8], which can be an element of interpretable fake news detection [9]. Assessing the veracity of a given statement has been approached in many ways, such as analysing the topology of the knowledge graph based on Wikipedia [10], searching for linguistic characteristics of untrustworthy text [11] or leveraging the comments made by social media users reacting to the claim [12]. Many solutions have been developed in the framework of the FEVER shared task [2], where the candidate system needs to identify suitable evidence from Wikipedia at the sentence level and assign a label describing the relation between the claim and the evidence ('supports', 'refutes', 'not enough info').

### B. Wikipedia: citations and credibility

The basic question about Wikipedia in the context of fact-checking and FRS is its reliability as a knowledge source. While initially it was perceived as a low-credibility source, this has changed over time and even professional journalists started to use it in their work [13]. The style and quality vary between Wikipedia articles: in-depth comparisons to printed sources in medical domains have shown favourable results for some topics [14] and much worse on others [15].

Here we are interested in the quality of references. It was unsatisfactory a decade ago [16], but later work has shown that citation rates in Wikipedia are often aligned with those in scholarly literature [17], [18] and the improvements continue to be made presently [19]. While we are only dealing with English, it is important to note that links between languages versions of articles also facilitate the spread of reliable citations [20].

The credibility of Wikipedia citations is likely to continue improving due to work in this field: development of algorithms for determining if a statement requires a citation [21] and analysis of history of the references [19]. Some of these efforts are performed within the *wikicite* initiative [22], aiming to study the references appearing on Wikipedia and create a bibliographic database based on various Wiki projects.

### C. Citation recommendation

The importance of citations in scholarly work analysis has been appreciated for a long time [23]. Developments in NLP and ML have enabled solutions that recommend the most appropriate citation for a given context in a manuscript. However, the FRS task is not compatible with every scenario considered in this domain.

Many researchers treat the problem as finding connections between the citing document (manuscript being written) and the cited ones (sources). Those associations can be explored through document clustering [24], SVD [25] or taking into account the manuscript structure [26]. But, in FRS no 'citing document' is available and we need to find a matching source based on a single sentence or claim. The approach, in which a recommendation is based on the immediate context, is known as *local* citation recommendation [4]. But, even in this area, most solutions are not applicable to FRS, since they rely on the contents of the cited documents, e.g. titles and abstracts [27]–[30]. In the FRS task, a source might be a web page or a book, and there is no database of abstracts for these. Nevertheless, recommending a source solely based on previous citation contexts is an active research direction. For example, Saier and Färber [31] check, which method of representing citation context yields the best recommendations. These methods are also evaluated within the FRS task (see Section IV-B).

When it comes to recommending citations for Wikipedia articles, the research is less abundant. It is worth mentioning the work of Fetahu et al. [32], who have constructed a classifier to detect whether statements need a news citation and investigated automatic news citation discovery problem (proposed by Peng et al. [33] as well); and of Jana et al. [34], who have developed a system for expanding references section with sources from related Wikipedia articles.

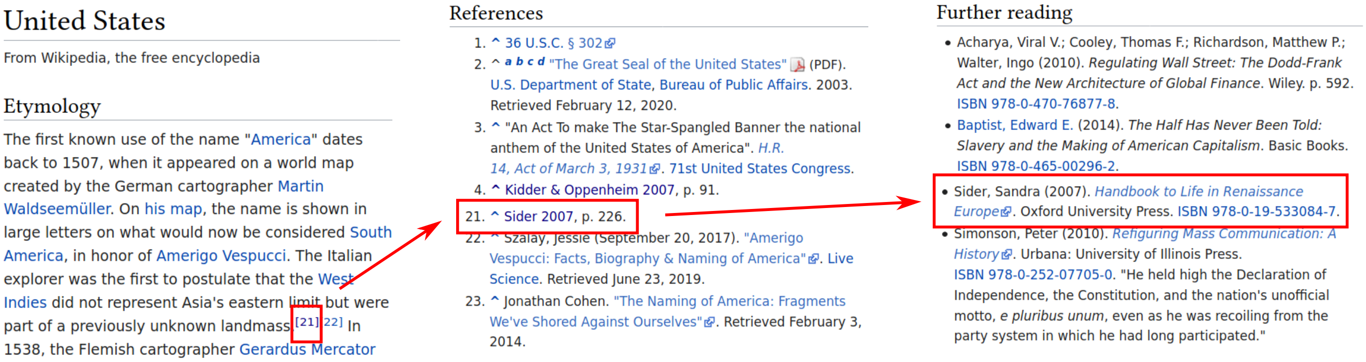
## III. Citation resources

In order to prepare resources for local citation recommendation beyond scholarly domains, we are converting the English Wikipedia into a training corpus. The process described here

[1] WCCC corpus: https://doi.org/10.5281/zenodo.6539054

[2] Evaluation datasets: https://doi.org/10.5281/zenodo.6539087

[3] Source code: https://github.com/piotrmp/finding_reliable_sources



United States

From Wikipedia, the free encyclopedia

Etymology

The first known use of the name "America" dates back to 1507, when it appeared on a world map created by the German cartographer Martin Waldseemüller. On his map, the name is shown in large letters on what would now be considered South America, in honor of Amerigo Vespucci. The Italian explorer was the first to postulate that the West Indies did not represent Asia's eastern limit but were part of a previously unknown landmass[21][22] In 1538, the Flemish cartographer Gerardus Mercator

References

1. ^ 36 U.S.C. § 302
2. ^ a b c d "The Great Seal of the United States" (PDF). U.S. Department of State, Bureau of Public Affairs. 2003. Retrieved February 12, 2020.
3. ^ "An Act To make The Star-Spangled Banner the national anthem of the United States of America". *H.R. 14, Act of March 3, 1931*. 71st United States Congress.
4. ^ Kidder & Oppenheim 2007, p. 91.

21. ^ Sider 2007, p. 226.
22. ^ Szalay, Jessie (September 20, 2017). "Amerigo Vespucci: Facts, Biography & Naming of America". Live Science. Retrieved June 23, 2019.
23. ^ Jonathan Cohen. "The Naming of America: Fragments We've Shored Against Ourselves". Retrieved February 3, 2014.

Further reading

- Acharya, Viral V.; Cooley, Thomas F.; Richardson, Matthew P.; Walter, Ingo (2010). *Regulating Wall Street: The Dodd-Frank Act and the New Architecture of Global Finance*. Wiley. p. 592. ISBN 978-0-470-76877-8.
- Baptist, Edward E. (2014). *The Half Has Never Been Told: Slavery and the Making of American Capitalism*. Basic Books. ISBN 978-0-465-00296-2.
- Sider, Sandra (2007). *Handbook to Life in Renaissance Europe*. Oxford University Press. ISBN 978-0-19-533084-7.
- Simonson, Peter (2010). *Refiguring Mass Communication: A History*. Urbana: University of Illinois Press. ISBN 978-0-252-07705-0. "He held high the Declaration of Independence, the Constitution, and the nation's unofficial motto, *e pluribus unum*, even as he was recoiling from the party system in which he had long participated."

Fig. 1. An example of three levels of Wikipedia elements included in the WCCC corpus: text with in-line **citations** (left), **references** including locations (middle) and full **sources** (right) with **identifiers** (ISBN).

starts with procedures that take a Wikipedia database dump as input and extract all citations, references, and sources marked with various WikiCode templates and conventions (Section III-A), resulting in Wikipedia Complete Citation Corpus (WCCC) (Section III-B). The corpus is then refined by removing the unnecessary content and simplifying its structure, resulting in Claim-Source Pairing dataset (Section III-C), used in further experiments.

### A. Processing Wikipedia

Broadly speaking, Wikipedia articles can use three types of elements to cite external sources, as illustrated on the example of *United States* entry in Figure 1:

- **citations** – numbered links in text, e.g. *[21]*,
- **references** – footnotes pointing to sources, often including location of relevant content, e.g. *Sider 2007, p. 226*,
- **sources** – full bibliographic entries describing source publications, possibly including **identifiers**, e.g. *Sider, Sandra (2007). Handbook to . . .* , identified by ISBN *978-0-19-533084-7*.

Each citation is linked to one reference, which in turn is linked to one or many sources.

Wikipedia offers a great variety of guidelines, styles, templates and other technical means for putting citations in text[4] and encourages referencing sources[5] without enforcing usage of any particular technique. The resulting diversity of citation representation poses challenges to an automatic process of extracting a citation corpus. We aimed to recognise as many citations as possible and represent them through the structure outlined above. The extraction process consists of the following steps:

1) Reading a database dump of English Wikipedia (from 01.02.2021) and parsing the wikicode using *mwparserfromhell*[6].
2) Interpreting the structure of citations, references and sources encoded using the following WikiCode tags end templates: `ref`, `refn`, `reflist`, `refs`, `wikicite`, `citation`, `cite *`, `harv*`, `sfn*`, `r`, `rp` and converting the rest into plain text,
3) Cleaning the data by removing sources that are not used in any references and references that are not used in any citations.
4) Extracting the identifiers describing sources (DOI, ISBN, arXiv ID or URL).
5) Storing the output for each article (citations, references, sources, identifiers and plain text) as text files.

### B. Wikipedia Complete Citation Corpus (WCCC)

The result of the process described above is called Wikipedia Complete Citation Corpus (WCCC). Table II shows the size of this resource and previous large open-domain citation corpora based on Wikipedia or scholarly literature. Compared to resources using Wikipedia [35], [36], WCCC includes more citations and identified sources. The latter is likely because those datasets only take into account DOI and ISBN as identifiers, while we accept URLs as well. Additionally, we include full document text and citation location, essential for local citation recommendation. These data were not included in previous resources, though they were considered a future work direction [36].

In order to compare WCCC to scholarly citation resources that match the open-domain character of FRS (challenge B), we focus on large open-domain corpora (for smaller resources see a review [4]). CiteSeerX has been extensively used for local citation recommendation experiments based on 400 characters-long contexts and was extended with DBLP linkage by Caragea et al. [37]. UnarXiv is a high-quality corpus extracted from arXiv by Saier and Färber [38], including linking to Microsoft Academic Graph (retired in 2021) and providing full-text for context analysis. These datasets are based on corpora of 1-2 million documents and include around 15 million citations. With 50 million citations to 24 million sources, WCCC is clearly a more comprehensive source. This is likely due to Wikipedia including references to sources of various genres that are rare in academic corpora, such as works of popular culture, news reports and other websites. While

[4] https://en.wikipedia.org/wiki/Wikipedia:Citing_sources
[5] https://en.wikipedia.org/wiki/Wikipedia:Verifiability
[6] https://github.com/earwig/mwparserfromhell

TABLE II
SIZE OF THE WCCC CORPUS COMPARED TO PREVIOUS LARGE OPEN-DOMAIN WIKIPEDIA AND SCHOLARLY CITATION CORPORA.

| | | documents | citations | sources (with ID) | context |
|---|---|---|---|---|---|
| Wikipedia | Halfaker et al. [35] | **6.751** M | 3.795 M | 1.631 M | No |
| | Singh et al. [36] | 6.069 M | 29.276 M | 2.059 M | No |
| Scholarly | CiteSeerX [37] | 1.957 M | 15.375 M | - | 400 chars |
| | unarXiv [38] | 1.043 M | 15.955 M | 2.746 M | **Full** |
| | WCCC (this work) | 4.815 M | **50.829 M** | **24.349 M** | **Full** |

not all of these are equally reliable, their presence reflects the features of FRS (challenge B).

Given that WCCC contains more citations and sources than any of the previous citation corpora, we hope it may prove useful for tasks other than FRS, such as general citation recommendation or analysis of citation habits of Wikipedia editors. Thus, we make it openly available for download.

### C. Claim-Source Pairing (CSP) Dataset

WCCC contains the complete snapshot of all the citations in Wikipedia, but in the FRS task only the immediate context is available (challenge A). To reflect this limitation, we prepare a collection of examples, each including (1) a fragment of text preceding a citation and (2) a list of identifiers of sources cited. We refer to this resource as *Claim-Source Pairing* dataset.

First, the three-level citation structure is flattened to obtain a list of identifiers of sources cited in a given location. Then, we discard sources without identifiers and their citations. We can also remove most of the text of articles, as only passages that include a citation will play a role in the learning process. In other words, we need to decide on what textual **context** is associated with each citation. Here, three variants of the CSP dataset are considered. In the basic **sentence** version, just the sentence including or immediately preceding a citation is included. For **title+sentence** variant, we also include the title of the Wikipedia article. Finally, **sentence+sentence** uses two sentences preceding the citation. Text is split into sentences using *spaCy* model `en_core_web_sm`.

The complete CSP dataset includes 32 million citations, each linking context text with one or more (1.3571 on average) of the 24 million available source identifiers.

## IV. LEARNING TO CITE

Here we describe how the CSP dataset is employed to predict the citation in a given context, establishing the first solution for the FRS task. In our approach, a context fragment (e.g. a sentence) is represented as a sparse (Section IV-A) or dense vector (Section IV-B). Next, we build search indices that hold these vectors and the associated sources (Section IV-C). At testing time, a query sentence is converted to a vector and its closest neighbours are retrieved from an index, with the ranked list of the associated sources returned as the result.

### A. Sparse context representation

To enable comparison with techniques used in local citation recommendation for scholarly text, we perform an analogous procedure to that of Saier and Färber [31]. This involves associating each source with a pseudo-document, created through concatenation of all contexts (sentences), in which it was cited, represented through:

- **words**, equivalent to bag of words (BoW) representation,
- **claims**, extracted from *PredPatt* [39] parse trees, based on Universal Dependencies,
- **noun phrases**, defined as maximal word sequences from a pre-computed dictionary.

Saier and Färber [31] used a noun phrase dictionary generated from an arXiv corpus. In order to adjust their approach to non-scholarly text, we extracted noun phrases (*noun chunks* in *spaCy*) from sentences in our training set based on Wikipedia (see Section V-B). By filtering out items that appear only once, we obtain a set of 4,629,696 noun phrases.

### B. Dense context representation

In order to check how modern neural meaning representation methods can help in our task, the following three methods are used to convert each citation context to a dense vector:

- *GloVe* [40] word embeddings (version trained on 6 billion tokens), averaged over all words in a fragment (300-dimensional),
- sentence embeddings from *Universal Sentence Encoder* [41], using the deep averaging network (DAN) variant (512-dimensional),
- sentence embeddings from *Sentence-BERT* [42] encoder, using the `paraphrase-distilroberta-base-v1` model (768-dimensional).

### C. Indexing and search

To obtain the sources relevant for a particular query sentence, we would like to find the nearest neighbours, i.e. those indexed contexts, which are the closest to the query. The large number and variety of available sources (challenge B) in FRS makes it necessary to consider trade-offs between the accuracy and time of retrieval.

For sparse representations, we use *Elasticsearch* to index pseudo-documents (consisting of words, claims or noun phrases) and the associated sources. During the search, the best match is found through the Okapi BM25 ranking function [43], an established baseline in information retrieval [44].

For dense representation, finding the closest neighbour among 32 million vectors is much more challenging and time-consuming. We rely on an approximate nearest neighbour (ANN) solution, where the optimal results are not guaranteed. Specifically, we use ANNG graph and tree index [45], [46],

TABLE III
PERFORMANCE OF FINDING RELIABLE SOURCES WITH RESPECT TO DIFFERENT CONTEXT MEANING REPRESENTATION METHODS AND TEST DATASETS.

| | | Wikipedia | | FEVER-FRS supports | | FEVER-FRS refutes | |
|---|---|---|---|---|---|---|---|
| representation | | **NDCG** | **MAP** | **NDCG** | **MAP** | **NDCG** | **MAP** |
| Dense | GloVe | 0.2360 | 0.2239 | 0.0265 | 0.0204 | 0.0174 | 0.0134 |
| | USE | 0.2524 | 0.2394 | 0.0336 | 0.0266 | 0.0244 | 0.0190 |
| | Sentence-BERT | 0.2694 | 0.2544 | **0.0652** | **0.0507** | **0.0483** | **0.0380** |
| Sparse | Words (BoW) | **0.2733** | **0.2553** | 0.0543 | 0.0412 | 0.0401 | 0.0301 |
| | Noun Phrases | 0.1032 | 0.0958 | 0.0060 | 0.0043 | 0.0055 | 0.0037 |
| | Claims | 0.1908 | 0.1804 | 0.0286 | 0.0213 | 0.0191 | 0.0144 |

implemented in the NGT library[7], achieving state of the art results in several ANN benchmarks[8]. The search process is divided into two steps: (1) finding $k$ approximate nearest neighbours using ANNG and (2) ranking these candidates according to normalised cosine similarity. Larger $k$ may return better candidate neighbours, but it carries a computational cost, problematic for the dataset sizes we are using here. For that reason, the experiments were performed with $k = 10$ for Wikipedia evaluation and $k = 100$ for FEVER evaluation. In Section VI-B we investigate the influence of $k$ on the results.

## V. Evaluation

The goal of the evaluation is to check how frequently the system is able to predict the correct sources for a given query context. Each query context (e.g. a sentence) is used to search for the most similar documents in the indices. Then, the sources (represented as identifiers) of the retrieved citations are combined into a single ranking list, ordered according to similarity value and with duplicates removed.

### *A. Measures*

We used the following ranking assessment measures: normalised discounted cumulative gain (NDCG) and mean average precision (MAP) [44]. NDCG is calculated as a sum of relevance values (1 for relevant source, 0 otherwise) for the top 10 positions, weighted by a logarithmic reduction factor and normalised to 0-1. MAP is computed as a mean of the precision scores for each of the positions when the recall increases, up to position 10.

### *B. Wikipedia evaluation*

For Wikipedia evaluation, the CSP citations are randomly assigned to training (80%) and test (20%) subsets, so that citations from the same article belong to the same subset. All training instances (25,589,764 citations) are indexed; the remaining ones are used for testing. If a test citation includes sources not covered by any of the training citations, these sources are not taken into account as relevant in evaluation.

### *C. FEVER evaluation*

For additional evaluation, the Wikipedia-based training citations are added to the index as previously, but a separate test set (FEVER-FRS) based on the FEVER fact-checking shared task [47] is created. The original dataset consists of claims, each associated with a list of evidence sentences from Wikipedia that either 'supports' or 'refutes' the claim. The presence of the latter category allows us to measure the impact of contradictions between claims and sources (challenge C) on performance. The claims labelled 'not enough info' and missing evidence are not taken into account.

To convert this dataset into the FRS task format, we find the evidence sentences in the CSP dataset and treat sources associated with them as relevant for the verification of the claim. Because our work is based on a newer version of Wikipedia, some of the sentences could have been removed or significantly modified. Therefore, from the original evidence, only the sentences that correspond (with possible minor modifications) to the sentences in our training subset of Wikipedia are selected. For this purpose, the Jaro similarity measure [48] is used with a minimum value of 0.8. Additionally, many of the original sentences contain no citations and thus cannot be used for our task. In the end, our FEVER-FRS test set consists of 10,769 'supports' claims (13% of FEVER 'supports' claims) and 4,188 labelled 'refutes' (14% of FEVER 'refutes'). The average number of sources per claim equals 1.5.

## VI. Results

We present the results of three experiments designed to explore the differences between FRS and scholarly citation prediction. In the first one (Section VI-A), we show how the different meaning representation techniques perform in this scenario, demonstrating the impact of refuted claims (challenge C). In the second experiment (Section VI-B), we check the role of the approximate nearest neighbour retrieval mechanism, which is necessary due to the large size of the open-domain collection (challenge B). The third experiment (Section VI-C) shows how the amount of available context (challenge A) influences the results. The computations involved in these experiments were performed in the *Poznan Supercomputing and Networking Center*[9].

### *A. Experiment 1: meaning representation*

Table III shows the results for different meaning representation methods. We can see that the performance measurements using NDCG and MAP are closely aligned. The Wikipedia test set is the easiest one, clearly due to the similarity between training and test data. The more complicated similarity

[7] https://github.com/yahoojapan/NGT

[8] https://github.com/erikbern/ann-benchmarks

[9] https://www.psnc.pl/

TABLE IV
FRS PERFORMANCE FOR THE SOLUTION BASED ON SENTENCE-BERT, USING DIFFERENT CONTEXT REPRESENTATION METHODS.

| **training context** / **test context** | sentence | | title + sentence | | sentence + sentence | |
|---|---|---|---|---|---|---|
| | **NDCG** | **MAP** | **NDCG** | **MAP** | **NDCG** | **MAP** |
| Wikipedia sentence | 0.2665 | 0.2516 | 0.2550 | 0.2390 | 0.2105 | 0.1903 |
| Wikipedia title + sentence | 0.2607 | 0.2442 | **0.2821** | **0.2633** | 0.2293 | 0.2068 |
| Wikipedia sentence + sentence | 0.2128 | 0.1950 | 0.2291 | 0.2092 | 0.2690 | 0.2487 |
| FEVER-FRS supports | 0.0652 | 0.0507 | **0.1011** | **0.0795** | 0.0675 | 0.0535 |
| FEVER-FRS refutes | 0.0483 | 0.0380 | **0.0809** | **0.0628** | 0.0515 | 0.0405 |

relationship between a claim and the contradictory evidence (challenge C) is demonstrated in the decreased performance for ‘refutes’ data. Interestingly, while word-based sparse representation performs the best in Wikipedia evaluation, Sentence-BERT beats it in FEVER-FRS evaluation. Clearly, a dense meaning representation manages to capture the similarities between sentences from different corpora more effectively. The more advanced sparse solutions (using claims and noun phrases) do not outperform the simple word baseline. This is in line with the experiments on scholarly citations [31], where the BoW baseline was providing the best (or nearly the best) NDCG as well.

### B. Experiment 2: candidates retrieved

The second experiment is meant to illustrate the impact of the number of approximate nearest neighbour retrieved ($k$) on our results. Due to high computational cost of search with large $k$, the experiment was carried out on randomly selected 22,399 instances of the test data. Figure 2 shows the value of NDCG and MAP using Sentence-BERT on Wikipedia evaluation for $k = 1, 2, 4, \ldots, 512$. The results of regular BoW (where approximate neighbour retrieval step is not necessary) are shown as horizontal lines. As expected, the results improve for larger $k$. The dense representation performs worse than BoW for low $k$, about the same around $k = 10$ (consistent with Table III), and much better when more candidates are available. This means that Sentence-BERT is actually better suited for the FRS task, but limited due to approximate nature of retrieval among so many sources (challenge B).

### C. Experiment 3: context generation

Table IV shows how the limited context available for each citation (challenge A) influences the results. We evaluate using Sentence-BERT representation and combine approaches for training context (text associated with sources in search index) and test context (text used to generate queries[10]). Understandably, for Wikipedia[11] the best results are obtained when the same method is used for training and test context. Moreover, the title+sentence variant easily outperforms others. This is consistent with the writing style of encyclopaedic articles, where the main subject is explicitly denoted in the title (e.g. as a person name), but frequently omitted from the text body. The same approach clearly provides the best performance in the FEVER-FRS test. Using just the single-sentence context indeed makes the task hard, and having even an indication of the general topic (article title) improves the performance.

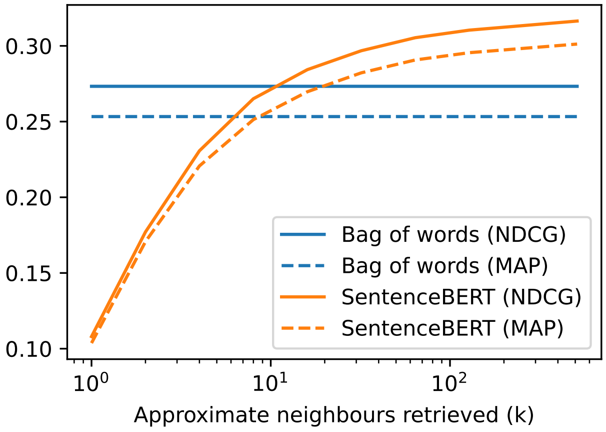


Fig. 2. Performance of the Sentence-BERT variant of our solution on Wikipedia with respect to number of candidate neighbours retrieved (x axis, logarithmic) compared to BoW.

## VII. Discussion

Finding reliable sources, as proposed here, might appear as a step backwards from the ambitious goals of automatic fact-checking. A solution to FRS does not answer whether a claim is true – instead, it points to credible sources and making use of the provided knowledge remains up to a human user. This corresponds to a usage scenario of fact-checking performed in cooperation between a human and a machine [49] and could help in applying crowdsourcing to veracity assessment [50].

Similarly to the classic fact-checking, it remains to be established what will be the best application area for the FRS task. It might involve professional fact-checkers looking for ways to accelerate their work; editors of community-generated content services (such as Wikipedia), concerned with the credibility of the information created; or regular readers of online articles and social media posts, looking for ways to verify the information encountered.

Despite more limited scope than fact-checking, our task still brings many challenges. Firstly, we can see how complex is the relationship between a source and the claims it can be relevant for. Even for a short publication, one can come up with a great variety of claims that could be verified using

[10] This does not apply to FEVER-FRS data, where the query is fixed.

[11] The Wikipedia tests were carried out using 10% of the test instances for performance reasons.

it. The model is expected to infer the relevance of a source, generalising from seeing just a handful of positive examples and no negative ones.

Secondly, the task formulation provided here does not impose any domain restrictions, which contributes to a large number of sources available, limiting the possible solutions through performance considerations. In future, it might be beneficial to limit the task to a restricted domain (e.g. health advice) and perform more in-depth analysis of possible approaches, for example different meaning representations.

Thirdly, the level of difficulty clearly depends on how different is the language used in claims to that of training citations. As visible from Experiment 3, the task is easiest when they come from the same resource (Wikipedia) and follow the same format (title+sentence). The problem becomes more challenging, when claims come from a separate source, as is the case with FEVER-FRS here. This more challenging scenario corresponds to verifying a short claim made without context, e.g. as a Twitter message.

Finally, an additional difficulty is introduced when the claim is untrue and reliable sources (and their citations) contain information contradictory to it. This becomes an obstacle for the applied meaning representation solution (Sentence-BERT), which was originally trained on paraphrases to make it assign similar vectors to sentences with the same meaning. For example, the sentences *Albert Einstein was born in France.* and *Albert Einstein was born in Germany* have different meanings, but could likely be verified with the same source.

Our results can inform a wider discussion about the advantages of the dense and sparse representations in document retrieval. Figure 2 shows that while the dense solutions could in theory provide more relevant sources for our task, they are limited by the approximate neighbour retrieval mechanism. Improving this element, so that all relevant results are returned within acceptable processing time, is a clear direction for future work.

To compare the FRS task to local citation recommendation for scholarly literature, we can look at the results of Saier and Färber [31]. In terms of internal evaluation, the scores are similar, e.g the NDCG for arXiv dataset was reported at 0.22, while the best score on Wikipedia is 0.27, which would suggest that non-scholarly language is not making the task harder. On the other hand, the NDCG values around 0.1 for FEVER-FRS data (see Table IV) clearly indicate that the claims generated externally, e.g. in social media posts, are much more challenging than the scholarly language.

Note that other ways to approach FRS, without learning from citations, are possible. For example, one might choose scientific studies relevant for a particular claim based on their metadata, e.g. title or keywords. Another alternative is to learn from social media posts, which do not contain explicit citations but links to external information relevant in the context of the discussion. However, working with Wikipedia has the advantage of a clearly defined and enforced policy of using reliable sources, which social media do not guarantee.

Finally, we would like to emphasise the major discrepancy between our approach and the typical efforts for countering disinformation. While the research in the area is rich in attempts to detect unreliable content (fake news, propaganda, bots, etc.), little attention is given to reducing the impact of misinformation by promoting reliable content. However, this direction can be currently observed on major content sharing websites (e.g. *YouTube*) that recommend material relevant to COVID-19 produced by credible sources to reduce misinformation impact. An FRS solution is a context-aware version of these attempts, as it could offer content relevant to what a user is viewing at any given time.

## VIII. Conclusions

In this study, we have introduced the task of finding reliable sources and explained how it can be approached by learning from citations added to Wikipedia articles. We have shown how a corpus with over 50 million citations could be obtained and converted to a dataset for this purpose. Finally, we have proposed solutions based on sparse and dense representations and evaluated them using an adapted fact-checking dataset. We expect that this work is just the beginning of the FRS task and the created resources encourage the development of other solutions to the problem. Ultimately, we hope these efforts will lead to a better understanding of the methods for choosing reliable sources for verifying claims.

## References

[1] C. J. Vargo, L. Guo, and M. A. Amazeen, “The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016,” *New Media & Society*, 2017.

[2] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, “The Fact Extraction and VERification (FEVER) Shared Task,” in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018.

[3] D. Shaw, “Wikipedia in the Newsroom,” *American Journalism Review*, 2008.

[4] M. Färber and A. Jatowt, “Citation recommendation: approaches and datasets,” *International Journal on Digital Libraries*, vol. 21, no. 4, pp. 375–405, 2020.

[5] A. Vlachos and S. Riedel, “Fact Checking: Task definition and dataset construction,” in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014, pp. 18–22.

[6] L. Graves, “Understanding the Promise and Limits of Automated Fact-Checking,” Reuters Institute, University of Oxford, Tech. Rep. February, 2018.

[7] N. Hassan, F. Arslan, C. Li, and M. Tremayne, “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*, ser. KDD ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1803–1812.

[8] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, and J. Glass, “Automatic fact-checking using context and discourse information,” *Journal of Data and Information Quality*, vol. 11, no. 3, pp. 1–27, jul 2019.

[9] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “dEFEND: Explainable Fake News Detection,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’19. New York, NY, USA: ACM, 2019, pp. 395–405.

[10] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, “Computational Fact Checking from Knowledge Networks,” *PLOS ONE*, vol. 10, no. 6, 2015.

[11] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2931–2937.

[12] L. Wu, Y. Rao, Y. Zhao, H. Liang, and A. Nazir, "DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL, 2020, pp. 1024–1035.

[13] M. Messner and J. South, "LEGITIMIZING WIKIPEDIA," *Journalism Practice*, vol. 5, no. 2, pp. 145–160, 2011.

[14] N. J. Reavley, A. J. Mackinnon, A. J. Morgan, M. Alvarez-Jimenez, S. E. Hetrick, E. Killackey, B. Nelson, R. Purcell, M. B. H. Yap, and A. F. Jorm, "Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources," *Psychological Medicine*, vol. 42, no. 08, pp. 1753–1762, 2012.

[15] S. J. Handler, S. E. Eckhardt, Y. Takashima, A. M. Jackson, C. Truong, and T. Yazdany, "Readability and quality of Wikipedia articles on pelvic floor disorders," *International Urogynecology Journal*, vol. 32, pp. 3249–3258, 2021.

[16] B. Luyt and D. Tan, "Improving Wikipedia's credibility: References and citations in a sample of history articles," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 4, 2010.

[17] X. Shuai, Z. Jiang, X. Liu, and J. Bollen, "A comparative study of academic and Wikipedia ranking," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*. New York, New York, USA: ACM Press, 2013, p. 25.

[18] M. Mesgari, C. Okoli, M. Mehdi, F. Å. Nielsen, and A. Lanamäki, ""The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 219–245, 2015.

[19] O. Zagovora, R. Ulloa, K. Weller, and F. Flöck, "'I Updated the <ref>': The Evolution of References in the English Wikipedia and the Implications for Altmetrics," *Quantitative Science Studies*, pp. 1–27, oct 2020.

[20] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Analysis of References Across Wikipedia Languages," in *International Conference on Information and Software Technologies (ICIST 2017)*. Springer, Cham, 2017, pp. 561–573.

[21] M. Redi, B. Fetahu, J. T. Morgan, and D. Taraborelli, "Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability," *CoRR*, vol. arXiv:1902, 2019.

[22] D. Taraborelli, L. Pintscher, D. Mietchen, and S. Rodlund, "Wikicite 2017 report," 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.5648233.v3

[23] J. Nicolaisen, "Citation analysis," *Annual Review of Information Science and Technology*, vol. 41, no. 1, pp. 609–641, 2007.

[24] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "ClusCite: effective citation recommendation by information network-based clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. New York, New York, USA: ACM Press, 2014, pp. 821–830.

[25] C. Caragea, A. Silvescu, P. Mitra, and C. L. Giles, "Can't see the forest for the trees?: a citation recommendation system," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*. New York, USA: ACM Press, 2013, p. 111.

[26] Y. Zhang and Q. Ma, "Dual Attention Model for Citation Recommendation," in *Proceedings of the 28th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 3179–3189.

[27] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proceedings of the 19th international conference on world wide web - WWW '10*. New York, New York, USA: ACM Press, 2010.

[28] Y. Lu, J. He, D. Shan, and H. Yan, "Recommending citations with translation model," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, USA: ACM Press, 2011, p. 2017.

[29] W. Huang, Z. Wu, P. Mitra, and C. L. Giles, "RefSeer: A Citation Recommendation System," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 371–374.

[30] P. Molloy, J. Beel, and A. Aizawa, "Virtual Citation Proximity (VCP): Empowering Document Recommender Systems by Learning a Hypothetical In-Text Citation-Proximity Metric for Uncited Documents," in *Proceedings of the 8th International Workshop on Mining Scientific Publications*. Wuhan, China: ACL, 2020, pp. 1–8.

[31] T. Saier and M. Färber, "Semantic Modelling of Citation Contexts for Context-Aware Citation Recommendation," *Proceedings of the European Conference on Information Retrieval (ECIR 2020)*, pp. 220–233, 2020.

[32] B. Fetahu, K. Markert, W. Nejdl, and A. Anand, "Finding news citations for wikipedia," in *International Conference on Information and Knowledge Management, Proceedings*, vol. 24-28-Octo. New York, NY, USA: ACM, 2016, pp. 337–346.

[33] H. Peng, J. Liu, and C.-Y. Lin, "News Citation Recommendation with Implicit and Explicit Semantics," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: ACL, 2016, pp. 388–398.

[34] A. Jana, P. Kanojiya, P. Goyal, and A. Mukherjee, "WikiRef: Wikilinks as a route to recommending appropriate references for scientific Wikipedia pages," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: ACL, 2018, pp. 379–389.

[35] A. Halfaker, B. Mansurov, M. Redi, and D. Taraborelli, "Citations with identifiers in Wikipedia," 2019. [Online]. Available: https://figshare.com/articles/dataset/Citations_with_identifiers_in_Wikipedia/1299540/1

[36] H. Singh, R. West, and G. Colavizza, "Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia," *Quantitative Science Studies*, pp. 1–19, 2020.

[37] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernández-Ramírez, H. H. Chen, Z. Wu, and L. Giles, "CiteSeerX: A Scholarly Big Dataset," in *Proceedings of the European Conference on Information Retrieval (ECIR 2014)*. Springer, Cham, 2014, pp. 311–322.

[38] T. Saier and M. Färber, "unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata," *Scientometrics*, vol. 125, no. 3, pp. 3085–3108, 2020.

[39] A. S. White, D. Reisinger, K. Sakaguchi, T. Vieira, S. Zhang, R. Rudinger, K. Rawlins, and B. Van Durme, "Universal Decompositional Semantics on Universal Dependencies," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: ACL, 2016, pp. 1713–1723.

[40] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: ACL, 2014, pp. 1532–1543.

[41] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal Sentence Encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: ACL, 2018, pp. 169–174.

[42] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: ACL, 2019, pp. 3982–3992.

[43] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[44] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008.

[45] M. Iwasaki, "Proximity search in metric spaces using approximate k nearest neighbor graph," *IPSJ Trans. on Database*, vol. 3, no. 1, pp. 18–28, 2010.

[46] ——, "Proximity search using approximate k nearest neighbor graph with a tree structured index," *IPSJ Journal*, vol. 52, no. 2, pp. 817–828, 2011.

[47] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification," *arXiv:1803.05355*, mar 2018.

[48] W. W. Cohen, P. Ravikumar, S. E. Fienberg, and Others, "A Comparison of String Distance Metrics for Name-Matching Tasks." in *IIWeb*, vol. 3, 2003, pp. 73–78.

[49] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino, "Automated Fact-Checking for Assisting Human Fact-Checkers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, mar 2021.

[50] K. Roitero, M. Soprano, S. Fan, D. Spina, S. Mizzaro, and G. Demartini, *Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background*. New York, NY, USA: ACM, 2020, pp. 439–448.