Received: 19 February 2021 | Revised: 17 July 2021 | Accepted: 12 September 2021

DOI: 10.1002/asi.24577

RESEARCH ARTICLE

JASIST WILEY

Studying effectiveness of Web search for fact checking

Maram Hasanain | Tamer Elsayed

College of Engineering, Qatar University, Doha

Correspondence
Maram Hasanain, College of Engineering, Qatar University, Doha, Qatar.
Email: mh081131@qu.edu.qa

Abstract

Web search is commonly used by fact checking systems as a source of evidence for claim verification. In this work, we demonstrate that the task of retrieving pages *useful* for fact checking, called *evidential pages*, is indeed *different* from the task of retrieving topically relevant pages that are typically optimized by search engines; thus, it should be handled differently. We conduct a comprehensive study on the performance of retrieving evidential pages over a test collection we developed for the task of re-ranking Web pages by usefulness for fact-checking. Results show that pages (retrieved by a commercial search engine) that are topically relevant to a claim are not always useful for verifying it, and that the engine's performance in retrieving evidential pages is weakly correlated with retrieval of topically relevant pages. Additionally, we identify types of evidence in evidential pages and some linguistic cues that can help predict page usefulness. Moreover, preliminary experiments show that a retrieval model leveraging those cues has a higher performance compared to the search engine. Finally, we show that existing systems have a long way to go to support effective fact checking. To that end, our work provides insights to guide design of better future systems for the task.

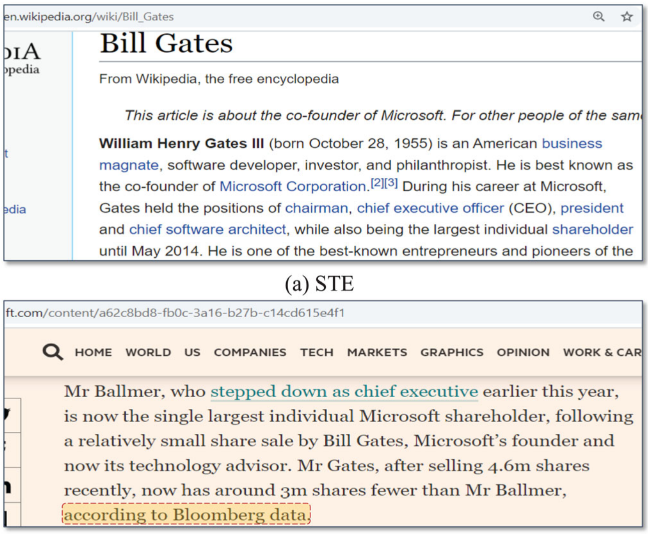
1 | INTRODUCTION

Evaluation of information retrieval (IR) systems usually focused on their ability to retrieve *topically relevant* documents, that is, documents that are "about" the "topic" representing the user's information need (Harman, 1992; J. Jiang et al., 2017; Saracevic, 2007). However, recent studies argued that other dimensions of relevance (e.g., document understandability) should be considered for better system development and evaluation (J. Jiang et al., 2017; Johnson et al., 2016; Unkel & Haas, 2017; Yigit-Sert et al., 2020). Building on earlier studies that examined the dimension of document utility or usefulness for a user searching the Web (Mao et al., 2016), in this work, we study this aspect of relevance within the domain of fact checking.

A crucial step in the fact checking pipeline, is the retrieval of information sources (e.g., Web pages) against which a claim can be verified (Cazalens et al., 2018). Recently, studies demonstrated the value of and need for extracting *evidence* snippets from identified information sources. *Evidence* is essential to justify or explain the system's veracity prediction and provide user with information to make further assessment and decision regarding the claim's veracity (Conroy et al., 2015; Ma et al., 2019). Despite the recognized importance of evidence-based fact checking, few studies have attempted to characterize such important pieces of information and the documents that contain them (Rinott et al., 2015). Furthermore, effectiveness of retrieving documents with evidence is rarely characterized or evaluated.

In this work, we aim to extensively analyze snippets constituting evidence for fact checking and the documents that contain them, and measure effectiveness of state-of-the-art systems designed to retrieve such documents. We focus on Web pages since many fact checking systems rely on searching the Web (e.g., Wikipedia) to extract evidence (Nie et al., 2019). Formally, we examine one dimension of page relevance, which we denote as *usefulness for claim verification*. We refer to a Web page that is useful for claim verification as an "evidential page" defined as *a topically relevant page that contains at least one objective self-contained source-based evidence*. Examples of evidence can be quotes, statistics, or mentions of sources, to name a few.

In our definition of an evidential page, we focus on one major type of evidence denoted as source-based evidence (SRE). "Evidence" in fact-checking research has been typically defined in terms of stance; studies assumed the evidence to be a snippet of text that supports or refutes the claim (Rinott et al., 2015; Thorne et al., 2018; Zhang et al., 2019). These stance-based evidence (STE) snippets are usually repetitions of the claim itself in addition to making other claims, that is, STE conveys information not supported by any source or reference but the document itself. Differently, we analyze the SRE type of evidence, which is objective and presented with a clear source of information. Figure 1 compares the two evidence types for the claim "Bill Gates was the largest individual shareholder of Microsoft." The STE snippet and claim are borrowed from FEVER dataset (Thorne et al., 2018). We notice in the SRE snippet that "Bloomberg," a specialized source on the subject, was cited to support the given claim. Such source offers great potential for verification as it can be further consulted, if needed, by the user. For the STE snippet, the page only repeats the claim itself offering no actual evidence other than the statement of the page's author. This situation becomes even trickier when the author is anonymous or unknown to the user, making trusting a mere repetition of the claim unreasonable.



Bill Gates

From Wikipedia, the free encyclopedia

This article is about the co-founder of Microsoft. For other people of the sam

William Henry Gates III (born October 28, 1955) is an American business magnate, software developer, investor, and philanthropist. He is best known as the co-founder of Microsoft Corporation.[2][3] During his career at Microsoft, Gates held the positions of chairman, chief executive officer (CEO), president and chief software architect, while also being the largest individual shareholder until May 2014. He is one of the best-known entrepreneurs and pioneers of the

(a) STE

ft.com/content/a62c8bd8-fb0c-3a16-b27b-c14cd615e4f1
HOME WORLD US COMPANIES TECH MARKETS GRAPHICS OPINION WORK & CAR

Mr Ballmer, who stepped down as chief executive earlier this year, is now the single largest individual Microsoft shareholder, following a relatively small share sale by Bill Gates, Microsoft's founder and now its technology advisor. Mr Gates, after selling 4.6m shares recently, now has around 3m shares fewer than Mr Ballmer, according to Bloomberg data.

(b) SRE

FIGURE 1 Web pages showing two types of evidence: Stance-based (STE) and source-based (SRE). Source of evidence in SRE is highlighted

We believe Web search systems tailored for evidence extraction can support two types of users: Normal users searching the Web to fact-check a claim, and assistive fact checking systems exploiting the huge amount of information on the Web. Consequently, a major objective of our study is to provide insights on how to improve Web search for the task of evidential page retrieval. We address our main objective through answering the following research questions:

RQ1. *To what extent topical pages are evidential, and how correlated is the effectiveness of retrieving these two types of pages?*

RQ2. *What types of evidence can be found in evidential pages?*

RQ3. *What textual features distinguish evidential and nonevidential pages?*

RQ4. *How effective are existing systems in retrieving evidential pages?*

We answer the above questions by analyzing the performance of a commercial search engine in two tasks: Topically relevant pages retrieval and evidential pages retrieval. Our study shows that pages (retrieved by a commercial search engine) that are topically relevant to a claim are not always useful for verifying it, and that the search engine performance is generally weakly correlated across the two tasks using two correlation coefficients (Kendall's τ and Pearson's r). Given the aforementioned finding, we investigate and identify characteristics or features specific to evidential pages. Furthermore, preliminary experiments show that effectiveness of a supervised evidential pages retrieval model that employs them has a 5.3% increased recall of evidential pages over the search engine.

Overall, our contribution in this study is fourfold.[1]:

1. We conducted the first in-depth comparative study of the performance of Web search for the tasks of

retrieving topically relevant versus evidential pages for verifying a given claim, showing that the two tasks are inherently different.
2. The study provides a thorough analysis of distinguishing characteristics of *evidence* appearing in *evidential* Web pages, which is rarely studied in existing literature. Furthermore, it shows that the identified characteristics, when leveraged in a supervised evidential pages retrieval model, lead to promising results.
3. The study quantifies the potential performance gain Web search systems can attain to better support the task of retrieving evidential pages for fact-checking.
4. We release an annotated dataset for the task of re-ranking of Web pages by usefulness for claim verification.[2] The dataset includes 2,641 Web pages that are potentially relevant to 59 claims and annotated by both dimensions of relevance (i.e., topical and evidential) compared in this study.

The remainder of this article is organized as follows. We first summarize related studies. Next, dataset construction process is described. We then discuss how we evaluate the task of evidential pages retrieval. We proceed to compare the performance of Web search in the tasks of retrieving topically relevant versus evidential pages. We then analyze evidential pages and identify distinguishing linguistic characteristics. Effectiveness of those characteristics is then examined. Before concluding, we demonstrate that a significant improvement in evidential page retrieval is attainable by search engines.

2 | RELATED WORK

Misinformation on the Web and social media encouraged research on approaches to battle this flood of false information. For a broad coverage of the state of automatic fact-checking, many surveys can be found in literature (Collins et al., 2020). This study is focused on a specific type of verification systems which are evidence-based. We summarize related studies developing such systems focusing on how evidence is identified. We also discuss studies that conducted linguistics analysis on documents in the fact-checking domain. Finally, we summarize studies that evaluated retrieval systems by usefulness.

2.1 | Evidence-based verification systems

In order to trust the verification system's decision or even verify it further, the system should provide interpretable decisions usually explained in terms of evidence used to make these decisions (Nguyen et al., 2018). Ma et al. (2019) proposed a hierarchical neural network using attention to capture sentence topical coherence and semantic entailment with respect to the claim. The DeClarE system is also based on a neural network that predicts claim veracity given related articles. Attention is used to capture article salient words with respect to the claim and present them as evidence (Popat et al., 2018).

As part of the recent FEVER challenge on evidence-based fact-checking, several systems have been proposed (Hanselowski et al., 2018; Malon, 2018). The task focused on Wikipedia articles only from which systems are required to extract STE. Another recent challenge is Task 2.A of the CheckThat! lab at CLEF2019 (Hasanain et al., 2019). Similar to the focus of our work, the task targeted SRE and systems were required to rank pages potentially related to a claim by usefulness. Proposed approaches included a BERT model to rank pages (Favano et al., 2019), and a learning-to-rank model using page credibility and similarity to the claim as features (Haouari et al., 2019). All aforementioned works, evaluated systems by effectiveness of performing the required task and did not clearly identify features most helpful in characterizing evidence. Moreover, most of evidence-based fact checking systems aimed to identify and use STE while we are interested in SRE.

2.2 | Analysis of verification Web pages

Several studies in the fact-checking domain analyzed language in documents. Work by Rinott et al. (2015) is the closest to ours. In their work, authors designed features that capture types of supporting evidence extracted from articles related to a given topic. Features depended on a manually crafted lexicon for each evidence type, patterns (e.g., presence of quotes), named entities and subjectivity words. Contrary to our work, the focus was on developing a system for evidence ranking. Moreover, effectiveness of the proposed features in characterizing evidence was not studied. Finally, experiments were limited to Wikipedia articles while we consider the general Web. Wang et al. (2018) designed features to characterize and classify documents as supporting or refuting a claim. Features were mainly textual similarity and entity-based features in addition to a manually crafted lexicon to detect contradicting discourse. Experiments were conducted on documents acquired by searching the Web using a commercial search engine. Evaluation of retrieval performance of the engine was done using recall of supporting documents which is a measure we also consider in this work. We furthermore show how evidential pages retrieval correlates to topical relevance retrieval. Also, evaluation was focused on system classification accuracy and not on the linguistic characteristics of the retrieved pages.

S. Jiang et al. (2020) identified linguistic patterns that can aid in extraction of the claim, claimant, and claim veracity from fact-checking articles. Although the work pointed out the importance of extracting evidence from fact-checking articles, identifying this component was left for future work. Additionally, the study was limited to articles from fact-checking websites only, while we consider Web pages in general regardless of their domain.

In a different line of work, several studies focused on identifying linguistic characteristics differentiating trusted and false news (S. Jiang & Wilson, 2018; Trielli & Diakopoulos, 2019). A clear difference between our work and these studies is that we aim to identify linguistic cues that can be used to extract evidence from Web pages as opposed to differentiating true and false documents.

2.3 | Evaluation of IR systems by usefulness

With the growing interest in more user-centric evaluation of IR systems and Web search engines specifically, several works studied evaluation of usefulness. In a recent work, Vakkari (2020) presented an elaborate survey on usefulness evaluation in the IR field. This survey found that research usually agrees on usefulness definition, where usefulness of retrieved results is defined as the extent to which information in retrieved documents contribute to performing a larger task. In this work, our larger task is claim verification, and useful documents are those that give evidence needed to fulfill this task. In a related work, Mao et al. (2016) compare relevance to usefulness evaluation and investigate, through a user study, how they correlate to user satisfaction. The study was carried over search tasks (topics) that are generally informational in nature (i.e., users trying to find information about a certain topic or for a larger task). Majority of existing work studied perceived usefulness as judged by real users or external annotators. Vakkari et al. (2019) took it a step further and studied actual usefulness by asking users to benefit from retrieved documents in doing a writing task. Their focus was on information gathering under some search topics. Differently from existing work, our focus is on measuring usefulness and contrasting it with topical relevance for a specific task, which is claim verification. Furthermore, we identify some of the content-distinguishing features of useful documents.

3 | DATASET

Several datasets that tackled evidence-based fact checking can be found in literature; however, they mostly either include artificially constructed claims such as FEVER (Thorne et al., 2018), provide unlabeled Web pages as evidence such as MultiFC (Augenstein et al., 2019) and WIKIFACTCHECK (Sathe et al., 2020), include a small set of claims and evidential pages as in Yasser et al. (2018), or include pages labeled for stance rather than usefulness such as EMERGENT (Ferreira & Vlachos, 2016). Differently, our study aims at understanding how "topical relevance" to a claim is different from "usefulness" for verifying that claim. To achieve this goal, we need a dataset that enables such study. To that end, we conduct our analysis on a dataset we recently constructed (CT19-T2) designed specifically for the task of predicting page usefulness for claim verification as part of CheckThat! lab at CLEF2019 (Elsayed et al., 2019; Hasanain et al., 2019). CT19-T2 includes general Web pages that are not limited to few domains. Claims were manually curated and coupled with a large set of manual annotations for both topical relevance and usefulness making the effectiveness comparison between the two tasks possible.

In CT19-T2, we define an evidential page (i.e., a page that is *useful* for verification) with respect to a given claim as *a page that is both topically relevant to the claim and it provides evidence to determine the claim's veracity*. Examples of evidence can be quotes, some statistics, or mentions of sources. The evidence must be supported by a mention to its source in the page. The dataset is composed of three components: (a) 59 Arabic claims (labeled by veracity); 30 of them are verified as true, and the rest are false. (b) 2,641 corresponding Web pages (labeled by usefulness); 661 of them are found evidential, and (c) 1,940 passages resulting from manual splitting of evidential Web pages (labeled also by usefulness); 737 of them are found evidential. This section presents the process of curating each of these components.

3.1 | Claims

We selected 59 claims from multiple sources including a pre-existing set of Arabic claims (Baly et al., 2018), a survey in which we asked the public to provide examples of claims they have heard of, and headlines from six Arabic news agencies that we rewrote into claims. The news agencies selected are well-known in the Arab world: AlJazeera, BBC Arabic, CNN Arabic, AlYoum AlSabea, AlArabiya, and RT Arabic. We note that the number of claims used in our study is in the range of that reported in many similar datasets (e.g., TREC Web search collections [Collins-Thompson et al., 2015]).

We manually categorized the collected claims into topics to ensure that the dataset is not skewed toward one type of claims. Table 1 demonstrates that the claims cover a variety of topical categories.

TABLE 1 Claims distribution in CT19-T2

Category	# Claims	Example (translated) claim
Politics and economy	21	CT19-T2-024: Egyptian President El-Sisi proposed expanding the Gaza Strip towards Sinai
Health	11	CT19-T2-055: Excessive consumption of sugar causes the growth and spread of cancer cells in the human body
Science and technology	10	CT19-T2-029: China announces iPhones sale ban in China starting from iPhone 6 through iPhone X
Arts and culture	6	CT19-T2-002: *Capernaum* made it to final nominations for the 2019 Golden Globe Awards in the category of Best Foreign Film
Sports	4	CT19-T2-021: Brazilian goalkeeper Alison Baker moved from Italian club Roma to Liverpool
Others	4	CT19-T2-033: Two express trains collided at the Marsandiz Station in Ankara, leaving dozens of deaths and injuries
Social	3	CT19-T2-030: Artist Amr Diab married artist Dina El-Sherbiny

3.1.1 | Labeling claims

We acquired the veracity labels for the claims in two steps. First, two graduate students labeled all claims independently. Then, they met to resolve any disagreements, and thus reached consensus on the veracity labels for all claims.

3.2 | Pages and passages

We depend on Web search to retrieve potentially related pages to the claims. We manually formulated a search query representing each claim, and issued it against Google to retrieve the top 50 Web pages for each claim. Retrieved pages that were not Arabic or for which we could not acquire the Hyper Text Markup Language (HTML) representation were discarded, leaving us with an average of 45 pages per claim. The pages were labeled following this pipeline:

1. *Topical relevance*: We first identified topically relevant pages. In order to speedup this labeling process, we hired two groups of annotators: Amazon Mechanical Turk crowd-workers and in-house annotators. Each page was labeled by *three* annotators, and majority voting determines the final label of the page.
2. *Usefulness*: Topically relevant pages were then given to in-house annotators to be labeled for usefulness (i.e., evidentiality) using a two-way classification scheme: *Evidential* and *not evidential*. Annotators were trained on the task and were instructed to closely look for the source of the evidence in the page. However, they were not asked to explicitly extract the source of evidence as part of the annotated dataset as this will over-complicate the annotation task for them. Each page was labeled by three annotators, and majority voting determines the final label of the page.
3. *Usefulness of passages*: In addition to identifying evidential pages, we are also interested in finding out which passages in these pages contain the evidence. We manually split the *evidential* pages into passages, as we found that automatic splitting techniques are not accurate enough. Finally, one of the authors labeled each passage as evidential or not.

3.3 | Verifying annotations quality

3.3.1 | Validating usefulness definition

When annotating the dataset, we hypothesized that an evidential page must be topically relevant to the claim, and thus a nonrelevant page cannot be useful. We examine the validity of this assumption in CT19-T2 by relabeling a set of nonrelevant pages for usefulness. For each claim, a single annotator (one of the authors) relabeled the 5 highest-ranked pages that received unanimous "nonrelevant" judgment from the original annotators. We stress full agreement to maximize the chance that selected pages are indeed nonrelevant. Only 260 nonrelevant pages for 56 out of the 59 claims matched that condition. Results show that *none* of the relabeled nonrelevant pages were eventually found evidential, reassuring the validity of the earlier assumption for the CT19-T2 dataset. This outcome is also in line with observations found in literature where relevance is shown to be a condition for usefulness (Mao et al., 2016). This suggests that it is very unlikely that a page which is not topically relevant to a claim will be useful in verifying it. That encourages search-based fact checking systems to focus on retrieving topically relevant pages first to downsize the pool of potentially evidential pages.

3.3.2 | Inter-annotator agreement

We next evaluate the label quality by computing inter-annotator agreement (IAA) among the three judges of topical relevance and also usefulness of pages using Fleiss Kappa (Fleiss, 1971). IAA describes the degree of consensus in judgments among annotators and has been found to be a reasonable measure of judgment quality (Damessie et al., 2017). We first compute Fleiss κ for the topical relevance labels over 2,641 Web pages. We found $\kappa = 0.7$, which is considered substantial agreement according to a widely adopted interpretation of Kappa values (Landis & Koch, 1977). Next, we compute κ over usefulness labels on *topically relevant documents*. Agreement on usefulness was moderate ($\kappa = 0.49$), which is lower than agreement on topical relevance. A possible justification is that usefulness is more complex to judge and requires good understanding of the fact checking process and what characterizes evidence in a Web page. Furthermore, explaining the concept of "evidence" to crowd-workers is quite difficult. Overall, agreement level for *both tasks* is comparable to or higher than those achieved in literature for relevance judgments (e.g., Alonso & Mizzaro, 2012).

4 | EVALUATING EVIDENTIAL RETRIEVAL

Before studying the quality of evidential pages retrieval, we need first to establish a suitable evaluation approach for the task. We argue that using typical precision-oriented ranked retrieval evaluation measures might not be enough for this task, due to the different objective we envision the fact checking user has.

We assume the evaluation approach simulates an *artificial user* interacting with the search engine (Wicaksono & Moffat, 2020) to retrieve Web pages useful in fact-checking a given claim. This allows us to put together a user model capturing the user interaction with the system and a corresponding evaluation measure. In designing this proposed user model, we benefit from existing models proposed for two related tasks: (a) Focused retrieval tasks (e.g., passage retrieval or question answering), where the system is expected not only to retrieve a ranked list of relevant documents, but also to identify relevant text snippets from these pages given an initial query (Pehcevski & Thom, 2007), and (b) the argument retrieval task, such as that described by Roitman et al. (2016), where the system should return a list of pages ranked by their potential of containing claims supporting or denying an argument. Differently from the latter task, our work only focuses on *factual evidence*, while in Roitman et al. (2016)'s work, retrieved claims can be opinionated or factual.

Let us assume that our user is trying to verify a claim by searching through a Web search engine using the claim as the input query. We hypothesize that the user's objective is to find as many *relevant* evidence from as few Web pages. More specifically, we hypothesize that:

- In order to verify the given claim, the user seeks to find as many evidential pages as possible, to help support or refute the claim. This means that the task is more *recall-oriented*.
- Due to the scale of claims a user might face each day, she has very limited time to verify claims; therefore, she is willing to spend some time looking for evidence for the given claim, but not so much. This calls for a cut-off point, k, at which the user stops looking at the retrieved pages. In this work, we set this cut-off point to a small value ($=10$), as in typical Web search (with 10 results per results page), users are not likely to switch to the next results page. For professional fact-checkers, that cut-off point can be set to a larger value; this is left for future work.
- Focusing more on the task of fact checking, the user is more lenient about the rank of the retrieved evidential pages within the ranked list, before reaching the cutoff point.

Based on the above user model, we adopt a recall-based evaluation measure (Pehcevski & Thom, 2007; Roitman et al., 2016). The proposed measure is *Recall@k* or *R@k* for short. We specifically chose this measure since it was shown through fidelity testing that it is able to model and evaluate focused retrieval tasks (Pehcevski & Thom, 2007). The measure captures the percentage of retrieved evidential pages, for a given input claim, within the top k retrieved pages. In our experiments, we set k to 10.

5 | TOPICAL RELEVANCE VERSUS USEFULNESS

In this section, we answer **RQ1**: *To what extent topical pages are evidential, and how correlated is the effectiveness of retrieving these two types of pages?* We conduct two studies addressing the following sub-questions:

1. How much does topical relevance imply usefulness for claim verification?
2. How effective is the search engine in evidential pages retrieval?
3. How correlated is retrieval of evidential pages to retrieval of topically relevant pages?

5.1 | How much does topical relevance imply usefulness for claim verification? (RQ1.a)

In this section, we test the hypothesis that usefulness is different from topical relevance by first examining the percentage of evidential pages from those topically relevant per claim.

Figure 2 shows that for about two thirds of the claims, the percentage of evidential pages out of the relevant ones is less than 75%. Moreover, for third of the claims, this percentage is lower than 30%. The average percentage of evidential pages out of the relevant ones is 55.7% per claim. This indicates that, for many claims, a small percentage of topically relevant pages are indeed useful for verification.

But is it the case that we can observe more evidential pages by getting more topically relevant pages? To answer this question, we also look at the correlation of number of topically relevant and evidential pages over all claims (Figure 3). We found that Pearson's Correlation r (Pearson, 1895) is 0.78 (significant with $p < .05$, and two-tailed paired t-test). High correlation is somewhat expected, since evidential pages are a subset of the relevant ones; however, the two sets are far from being equal or even close. As the figure shows, more topically relevant pages among search results do *not* always imply more evidential pages. In fact, we observe that only few claims have equal number of relevant and evidential pages. We also observe some extreme cases. At one end, claims 20 and 2, for example, had many topically relevant pages *and* almost all were found evidential. At the other end, claims 43 and 53 had many topically relevant pages too *but* very few were found evidential. We also observe that the topical category of the claim does not generally influence correlation between topical relevance and evidentiality. We conclude that topical-relevance is indeed not equivalent to usefulness for verification.

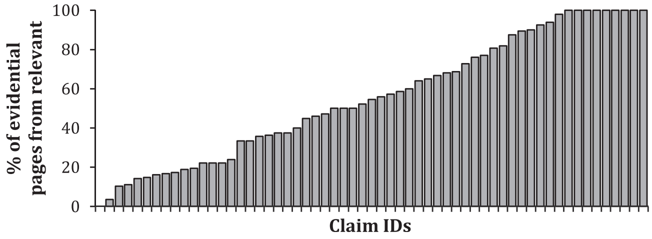


FIGURE 2 Percentage of evidential Web pages out of relevant per claim

5.2 | How effective is the search engine in evidential pages retrieval? (RQ1.b)

Search engines have been proven to have great influence on users' opinions and reshaping their perceptions (Trielli & Diakopoulos, 2019). To assist human fact-checkers who use search engines to form an opinion about the veracity of a claim, the engine, given a claim, should optimize retrieval to present evidential pages. To assess how effectively a search engine achieves that goal, we evaluate the commercial search system (i.e., Google) on the task of evidential pages retrieval using the aforementioned recall measure (R@10). For that task, evidential pages have a label of 1, while the remaining nonrelevant and nonevidential pages get a label of 0. We find that, using our dataset, the engine achieves R@10 = 0.54, which is very far from the maximum possible value of 1. We believe this is the case because Web search engines are usually more optimized for precision than recall. In terms of topical relevance, we evaluate the engine using measures typically used for ranked retrieval, namely, precision and average precision @ rank k (P@k and AP@k, respectively) setting k = 10. The same engine achieves AP@10 = 0.77 and P@10 = 0.72 in topical relevance retrieval, which is expected from a powerful commercial search engine typically optimized for the task of topical relevance retrieval. This shows a big gap in how

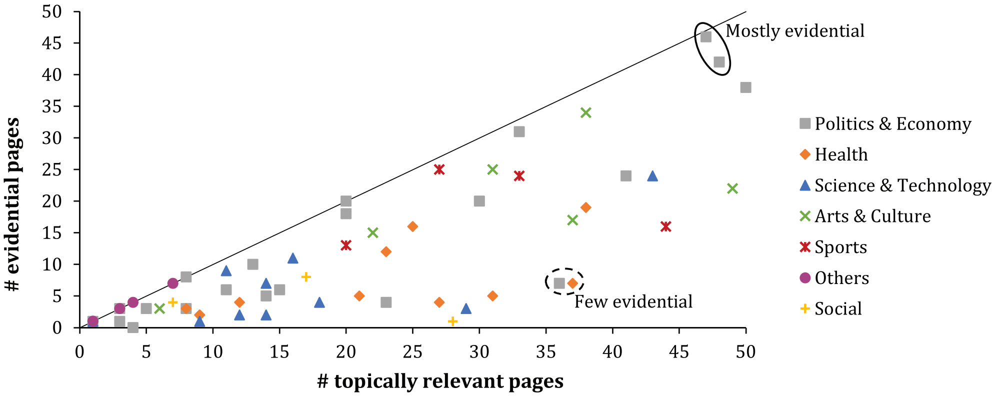


FIGURE 3 Correlation between number of evidential and topically relevant pages per claim. Line shows ideal case where all relevant pages are evidential

the engine is optimized across the two tasks, reflected in the estimated performance, hence the estimated user's satisfaction, in each. We also note that, in terms of *P*@10, the system performs much better in the relevance retrieval task compared to earlier work on the Arabic Web (Tawileh et al., 2010). This can be due to the fact that the existing work is a decade old and commercial search engines are expectedly better now.

5.3 | How correlated is retrieval of evidential pages to retrieval of topically relevant pages? (RQ1.c)

Since the two tasks are assumed to have different user models, we find that direct comparison of system performance in the two retrieval tasks using the same evaluation measure is not meaningful. Instead, we opt to characterize the difference between the two tasks using correlation between the system performance per claim for both tasks. If the search engine sees the two tasks very similar, we expect perfect correlation. Correlation is a standard approach used in the IR field to compare pairs of systems (Carterette, 2009; Yilmaz et al., 2008). We first compute Kendall's τ correlation coefficient (Kendall, 1938) between the *rankings* of claims by the effectiveness of relevance retrieval (measured by *P*@10 or *AP*@10) and by the effectiveness of usefulness retrieval (measured by *R*@10). We also report a linear correlation measure, Pearson's *r* (Pearson, 1895), between scores of both tasks.

Results in Table 2 demonstrate that search system performance across the two retrieval tasks is consistently different. We also observe that the correlation is generally low in 3 out of the 4 scores from the table. Interestingly, we observe a negative correlation between recall of evidential pages and precision of relevance ranking. We further investigate this observation by plotting this correlation in Figure 4. The figure shows that for many claims with perfect or near-perfect precision in retrieving relevant pages in the first results page, recall of evidential pages varies a lot. In fact, out of the 14 claims for which the system got perfect *P*@10 (the right most vertical line), 8 claims of them featured less than 50% of total evidential pages among the top 10 ranks. Similarly, for claims with perfect recall in retrieving evidential pages in the first results page (the top horizontal line), precision of relevant pages greatly varies across the board. The figure correlating *R*@10 and *AP*@10 is omitted since we observe a similar pattern to Figure 4.

Overall, experiments presented so far demonstrated that (a) topically relevant pages are *not* all evidential, that is, they are not all useful for fact checking, and (b) performance of retrieval of evidential pages is not correlated with that of retrieval of topically relevant pages. This suggests that a Web search system used in a fact-checking setting need to be optimized to retrieve evidential pages to better support claim verification.

6 | CONTENT ANALYSIS

Establishing that retrieval of evidential and relevant pages is different through a statistical analysis has been insightful. However, understanding distinctive linguistic features in evidential pages can have more direct contribution to improving search system design for claim verification. We first develop a categorization of evidence types by manual inspection of evidential pages. Then, we characterize differences between evidential and non-evidential pages through a study of their lexical features.

6.1 | What types of evidence can be found in evidential pages? (RQ2)

Studies from the argumentation theory domain identify three types of evidence that one can use to support a claim (Hoeken & Hustinx, 2009): (a) Anecdotal (giving examples), (b) statistical (providing statistics and results),

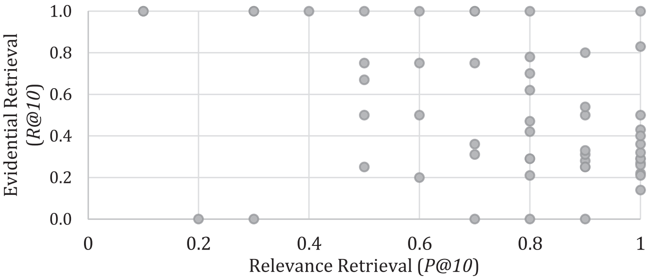


FIGURE 4 Correlation between evidential retrieval performance (*R*@10) and topical-relevance retrieval (*P*@10) per claim

TABLE 2 Correlation between retrieval performance of evidential pages (measured by *R*@10) and topically relevant pages (measured by *P*@10 or *AP*@10)

Usefulness measure	Relevance measure	Kendall τ	Pearson *r*
R@10	P@10	−0.24	−0.42
R@10	AP@10	0.13	0.15

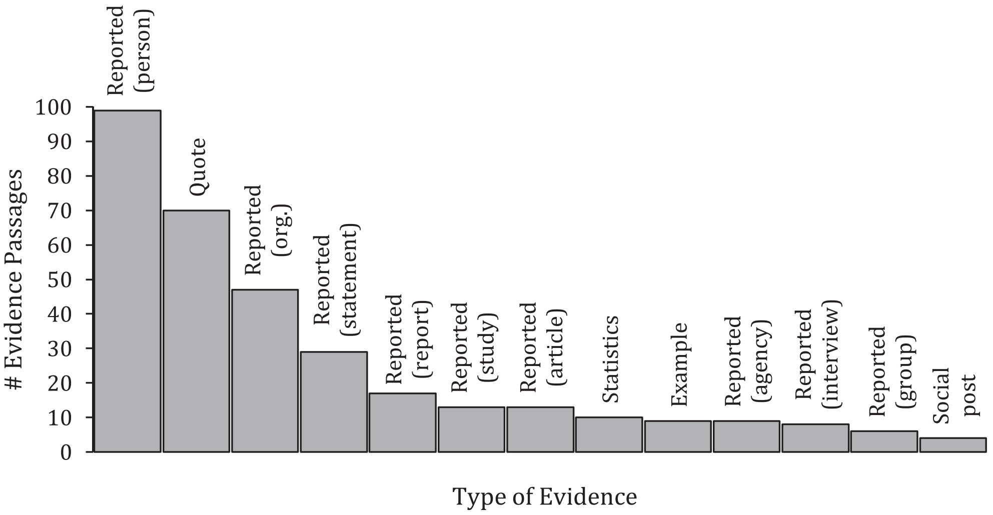


FIGURE 5 Distribution of evidence passages by type of evidence

and (c) testimonial (quoting testimony from a source, e.g., person or a group). Starting from this high-level taxonomy, we identify a set of common patterns or types of evidence in the dataset. For each claim, the 5 highest-ranked evidential pages that got unanimous annotation were inspected, resulting in 163 evidential pages with 334 evidential passages covering 44 claims.

One annotator (first author) annotated each passage by the category of the main evidence in a passage, adding new categories while annotating. More precisely, the annotator read each evidence, answered a question on whether the evidence type is anecdotal, statistical, or testimonial. Then, a sub-category was given whenever possible, for example, for testimonial (reported) evidence, she mentioned the source (e.g., person, organization, etc.) of the testimony. She then revisited all the labeled sub-categories, grouped or re-defined some of them, and then reached a final taxonomy of 13 types. Figure 5 shows the distribution of passages across types of evidence. Most evidential passages reported evidence as stated in or from a source, for example, a person, an organization, an official statement, etc. Moreover, directly quoting from a source was the second commonly used type of evidence. To clarify how SRE evidence was present in annotated paragraphs, consider this claim "*CT19-T2-050: Steve Jobs was the son of a Syrian immigrant*." An example evidential paragraph reporting evidence from a person is: "*Banksy, who is a famous artist, said: We are often tempted to believe that immigration is a drain of the country's resources, but Steve Jobs was the son of a Syrian immigrant*." This example paragraph reported both the claim and that Banksy is a source of evidence.

In conclusion, the study under **RQ2** suggests that features capturing reported and quoted speech can help retrieve evidential pages. We further study this observation next.

6.2 | What textual features distinguish evidential pages? (RQ3)

We examined several lexical and stylistic features in evidential and nonevidential (but topically relevant) pages. We run the analysis on a sample of the labeled pages since we had to manually extract content from the live version of some pages due to poor HTML structure. The random sample covers *all claims* from both evidential and nonevidential (but topically relevant) pages. For each category of pages, each claim was covered by the minimum of five pages (if any). Eventually, we analyzed 234 evidential and 220 nonevidential pages (representing 35 and 39% of pages per category, respectively). Pages were tokenized, and stop words and URLs were removed. Features computed per page are listed below.

- *Length*: Tokens counts
- *Quotes*: Number of quoted statements, since many evidential pages contained quotes as evidence.
- *Statements*: Number of (Arabic) reported speech words, as addressing **RQ2** showed that most evidence passages had reported statements. We compiled a list of 50 words frequently used to convey or report a statement made by others, such as (translated to English) "said," "reported," and "announced."
- *Unique tokens*: Number of unique tokens. We use this feature to capture lexical diversity in text.
- *Claim frequency*: Total frequency of claim words. We use the prevalence of claim words in a page as an indication of its topical focus.
- *Entities*: Number of named entities, extracted using a multilingual named-entity recognition tool (Al-Rfou et al., 2015). This aims at capturing the frequency of mentioning names of sources.

- *Numbers*: Count of numbers. It captures usages of statistics as evidence.
- *Sentiment frequency*: Count of words with positive polarity (e.g., "holy") and negative polarity (e.g., "corruption") identified using a large-scale multilingual sentiment lexicon (Chen & Skiena, 2014). We hypothesize that evidential pages, by definition, contain *objective* evidence and thus, will show less sentiment.
- *Exclamation frequency*: Number of characters associated with conveying emotions such as "!" and "?." We hypothesize that evidential pages will show less emotions.
- *POS*: Counts of Part-of-Speech tags using Stanford tagger (Toutanova et al., 2003).

Overall, we compute 9 features and 31 POS tags features. For each feature, we compute per-page count, normalize it by page length, and compute average per class. Table 3 reports ratio of averages between evidential and nonevidential pages. Ratios >1 indicate features more prevalent in evidential pages, while ratios <1 denote features more prevalent in nonevidential pages. *Note that the table only shows features with values that are significantly different across the two classes* ($p < .05$ using two-tailed *t*-test). Additionally, the table shows *translated* examples from the Arabic pages.

We gauge power of each feature in discriminating among evidential and nonevidential classes by computing Kendall's τ correlation. τ is computed between two lists: Feature value and page label for all pages. Rank of pages in the search result list per claim was used to break ties in both lists of scores.

Results in Table 3 strengthen our conclusion in **RQ2**. Features capturing reported speech, named entities, and quotes were most indicative of page usefulness. POS tags also exhibited similar trend. Participles describing actions (e.g., "saying") and past-tense verbs (e.g., "stated") were more prevalent in evidential pages showing a tendency to refer to and report information. We also observe that language in nonevidential pages was more subjective or opinionated with more use of comparative adjectives, which has been shown to be a strong indicator of opinions (Liu, 2010). However, this feature is not correlated to page label. Stronger correlation was found between noun quantifiers and page evidentiality. Interestingly, nonevidential pages are longer on average, also showing more lexical redundancy with less unique words. Closer inspection of nonevidential pages showed that several of them were actually directory pages that list a summary of many pages including one that is relevant to the claim. Some pages were forum pages with long discussion threads. Other pages were long articles covering a very general topic, in which the claim's topic is a subtopic.

7 | A PROOF-OF-CONCEPT: EVIDENTIAL PAGES RETRIEVAL MODEL

Following the prior identification of features that can characterize evidential pages, we now study their effectiveness by implementing a ranking model, as a proof-of-

TABLE 3 Relationship of page usefulness and linguistic features. Ratios indicate how frequently a feature appears in evidential pages compared to nonevidential

Feature	Ratio	τ	Example
POS-active & passive participles	1.49	0.14	Prince Mohammad pointed out … saying, "But we have learnt from previous experience …"
Quotes	1.36	0.15	Described as "a new Hitler in Middle East"…
Statements	1.34	0.12	He announced the news on Twitter saying: …
Entities	1.19	0.12	HFPA has announced … Golden Globe Awards
POS-verb (past)	1.10	0.03	Macron pledged economic reforms …
Unique tokens	1.06	0.09	—
POS-verb (present)	0.90	−0.07	… enables the immune system to …
POS-adjective (comparative)	0.69	−0.03	Most common question is why…
Length	0.66	−0.13	—
POS-noun quantifier	0.66	0.13	Most operating systems uses a GUI…
POS-verb (command)	0.26	0.41	Put half a teaspoon of olive oil and mix.

Note: Examples show translated text from the pages matching features.
Abbreviation: POS, part-of-speech.

concept, for evidential pages retrieval that employs those features. The model re-ranks the same documents returned by the search engine allowing for comparison between the two scenarios.

7.1 | Features and classifiers

The supervised model integrates features from Table 3 using traditional machine learning models. Additionally, we experimented with the rank returned by the search engine as a feature that somewhat indicates the relative page relevance.

We experiment with three models: Random forest (RF), logistic regression, and multilayer perceptron. For all classifiers, prediction probability of the positive class (i.e., probability that the page is evidential) is used to rank pages per claim.

7.2 | Dataset

We train models over the CT19-T2 dataset extended with 10 claims (and corresponding annotated pages) of the train set from the CheckThat! lab (Elsayed et al., 2019; Hasanain et al., 2019) to increase the number of training examples. We only consider the topically relevant Web pages from the ground truth, since we previously carried our analysis on relevant pages which resulted in proposing features that differentiate between evidential and nonevidential *but topically relevant* pages. We run experiments using 69 claims and 1,314 relevant pages (half of which are evidential).

7.3 | Experimental setup

We use the default parameters provided by scikit-learn Python library[3] for the classifiers. We follow a leave-one-claim-out cross-validation setup; we elected to do so due to the relatively small dataset size. We report average performance over folds (i.e., claims). Runs were evaluated using the aforementioned recall measure. We also report precision and average precision at the top 10 ranks to support other potential user models.

7.4 | Results

Table 4 shows performance results of seven models:

- The search engine (SE) (i.e., Google) baseline.
- The three traditional learning models using our 11 features.
- The three traditional learning models using our 11 features and the original rank feature (given by the search engine). We hypothesize this feature can capture relative topical relevance of the page following Google's scoring model.

Table 4 shows all the models only employing our proposed features had comparable or higher performance scores compared to the search engine, with and increase up to 2.5 and 6% by RF model measured by R@10 and MAP@10, respectively. Interestingly, adding the SE rank feature to the proposed features shows further increase in scores over using the features alone, with an increase up to 5.3 and 9.5% over the search engine performance measured by R@10 and MAP@10, respectively. This suggests that the proposed features are better coupled with features capturing Web page topical relevance. While these results demonstrate potential effectiveness of the proposed features, the difference in performance over the search engine model was not statistically significant ($p < .05$, two tailed paired t-test), indicating that more effective features are needed to attain higher improvements.

TABLE 4 Evidential retrieval model performance

Model	Features	R@10	P@10	MAP@10
SE	SE rank	0.599	0.537	0.539
LR	11 features	0.597	0.543	0.561
MLP	11 features	0.600	0.544	0.559
RF	11 features	0.614	0.548	0.571
LR	11 features + SE rank	0.589	0.548	0.573
MLP	11 features + SE rank	**0.631**	**0.560**	**0.588**
RF	11 features + SE rank	0.618	0.562	0.590

Note: Results for best model by R@10 are boldfaced.

Abbreviations: LR, logistic regression; MLP, multilayer perceptron; RF, random forest; SE, search engine.

8 | ARE WE THERE YET?

Previous sections provided insights on some features to consider when designing systems for retrieving evidential pages. We wonder if current systems, designed for reranking Web pages for usefulness, are good enough (**RQ4**). Although the problem of re-ranking pages by usefulness for claim verification is relatively new to the automated fact-checking domain (Yasser et al., 2018), there is already some effort in literature to design such systems, in particular in Subtask A of Task 2 of the CheckThat! lab at CLEF 2019. These systems were also evaluated using the dataset we are proposing in this work: CT19-T2 (Hasanain et al., 2019).[4] It is worth noting here that in Subtask A, the task required systems to rank *potentially relevant* Web pages by usefulness which is slightly different from the task as defined in the previous section where the system ranked *relevant* Web pages (given ground truth). Table 5 compares performance of the best run submitted to the lab (CLEF-Best) against two other runs. The first is the original ranking returned by Google (SE), representing the performance of existing search engines for the task. The other is an Oracle run that perfectly re-ranks pages retrieved by SE by placing the evidential pages at the top of the list. This Oracle run is indeed a "cheating" run that knows the labels of the pages and orders pages using these labels. The goal of that run is to establish an upper bound for usefulness-oriented retrieval systems on *this dataset*.

Systems were evaluated using recall, precision, and average precision at 10. We also report statistical significance of performance difference between the Oracle and the other runs ($p < .05$, two-tailed paired t-test).

Results demonstrate that a significantly large performance improvement can potentially be attained by search engines or existing fact-checking systems. In the modern age of rapid spread of fake news, efficient fact-checking is a primary goal (Sharma et al., 2019). More emphasis should be given to designing systems that provide the users with a short but highly effective list of evidential pages. Such list will help the user (and a fact-checking system) reach a fact-checking decision faster since she only need to look at few documents to make a decision as opposed to having a longer list of topically relevant but not fully evidential documents.

TABLE 5 Performance of retrieving evidential pages

Run	R@10	P@10	MAP@10
CLEF-Best	0.48	0.40	0.45
SE	0.54	0.42	0.49
Oracle	$0.80^{a,b}$	$0.63^{a,b}$	$1.00^{a,b}$

[a] Oracle scores indicating statistically significant difference from search engine.

[b] Oracle scores indicating statistically significant difference from CLEF-Best.

9 | CONCLUSION AND FUTURE WORK

In this analytical study, we employed several features to characterize differences between evidential and non-evidential Web pages in the context of fact checking. Furthermore, we showed that those features have some potential by leveraging them in a learning model for evidential pages retrieval. We also examined the performance of existing search systems in retrieving such pages. Our main aim was to provide insights on how to better design usefulness-oriented search systems for claim verification. Our study has showed that: (a) Topically relevant pages retrieved by a search engine do not always contain evidence needed to verify the given claim, (b) performance of an effective commercial search engine is different in usefulness retrieval compared to topical relevance retrieval and the system performance is weakly correlated, (c) most evidential pages include reported statements from sources, quotes, and entities; these linguistic cues are strong predictors of page usefulness, and (d) Significantly large performance improvements can be attained to better support evidential page retrieval.

There are several potential directions for future work. We are interested in a more thorough textual analysis using more sophisticated features such as subjectivity (Abdul-Mageed et al., 2014). Investigating other aspects of evidence, such as reliability, is another interesting direction.

ACKNOWLEDGMENTS

This work was made possible by NPRP grant# NPRP 11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. Open access funding is provided by the Qatar National Library.

ENDNOTES

[1] This work is a significant extension to an earlier work (Hasanain et al., 2019). The current study has a completely different objective which is to study effectiveness of Web search engines in retrieving evidential Web pages in the fact-checking domain. In this work, we also provide a thorough analysis of the quality of the dataset that was initially constructed in our earlier work.

[2] http://qufaculty.qu.edu.qa/telsayed/datasets.

[3] https://scikit-learn.org/.

[4] Summary on these systems is presented in the related work section.

REFERENCES

Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, *28*(1), 20–37.

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for trec relevance assessment. *Information Processing & Management*, *48*(6), 1053–1066.

Al-Rfou, R., Kulkarni, V., Perozzi, B., & Skiena, S. (2015). Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of SDM'15*.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Grue Simonsen, J. (2019). Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *Proceedings of EMNLP'19*.

Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., & Nakov, P. (2018). Integrating stance detection and fact checking in a unified corpus. *Proceedings of NAACL-HLT'18*.

Carterette, B. (2009). On rank correlation and the distance between rankings. *Proceedings of SIGIR'09*.

Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., & Tannier, X. (2018). A content management perspective on fact-checking. *Companion Proceedings of WWW 2018*.

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. *Proceedings of ACL'14*.

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2020). Fake news types and detection models on social media a state-of-the-art survey. In P. Sitek, M. Pietranik, M. Krótkiewicz, & C. Srinilta (Eds.), *Intelligent information and database systems*. Springer.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2015). Trec 2014 web track overview. *Proceedings of TREC'14*.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of ASIS&T'15*.

Damessie, T. T., Nghiem, T. P., Scholer, F. & Culpepper, J. S. (2017). Gauging the quality of relevance assessments using inter-rater agreement. *Proceedings of SIGIR'17*.

Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., & Atanasova, P. (2019). Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims. In *Experimental IR meets Multilinguality, multimodality, and interaction* (pp. 301–321). Springer.

Favano, L., Carman, M. & Lanzi, P. (2019). TheEarthIsFlat's submission to CLEF'19 CheckThat! challenge. *CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org*.

Ferreira, W., & Vlachos, A. (2016). Emergent: A novel data-set for stance classification. *Proceedings of NAACL-HLT'16*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). Ukp-athene: Multi-sentence textual entailment for claim verification. *Proceedings of FEVER Workshop 2018*.

Haouari, F., Ali, Z. & Elsayed, T. (2019). bigIR at CLEF 2019: Automatic verification of Arabic claims over the Web. *CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org*.

Harman, D. (1992). Overview of the first text retrieval conference (TREC-1). *Proceedings of TREC 1992*.

Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeno, A. & Nakov, P. (2019). Overview of the CLEF-2019 CheckThat! Lab on automatic identification and verification of claims. Task 2: Evidence and factuality. CLEF 2019 Working Notes. *CEUR Workshop Proceedings, CEUR-WS.org*.

Hoeken, H., & Hustinx, L. (2009). When is statistical evidence superior to anecdotal evidence in supporting probability claims? The role of argument type. *Human Communication Research*, *35*(4), 491–510.

Jiang, J., He, D., & Allan, J. (2017). Comparing in situ and multidimensional relevance judgments. *Proceedings of SIGIR '17*.

Jiang, S., Baumgartner, S., Ittycheriah, A., & Yu, C. (2020). Factoring fact-checks: Structured information extraction from fact-checking articles. *Proceedings of The Web Conference 2020*.

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of HCI'18*.

Johnson, F., Rowley, J., & Sbaffi, L. (2016). Exploring information interactions in the context of google. *Journal of the Association for Information Science and Technology*, *67*(4), 824–840.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau, *Handbook of Natural Language Processing* (Vol. 2, pp. 627–666). Chapman and Hall/CRC.

Ma, J., Gao, W., Joty, S., & Wong, K.-F. (2019). Sentence-level evidence embedding for claim verification with hierarchical attention networks. *Proceedings of ACL'19*.

Malon, C. (2018). Team papelo: Transformer networks at FEVER. *Proceedings of FEVER Workshop 2018*.

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., & Luo, H. (2016). When does relevance mean usefulness and user satisfaction in web search? *Proceedings of SIGIR'16*.

Nguyen, A. T., Kharosekar, A., Lease, M. & Wallace, B. (2018). An interpretable joint graphical model for fact-checking from crowds. Proceedings of AAAI'18.

Nie, Y., Chen, H., & Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. *Proceedings of AAAI'19*.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Pehcevski, J. & Thom, J. A. (2007). Evaluating focused retrieval tasks. SIGIR 2007 Workshop on Focused Retrieval.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. *Proceedings of EMNLP'18*.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., & Slonim, N. (2015). Show me your evidence-an automatic method for context dependent evidence detection. *Proceedings of EMNLP'15*.

Roitman, H., Hummel, S., Rabinovich, E., Sznajder, B., Slonim, N. & Aharoni, E. (2016). On the retrieval of Wikipedia articles containing claims on controversial topics. *Proceedings of WWW'16 Companion* (pp. 991–996).

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144.

Sathe, A., Ather, S., Le, T. M., Perry, N., & Park, J. (2020). Automated fact-checking of claims from wikipedia. *Proceedings of LREC'20*.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, *10*(3), 21:1–21:42.

Tawileh, W., Mandl, T., Griesbaum, J., Atzmueller, M., Benz, D., Hotho, A., & Stumme, G. (2010). Evaluation of five web search engines in arabic language. *Proceedings of LWA2010* (pp. 221–228).

Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *Proceedings of NAACl-HLT'18*.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of NAACL-HLT'03*.

Trielli, D., & Diakopoulos, N. (2019). Search as news curator: The role of google in shaping attention to news information. *Proceedings of CHI'19*.

Unkel, J., & Haas, A. (2017). The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology*, *68*(8), 1850–1862.

Vakkari, P. (2020). The usefulness of search results: A systematization of types and predictors. *Proceedings of CHIIR '20*.

Vakkari, P., Völske, M., Potthast, M., Hagen, M., & Stein, B. (2019). Modeling the usefulness of search results as measured by information use. *Information Processing & Management*, *56*(3), 879–894.

Wang, X., Yu, C., Baumgartner, S. & Korn, F. (2018). Relevant document discovery for fact-checking articles. *Companion Proceedings of WWW 2018*.

Wicaksono, A. F., & Moffat, A. (2020). Metrics, user models, and satisfaction. *Proceedings of WSDM '20* (pp. 654–662).

Yasser, K., Kutlu, M., & Elsayed, T. (2018). Re-ranking web search results for better fact-checking: A preliminary study. *Proceedings of CIKM'18*.

Yigit-Sert, S., Altingovde, I. S., Macdonald, C., Ounis, I., & Ulusoy, O. (2020). Explicit diversification of search results across multiple dimensions for educational search. *Journal of the Association for Information Science and Technology*, *72*(3), 315–330.

Yilmaz, E., Aslam, J. A., & Robertson, S. (2008). A new rank correlation coefficient for information retrieval. *Proceedings of SIGIR'08*.

Zhang, Y., Ives, Z., & Roth, D. (2019). Evidence-based trustworthiness. *Proceedings of ACL'19*.

How to cite this article: Hasanain, M., & Elsayed, T. (2022). Studying effectiveness of Web search for fact checking. *Journal of the Association for Information Science and Technology*, *73*(5), 738–751. https://doi.org/10.1002/asi.24577