Using a Credibility Classifier to Improve Health-Related Information Retrieval

by

Fuat Can Beylunioglu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Fuat Can Beylunioğlu is the sole author for Chapter 1, 2, and 5, which were written under supervision of P. Robert Duimering and Mark D. Smucker.

Chapter 3 and Section 4.1 contain materials from a study presented in Text Retrieval Conference (TREC) Decision Track 2019, which was co-authored with Mustafa Abualsaud, Mark D. Smucker and P. Robert Duimering (Abualsaud et al., 2019).

Exceptions to sole authorship of materials are as follows:

ClueWeb12-B13 collection used in this thesis was indexed by Mustafa Abualsaud, who also computed the baseline BM25 retrieval scores. The algorithms proposed in this thesis employ supervised classifiers to modify the baseline retrieval scores, which was trained and tested on two respective annotated corpora. Mustafa Abualsaud carried out the necessary computation to retrieve the unannotated documents contained in these corpora, modified the software (HiCAL) to be used to annotate raw web documents and administrated the server that ran the software.

Abstract

In this thesis, we address improving the credibility and correctness of information retrieved by search engines in health-related searches. Health misinformation presented in the search engine results pages (SERPs) is a challenging problem to search engines whose successes have been measured with the number of URLs in the SERPs relevant to the user's query. However, research shows that relevant but inaccurate information can lead to wrong decisions, which is a challenge to the current search engines. Although existing studies have proposed different ways to help to make better health decisions, there is not much done in the information retrieval context.

In our study, we proposed algorithmic methods to improve correct and credible information presented in the results pages. The algorithms are motivated by the hypothesis that credibility of a document correlates with its correctness. Therefore, we trained classifiers to predict the credibility of documents retrieved by a search engine and adjust their ranks based on the credibility and spaminess scores. To test the performances of the algorithms, we have conducted an experiment as a part of our participation in TREC Decision Track 2019. As we show in this study, we can significantly improve the baseline BM25 algorithm in credibility and correctness tasks. We also present an analysis of the credibility and correctness judgments produced for the track to give insight into the distribution of credibility and correct documents retrieved in health-related tasks. Our analysis suggests that credibility can help to reach accurate information when the underlying treatment is ineffective, but there is a limit to its contribution to users' search experience.

Acknowledgements

First and the foremost, I would like to thank my supervisors, P. Robert Duimering and Mark D. Smucker, for their limitless patience and priceless effort in shaping my understanding and my study. I truly appreciate their time for all lengthy discussions that brought me more scientific and intellectual point of view, though this thesis presents only a humble proportion of it. I cannot thank enough for their tolerance for my persistence in repeating the same mistakes.

I would like to thank Mehrdad Pirnia and Olga Vechtomova for their valuable feedback. I appreciate their efforts to read my lengthy study carefully during this these unusual times.

I would like to express my gratitude to my beloved friends, Elif and Yusuf Altındal, for their support during my long stay at their home prior to the TREC submission, which is central to my thesis. I appreciate their hospitality and feel thankful to them for unconditionally opening all of their resources knowing that I would consume all coffee reserves and damage their properties. I owe them lots of thanks and a french press.

I also want to thank my friends and their families living in KW/Guelph region who supported me when adapting to this new environment. I truly feel thankful to them for their succour.

Lastly, I want to acknowledge Natural Sciences and Engineering Research Council of Canada for financial support (Grant RGPIN-2014-03642).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

NIST National Institute of Standards and Technology

TREC Text Retrieval Conference

nDCG Normalized Discounted Cumulative Gain

MAP Mean Average Precision

Chapter 1

Introduction

Search engines are traditionally engineered to satisfy users' information needs by retrieving relevant information to the user's query, and today's search engines are quite successful in this task. Accordingly, the algorithms via some complex content analysis, find and rank documents based on their similarities with the user's query (e.g. BM25), modifies the ranks with measured importance of documents via link analysis (e.g. PageRank) and then with other users' search history (i.e. clickthrough behaviour).

This very well engineered system has proven to be successful and satisfy users' health-related information needs. According to a 2013 Pew research, 59% of Americans reported using the internet for health information, and 77% of online health seekers start searching with one of the major search engines (Fox and Duggan, 2013). Health information seekers use search engines for learning more about the diseases, self-diagnosis, seeking advice and looking for treatment options (De Choudhury et al., 2014). Research also shows that users were more confident when communicating their issues to medical experts after an internet search. On the other hand, some studies found that reading online health information can contribute to confusion, anxiety and depression (Samal et al., 2011; Bessière et al., 2010; Helft et al., 2005), which may lead to overestimating the outcomes of symptoms and cause negative emotional outcomes (White and Horvitz, 2009).

Search engine interaction literature presents evidence that users are not interpreting the search engine result pages in the same way its engineer interprets. When given too much information to process, the users naturally enact cognitive heuristics, which may or may not be suitable for the given situation. For example, research shows that users attribute the top results superficial importance, their decisions are correlated with the order and frequency of the information presented, and they overinterpret the information that they could reach via search engines (Lau and Coiera, 2007; Pogacar et al., 2017; White, 2014).

How to control such influence on health information seeker to avoid undesired outcomes still remains an unsolved problem. One approach could be triggering the searcher's skepticism or accuracy mechanism to help to make unbiased decisions. The studies in confirmation bias and selective exposure theory highlight some useful aspects of human decision making, yet the literature is far from giving clear guidelines (Hart et al., 2009). Nevertheless, research shows in that searcher experience can be improved by promoting the ranks of credible and correct information retrieved by search engines (Pogacar et al., 2017). On the other hand, the studies in IR literature has focused on removing spam to improve searcher experience Davison (2006); Cormack et al. (2011). The studies in the last decade have examined the spreading of misinformation (Vosoughi et al., 2018) and classify credible sources (Castillo et al., 2011; Gayo-Avello et al., 2013), fake news classification

(Shu et al., 2017; Conroy et al., 2015) but, to the best of our knowledge, there has not been much done to improve correctness and credibility of information presented in search engine results pages (SERPs).

In this thesis, we propose algorithmic methods to improve the ranks of the credible and correct health information and lower the positions of inaccurate information presented in the SERPs. Our methods are motivated by the hypothesis that credibility correlates with the correctness of the information, and by targeting credible documents, it is possible to improve the accuracy of the information presented in the results page. Therefore, we trained classifiers to measure the credibility of documents and adjusted the ranks based on their credibility scores.

We tested these methods' performance by participating in the Text Retrieval Conference (TREC) Decision Track 2019, a workshop that aims to improve users' experience in health searches. In the track, participants ran their algorithms for a given set of queries and a collection of documents, and the organization labelled documents retrieved by the participants' runs based on their topic relevance, correctness and credibility. The results indicate that we can improve our baseline BM25 run significantly in retrieving the correct and credible information. In this thesis, we also present further analysis of the labelled collection to give insight into the correlation between credibility and correctness of the information and the present state of the health-related information online.

We contribute the previous literature by;

- developing algorithms that can significantly promote credible and correct information in the SERPs,

- measuring how much the credibility of information correlates with its correctness,
- giving insight to assessment errors and its impact over the research,
- providing an analysis of the judgments produced for the TREC Decision Track.

We will start by reviewing the literature in the following chapter, which addresses the credibility and correctness in health searches from different domains, including Information Retrieval, Communication, Psychology and Human-Computer Interaction. In the third chapter, we will outline the methodology by giving the details of the experiment and the algorithms that we proposed. We will present the algorithms' performances, analyze the judgments produced for the track, and the assessment errors in the fourth chapter. We will then discuss our findings and conclude in the fifth chapter.

Chapter 2

Background & Related Work

In our study, we aim to improve search engines by promoting credible information under the hypothesis that credibility correlates with the accuracy of the information. Therefore, our research lies in between the watershed of several research fields, including communication and psychology, decision sciences and computer science. Related studies to our thesis highlight how the user perceives the credibility of the information presented online, the complexities between the user and the search engine and bias during the health searches. We will review these studies under five sections and conclude this chapter with a summary of the literature.

2.1 Credibility of Web Content

Traditional approaches to credibility view it as what is known today as *source credibility*, that is, the believability of the source, which relies upon two notions (i) trustworthiness and (ii) expertise (Hovland et al., 1953; Wierzbicki, 2018). Practically, it requires a unique source, e.g. author that is supposed to be knowledgeable enough and trusted not to deceive the information receiver. This definition functioned well in the 20th-century environment where information production was expensive, limited to very few mainstream sources, and therefore the content could easily be associated with one and a unique source. The credibility of sources could be maintained by ensuring proper training of expertise and investigating conflict of interest.

Researchers argue that in the age of networked technologies, traditional ways will not be a solution to credibility (i.e. by examining source credibility) (Wierzbicki, 2018; Warnick, 2004). First, there is an enormous number of content writers online, and gatekeeping information is not feasible (Metzger and Flanagin, 2013; Rieh and Danielson, 2007). Second, the information cannot be attributed to a single source in today's internet. The content writers are often not the source of the work but are the ones who gather information into a single body of text (Warnick, 2004). The content is also developed interactively.

As a more compatible definition, the literature refers to the three types of credibility also proposed by Hovland and Weiss (1951), that is source credibility, reputation (or also known as medium credibility) and message credibility (Kalbfleisch, 2003; Wierzbicki, 2018). Medium credibility refers to the degree of trust in the medium that conveys the information, such as newspapers, web-portal or social media. Message credibility, on the other hand, refers to the information quality, including technical usage of language, completeness of information, etc. Wierzbicki (2018) modifies message credibility definition by Hovland and Weiss (1951) with a new definition influenced by Information Theory by Shannon (1948). Accordingly, credibility is "a signal that may make a receiver believe that the information is true" (page 41). The receiver wants the information to be *true*, and the sender wishes it to be accepted or consumed.

While the above distinctions can be elaborate, it may not be practical. Warnick (2004), by following Barthes (1977), suggests distinguishing *work*, a piece that can be associated to a person, from *text*, that is "a multi-dimensional space in which a variety of writings, none of them original, blend and clash" (Barthes, 1977). Warnick (2004) argues that the latter is more appropriate in case of web-credibility, in which the content may or may not be traced back to its origin, whereas the former could be used for traditional information view of credibility. Similarly, Fogg and Tseng (1999) define credibility of online information as a perceived quality which "does not reside in an object, a person, or a piece of information". Fogg and colleagues mention perceived credibility as measured in multiple dimensions, but reduce the dimensionality back to the two concepts of trustworthiness and expertise by Hovland et al. (1953). Here, trustworthiness refers to "well-intentioned, truthful, unbiased and so on", and expertise is defined by "knowledgeable, experienced, competent, and so on" (Fogg, 2003).

In this thesis, we will adopt a conception of credibility similar to Fogg, Warnick and Wierzbicki. The credibility as a property inherited in the document independent of the user is not achievable, instead, it is perceived by processing the text presented online (Fogg, 2003). The source information (e.g. author's credentials, website owner, domain, etc.) may not be accessible to information seekers; instead, they perceive a specific web document with all sorts of credibility signals embedded. The content writer wants the information to be consumed and therefore modifies the layout, design, language intensity, content, etc. for the message to be accepted.

2.2 Perceived Credibility Judgments

How information consumers evaluate credibility has been studied by a few researchers since the early 2000s. With the earlier concept of source credibility in mind, researchers were recommending users to evaluate accuracy (truthfulness of information), authority (expertise, credentials), coverage (comprehensiveness of information), objectivity and scope of the information (i.e. whether the information is up to date) (Metzger, 2007; Metzger and Flanagin, 2013).

An extensive telephone survey by Princeton Survey Research Associates (WebWatch, 2002) reported that the majority of laypeople evaluated the trustworthiness of information presented on the page, whether the information is relevant, is up to date, and the information sources were referenced. However, Fogg (2003) found that people mostly pay attention to webpage design and layout, rather than content. In this study, the researchers asked the participants to evaluate the credibility of 100 webpages and write down their opinion. The analysis of these comments suggests that people paid more attention to appearance (46.1%) and information design (28.5%) than perceived information focus/scope of the site (25.1%).

Other studies verify that users focus more on the feel and look of the website than the content. In an in-depth interview study with 17 women, the participants were asked to decide whether or not to receive hormone replacement therapy after an online search, and to write down the reasons why they chose or rejected a particular website (Sillence et al., 2007). The results show that they rejected the pages mainly (94%) due to design cues (poor interface, corporate feeling and pop-up/banner ads, confusing displays), or language intensity; they trusted the websites mainly (87%) due to content features (expertise and trustworthiness, consistency, accessibility of knowledge). In the study, participants preferred charitable websites over governmental and pharmaceutical ones, despite the latter are often thought more credible. Participants perceived pages as particularly credible based on unbiasedness and expertise. A group of studies by Metzger and colleagues suggest users do not verify the information they encounter online (Metzger, 2007), or superficially verify the information by visiting an insufficient number of websites that are not enough for verification (Metzger et al., 2010). Thon and Jucks (2017) examined the effects of author credentials and language use on the perceived credibility of webpages containing medical information. They showed that users evaluated a web page as credible if the content writer was a medical expert. However, when the language used was too technical to process the information, the page could not communicate credibility to the user and was perceived as non-credible.

Stanford et al. (2002) compared how experts and laypeople evaluated the credibility of health and finance pages. They found that medical experts paid more attention to the reputation of the website and author affiliation (43.9%), information source (25.8%) and website organizer (22.7%). Only 7.6% of the experts in the study reported that they looked to the design of the page (layout, colour schemes, typology) compared to 41.8% of laypeople. Similarly, Liao (2010) compared younger and older adults and found younger adults are better in finding content cues (weak vs. strong arguments), whereas older adults tend to focus on design, third party endorsements, etc.

Evaluation of information is content-dependent. It depends on the characteristics of the search, the person who is performing the search or the medium through which the information is received. Researchers argue that the criteria used to assess credibility depends on the type of website, and the "one-size-fits-all" approach does not work (Warnick, 2004). Hu and Sundar (2010) show that the perceived credibility of health information changes with respect to the information sources (e.g. Web-site, personal blog, social media, search engine, etc.) and whether the information is gathered by a layperson from different sources or is contained in the original source. Accordingly, webpages and bulletins are perceived to be more credible than blogs, but this is mediated by whether the source is original. Also, pages that present topics in complete detail are considered more credible. Lucassen et al. (2013) examined credibility evaluation of Wikipedia articles based on topic familiarity and information skills (level of education) and found that domain experts focused more on content than non-experts and users with better information skills paid more attention to surface features than to semantic features.

There are a number of theories proposed to model users' perceived credibility. Fogg and colleagues at Stanford Web Credibility Project conducted experiments (Fogg and Tseng, 1999; Fogg et al., 2001; Fogg, 2003) and provided guidelines for web developers as to designing websites that will be perceived as credible[1]. These guidelines summarize their findings, e.g. "Make it easy to verify the accuracy of information", "Show the organization behind your site", "Avoid spellchecks", "Update your website often" and so on. As discussed above, this approach is not appealing since it fails to capture variability among different types of websites and the underlying tasks.

Wierzbicki (2018) reviews several multilayered models, which are based on evaluating

[1] The guidelines were displayed at http://credibility.stanford.edu/ but currently not available. It can be found in Wierzbicki (2018), page 45.

the design and layout cues vs. content, source vs. message credibility, and peripheral cues vs. central cues. The Iterative model by Wathen and Burkell (2002) asserts that users evaluate webpages iteratively, first based on surface cues, then based on source and message credibility, and lastly, by checking whether the information is compatible with their own beliefs. The Prominence - Interpretation Theory by Fogg (2003) suggests that users evaluate perceived credibility in two recursive steps. In the first step (prominence), users pick signals based on page design and their involvement in the content; then in the second step, they interpret these signals at a deeper level. The theory suggests that these steps are recursive and simultaneous, so that the reader may find new cues (first step), re-interpret them and adjust their justifications (second step). The Dual Processing Model by Metzger (2007) modify Prominence-Interpretation by taking account of cognitive processes. Accordingly, in the first step (Heuristic/Peripheral), users focus on peripheral cues (design & layout) by using heuristic strategies (we will cover below) and then they evaluate these cues more deeply in the second step (Systematic/Central). However, if the user does not have enough motivation or involvement in the content, the second step is aborted. Wierzbicki's review also includes the MAIN Model (Sundar, 2008) and Ginsca Model (Ginsca et al., 2015), which summarize various evidence regarding user credibility judgments from the literature, which they combine into a few latent features.

An important point that deserves attention has to do with the relationship between perceived credibility and the relevance of a web document to the information need. The studies in the literature found that relevance is a factor of credibility evaluation that is embedded in peripheral and deeper judgments. In the studies by Fogg (2003), Sillence et al. (2007), Kalbfleisch (2003), and Metzger (2007), the users dismissed pages as noncredible if the information was not fully relevant to their information need, incomplete or if they did not cover all aspects of the issue. Therefore, the above models include relevance as a part of surface cues or deeper level interpretation cues used to evaluate credibility. Related to this, the Predictive and Evaluative Model by Rieh (2002) asserts users anticipate (predictive step) that the page will satisfy their need and choose to visit the page, and then evaluates the credibility of the content. Here, the way *need* is defined does not separate relevance from credibility. This is because the users that seek help for a decision hope that the content will not deceive them into the wrong action and satisfying the information need requires the information to be credible.

Although there are differences in nuances, the preceding studies and models highlight similar aspects of user behaviour, which is vital for our goal. (1) The literature emphasizes field and case dependency of credibility evaluations and warns against the rule-based approaches (such as the guidelines given by Fogg and colleagues). Also, the studies treat credibility as a latent variable that can be constructed as a combination of different cues. Lastly, relevance and credibility are tightly coupled so that users do not evaluate them separately, but as latent cues for their choices. These aspects favour a case-dependent method that can capture various cues embedded in the document. In this study, we use logistic regression to capture these signals, measure credibility as a latent variable and blend it with retrieval scores.

2.3 Misinformation, Cognitive Heuristics and Biases

Overall the perceived credibility research found that the information searchers tend to give quick decisions based on superficial features of the website and rarely evaluate source and message credibility in depth. The studies even show that users rate the page's visual appeal in as little as 50 milliseconds (Lindgaard et al., 2006; Tractinsky et al., 2006). This fact led researchers to consider this role as *heuristics* or cognitive shortcuts, to reduce the effort and time in decision making. The concept was first proposed by Simon (1955), who suggested that human decision-makers are not perfectly rational while making decisions and distinguished between *economic man* and *administrative man*. The former describes a perfectly rational system that exhaustively evaluates all available information and stops when an optimal solution is reached, whereas the latter is subjectively, and boundedly, rational, operates over small samples of information and stops searching when a good enough solution is achieved. People behave like administrative man due to cognitive limitations of the information processing constraints of the problem environment (Simon, 1956). So we use cognitive shortcuts, or *heuristics*, to find solutions and make decisions in complex task environments.

In the modern age, when there is too much information to be evaluated online, which is growing and changing rapidly, source and message credibility evaluation may be unfeasible. Sundar (2008) proposed that this leads searchers to use cognitive shortcuts, and a blend of different strategies to find out whether the text is credible with less effort and time. These strategies do not necessarily focus on superficial cues; depending on the user's ability and involvement, she may find it easy to evaluate deeper cues (Metzger, 2007).

Metzger and Flanagin (2013) outline the evidence of cognitive heuristics that operate while making credibility decisions. Reputation heuristic refers to developing insights about credibility by evaluating the reputation of the content writer. People tend to trust familiar sources/authors or authoritative ones more than unfamiliar ones (recognition heuristic (Gigerenzer and Todd, 1999; Todd and Gigerenzer, 2000; Koh and Sundar, 2010)). Endorsement heuristic describes people's tendency to trust a page if it is recommended by someone they know (Metzger et al., 2010), or trusted by the people they like (Chaiken, 1987). Consistency heuristic refers to the user's tendency to trust if she can verify the information with other sources. However, in practice, users are satisfied after visiting very few websites (Metzger et al., 2010).

Self-confirmation heuristic is the tendency to believe information that supports their own prior beliefs. Researchers found evidence for a similar cognitive shortcut, expectancy violation heuristics whereby people tend to judge a website that violates their expectations (Metzger, 2007). The expectations can be due to finding more information on the website than expected, typo/spelling errors, design, etc. (Metzger et al., 2010; Fogg, 2003). Effort heuristic (Kruger et al., 2004) refers to people's tendency to trust information if they believe that the content writer put much effort into producing information. On the other hand, research shows that users dismiss websites if they find cues that indicate the website is biased (e.g. commercial information, ads, etc.) (Metzger et al., 2010).

Notice that the above strategies significantly reduce the time and effort needed to evaluate credibility by filtering alternatives, delegating judgment practices to people/organizations they trust, or simply by accepting the consensus view. Similarly, a user's previous beliefs can play as a knowledge-base, or she may find a page unlikely to be non-credible if the content writer has put a good deal of effort. On the other hand, it is not hard to see that a habit of overlooking details may lead to incorrect evaluations and poor decisions. A large body of literature shows how ready-made heuristics can transform into biases. Following Simon's works on bounded rationality, Kahnemann and Tversky showed heuristics do not always lead to "good" decisions, but may lead to erroneous judgments called *heuristic biases* including biases such as representativeness (Kahneman and Tversky, 1972), availability and anchoring (Tversky and Kahneman, 1973, 1974). For example, in one study (Tversky and Kahneman, 1974) had participants spin a fixed roulette wheel that stopped at either 10 or 65; and asked them to estimate the percentage of UN members in Africa. Participants showed systematic errors in their estimates, which were anchored the number that had tuned upon the roulette wheel. The median estimates for UN members were 25 and 45 percent if the roulette outcomes were 10 and 65, respectively. Here, notice that the roulette wheel and UN members in Africa have no relation, so systematic errors in outcomes are a product of malfunctioning humans' heuristics.

Credibility researchers also discuss similar effects. As Metzger and Flanagin (2013) discuss, the reputation heuristic is deeply connected to the *ad verecundiam fallacy*, that is trusting the opinion of an authority even if the topic is out of her level of expertise. Moreover, similar to a well-known conformity experiment by Asch (1951), users tend to trust the information (bandwagon heuristic Sundar (2008)) or adjust their credibility insights (Metzger et al., 2010) if they perceive that many others believe it to be correct. However, as Tversky and Kahneman (1973) show, making decisions based on immediate examples can easily lead to biased decisions (availability heuristic). Moreover, Ross et al. (1977) show that people tend to believe many other people share their beliefs, and therefore they are right.

Self-confirmation bias has been studied widely in an attempt to understand whether searchers perceive congenial information more favourable than information that contradicts their beliefs. There is evidence that people choose pre-decision supportive information when there are limits to information access (Fischer et al., 2005), there are many documents to evaluate (Fischer et al., 2008), and instructions are positively framed (Kastenmüller et al., 2009). A meta-analysis shows that people choose congenial information twice as often as conflicting information one on $odds\ ratio = 1.92, d = 0.36$ (Hart et al., 2009). Research also shows that defence mechanisms (a tendency to refute dissonant information) contribute to this bias, whereas accuracy motivation (the will to be truthful) did not reduce the bias and even increased it in some cases.

However, Hart et al. (2009) point out the complexity of cognitive biases and discuss inconsistencies among studies. In an attempt to explain the differences, Fischer and Greitemeyer (2010) argued that human information processors seek shortcuts to be able to process all available information. As the number of documents increases, users prefer "good looking" information and overlook others, but when there are not many alternatives, users do not have to put the effort that exceeds their limits and can make unbiased evaluations. Lewandowsky et al. (2012) address a similar question about why misinformation spreads reveal that people tend to accept the information that does not contradict with personal worldview or require extra effort, but "slips" into ones mind smoothly. Once misinformation is accepted, correction via retraction is ineffective because it causes missing links in the established memory; however, warning about misinformation prior to the information session motivates people to be more skeptical and guarded against misinformation.

Overall, the cognitive heuristics literature indicates the complexity of several mechanisms operating when evaluating information. Here cognitive heuristics can be regarded as ready-made strategies that humans learn over time or naturally have. Difficulties arise when such strategies are not appropriate to the given situation and lead to false impressions and judgments. Nevertheless, the literature indicates that internet users overlook details in cases when there is information overload, when the task is difficult, and when more in-depth evaluation requires too much effort. By combining the above literature with Metzger's (2007) Dual Processing Model, it would not be wrong to conclude that the users are intelligent enough to process information correctly. However, they do not process information in sufficient depth to evaluate central features but focus on peripheral cues if their skills or topic involvement do not reduce the effort, the page design is messy, or there is an information overload.

In the online health search context, there is no question that these biases will occasionally emerge. However, maintaining a better search environment for the user is possible with a proper IR system and a better design. Below we will review studies about cognitive biases occurring during search engine interaction and discuss what must be done for a better search experience.

2.4 Biases in Web Search

Today's search engines are quite successful in finding the most relevant information to satisfy users' information needs. Accordingly, the algorithms use complex content analysis to find and rank documents based on their similarities with the user's query (e.g. BM25), modify the ranks based on each document's measured importance via link analysis (e.g. PageRank) and based on other users' search history (i.e. clickthrough behaviour). This very well engineered system has proven to be quite successful and to satisfy users. However, the search engine interaction literature presents evidence that users do not always interpret the search engine result pages in the same way the engineer interprets them. For example, the top result's retrieval score can be 22.532, and the second's score can be 22.122, but users who do not see these scores may attribute the top result much more superficial importance.

A series of studies by White and colleagues (White, 2013, 2014; White and Hassan, 2014; White and Horvitz, 2015) show that in health searches, the search engine result pages are biased towards confirming treatment efficacy, and users who interact with search engines adjust their beliefs and decisions in favour of the treatment efficacy regardless of the underlying health issue or treatment. White and Hassan (2014) give three sources of this bias; (i) users favour information confirming intervention effectiveness over non-confirming (user behaviour), (ii) the majority of online health documents, prior to ranking, claim treatment efficacy (creators of page content), and (iii) indexing algorithms of the search engines embed these biases (system bias). White (2013) and Novin and Meyers (2017) also note that cognitive biases and search engine biases reinforce each other; the indexing algorithm learns from searchers' choices, and searchers incline to click on URLs at higher positions.

The literature has identified the various cognitive heuristic biases during the search. Studies report evidence of the *anchoring effect* as discussed in the previous section. White (2014); White and Horvitz (2015); Lau and Coiera (2007) found a significant correlation between the participants' pre-search beliefs and their post-search attitudes towards the treatment and showed that they adjust their beliefs rather than changing it. Closely connected to anchor bias, studies also showed that user's decisions are affected by order of the information accessed (Lau and Coiera, 2007; Novin and Meyers, 2017) (also known as order effect) so that documents displayed earlier in a session have more impact on user's decision. Moreover, the URL presented at higher positions in SERPs have relatively more weight on the user's evaluation (Lauckner and Hsieh, 2013; Pogacar et al., 2017).

Self-confirmation heuristics and selective exposure also operate during searches. Liao and Fu (2013) examined the effect of perceived threats and topic involvement on users' click behaviour and after-session decisions. When participants were shown disturbing images (e.g. a bleeding child) before the session, they selected information that was consonant with their attitudes; however, the effect was moderated if the participant had higher involvement in the underlying topic.

On the other hand, research also shows that search engines can influence the user towards correct information. In the study by Pogacar et al. (2017), when the search engine was biased towards correct information, the proportion of harmful decisions dropped significantly from 20% to 6%, and correct decisions increased from 43% to 70%. Another study manipulated a knowledge box appearing on the right-hand side of the SERPs with short information about vaccines and warning messages to improve users' knowledge about vaccines (Ludolph et al., 2016). When users were given easily comprehensible messages in knowledge boxes (i.e. from Wikipedia), their vaccine knowledge was higher than when they were given messages in a more technical language (e.g. WHO) or merely presented warning messages.

Overall, the search engine bias literature presents evidence about the presence of heuristic biases in searches. Research shows that users do not process all available information, but pay more attention to higher-ranked results, which have more relative impact on their decisions. Information displayed earlier in a session sets the mind and anchors how new information will be displayed. As in misinformation studies, it is hard to revise user interpretations afterwards, so giving warning signs with brief information about the issue can contribute to the user's knowledge.

Combined with other studies on heuristic bias, three important points come to light for developing better search engines. First, the information presented earlier in a session has relatively more impact on decisions, and it is indicative of how further information will be processed by anchoring to a particular state of knowledge. Second, cognitive heuristics come into play when there is too much information to be processed, the language is technical, or the page has cues indicating the information is decisive (ads, commercial motives). Lastly, the users are intelligent enough to process the information fairly, and their knowledge state can be improved under certain conditions.

In light of this knowledge, search engines can improve decisions by filtering low-quality information and improving ranks of easily processed correct information. Although there are few similar studies, the knowledge box can play an important role by setting up the state of knowledge before the user searches through natural search engine results.

2.5 Detecting Credibility

Early studies in IR addressed detecting a particular form of non-credible documents, spam, which manipulates not the user but the search engine itself. Spam can be classified into two main groups based on how they attempt to manipulate search algorithms, (i) content-based spam such as adding as many keywords as possible to a document, or creating a document by copying random content from various other sources, (ii) link-based spam is created by using link farms to inflate the documents PageRank score using phony links from other phony web pages, that is artificially made pages that are referenced from several other pages to propagate the PageRank score. These two types are usually orthogonal (Becchetti et al., 2008) so that combating one type does not affect the removal of the other.

Eiron et al. (2004) report that the top 20 PageRank results for each query contained around 11 adult content, which they suspect to be a result of link farms. To detect these superficial pages Davison (2000) proposed a machine learning algorithm that detects the link farms by training over 75 manually determined features. Castillo et al. (2007); Becchetti et al. (2008) statistically showed the link properties of natural and unnatural pages, domains and hosts, and developed classifiers to detect link farms. On the other hand, Gyöngyi et al. (2004) proposed a novel method, an algorithm called TrustRank, which is a particular form of biased PageRank that propagates information about the trustability of a webpage, from a list of seed trustworthy web pages to other topologically close pages. Developers inverted this principle to detect link spam, a method which was later called BadRank. Benczur et al. (2005) proposed a similar method to detect link spam that does not involve human intervention to determine seed documents. Wu et al.

(2006) showed that TrustRank overrepresents the seeds in the results and used a modified algorithm to overcome this bias. Davison (2006) combined link-based models to effectively remove link-based spam and promote trustworthy pages in SERPs.

Content-based spam is also addressed in the literature. Fetterly et al. (2005) proposed a method to detect a common type of web-spam generated by stitching texts from random websites together. Cormack et al. (2011) proposed a model based on previous work by Cormack (2007), which employs a logistic regression trained on byte 4-grams to reduce computational cost. The method later was used to give a spaminess score for each document in the ClueWeb12 collection[2].

In the last decade, these methods were applied to spam detection on social media, and some other works focused on classification. These studies include detecting spamers and link farms on Twitter (Benevenuto et al., 2010; Grier et al., 2010; Lee et al., 2010; Ghosh et al., 2012), distinguishing good quality user-generated content from abuse or spam (Agichtein et al., 2008), fake news detection (Zhou and Zafarani, 2018; Shu et al., 2017), controversy detection (Dori-Hacohen et al., 2015; Dori-Hacohen, 2017; Mejova et al., 2014), and detecting credibility of event news on Twitter (Castillo et al., 2011; Castillo Ocaranza et al., 2013; Morris et al., 2012).

2.6 Summary of the Literature

Studies in health IR and credibility indicate the gap between search engine developers' and user's influence from the SERPs. Traditionally, IR has been motivated by satisfying user's

[2] The scores can be downloaded from https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/.

information needs to fulfill the research goals. In the health context, it is finding a solution to a health issue, and verifying whether a treatment is a solution to the issue. However, when the only motivation is to retrieve documents relevant to the query (topic relevance), search engines fail to satisfy the need, which is to find accurate information.

The studies in search engine bias during health searches employ a common research design. The participants are asked their knowledge about the efficacy of a treatment for a health issue before and after interacting with a search engine. The evidence shows that users' decisions are primarily influenced by the URLs presented higher in the SERPs; they attribute superficial importance to the high-ranked documents. The outcome of this effect can be harmful and lead to undesired outcomes when the retrieved information is not accurate. Nevertheless, the effect is two-sided; when the users are presented with accurate information about a treatment's efficacy, the decisions can also be improved compared to the decisions before the interaction.

The reasons as to why users find the information presented in a webpage trustworthy is also studied by communication and psychology scholars. The studies show that the perceived credibility of a web document is affected by a number of factors from the user's level of expertise to the document type. However, many studies indicate that they do not evaluate the credibility of the information not through an in-depth analysis of the content, the authors' credentials and the organization behind the webpage, but by evaluating superficial cues such as page's design and layout, usage of language and the presence of ads. The models for perceived credibility suggest that users evaluate information in two recursive steps, first by noticing the cues for decision and then by deeper examination of these cues. However, depending on the user's information processing capacity, they may not go through the second step.

How deeper the information is processed, and thus the decision outcome, depends on the user's cognitive abilities and motivation. If they are presented with too much information than they can process, they use cognitive shortcuts to come up with a decision. Research shows that these shortcuts often are not compatible with the decision environment. For example, they prefer congenial information over uncongenial, overinterpret the friends' opinions, and verify the information by comparing it with an insufficient amount of other sources. Similarly, when they are presented 10 URLs in a SERP and buttons to browse other results, they tend to visit only the higher results and attribute superficial importance to this information.

These findings in the literature imply that search engines could be improved by increasing the density of good-quality information presented higher in the SERPs and demoting inaccurate or false information from the results. However, there is no simple method to ensure the correctness of retrieved documents. In this thesis, we are motivated by the hypothesis that one way to reach correct information is targetting credible documents. To the best of our knowledge, there is no study measuring the correlation between the two concepts, but only philosophical discussions by Wierzbicki (2018) who outline the conceptual relationship between credibility and the truth.

Lastly, we want to point out that many studies we outlined are not specifically targeting health-related web content, or usually are not about any specific domain. Therefore, some generalizability concerns may arise. In perceived credibility literature, there are some health specific studies (Stanford et al., 2002; Sillence et al., 2007; Thon and Jucks, 2017) and their findings are quite similar to the rest of the literature's. Our discussion about misinformation and cognitive biases, on the other hand, do not focus on a specific domain, whereas the search engine bias literature that we covered focuses only on health-related searches. The reader should note that the latter inherits the former and provides more evidence from another domain. Therefore, we believe it would not be wrong to suggest that these findings are not specific to any domain and generalizable to health-related information studies.

Chapter 3

Methodology

Our study is based on an experiment that we conducted as a part of Text Retrieval Conference (TREC) Decision Track 2019. The author of this study has involved during the design of the experiment and participated in the workshop. In this chapter, we will first outline the TREC workshop and give details about the 2019 Decision Track. We will then present our methods to improve search engines.

3.1 TREC

TREC is a conference involving a series of workshops to support developing and improving IR systems operating on gigantic document collections and is organized by the National Institute of Standards and Technology, NIST, an agency of the United States Department of Commerce. Traditionally, the main goal has been retrieving the most relevant information from a gigantic collection of documents (hundreds of millions) to help searchers satisfy their information needs. However, as with the increasing variability of the data that contains the information, the organizers added new tasks, known as tracks, to fulfill the goals in other data types such as Video Track, Question Answering Track, Web Track and News Track.

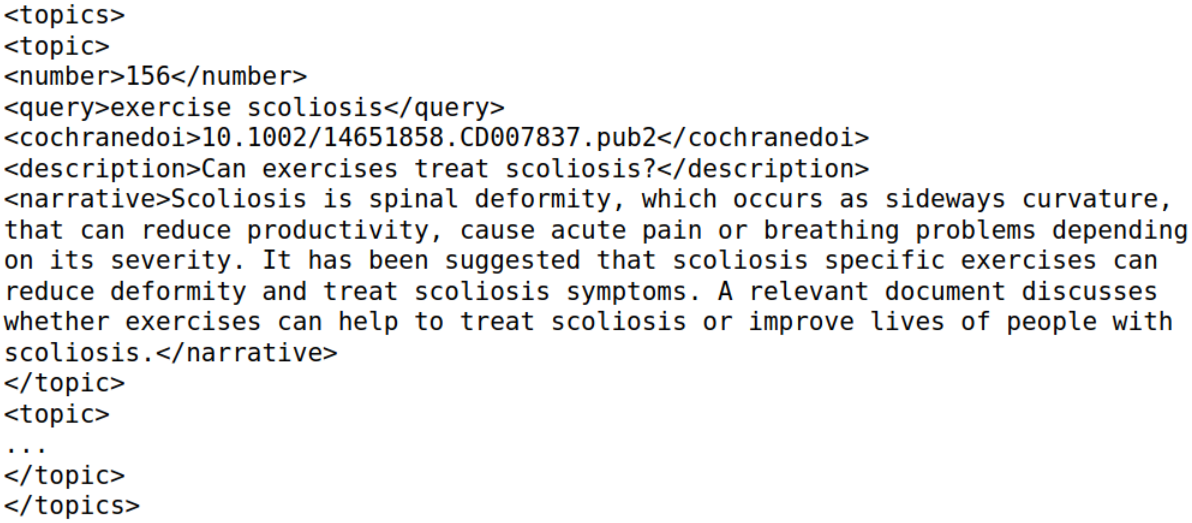
As described by Voorhees and Harman (2002), the prototypical task is similar to the case of a researcher performing a general literature search in a library who does not have a particular topic to be investigated. The task is to find the documents from the library's collection that satisfies the information need, ranked by their likelihood to be relevant to the query.

Although there are differences between the tracks, the main structure is as follows: (i) the organizers determine a collection of documents tagged with document IDs, (ii) then determine a list of topics to be searched, (iii) the participants submit files (or *runs*) involving document IDs ranked by relevance to the topics in decreasing order, (iv) the organizers pool the DOCIDs submitted by participants and use human assessors to label these documents (e.g. as relevant or not relevant to the topic), and (v) evaluate the runs submitted by the track participants based on the labelled documents. We will discuss the details of these steps below.

Collection: The corpus must reflect the task's nature in terms of the document types, language and distribution of documents relevant to the topics. Each document is labelled with a unique document id and is not cleaned or changed, even if they contain malware (e.g. SPAM Track collection). There are several datasets available on the TREC website (`https://trec.nist.gov/data.html`). However, TREC also used external collections for some of their tracks, for example, the ClueWeb12 dataset containing 733 million web documents used by different tracks, including the TREC Decision Track.

Topics: The information need of a searcher may vary in content, and the search engine must be able to respond to the variety. Traditionally topics of a track are listed in an XML file listing the details in four fields: Topic number, title, one-sentence description and narrative that details description in a small paragraph. From these narratives, the participants can generate their queries to search for the most relevant documents that can satisfy the information need. However, some tracks may give more details such as the specific query to be used for search sessions (e.g. in Decision Track). Figure 3.1 presents a sample XML file from Decision Track 2019[1].

Figure 3.1: An illustration of topic XML file from Decision Track 2019 website



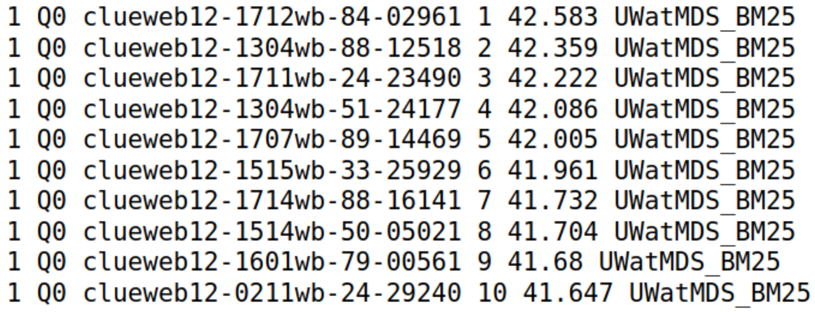
```
<topics>
<topic>
<number>156</number>
<query>exercise scoliosis</query>
<cochranedoi>10.1002/14651858.CD007837.pub2</cochranedoi>
<description>Can exercises treat scoliosis?</description>
<narrative>Scoliosis is spinal deformity, which occurs as sideways curvature,
that can reduce productivity, cause acute pain or breathing problems depending
on its severity. It has been suggested that scoliosis specific exercises can
reduce deformity and treat scoliosis symptoms. A relevant document discusses
whether exercises can help to treat scoliosis or improve lives of people with
scoliosis.</narrative>
</topic>
<topic>
...
</topic>
</topics>
```

Runs: With the collection and topics, the participants develop methods to find the

[1] https://trec-health-misinfo.github.io/2019.html.

documents that are most likely to be about these topics. The documents themselves are not submitted, but a file contains their unique identifiers (DOCID) sorted by their retrieval scores. These files contain the topic ID, iteration (always Q0 and is unused), document ID, likelihood score, rank and submission name in space-delimited format. Participants can submit multiple runs, and each will be evaluated separately. Figure 3.2 presents top 10 rows of a run.

Figure 3.2: Top 10 rows of a sample run

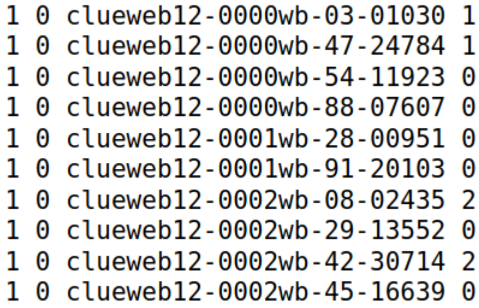


```
1 Q0 clueweb12-1712wb-84-02961 1 42.583 UWatMDS_BM25
1 Q0 clueweb12-1304wb-88-12518 2 42.359 UWatMDS_BM25
1 Q0 clueweb12-1711wb-24-23490 3 42.222 UWatMDS_BM25
1 Q0 clueweb12-1304wb-51-24177 4 42.086 UWatMDS_BM25
1 Q0 clueweb12-1707wb-89-14469 5 42.005 UWatMDS_BM25
1 Q0 clueweb12-1515wb-33-25929 6 41.961 UWatMDS_BM25
1 Q0 clueweb12-1714wb-88-16141 7 41.732 UWatMDS_BM25
1 Q0 clueweb12-1514wb-50-05021 8 41.704 UWatMDS_BM25
1 Q0 clueweb12-1601wb-79-00561 9 41.68 UWatMDS_BM25
1 Q0 clueweb12-0211wb-24-29240 10 41.647 UWatMDS_BM25
```

Pooling and Judgments: Once the participants submit their runs, TREC organizers have lists of documents to be assessed as to whether they are relevant to the topic. However, as the collection size has been too large, the judgment has been infeasible, with, e.g. 800000 documents to be assessed. Therefore the organizers prepare test collection via a technique known as *pooling*; that is selecting the top N documents (usually $N = 100$) per topic for each of the submitted runs (Voorhees and Harman, 2002). Once the test collection is prepared, the documents are distributed to judges to be assessed for relevance based on the instructions that are given to them.

Evaluation: After the assessment, the judgments are collected in a file known as *qrels*. This file contains all the information needed for measuring the performance of the runs, i.e. topic ID, iteration (almost always zero and is disregarded), document ID and relevance judgment. Figure 3.3 presents a sample qrels file.

Figure 3.3: Top 10 rows of a sample qrels file



```
1 0 clueweb12-0000wb-03-01030 1
1 0 clueweb12-0000wb-47-24784 1
1 0 clueweb12-0000wb-54-11923 0
1 0 clueweb12-0000wb-88-07607 0
1 0 clueweb12-0001wb-28-00951 0
1 0 clueweb12-0001wb-91-20103 0
1 0 clueweb12-0002wb-08-02435 2
1 0 clueweb12-0002wb-29-13552 0
1 0 clueweb12-0002wb-42-30714 2
1 0 clueweb12-0002wb-45-16639 0
```

The evaluation is made via trec_eval package that can be downloaded from TREC website[2]. The package evaluates a given run by comparing it with the qrels file based on a given measure. In particular, it reports performances based on various IR measures, including precision, recall, nDCG and MAP. We will give the details in the below section when we discuss the evaluation procedure of the Decision Track.

3.2 Decision Track

In 2019, researchers from the University of Waterloo, University of Queensland and University of Copenhagen, organized the TREC Decision Track to help develop IR systems that (i) promote correct, credible information and (ii) help users to make better decisions. Although the main goal was generic, the focus of the track in 2019 was health topics, i.e. improving the correctness and credibility of health information searches. The structure

[2] https://trec.nist.gov/trec_eval/

of the track is similar to traditional tracks as described above, but the submissions were evaluated not only for their topic relevance performances but also for their credibility and correctness. Below we will detail the topic, collection, judgments and evaluation steps.

3.2.1 Topics

51 health topics were determined, each in the form of [treatment] for [health issue] in [target group (if applicable)]; for example, "traction for lower back pain", "surgery for obesity", "antibiotics for wet cough in children". The topics were all positively framed, i.e. about whether or not a medical intervention *treats* a health issue. On the other hand, there was not any topic about potential harms, e.g. "does smoking kill", "does over consumption of eggs cause heart attack" etc. This format was chosen to be consistent with common research designs in the literature (see White (2013, 2014); Pogacar et al. (2017)).

The topics were selected from the Cochrane Review database[3] or from previous studies that used major search engine logs (White, 2013, 2014). A candidate topic list was created, which contained 431 topics. As with some previous studies, three types of topics were considered in terms of treatment efficacy, i.e. unhelpful, inconclusive and helpful. To represent these three types evenly, 17 topics per treatment efficacy type were identified. This list was determined based on current information available online, i.e. whether or not current major search engines can retrieve relevant documents and whether the topic is understandable by the layperson (e.g. excluded topics such as "surgery for small asymptomatic abdominal

[3] Cochrane Review database consists of medical article reviews written by experts; each is a detailed report about a treatment's efficacy for a given health issue. These reviews are highly reputable and considered to represent the current medical consensus.

aortic aneurysms").

For each of the 51 topics, underlying treatment efficacy was determined based on the Cochrane Review report discussing the topic. Each review has a summary of the review published on the Cochrane Review website, which includes an overview of the topic, the author's conclusion and the abstract of the report[4]. The scientific consensus about each treatment's efficacy was determined by reading the Main Results and Author's Conclusion sections. Accordingly, each topic was labelled as helpful or unhelpful if the underlying treatment is effective or ineffective, respectively, and labelled as inconclusive if there is no evidence or the evidence is not sufficient to decide. Table 3.1 shows the resulting topics together with the assigned intervention efficacy, unhelpful, inconclusive or helpful. Note that one of the topics (Topic 14) was excluded from the experiment by NIST due to unexpected challenges in assessments. In the rest of the study, we will refer to these 50 topics.

The topic XML file was prepared based on the determined list of topics and all supplementary information. The file contained a topic number, description, and narrative for each topic, along with the DOI number of the Cochrane Report discussing the treatment. The file also included the query to be used by participants to retrieve the documents. A sample of topics.xml file is presented in Figure 3.1.

[4] The reader can find an example from the following link: https://www.cochrane.org/CD005062/EPILEPSY_acupuncture-for-epilepsy.

Table 3.1: TREC Decision Track Topics

Query	Topic ID	Cochrane Conclusion	#Rel	#Cred	#Cor
cranberries for urinary tract infections	1	U	100	55	8
acupuncture for epilepsy	3	U	26	18	0
amygdalin for laetrile cancer	6	U	86	28	20
aspirin for vascular dementia	7	U	10	9	0
antidepressants for low-back pain	13	U	49	48	3
magnesium for muscle cramps	16	U	144	20	3
lumbar supports for lower back pain	17	U	72	32	0
electrical stimulation for male urinary incontinence	18	U	18	18	0
hydroxyzine for generalized anxiety disorder	22	U	8	1	0
ginkgo biloba for tinnitus	32	U	50	43	0
hypnotherapy for quit smoking	33	U	305	12	0
traction for lower back pain	38	U	98	57	48
cinnamon for diabetes	40	U	101	31	3
probiotics for eczema	42	U	72	45	1
vitamins for epilepsy	44	U	39	20	4
insoles for back pain	47	U	60	12	3
dehumidifiers for asthma	51	U	24	6	4
acupuncture for insomnia	2	I	220	195	0
honey for wound	4	I	111	65	6
ear drops for ear wax removal	9	I	39	35	7
gene therapy for sickle cell	10	I	43	42	31
breathing exercises for children with asthma	14	I	0	0	0
probiotics for bacterial vaginosis	15	I	52	5	13
acupuncture for vascular dementia	21	I	114	99	57
insulin for gestational diabetes	23	I	85	85	1
yoga for epilepsy	24	I	22	17	0
fish oil for ulcerative colitis	25	I	45	21	15
vaccine for common cold	26	I	11	9	3
aloe vera for wounds	30	I	116	106	0
exercise for hot flashes night sweats menopause	31	I	36	28	1
valerian for anxiety disorder	35	I	66	43	0
compression stockings for varicose veins	43	I	82	23	45
feverfew for migraines	46	I	72	9	44
acupuncture for asthma	48	I	137	10	12
acupuncture for migraine	5	H	85	34	56
melatonin for jet lag	8	H	107	38	91
exercise for lower back pain	11	H	190	112	181
circumcision for hiv	12	H	151	139	101
honey for cough in children	19	H	118	65	115
steroids for spinal cord injury	20	H	16	15	10
antibiotics for wet cough in children	27	H	5	2	3
antibiotics for whooping cough	28	H	78	50	26
antibiotics for children with pneumonia	29	H	108	79	94
exercises for female incontinence	34	H	95	30	60
dental sealants for cavities	36	H	166	86	166
laxatives for hemorrhoids	37	H	56	29	36
muscle relaxants for back pain	39	H	85	26	76
benzos for alcohol withdrawal	41	H	66	51	40
caffeine for asthma	45	H	40	18	29
sulfasalazine for rheumatoid arthritis	49	H	74	51	13
surgery for obesity	50	H	212	157	143

3.2.2 Collection and Runs

The document set used in the track was a widely used ClueWeb12-B13 that consists of 50 million web pages crawled in 2012 between February 10 and May 10. It can be requested from Lemur Project website[5]. Participants prepared and submitted their runs based on the topics.xml file and the document set. Each run file contains 1000 documents for each of the 50 topics in the same form as the sample file displayed above (Figure 3.2).

3.2.3 Assessment of the Documents

After runs were submitted, the organizers pooled documents as described above and constructed the collection of documents to be evaluated. Each document in this collection was evaluated for its relevance to the topic, credibility and the efficacy claim. The assessors were given the following instructions:

Relevance: According to the assessment guidelines given to NIST assessors, a document is Highly Relevant if it directly addresses the topic; for example, if the topic is "can cinnamon help to improve symptoms of diabetes", a document is "highly relevant" if it is dedicated to the topic and have a discussion as to whether cinnamon is helpful. On the other hand, a document is "relevant" if there is any piece of information that helps answer the question, e.g. a document discussing the benefits of cinnamon and has one sentence about its effect on diabetes. Finally, it is "not relevant" if no information can help to satisfy the information need, or the information is not displayed well.

[5] https://lemurproject.org/clueweb12/

Efficacy: The document is “effective” if it claims that the intervention can improve the symptoms of the health issue, and “ineffective” if the document states it is harmful/not effective or the overall decision of the content writer is against the efficacy. The document is “inconclusive” if either the document discusses that current scientific consensus neither supports nor rejects the efficacy, or it discusses both options but does not favour one over another. Lastly, a document should be assessed as “no information” if it is found to be relevant but does not discuss the health issue.

Credibility: Assessment guidelines partially control credibility by outlining some cases to determine if the document should be considered as credible or not credible. However, it also gives the assessor some flexibility to judge based on their perceived credibility. According to the guidelines, a document is “credible” if it is written by an expert (e.g. a Medical Doctor), has proper citations, or is a web document of a hospital, university, well-known newspaper or a well known medical website. A document is labelled “noncredible” if it is written by non-experts such as bloggers, if it promotes a product, or if it is spam. On the other hand, documents that do not fall under these categories are left for the judge’s initiative; for example, documents with an authoritative tone without any information about the content writer may be labelled as credible or noncredible.

Once the assessment was completed, the success of the submitted runs by participants could be evaluated based on these three criteria. As the judges did not evaluate correctness but efficacy, the organizers mapped efficacy to correctness by matching the document’s efficacy claim with the scientific consensus. This whole process resulted in an annotated collection of documents, whose judgments are to be made publicly available to future researches after one year.

Assessment guidelines reduce the work on the human judges by not requiring them to judge the documents based on credibility and efficacy claim if it is not relevant to the topic. This assumption is reasonable because if the document does not discuss the topic, it cannot claim effectiveness. Nevertheless, when the document is irrelevant to the topic, it may still be credible. Although this might be very well the case, the assumption is still meaningful because the information will have little or no impact on the searcher if it is not relevant to the goal. Moreover, the literature discusses a deeper connection between credibility and information needs. Specifically, Fogg (2003) outlines the dynamics of the user's perception about page credibility in two recursive steps, prominence and interpretation. The first step refers to the signals that the user notices at first glance, such as page design, topic and personal involvement; the second step refers to their credibility interpretation based on these signals. So, even though the credibility can somehow be abstracted from the topic, it relies largely on the topic relevance in the search context. We can clarify this idea with an example: If a person is searching for whether garlic can help his child's influenza, he will not find a web page useful if it gives details of growing garlic. The page may be written by an agriculture expert who has specialization in herbs and nutrition, but this fact does not ensure the credibility of the source for medical purposes.

3.2.4 Evaluation Metrics

As discussed above, the runs were evaluated based on their correctness, credibility, relevance and some combinations of these. In the TREC Decision Track, only four metrics (MAP, nDCG@10, NLRE, CAM) were calculated to measure the runs' performance, but we will also report different measures. As in most of the tracks, trec_eval software was used to calculate these scores. Below we detail these metrics:

MAP (relevance): The first measure is mean average precision (MAP) based on topic relevance, which is calculated by comparing the run with the qrels file containing the relevance of the documents to the topic. The MAP score is calculated using precision cut-off k, that is the fraction of relevant documents in the top k:

$$Prec(k) = \frac{1}{k}\sum_{i=1}^{k} Rel(D_i),$$

where D_i is the document at rank i and $Rel(\cdot)$ is a binary function that returns 1 if D_i is relevant and 0 otherwise. Average precision for a given topic is calculated as:

$$AveP = \frac{1}{|R|}\sum_{d \in R} Prec(Rank(d)),$$

where $Rank(\cdot)$ returns the rank of the document d, R is the set of relevant documents and $|R|$ is the number of relevant documents. If the document d is not retrieved, $Prec(Rank(d)) = 0$. Then, for T number of topics, MAP is calculated as:

$$MAP = \frac{\sum_{t=1}^{T} AveP(t)}{T}.$$

nDCG@10 (relevance): Another measure that is commonly used in IR literature is normalized discounted cumulative gain. It quantifies how much the search engine satisfies the information need. Since the searchers do not to browse all results, normalized discounted cumulative gain top 10 (nDCG@10) can be a more intuitive and practical measure than MAP. DCG@10 and nDCG@10 can be calculated as:

$$DCG@10 = \sum_{i=1}^{10} \frac{relevance_i}{\log_2(i+1)} \quad \text{and} \quad nDCG@10 = \frac{DCG@10}{IDCG@10},$$

where $IDCG@n$ is the ideal (upper limit of) $DCG@n$ obtained by calculating $DCG@n$ over a sorted list based on relevance. Depending on the design, $relevance_i$ can be binary or categorical. In the TREC Decision Track there were three levels of relevance (i.e. not relevant, relevant and highly relevant) and $relevance_i \in \{0, 1, 2\}$.

NLRE: Normalized local rank error

$$LRE = \sum_{i=1}^{n-1} \frac{1}{\log_2(1+i)}((\mu + \epsilon^r)(\nu + \epsilon^c) - \mu\nu), \tag{3.1}$$

where μ and ν are parameters controlling the penalty. ϵ^r and ϵ^c are relevance and credibility ranking errors and

$$NLRE = 1 - \frac{LRE}{C_{LRE}},$$

and

$$C_{LRE} = \frac{(n-2j-1)^2 + (\mu+\nu)(n-2j-1)}{1 + \log_2(1+f)}.$$

CAM: Convex aggregating measure can be calculated as below:

$$\lambda_{rel} M^{rel} + \lambda_{cre} M^{cre} + \lambda_{cor} M^{cor},$$

for M^r and M^e are any measure of credibility and correctness and $\lambda_{rel} = \lambda_{cre} = \lambda_{cor} = 1/3$.

MAP and nDCG@10 (combined criteria): The goal of the track was to reward search engines with more correct and credible results. In the TREC Decision Track, the organizers used the above combined measures to calculate performance scores. However, as we will discuss later, they are not suitable for the track. By following Abualsaud et al. (2019), we will use MAP and nDCG@10 to measure the runs' performances in credibility and correctness tasks. To this end, we will manipulate the qrels file by replacing the relevance with the below:

Credible	1 if the document is both credible and relevant, 0 otherwise
Correct	1 if the document is both correct and relevant, 0 otherwise
Correct and Credible	1 if the document is credible, correct and relevant, 0 otherwise

and based on these combined values, we will calculate MAP and nDCG@10 using trec_eval software. Notice that for a document to be considered as credible or correct, it must first be relevant. This is a necessary assumption that we will discuss later.

In this thesis, we will refer to the above scores, but our discussion will focus only on nDCG@10 scores calculated using relevance, credibility, correctness and all combined.

3.3 Algorithms

The primary purpose of Decision Track is to improve health-related decision making. Our literature review indicates the complex interaction between the searcher and the search engine; however, it also implies that one way to improve decisions is to promote the ranks of documents containing correct information in SERPs. Besides, as discussed previously, we are motivated by the presumption that correct information can be reached by targetting credible documents. To this end, we will score credibility of the documents and their relevance to a given topic, and blend them into a single score for ranking. Overall, our runs combine three types of scores on (i) relevance, (ii) credibility and (iii) spaminess.

For the retrieval scores, we used the BM25 algorithm with default parameters as implemented in Anserini[6]. Each document in ClueWeb12-B13 was stemmed and indexed via Anserini, and 10000 documents per each topic were retrieved using the queries shared in topics.xml file. To assess credibility, we trained a logistic regression classifier on a health corpus subsetted from the ClueWeb12-B13 dataset. Lastly, we filtered spam using spaminess scores proposed by Cormack et al. (2011). While spaminess scores capture features that signal whether or not a document is spam, the credibility classifier aims to capture the tone that signals whether the document is trustworthy. The spam and credibility scores were used to adjust the relevance rankings by elevating credible documents in the SERPs.

Below we will outline the scores and detail the process.

3.3.1 Retrieval Scores

BM25 (Robertson et al., 1995) has several different implementations with different parameters. By default, Anserini uses Lucene's accurate implementation of BM25, which can be calculated, for a given document D, query $Q = [q_1, \ldots, q_n]$ and query term, q_i, as below:

[6] https://github.com/castorini/anserini

$$BM25(D,Q) = \sum_{i=1}^{n} \log\left(\frac{N - df(t_i) + 0.5}{df(t_i) + 0.5}\right) \frac{tf(t_i, D)}{tf(t_i, D) + k_1\left(1 - b + b\frac{L(D)}{L_{avg}}\right)},$$

where L_{avg} is the average document length (number of tokens) and $L(D)$ is the length of the document D; $tf(t_i, D)$ is the count of t_i in D (term frequency) and $df(t_i)$ is the number of documents containing t_i (document frequency).

The parameters by the original work are not appropriate for large sets of collections. Following Trotman et al. (2012), Anserini inputs $k1 = 0.9$ and $b = 0.4$ by default. In our work, we used these default settings.

3.3.2 Spaminess Scores

As discussed in the literature, spam is a particular form of a noncredible document. It aims to manipulate search engines to occupy higher positions in the results page, regardless of the quality of the document's content or the degree of its relevance to the query. Therefore, removing spam from the results can improve not only the relevance of the results but also the credibility and correctness of the information presented to the user, as these documents were written to reach to the user regardless of their value.

To detect the spam, Cormack (2007) developed a spam model on the ClueWeb09 dataset, which was later used to generate spam scores for ClueWeb12 collection[7] (Cormack et al., 2011). The dataset contains spaminess percentiles ranging between 0 to 99

[7] Spam scores are available at `https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/`.

with 0 being the most *spamy*. We used these scores to evaluate the spaminess of each document.

3.3.3 Credibility Scores

To calculate the credibility of a document, we trained a logistic regression model on raw documents containing complete HTML code, CSS and scripts. For the training corpus, we subsetted documents from 25 health topics not included in TREC Decision Track. This corpus was labelled interactively during construction using HiCAL. We will present the details below.

Training and Test Corpora

We prepared two different corpora subsetted from ClueWeb12-B13 for (i) training the credibility classifier, and (ii) and measuring the performance of the methods that will be used for final runs. For the former, we determined 25 topics, and a number of queries to retrieve documents from ClueWeb12 for each topic. The topics were similar to the track's topics but were about different medical interventions and/or treatments. We did not follow a rule to determine the topics but arbitrarily chosen based on Cochrane entries and previous researches.

The topics cover a variety of health issues from cancer to diabetes and scoliosis, and were chosen based on different levels of controversy, from lower (e.g. exercise for scoliosis) to higher (e.g. vaccines for hepatitis B) and different target groups (such as vinpocetine for dementia, antioxidants for female subfertility). We then determined a set of queries in the form of “[treatment] for [issue]” and its variation using synonyms and different modifiers (e.g. “antidepressants for tinnitus”, “can antidepressants help tinnitus”, “antidepressants for ringing in the ear”). Then, the queries were inputted to Anserini to retrieve ranked documents per query using the default BM25 algorithm implemented in the software. For each query, we retrieved the top 1000 documents, which may or may not be related to the given query. This subset was then cleaned from malicious pages with open-source anti-virus software, ClamAV, resulting in 40753 unique documents.

For the second corpus, we selected topics based on their popularity to ensure a sufficient number of credible and noncredible content (e.g., “acupuncture for autism”, “antibiotics for otitis media”, “pilates for lower back pain”, “lycopene for prostate cancer”, “green tea cancer”). We retrieved 1000 documents per topic using the procedure described above.

Annotating the Corpora

The above collections were created using a set of topics and queries, but they involve many relevant and nonrelevant documents to the chosen topics, and even to healthcare, such as documents about garlic planting, benefits of certain herbs for a divine spirit. To train a credibility classifier, we needed all documents written in the health domain, but not, e.g. herbal remedies for a stronger spirit. While annotating the corpus, we labelled the documents accordingly as credible, noncredible or nonrelevant.

To prepare an annotated corpus, we used HiCAL (Abualsaud et al., 2018), a system for high-recall retrieval, by assessing documents’ credibility. The software takes a set of documents as input, ranks them based on their relevance to an initial query, then presents each document in a row. The user is given three choices, “nonrelevant”, “relevant” and “highly relevant,” as the session continues, an internal classification algorithm simultaneously learns the patterns that are in line with the user’s preferences and re-ranks the documents to increase the priority of the ones that are similar to the user’s choices. We imported the above corpus with 40753 documents and set the system following the instructions outlined in its webpage[8].

For each topic, we started a new session with an initial query in the form of “[treatment] [issue]” related to the topic, which we called, *seed query*. As described above, the software ranked the documents based on their relevance to the query[9] and re-ranked them based on the annotator’s choices. In each session, we chose either credible or noncredible documents. As the assessment continued, since the algorithm searched for *common* patterns in the chosen documents, it prompted any page somehow related to the previous judgments, independent of topic relevance and based solely on credibility. Therefore, for example, if the seed query was “exercise scoliosis”, the session most likely started with documents about scoliosis but continued with, e.g. otitis media or dementia, and visited a wide range of topics. After each session, we extracted the session information to separate files that contain the document IDs and the information about whether the document was chosen as relevant to the information need (1) or not relevant to the need (-1) [10]. We later merged these files by reverting signs of non-credibility assessments. The process yielded 2452 noncredible and 1081 credible non-duplicating documents.

Our interaction was informative in terms of detecting credibility. There are certain cues

[8] https://github.com/hical/HiCAL

[9] HiCAL employs BM25 to rank the documents initially.

[10] Note that in this context relevance to the need is credibility or non-credibility depending on the session.

in the documents that can be detected with an exhaustive investigation, which can easily signal credibility, such as explicit phrases in the HTML code including health on the net stamp, "<img src='/images/imgHonCode_petit.jpg'>", well-known newspaper, hospital or university names (e.g. New York Times, University of Oxford). Similarly, we effortlessly detected certain noncredible pages such as forums, blogs or spam that is algorithmically generated by filling in an HTML template. Some of these cues are easily detected by the human eye, but some require an in-depth examination. As HiCAL searches patterns embedded in previously judged documents, we could construct a credibility corpus quite easily.

For the set of five topics, however, we evaluated both the credibility and topic relevance of documents to be able to test the runs' ranking performance. If a document is nonrelevant, it was labelled with 0; if not, then it was given an integer between 0 to 2, with 0 for "noncredible", 1 for "cannot decide" and 2 for "credible". For each topic, we assessed at least 200 out of 1000 documents using HiCAL, which yielded 187 / 5000 relevant instances. The only exception was Topic 5 for which, instead of using HiCAL, we assessed the top 200 out of 1000 documents based on BM25 scores.

We want to make an important note about the first collection here. As the author of this study is also a member of the team who designed and prepared the Decision Track, there might be similarities between a few topics in our training collection and the track's collection. The training collection was prepared to include a wide range of health documents to be able to represent the task. Although our labelling procedure ignored topic relevance and thus is not restricted to annotate the credibility of documents from any particular topic, such influence might have improved the classifier's accuracy in some cases.

Credibility Classifier and Scores

The supervised classifier aims to detect patterns in documents that signal credibility independent from the topic. These patterns can be a combination of features that give all kinds of information from page design to colours to keywords that indicate source credibility. We trained a logistic regression classifier on raw documents by first converting text to lower case and then tokenizing it into all sequential character 4-grams. For example the word "HonCode" is parsed into "honc", "onco", "ncod" and "code". We trained the classifier with a binary setting, i.e. 1 if the feature is present, 0 if absent.

Each document D_i in the training collection with size M was converted into a sparse binary vector of N dimensions, $\mathbf{X_i} = [1, X_{i1}, \ldots, X_j]^T$ for $X_{ij} \in \{0, 1\},\ j \in \{1, \ldots, N\}$ and for $\{i \in 1, \ldots, M\}$ documents. Here there are N features corresponding to each character 4-grams appeared in the collection. We fit the standard logistic regression model:

$$P(D_i \ is \ Credible) = \frac{1}{1 + e^{-Z_i}}, \text{ for } Z_i = \mathbf{w}^T \mathbf{X}_i,$$

for $\mathbf{w}^T = [w_0, w_1, \ldots, w_N]$. To train and run the model, we used Python 3, sklearn package, which is available PyPI repository. We used two functions defined in the package CountVectorizer[11] with *binary=True* and *analyzer='char'* and *ngram_range = (4,4)* to create above $\mathbf{X}_i$, and LogisticRegression[12] to train the model.

The above model was trained on the full collection of 40753 documents. This model

[11] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[12] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

was saved to disk and used for computing probabilities, $P(D_i \ is \ Credible)$, for the test runs and submitted runs.

However, during hyperparameter tuning with the test runs (we will describe below), we realized that strict filtering of the documents based on the above probabilities might not be appropriate for IR tasks. Namely, we wanted to reduce the impact of false negatives not by filtering but by re-ranking based on the credibility scores extracted from the probabilities. To this end, we transformed the probabilities to Z-scores using logit function as below

$$p = P(D_i \ is \ Credible) = \frac{1}{1+e^{-Z}}, \text{ and } \ Z = \ln\frac{p}{1-p}, \tag{3.2}$$

and used them to combine with the other scores linearly.

3.3.4 Combining Scores

We submitted seven runs to TREC Decision Track, (i) one is the default ranking with BM25, (ii) three are re-ranking the BM25 scores adjusting with the above credibility (Z) scores, and (iii) three are filtering the list of documents initially retrieved by using BM25 algorithm via spaminess scores and credibility probabilities. We determined the list of runs by tuning transformation parameters on the test corpus (of 5 topics). Below we present the list of submitted runs and calculation rules that we will detail later in this section.

1. `UWaterMDS_BM25`,

2. `UWatMDS_BM25_ZS = IF SPAM > 10,` $BM25 * (1 + Z)$ `ELSE 0`,

3. UWatMDS_BM25_Z = $BM25 * (1 + Z)$,

4. UWatMDS_BMZBS10 = IF SPAM > 10, $BM25 \times (1 + Z * 2)$ ELSE 0,

5. UWatMDS_BMF_C90 = IF $1 - P(d\ is\ Credible) < 0.90$ BM25, ELSE 0,

6. UWatMDS_BMF_C95 = IF $1 - P(d\ is\ Credible) < 0.95$ BM25, ELSE 0,

7. UWatMDS_BMF_S30 = IF SPAM > 30 BM25, ELSE 0.

Baseline Run: UWaterMDS_BM25

This run is the default output of Anserini. We indexed the whole ClueWeb12-B13 and ran the BM25 algorithm with default parameters implemented in Anserini. We retrieved the top 10000 results for 51 topics and selected the top 1000 for this run. We also used this indexed collection to retrieve the documents for the training corpora we mentioned above.

Runs with Reordering: UWatMDS_BM25_ZS**,** UWatMDS_BM25_Z **and** UWatMDS_BMZBS10

In the test runs, using the credibility classifier probabilities for filtering out documents or adjusting their ranks directly (e.g. by computing $BM25 * P(d \in Credible)$) resulted in discarding many useful documents. Besides, the probabilities were not appropriate for linear transformations for re-ranking documents. Therefore we calculated the scores, which we will refer to as Z scores, using the formula we presented in the equation 3.2 and combined with the retrieval scores via linear transformations.

To combine with the BM25 scores, we rescaled Z to 0-1. As we want to improve the BM25 scores in favour of credible documents, we added credibility reward proportionate to relevance as $BM25 * (1 + Z)$. If Z is close to 0 (noncredible), then the score is equal to the BM25 score, if it is close to 1, it doubles the score. Using this transformation, we could prevent a nonrelevant but credible document from occupying a higher position on the results page. We also added a parameter, β to $Z-$score to control the weight of credibility judgments in the final score:

$$BM25 * (1 + \beta Z).$$

Further, we used spam scores (SPAM) as filters. In the test runs, it filtered out too many necessary documents when the threshold was set to 70. The best performances were reached when $10 < SPAM < 40$ varying depending on the test topic. Hence we used the spam filter with a threshold of 10 to filter out "junk" documents and adjusted scores using the above rule.

We generated test runs on the five test topics using different combinations of SPAM and β values. After a series of trials with $SPAM \in \{10, 20, \ldots, 70\}$ and $\beta \in \{1, 1.1, \ldots, 2\}$ we decided on soft filtering with spam scores (i.e. filtering out the documents with $SPAM < 10$) and the values $\beta = 1$ and $\beta = 2$. We chose these values based on their MAP performances on five test topics. The combination of these determined the rules for the three runs, which we listed early in this subsection. We also present MAP scores of these runs are below in the "Test Performances" subsection.

Runs with Filtering: `UWatMDS_BMF_C90`, `UWatMDS_BMF_C95` and `UWatMDS_BMF_S30`

When combined with retrieval scores as filters, credibility classifiers and spam scores also improved the performances in the test corpus. To determine parameters, we calculated test scores for $SPAM \in \{10, 20, \ldots, 70\}$ and $1 - P(d\ is\ Credible) < p_0$ for $p_0 \in \{0.89, 0.90, \ldots, 0.99\}$. Based on their MAP scores on five test topics, we decided on $p_0 = \{0.90, 0.95\}$ and $SPAM = 30$. We present the MAP scores of these runs in the below subsection.

Test Performances

Table 3.2 presents the performance of automatic runs on 5 test topics based on the Mean Average Precision (MAP) measure. As discussed above, we annotated the test corpus based on topic relevance and credibility. Namely, a document was counted as successful if it is both credible and relevant to the topic. Therefore, for a run to score high, it must improve the ranks of such instances. We did not evaluate correctness (or document's efficacy claim) due to time constraints.

As shown in the table, the credibility classifier improved the baseline BM25 scores in the test runs. BM25 scores filtered with credibility with 90% level (`UWatMDS_BMF_C90`) yielded inconsistent results, outperforming all other runs for Topic 3 and improved the baseline for Topic 1. The combinations of BM25, Z and SPAM (`UWatMDS_BM25_Z`, `UWatMDS_BM25_ZS`, `UWatMDS_BM25ZBS10`) performed better than BM25 baseline, and even doubled its precision for some topics.

For the 5th topic, other algorithms performed worse than `UWatMDS_BM25`. Assessment of this topic gives the baseline BM25 an advantage because, as mentioned before, the documents were labelled by only evaluating the top 200 documents retrieved by the baseline BM25 test run. More clearly, documents in the top 200 were labelled as either 0 or 1 based

Table 3.2: Mean average precision (map) and geometric map of the methods on our self-created tuning topics.

Topic Rel / Ret	1 9 / 1000	2 70 / 1000	3 17 / 1000	4 14 / 1000	5* 77 / 1000	All 187 / 5000	All (gm_map)
BM25	0.160	0.377	0.065	**0.016**	**0.514**	0.226	0.126
BM25_Z	0.345	0.456	0.132	0.013	0.470	0.283	0.165
BM25_ZS	**0.346**	0.462	0.141	0.015	0.393	0.271	0.167
BMZBS10	0.336	**0.463**	0.140	0.014	0.384	0.268	0.165
BMF_C90	0.255	0.298	**0.157**	0.005	0.170	0.177	0.100
BMF_C95	0.362	0.302	0.141	0.004	0.146	0.191	0.097
BMF_S30	0.163	0.380	0.083	0.010	0.321	0.191	0.111

on their topic relevance and credibility, but the rest 800 were labelled as 0 regardless of their relevance or credibility. Therefore, for a run to improve the baseline, it must precisely reduce only the ranks of noncredible or nonrelevant documents. As Table 2 shows, spam and credibility classifiers reduced precision sharply when used as filters (`UWatMDS_BMF_C90`, `UWatMDS_BMF_C95`, `UWatMDS_BMF_S30`), indicating that these algorithms filtered out many true positive documents. On the other hand, `UWatMDS_BM25_Z` caused the least distortion to the overall position of successful documents.

As mentioned in previous subsections, the above runs are selected among different values of $SPAM$, β and p_0. We ran many combinations of $SPAM \in \{10, 20, \ldots, 70\}$, $p_0 \in \{0.89, 0.90, \ldots, 0.99\}$ and $\beta \in \{1, 1.1, \ldots, 2\}$, and selected the above seven because they performed better and/or more consistent compared to the others.

Chapter 4

Results

In this chapter, we will present the results of the experiment detailed in the previous section. We will start by discussing the performance of runs and then make some changes to improve the performances. Later, we will present an analysis of the qrels file containing the judgments produced by NIST assessors to give insight into the challenges in developing systems to support health-related decisions.

4.1 TREC Results

Table 4.1 shows the performance of our runs using the track's evaluation measures. The first two columns, MAP and nDCG@10, are measuring topic relevance only, and the last two columns, NLRE and CAM, are combined measures for all three variables. The runs performed poorly compared to the baseline, `Uwater_MDS_BM25`, based on both of the relevance measures. MAP scores indicate that the best performing run is `UWatMDS_BM25_Z`, which softly adjusts the benchmark, and the worst-performing run is `UWatMDS_BMF_C90` which filters documents with a relatively high threshold. The only improvement in relevance is achieved in nDCG@10 scores by `UWatMDS_BMF_S30`, which filters highly spamy documents.

NLRE scores do not distinguish between performances of runs and do not seem to be an appropriate measure for this type of research. On the other hand, CAM scores present very similar results to MAP. Note that according to the track's guidelines, for a document to be considered as correct or credible, it must first satisfy the relevance criteria. Therefore, there is some redundancy in CAM scores, which are calculated by harmonically averaging relevance, correctness and credibility.

As the NIST measures do not distinguish performances of runs, Abualsaud et al. (2019) proposed additional measures that modify MAP and nDCG@10 to measure the performance of runs in credibility and correctness, which we have outlined in the previous chapter (see Section 3.2.4. We will not report MAP scores because it is not practical for the underlying task. We will only report nDCG@10 scores, which measure performances of runs at the top 10 in the results page.

Table 4.2 present the performances using these measures. Notice that the table includes separate columns for relevance, correctness, credibility, and relevance columns present the same numbers in Table 4.1. As discussed in the previous chapter, the correctness and credibility columns report performances in combined measures (with relevance), and the column entitled "All" is the most strict case, which counts a document as successful only if it is relevant to the topic, credible and correct.

Each run's nDCG@10 performance was calculated separately for each topic. The numbers presented in the tables below are the mean scores (of $N = 50$ topics). Using each of the 50 nDCG@10 scores, we also tested the significance of the runs; to be more clear, we ran two-tailed paired t-tests to compare each run's performance with the baseline run, and verify that the population means are significantly different. Here, the null hypothesis is "the population means are equal" versus the alternative "they are unequal". In the tables, the stars indicate significance at 0.1 (*), 0.05 (**) and 0.01 (***) levels. However, we did not report the runs that performed significantly worse than the benchmark as they are not of our interest.

The results are favouring runs with filters. Spam and credibility filtering improved correctness from 3.15% to 11.59%, however, none of these improvements were significant. In "Credibility" and "All" columns, we can see that all approaches improved the top 10 results. The largest improvement over the baseline credibility was achieved by `UWatMDS_BMF_C95`, which is also significant at 0.1 level. This run also performed the best when all criteria are combined, but the only significant improvement was achieved by `UWatMDS_BMF_S30`.

As we will show later, there are some topics for which there are not enough correct documents in the collection. For 14 topics, NIST assessors did not found any correct document (11) or correct and credible document (3). There are also 11 other topics for which the number of correct documents is less than 10. It would be unfair to expect improvements on these topics as nDCG@10 can be greater than 0 if there is at least one document that satisfies the given condition. Also, to be able to present only the correct information in the top 10 results, there must be at least ten correct documents available to retrieve. Therefore, we recalculated the average nDCG@10 scores for two more cases, with

Table 4.1: Performance of runs based on NIST evaluation measures

Run	MAP	nDCG@10	NLRE	CAM
UWatMDS_BM25_Z	0.345 (-8.40%)	0.443 (-11.15%)	0.997 (+0.10%)	0.547 (-0.18%)
UWatMDS_BM25_ZS	0.310 (-17.51%)	0.430 (-13.72%)	0.997 (+0.11%)	0.510 (-6.96%)
UWatMDS_BMF_C90	0.156 (-58.50%)	0.425 (-14.78%)	0.999 (+0.33%)	0.309 (-43.60%)
UWatMDS_BMF_C95	0.170 (-54.86%)	0.445 (-10.75%)	0.999 (+0.33%)	0.334 (-39.04%)
UWatMDS_BMF_S30	0.285 (-24.15%)	0.500 (+0.28%)	0.998 (+0.20%)	0.456 (-16.74%)
UWatMDS_BMZBS10	0.283 (-24.89%)	0.392 (-21.36%)	0.997 (+0.13%)	0.492 (-10.21%)
UWaterMDS_BM25	0.376	0.499	0.996	0.548

Bottom row presents the performance of baseline run which performed the best in all measures except for NLRE for which all runs performed almost the same. The percentages in parentheses are the percentage difference over the baseline.

Table 4.2: nDCG@10 scores calculated with relevance, correctness and credibility.

Run	Relevance	Correctness	Credibility	All
UWatMDS_BM25_Z	0.443 (-11.28%)	0.180 (-4.06%)	0.452 (+14.04%)	0.175 (+29.92%)
UWatMDS_BM25_ZS	0.430 (-13.84%)	0.181 (-3.31%)	0.446 (+12.60%)	0.177 (+31.25%)
UWatMDS_BMF_C90	0.425 (-14.90%)	0.193 (+3.15%)	0.444 (+12.12%)	0.178 (+32.37%)
UWatMDS_BMF_C95	0.445 (-10.88%)	0.201 (+7.26%)	**0.463* (+16.86%)**	**0.183 (+36.01%)**
UWatMDS_BMF_S30	**0.500 (+0.14%)**	**0.209 (+11.59%)**	0.426 (+7.62%)	0.161* (+19.52%)
UWatMDS_BMZBS10	0.392 (-21.47%)	0.163 (-12.76%)	0.418 (+5.50%)	0.162 (+20.49%)
UwaterMDS_BM25	0.499	0.187	0.396	0.135

The column entitled "All" is the most strict case in which a document must relevant, correct and credible to be regarded as valid. The bold values indicate best scores except the baseline's, which is presented in the bottom row. The numbers in parentheses indicate the percentage contribution over the baseline. The stars (*) indicate significance at 0.1 level.

Table 4.3: nDCG@10 scores calculated with relevance, correctness and credibility based on 36 topics.

Run	Relevance	Correctness	Credibility	All
UWatMDS_BM25_Z	0.462 (-14.24%)	0.247 (-3.91%)	0.461 (+16.52%)	0.243 (+29.95%)
UWatMDS_BM25_ZS	0.449 (-16.57%)	0.249 (-2.99%)	0.458 (+15.59%)	0.246 (+31.26%)
UWatMDS_BMF_C90	0.481 (-10.66%)	0.268 (+4.37%)	0.479* (+20.91%)	0.248 (+32.39%)
UWatMDS_BMF_C95	0.484 (-10.14%)	0.279 (+8.57%)	**0.484* (+22.3%)**	**0.254 (+36.05%)**
UWatMDS_BMF_S30	**0.544 (+1.16%)**	**0.288 (+12.22%)**	0.439* (+10.77%)	0.224* (+19.54%)
UWatMDS_BM_ZBS10	0.405 (-24.82%)	0.227 (-11.7%)	0.425 (+7.3%)	0.225 (+20.52%)
UwaterMDS_BM25	0.538	0.257	0.396	0.187

nDCG@10 scores recalculated based on runs' performance on 36 out of 50 topics for which there are at least 1 correct and credible document in the TREC collection. The stars (*) indicate significance at 0.1 level.

Table 4.4: nDCG@10 scores calculated with relevance, correctness and credibility based on 25 topics.

Run	Relevance	Correctness	Credibility	All
UWatMDS_BM25_Z	0.484 (-15.86%)	0.334 (-6.79%)	0.488 (+24.79%)	0.317 (+32.29%)
UWatMDS_BM25_ZS	0.462 (-19.59%)	0.336 (-6.14%)	0.479 (+22.42%)	0.320 (+33.52%)
UWatMDS_BMF_C90	0.495 (-13.93%)	0.382 (+6.61%)	0.503* (+28.59%)	0.351** (+46.23%)
UWatMDS_BMF_C95	0.517 (-10.09%)	0.397 (+10.96%)	**0.528**(+35.00%)**	**0.361**(+50.34%)**
UWatMDS_BMF_S30	**0.580 (+0.98%)**	**0.399 (+11.55%)**	0.460** (+17.5%)	0.290 (+20.78%)
UWatMDS_BM_ZBS10	0.411 (-28.51%)	0.306 (-14.63%)	0.437 (+11.82%)	0.298 (+24.08%)
UwaterMDS_BM25	0.575	0.358	0.391	0.240

nDCG@10 scores recalculated based on runs' performance on 25 out of 50 topics for which there are at least 10 correct documents in the TREC collection. The stars indicate significance at 0.1 (*) and 0.05(**) levels for runs improving the benchmark.

a subset of 36 topics for which there is at least one correct and credible document and 25 topics for which there are at least ten correct documents. We presented these results in Table 4.3 and 4.4.

Table 4.3 presents the results without the 14 topics. In this case, the performances improved overall, but the performances relative to the benchmark did not change dramatically. The only exception is for credibility for which the baseline run scores the same as the previous case (0.396), and thus other runs show larger relative improvements.

After removing the 25 topics, the performances of all runs almost doubled compared to the case with all topics. In Table 4.4, similar to the previous cases, the best increase in the task to improve correctness was achieved by `UWatMDS_BMF_S30`, but percentage improvement over the baseline remained the same. When all criteria are combined, the best performing run, `UWatMDS_BMF_C95`, increased from 0.183 to 0.361, improving the baseline 50% significant at 0.05 level. Moreover, credibility increased from 0.463 to 0.528, and this improvement is also significant.

4.2 Performance of Credibility Classifier

Our results indicate some significant improvements when the task was to promote credible documents, and credible & correct documents. It is natural to ask how strong is the classifier in predicting the credibility of a document. After the TREC Decision Track workshop, the organizers shared the qrels file containing NIST assessors' judgments, including credibility and correctness assessments for each document found relevant or highly relevant

Table 4.5: Inter-rater Reliability & Classifiers' Performances

	Accuracy	Min	Max
Inter-rater reliability	0.7143	-	-
10-fold	0.9454	0.9284	0.9614
10-fold (topic-wise)	0.6727	0.2193	0.9188
Test Accuracy	0.5976	-	-
10-fold (topic-wise, NIST)	0.9358	0.741727	1.0000

Table presenting accuracy and inter-rater reliability. The first two "10-fold" are based on classifiers trained and validated on our own judgments. "Test Accuracy" is accuracy of the classifier trained on our judgments and tested on NIST judgments. The last is the accuracy of the classifier trained and tested on NIST judgments. The min and max columns are minimum and maximum accuracies among each 10 iterations.

to the topic. Here in this section, we will measure inter-rater reliability between our and NIST assessors' judgments, and the performance of credibility classifier. We will cover these aspects below in three subsections.

4.2.1 Inter-rater Reliability

As we discussed before, to train the classifier, we subsetted health-related documents from ClueWeb12-B13. To this end, we determined 25 topics, which were not included in TREC Decision Track topics, retrieved 1000 documents using BM25 retrieval algorithm and labelled 3533 of the retrieved documents based solely on their credibility, ignoring their relevance to any topic. Therefore, even though the topics were distinct, some documents that we assessed were relevant to TREC Decision Track topics, and were also assessed by NIST. More clearly, there are 77 documents that were labelled by both us and NIST.

Based on this small set, we calculated the inter-rater reliability between our judgments and NIST's. The below confusion matrix compares our judgments (the rows) and the NIST assessors' judgments (columns):

	NIST (NC)	NIST (C)
Our (NC)	9	18
Our (C)	4	46

Here, the inter-rater reliability is 0.7143. As seen in the below confusion matrix, our judgments tend to be more conservative in judging a document as credible.

4.2.2 Classifiers Accuracy (10-fold Validation)

Based on our judgments, we trained the logistic regression classifier to be employed in our algorithms. We calculated the 10-fold validation accuracy of our classifier in two different ways. Firstly, we shuffled all the documents and split into 10 subsets. At each iteration, we selected one of the subsets as the validation set and the rest as the training set, then trained the classifier and tested on the validation set. The average of the 10 numbers (10-fold validation score) is 0.9454 and the scores range between 0.9284 to 0.9614.

This method to validate the classifier's accuracy may not be convincing for the reader. A better method could be splitting the topics to 10 subsets, determining training and validation document sets based on their relevance to the topics in respective sets. This method is more realistic because the classifier is expected to function well enough when it is unfamiliar to any topic. However, as we described earlier, the training corpus of the credibility classifier is assessed based solely on documents' credibility and topic relevance was ignored. Therefore, such a method does not ensure that the topics covered by the documents in the training set is distinct from the validation sets.

Nevertheless, we calculated the 10-fold validation accuracy following this procedure. The mean of the 10 scores is 0.6727, but the numbers are ranging between 0.2193 to 0.9188. Although reason behind this variance is not clear, this may be due to the some certain topics for which the assessment sessions were challenging and thus the judgments are not reliable.

4.2.3 Classifiers Performance on NIST Judgments

After the TREC Decision Track 2019 workshop, NIST shared qrels, the file containing their judgments. We calculated the performance of the classifier based on these judgments. Below we present the confusion matrix between predicted judgments (by the classifier) and the actual (by NIST):

	Actual (NC)	Actual (C)
Predicted (NC)	1486	1282
Predicted (C)	103	571

The accuracy of the classifier is calculated to be 0.5976. Notice that the classifier tends to be more conservative compared to the NIST assessors in judging a document as credible. These false negatives are consistent with the ones we presented earlier when we discussed inter-rater reliability, and can be attributed to our credibility perception. These numbers may not be presenting the classifier's own accuracy, but be dominated by the dissimilarities between our and NIST assessors' perceived credibility. Therefore, we trained and tested our classifier on NIST judgments using the same framework as above when we calculated 10-fold validation accuracy based on the topics.

To validate the classifiers accuracy on NIST judgments, we shuffled the 50 topics and split into ten groups. We then determined the training and validation document sets based on their relevance to the topics in respective groups. Some documents were found to be relevant to more than one topic by the NIST assessors; we ensured these documents were not contained in both training and validation document sets by removing such instances from the training set. After predicting the credibility of each documents in the validation sets, we combined these predictions, which yielded prediction for all documents[1]. Below we present the confusion matrix between predicted judgments (by the classifier) and the actual (by NIST):

	Actual (NC)	Actual (C)
Predicted (NC)	1411	42
Predicted (C)	179	1812

The average accuracy is calculated to be 0.9358. We can see, from the above matrix, that false negatives and false positives are more symmetrical compared to the previous case and thus the overall accuracy has improved.

[1] Note that the total number of documents here in the confusion matrix is less than the number of documents assessed for credibility in the qrels file. This is because we did not calculate these scores for all documents in the qrels file, but its intersection with the documents that we previously retrieved for our baseline BM25 run due to the time constraints. This yielded 3442 documents, and the remaining 688 documents in the qrels file were retrieved by other participants.

4.3 Further Improvement

In the previous section, we showed that the accuracy of the classifier that we used for the runs is as low as 0.5976. This may be because the credibility classifier was trained over a corpus annotated by the author of this thesis, but evaluated using the NIST assessors' judgments. Although, the runs we presented in the Section 4.1 improved the baseline credibility and in the case when all criteria are combined, the performances in the correctness task were relatively poor.

On the other hand, we also showed that when the classifier was trained on the NIST judgments, the accuracy has increased to 0.9358. In this section, we recalculated the runs by employing this classifier to translate this large increase into further improvements in runs' performances.

Below, we present performance of the runs using the new classifier's probabilities and scores, and compare them with the previous results. As detailed in the previous section, there are 10 classifiers, each trained on documents determined by 10-fold validation, i.e. trained on the collection of documents that are relevant to topics other than the validation topics. Here, instead of calculating accuracy, at each iteration, we combined retrieval scores, spaminess scores and credibility score/probability to determine the runs. Note that at each iteration, we computed combined scores for 5 target topics, thus after 10 iterations we had runs containing relevance scores for all 50 topics. Tables 4.6-4.8 compare nDCG@10 scores of the runs and their 10-fold versions which were labelled with "(K)" for each type of run. We removed `UWatMDS_BMF_C90` from the results to save space because its results were worse or not dramatically different from `UWatMDS_BMF_C95`.

In Table 4.6, the difference between 10-fold and original versions are notably different. UWatMDS_BM25_ZS performed the best, and together with UWatMDS_BM_ZBS10, they scored higher than UWatMDS_BM25_Z when trained using NIST judgments. Contrary to the submitted runs that generally performed poorly compared to the baseline, all 10-fold versions improved correctness and increased the score of the best performing UWatMDS_BMF_C95 run. Moreover, the 10-fold version of UWatMDS_BM25_ZS added 22.67% over the baseline run's credibility score, and this improvement is significant at 0.01 level. Notice also that the p-values of k-fold versions are higher than the submitted runs', and the improvements in "Credibility" and "All" columns are significant overall. Another notable difference is the k-fold version of UWatMDS_BMF_C95 improved the baseline relevance by 5.69%.

As in the previous section, we recalculated the scores using two subsets of 50 topics. When the 14 topics were removed, all scores increased, but relative contribution over the baseline did not improve except for some cases in relevance and credibility tasks. As in the previous table, the best performing run is still UWatMDS_BM25_ZS except for relevance. The improvement relative to the credibility baseline increased overall, and performances of two runs, UWatMDS_BM25_Z and UWatMDS_BM25_ZS, were significant at 0.01 level. Besides UWatMDS_BMF_C95 (K) to benchmark relevance increased to 7.76%.

When we removed the 25 topics for which the NIST assessors found less than ten correct documents, the baseline scores, relative increases over the baselines, and the p-values increased. Compared to the scores calculated over all of the topics, the correctness task scores have doubled. The best performing run in this task, UWatMDS_BM25_ZS (K), scored 0.412, but its contribution over baseline correctness reduced from 19.66% to 15.17%. There are notable increases over the baseline run's credibility score by four runs that performed significantly better than the baseline at 0.05 and 0.01 levels, and the best performing run among them `UWatMDS_BM25_ZS` (K) improved the baseline 40.45%. One unexpected difference with the previous results is that in the "All" column, `UWatMDS_BMF_C95` that were trained on our credibility judgments adds 50.34% over the baseline's score, which is greater than the score of `UWatMDS_BM25_ZS` (K) that improves the baseline 40.12%. Lastly, `UWatMDS_BMF_C95` trained on NIST credibility judgments performed distinguishably better than baseline relevance.

Table 4.6: NDCG@10 Results

Run	Relevance	Correctness	Credibility	All
`UWatMDS_BM25_Z`	0.443 (-11.28%)	0.180 (-4.06%)	0.452 (+14.04%)	0.175 (+29.92%)
`UWatMDS_BM25_Z` (K)	0.510 (+2.04%)	0.217 (+15.96%)	0.472** (+19.09%)	0.177** (+31.70%)
`UWatMDS_BM25_ZS`	0.430 (-13.84%)	0.181 (-3.31%)	0.446 (+12.60%)	0.177 (+31.25%)
`UWatMDS_BM25_ZS` (K)	0.504 (+1.04%)	**0.224 (+19.66%)**	**0.486*** (+22.67%)**	**0.190** (+41.08%)**
`UWatMDS_BMF_C95`	0.445 (-10.88%)	0.201 (+7.26%)	0.463* (+16.86%)	0.183 (+36.01%)
`UWatMDS_BMF_C95` (K)	**0.528 (+5.69%)**	0.215 (+14.84%)	0.417 (+5.30%)	0.157 (+16.41%)
`UWatMDS_BMZBS10`	0.392 (-21.47%)	0.163 (-12.76%)	0.418 (+5.50%)	0.162 (+20.49%)
`UWatMDS_BMZBS10` (K)	0.486 (-2.60%)	0.220 (+17.40%)	0.478** (+20.64%)	0.187** (+39.03%)
`UWatMDS_BMF_S30`	0.500 (+0.14%)	0.209 (+11.59%)	0.426 (+7.62%)	0.161* (+19.52%)
`UwaterMDS_BM25`	0.499	0.187	0.396	0.135

nDCG@10 scores recalculated based on runs' performance on 50 topics. The stars indicate significance at 0.1 (*),0.05 (**) and 0.01 (***) levels for runs improving the benchmark.

4.4 Analysis of the TREC Judgments

The qrels file shared by NIST contains the judgments produced by TREC Decision Track 2019 and consists of labels of 22842 documents from ClueWeb12-B13. As described above, each document was first judged based on relevance to the topic, then its credibility and its treatment efficacy claim, and no judgment is available if the document was not found to

Table 4.7: Comparison of nDCG@10 scores with 10-fold validation results with respect to relevance, correctness and credibility based on 36 topics.

Run	Relevance	Correctness	Credibility	All
UWatMDS_BM25_Z	0.462 (-14.24%)	0.247 (-3.91%)	0.461 (+16.52%)	0.243 (+29.95%)
UWatMDS_BM25_Z (K)	0.564 (+4.79%)	0.299 (+16.16%)	0.503*** (+27.12%)	0.246** (+31.69%)
UWatMDS_BM25_ZS	0.449 (-16.57%)	0.249 (-2.99%)	0.458 (+15.59%)	0.246 (+31.26%)
UWatMDS_BM25_ZS (K)	0.548 (+1.78%)	**0.307 (+19.58%)**	**0.511*** (+29.07%)**	**0.264** (+41.08%)**
UWatMDS_BMF_C95	0.484 (-10.14%)	0.279 (+8.57%)	0.484* (+22.30%)	0.254 (+36.05%)
UWatMDS_BMF_C95 (K)	**0.580 (+7.76%)**	0.299 (+16.20%)	0.433 (+9.45%)	0.218 (+16.42%)
UWatMDS_BM_ZBS10	0.405 (-24.82%)	0.227 (-11.70%)	0.425 (+7.30%)	0.225 (+20.52%)
UWatMDS_BM_ZBS10 (K)	0.527 (-2.10%)	0.301 (+17.16%)	0.496** (+25.37%)	0.260** (+39.03%)
UWatMDS_BMF_S30	0.544 (+1.16%)	0.288 (+12.22%)	0.439* (+10.77%)	0.224* (+19.54%)
UwaterMDS_BM25	0.538	0.257	0.396	0.187

nDCG@10 scores recalculated based on runs' performance on 36 out of 50 topics for which there are at least 1 correct and credible document in the TREC collection. The stars indicate significance at 0.1 (*),0.05 (*) and 0.01 (***) levels for runs improving the benchmark.

Table 4.8: Comparison of nDCG@10 scores with 10-fold validation results with respect to relevance, correctness and credibility based on 25 topics.

Run	Relevance	Correctness	Credibility	All
UWatMDS_BM25_Z	0.484 (-15.86%)	0.334 (-6.79%)	0.488 (+24.79%)	0.317 (+32.29%)
UWatMDS_BM25_Z (K)	0.599 (+4.23%)	0.405 (+13.08%)	0.535*** (+36.86%)	0.317** (+32.06%)
UWatMDS_BM25_ZS	0.462 (-19.59%)	0.336 (-6.14%)	0.479 (+22.42%)	0.320 (+33.52%)
UWatMDS_BM25_ZS (K)	0.583 (+1.42%)	**0.412 (+15.17%)**	**0.549***(+40.45%)**	0.336** (+40.12%)
UWatMDS_BMF_C95	0.517 (-10.09%)	0.397 (+10.96%)	0.528** (+35.00%)	**0.361** (+50.34%)**
UWatMDS_BMF_C95 (K)	**0.627 (+9.13%)**	0.402 (+12.35%)	0.427 (+9.22%)	0.272 (+13.4%)
UWatMDS_BM_ZBS10	0.411 (-28.51%)	0.306 (-14.63%)	0.437 (+11.82%)	0.298 (+24.08%)
UWatMDS_BM_ZBS10 (K)	0.573 (-0.27%)	0.410 (+14.52%)	0.544*** (+39.07%)	0.340** (+41.95%)
UWatMDS_BMF_S30	0.580 (+0.98%)	0.399 (+11.55%)	0.460** (+17.5%)	0.290 (+20.78%)
UwaterMDS_BM25	0.575	0.358	0.391	0.240

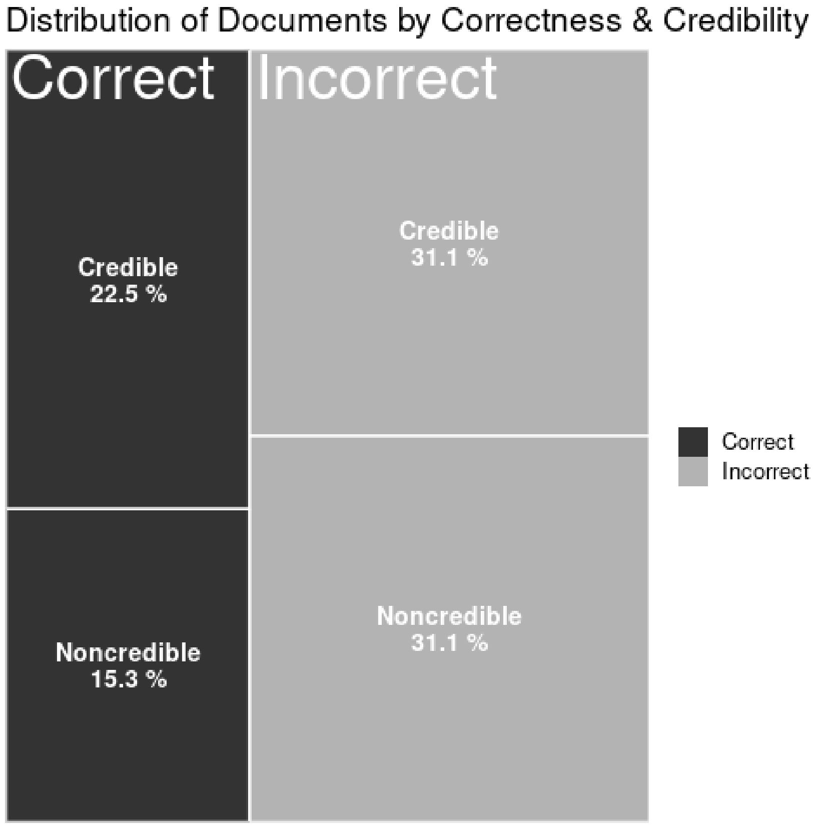
nDCG@10 scores recalculated based on runs' performance on 25 out of 50 topics for which there are at least 10 correct documents in the TREC collection. The stars indicate significance at 0.1 (*), 0.05 (**) and 0.01 (***) levels for runs improving the benchmark.

be relevant. There are 18677 nonrelevant documents, together with another 100 with no information about the treatment's efficacy or that were not judged by mistake. After these documents were removed, 4065 were left for further analysis. Each of these documents is fully labelled with judgments on their topic relevance (3-scale), credibility (binary) and correctness (binary).

The qrels file consists of labels of web pages from 50 topics (17 helpful, 17 unhelpful and 16 inconclusive), each written in form of [treatment] for [issue] with [target group, (if applicable)]. The topics are all positively framed, i.e. about whether or not a medical intervention *treats* a health issue, e.g. exercises for muscle cramp, antibiotics for whooping cough, honey for common cold. On the other hand, there is no topic about potential harms, e.g. "does smoking kill", "does over egg consumption cause heart attack." Although researches with such positively framed questions are more common in the literature, we believe this might affect the distribution of documents' efficacy claim, which we left for further studies.

An overall analysis of the judgments of documents labelled as relevant (and highly relevant) shows that 53% are credible documents, and 39% are correct documents. For 11 topics (7 unhelpful and 4 inconclusive) out of 50, there is no correct document found by the NIST assessors. When these topics are excluded, the overall percentage of correct documents increases to 50% and the distribution improves in favour of credible documents.

As displayed in Figure 4.1, overall, 74.4% of the relevant subset of the collection claims that the intervention is effective independent of the underlying treatment and issue, whereas 4% reject the claim and the remaining 21.6% does not provide any conclusive



Distribution of relevant documents conditioned on credibility and correctness. The percentages sum up to 1.

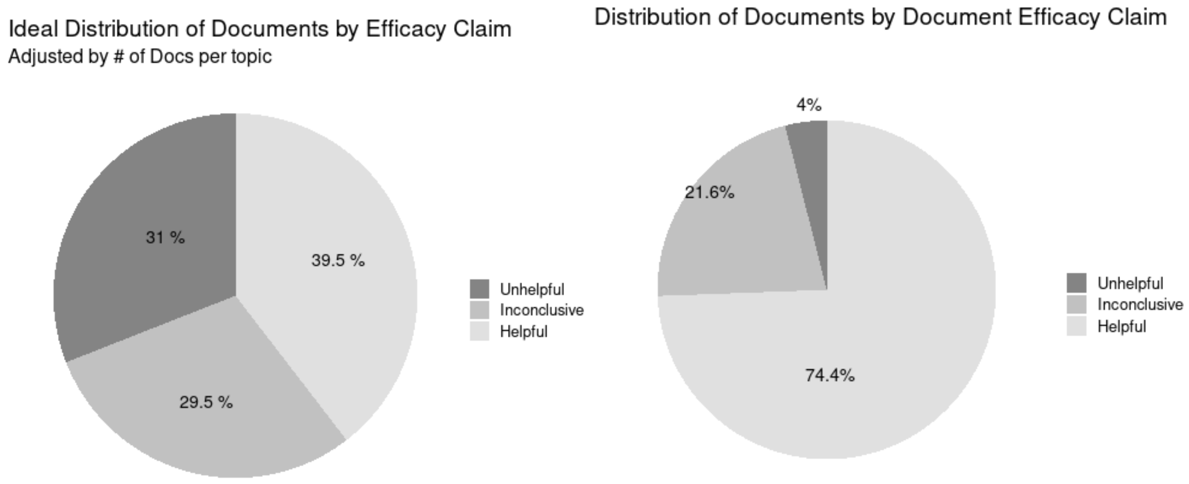


Figure 4.1: Distribution of Relevant Documents by Treatment Efficacy

statement. If all documents presented accurate information, then the distribution would be 31.0% unhelpful, 29.5% inconclusive, and 39.5% helpful. This biased distribution is similar to the distribution reported by White and Hassan (2014). In their study, they sampled logs of sessions with question queries issued by 10 million users of a major search engine and subsetted a collection of documents from the top 1000 of each search results. They pooled the documents resulted in top 10 and documents at every 50 ranks for each SERP of each query, i.e. $r = \{1, 2, \ldots, 10, 50, 100, 150, \ldots 1000\}$. The collection of documents was assessed by crowdsourced workers. In the pool of documents whose rank is up to 100, they found 76.7% of the web documents were towards helpful, 15.1% inconclusive, and 8.2% unhelpful. These statistics, along with the number of correct documents in Table 3.1, indicates that it is harder to find the correct information when treatment is unhelpful than the case when it is helpful.

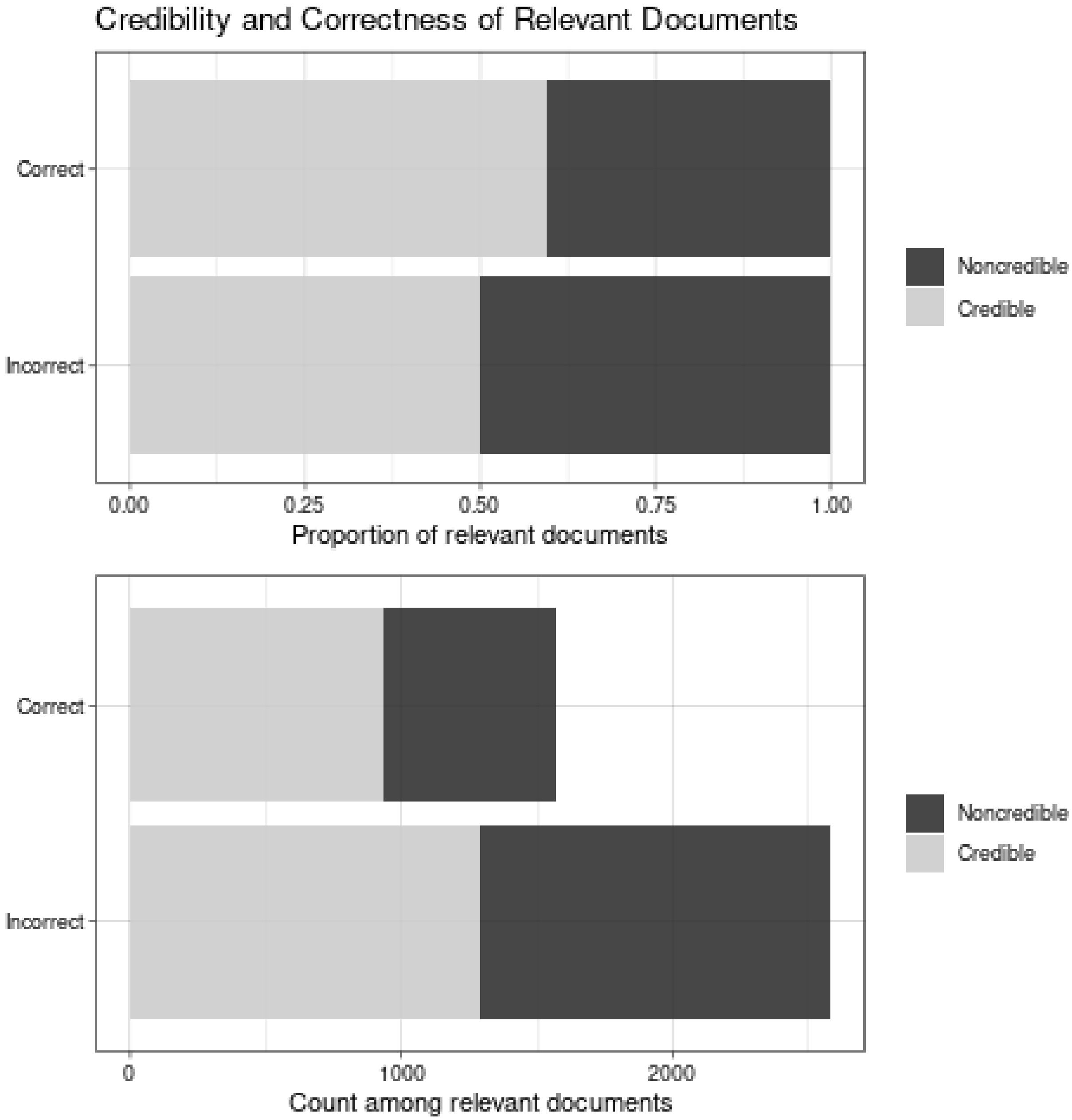


Figure 4.2: Distribution of Correct Documents by Credibility

Table 4.9: Distribution of Correct Documents

Treatment Eff.	Whole Set	Cred.	NonCred.	Whole Set*	Cred.*	NonCred.*
Unhelpful	7.7 %	17.0 %	2.5 %	12.5 %	23.9 %	4.4 %
Inconclusive	18.9 %	16.1 %	23.6 %	28.4 %	29.5 %	27.2 %
Helpful	75.1 %	74.3 %	76.3 %	75.1 %	74.3 %	76.3 %
All topics	37.8 %	42.0 %	33.0 %	48.3 %	53.8 %	42.1 %

Percentage of correct documents in different subsets defined by the combination of row and column names. First column presents the percentage of correct documents in the among topics for which treatment is unhelpful, inconclusive and helpful. The second column is the percentages among credible subset and third is among noncredible subset. The columns marked with '*' present adjusted numbers after removing the 11 topics for which no correct documents were found by the NIST assessors.

Further analysis gives insight into the role of credibility. Table 4.9 presents the percentages of correct documents in the set of all labelled documents, in the credible/noncredible subsets, and among different subsets of topics. It can be seen that credibility does not have a high impact on correctness overall. The difference between the percentages of correct documents within the collection and its credible subset is 4.5%. When the dataset is split into three classes of underlying treatment efficacy, the only notable difference is seen in the subset unhelpful topics; in other words, credibility can be promising when the underlying intervention is ineffective. This may be due to the skeptical voice of credible documents that become critical in reflecting the facts or the ads promoting unhelpful treatments.

The difference is noticeable when the distribution of correct documents is compared within the credible and noncredible subsets of the dataset. As presented in Table 4.9, the proportion of correct documents in the credible subset is 9% more than the noncredible subset. The main contribution is made when the treatment is unhelpful; among these topics, the proportion in the noncredible subset is 2.5%, compared to 17% in the credible subset. In case when the scientific consensus about the underlying intervention is inconclusive, the proportion of correct documents in the credible subset is less than the proportion in the noncredible subset. Although this can be attributed to the skeptical tone of credible documents, as we will discuss later, it may also be due to the assessment guidelines.

When the 11 topics were removed, the distribution changed fundamentally in favour of credible documents with an overall 11.8% increase. In the subset of credible documents, and in the case when the underlying intervention is unhelpful, the percentage of correct documents increased from 17% to 23.9%, and when it is inconclusive, it increased from 16.1% to 29.5%. The adjustments for noncredible documents were relatively smaller.

These more substantial changes for credible documents can be associated with the difference between organizers' interpretation of medical reports and experts in the field who are writing the web content. As presented in Table 3.1, for the topic "acupuncture for epilepsy" 18 documents (out of 26 relevant) were judged as credible, and for "gingko biloba for tinnitus" it is 43 out of 50. However, none of them were labelled as correct, which is very counter-intuitive and demands for further examination. We leave this discussion to the next section and continue our analysis with the remaining 39 topics.

Notice that correctness is a very strict case when the document's efficacy claim perfectly matches with medical consensus; the case that corresponds to the diagonals in Table 4.10. Ideally, the percentages in the diagonals must be 100%, and all else must be zero. In Table 4.10, however, we can see it heavily biased towards helpful. In the subset of 39 topics, 70.3% of the collection already claims that the intervention is helpful regardless of the ground truth, the percentage is 24.6% and 5.1% respectively for inconclusive and unhelpful topics.

This distribution does not change noticeably among the credible and noncredible subsets (see the last column of Table 4.10). The other columns also present similar distributions except for the case when the treatment is unhelpful, in which the credible documents more likely report that treatment is unhelpful, whereas the noncredible documents tend to remain inconclusive.

These results show that a search engine that can successfully retrieve all relevant information to the user's query, but has no motivation to improve credibility or correctness, would be quite successful in correctness task as overall 77% of the relevant documents are correct. However, for the subset of topics for which there is no medical consensus about treatment efficacy, this percentage drops to 30%. In both cases, there is no strong clue that increasing the position of URLs of credible pages in SERPs would have any impact on the correctness of the information retrieved by the search engine. However, when the user queries about the interventions that are unhelpful for the health issue, targeting credible documents may have a positive impact on correctness.

4.4.1 Topicwise Analysis

The cases when credible documents could not reflect the ground truth challenges the hypothesis that the credibility of a document correlates with its accuracy and thus deserves an in-depth examination. In this subsection, we analyze the cases when credible documents did not reflect the ground truth and give insight into the assessment process.

In Table 4.11 we highlighted the above mentioned 11 topics with light grey. Additionally, three more topics were also highlighted for which there are correct documents, but

Table 4.10: Distribution of document efficacy claim by the underlying treatment efficacy and credibility

Credible					
		Treatment efficacy			
		Unhelp.	Inconc.	Helpf.	Overall
	Unhelpful	24.0 %	3.0 %	2.6 %	6.8 %
Document's	Inconclusive	18.4 %	32.1 %	20.6 %	22.9 %
claim	Helpful	57.6 %	64.9 %	76.8 %	70.1 %
	Total	**321**	**396**	**949**	**1666**
NonCredible					
		Treatment efficacy			
		Unhelp.	Inconc.	Helpf.	Overall
	Unhelpful	4.5 %	1.0 %	3.4 %	3.1 %
Document's	Inconclusive	36.1 %	27.8 %	19.3 %	26.6 %
claim	Helpful	59.5 %	71.2 %	77.3 %	70.4 %
	Total	**449**	**385**	**657**	**1491**
All					
		Treatment efficacy			
		Unhelp.	Inconc.	Helpf.	Overall
	Unhelpful	12.6 %	2.0 %	2.9 %	5.1 %
Document's	Inconclusive	28.7 %	30.0 %	20.0 %	24.6 %
claim	Helpful	58.7 %	68.0 %	77.0 %	70.3 %
	Total	**770**	**781**	**1606**	**3157**

Comparison of page efficacy claim and treatment efficacy among credible and non-credible subsets. The percentages are calculated over 39 topics.

Table 4.11: Number of correct documents per topic

Query	Topic ID	Topic Efficacy	#Rel.	#Cred.	#Cor. (C)	#Cor. (NC)
cranberries for urinary tract infections	1	U	100	55	0	8
dehumidifiers for asthma	51	U	24	6	0	4
acupuncture for epilepsy	3	U	26	18	0	0
aspirin for vascular dementia	7	U	10	9	0	0
lumbar supports for lower back pain	17	U	72	32	0	0
electrical stimulation for male urinary incontinence	18	U	18	18	0	0
hydroxyzine for generalized anxiety disorder	22	U	8	1	0	0
ginkgo biloba for tinnitus	32	U	50	43	0	0
hypnotherapy for quit smoking	33	U	305	12	0	0
cinnamon for diabetes	40	U	101	31	2	1
insoles for back pain	47	U	60	12	2	1
probiotics for eczema	42	U	72	45	1	0
vitamins for epilepsy	44	U	39	20	3	1
antidepressants for low-back pain	13	U	49	48	3	0
magnesium for muscle cramps	16	U	144	20	3	0
amygdalin for laetrile cancer	6	U	86	28	18	2
traction for lower back pain	38	U	98	57	45	3
probiotics for bacterial vaginosis	15	I	52	5	0	13
acupuncture for insomnia	2	I	220	195	0	0
yoga for epilepsy	24	I	22	17	0	0
aloe vera for wounds	30	I	116	106	0	0
valerian for anxiety disorder	35	I	66	43	0	0
feverfew for migraines	46	I	72	9	6	37
compression stockings for varicose veins	43	I	82	23	9	36
honey for wound	4	I	111	65	2	4
acupuncture for asthma	48	I	137	10	6	6
vaccine for common cold	26	I	11	9	2	1
ear drops for ear wax removal	9	I	39	35	5	2
insulin for gestational diabetes	23	I	85	85	1	0
exercise for hot flashes night sweats menopause	31	I	36	28	1	0
fish oil for ulcerative colitis	25	I	45	21	12	3
acupuncture for vascular dementia	21	I	114	99	53	4
gene therapy for sickle cell	10	I	43	42	30	1
melatonin for jet lag	8	H	107	38	26	65
exercises for female incontinence	34	H	95	30	17	43
muscle relaxants for back pain	39	H	85	26	25	51
acupuncture for migraine	5	H	85	34	21	35
antibiotics for wet cough in children	27	H	5	2	1	2
caffeine for asthma	45	H	40	18	15	14
dental sealants for cavities	36	H	166	86	86	80
honey for cough in children	19	H	118	65	63	52
sulfasalazine for rheumatoid arthritis	49	H	74	51	8	5
exercise for lower back pain	11	H	190	112	109	70
antibiotics for whooping cough	28	H	78	50	17	9
laxatives for hemorrhoids	37	H	56	29	25	11
antibiotics for children with pneumonia	29	H	108	79	66	27
surgery for obesity	50	H	212	157	106	37
steroids for spinal cord injury	20	H	16	15	10	0
circumcision for hiv	12	H	151	139	95	6
benzos for alcohol withdrawal	41	H	66	51	39	1

The rows highlighted with dark gray are the topics for which no credible and correct information were found and light gray are the ones for which no correct information were found by the NIST assessors.

none of them were judged as credible. These cases are of our particular interest since both give insights about the challenges when developing support systems. Besides, given that the search engines are strong enough to retrieve all relevant information, this case limits the capacity of the credibility of the information.

The topics for which no correct documents found by the assessors are the ones with unhelpful treatments and the ones that there is no medical consensus ("inconclusive"). There are also many others for which, relative to the number of relevant documents, the NIST assessors found very few correct documents. On the other hand, there are 2 topics out of 17 topics with unhelpful treatments support the hypothesis that the credibility can be useful to target correct documents. This number is 3 out of 17 among topics with "inconclusive" treatments.

We are unsure as to why credible documents could not represent the ground truth, however, some topics give hints. For example, "cinnamon for diabetes" has been a controversial topic with mixed results (Rafehi et al., 2012; Allen et al., 2013) and there are recent studies have been reporting that cinnamon improves diabetes symptoms (Maierean et al., 2017; Kizilaslan and Erdem, 2019). It is possible that the scientific consensus known to the content writers was supporting the efficacy on the date when the content was written. As the ground truth about this topic in our study depends on the Cochrane Review written in September 2012 (Leach and Kumar, 2012), and the documents in ClueWeb12 were crawled early in 2012, the scientific consensus may have very well changed over time.

On the other hand, it would not be wrong to claim that "amygdalin for laetrile cancer" is a less controversial topic as, the to the best of our knowledge, there is not yet any supplementary product that can improve cancer symptoms, although we cannot support our claim with medical researches. Nevertheless, it may be easy for credible content writers to represent the ground truth for some topics.

For the rest of our discussion, we will focus on the three topics for which there were no correct and credible document. We will present our observations from in-depth reading, and give insight into the assessments.

In-depth Reading

For Topic 51, "dehumidifiers for asthma", the treatment was decided to be unhelpful according to the organizers' interpretation of Cochrane medical reports. In the dataset, 4 of 6 documents assessed as credible were assessed as "helpful" and the remaining 2 as "inconclusive". Among the four helpful documents, 1) was a newspaper article stating that "A good air-filtration system can make a difference for people with severe allergies or asthma, ... Keep in mind that studies have not proved that any filters dramatically reduce allergy or asthma symptoms"; 2) university page that suggested humidity control could reduce mould growth and, thus, the asthma episodes; 3) was a blog; and 4) was a web page selling air control devices, and were both assessed as credible. We disagree with the credibility judgments of these two documents.

One of the two inconclusive credible pages is a newspaper giving air control tips and hoping that it will help to keep "asthma and allergy symptoms under control". Another document promoting a product, which contained the phrase "DeLonghi dehumidifiers ... help prevent water damage, control odors, mildew, mold, allergies and asthma" was judged as inconclusive. Considering the topic's description ("Can dehumidifiers be used to control asthma?"), we believe these documents both were misjudged.

Among the four noncredible and correct (assessed as "unhelpful") documents, three were websites selling asthma products, which gave details about dehumidifiers but did not mention asthma in the text. The last was a Q&A forum advising on asthma control, "If you find that humidity is a bother, you may want to check into getting a good dehumidifier", also assessed as unhelpful. According to the assessing guidelines[2], the first can be labelled as relevant but we disagree that any of these claim that dehumidifiers are "unhelpful" for asthma.

Among the 52 relevant documents of Topic 15 ("probiotics for bacterial vaginosis"), five were labelled as credible, but all of the five documents' efficacy claims were judged as "helpful," whereas according to the medical report, the treatment efficacy is inconclusive. On the other hand, 13 noncredible documents were judged as inconclusive, therefore correct.

All of the five credible pages have phrases claiming efficacy. One among the 13 noncredible correct ("inconclusive") documents was selling probiotics products and claims that "Probiotic treatment ... may be helpful for such common female urogenital problems as bacterial vaginosis, ...". Two documents were blog and one was a forum where people share their personal experiences with BV and probiotic products without claiming efficacy. Another page had links to other pages with positively framed titles such as "Utilizing Probiotic Yogurt For Treating BV" and phrases such as "This can be done with probi-

[2] The guidelines can be downloaded from https://github.com/trec-decision/trec-decision.github.io/raw/master/docs/AssessingGuidelines.pdf

otic supplements such as Lactobacillus...". Another was a document about the probiotic, Lactobacillus Acidophilus, and the only relevant paragraph claims, "several studies have assessed the potential of lactobacilli in the prevention or treatment of certain genitourinary tract infections such as bacterial vaginosis, ...". Four documents were not loaded properly but contain phrases such as "You really need to ... build you own bodies natural bacteria balance to help prevent bacterial vaginosis". Another page was dedicated to natural remedies, and the only phrase relevant to the topic is "Some experts suggest inserting unflavoured yogurt into the vagina (directly or soak a tampon) to help restore the lactobacillus population.". One said "[probiotics] don't target the underlying cause of BV, but simply providing temporary relief". There is only one page claiming "there isn't enough evidence yet to recommend it over conventional approaches".

The documents discussing Topic 1 are all incorrect except eight that were judged as noncredible. The topic discusses the efficacy of cranberries for urinary tract infections, which the organizers labelled as "unhelpful" by reading the medical reports. Six of eight correct documents explicitly support efficacy claim with phrases such as "cranberry juice has also been recommended by doctors to help prevent occurrences of a Urinary Tract Infection". One does not contain any information about cranberries. Only one out of eight discusses that there is no evidence to support the efficacy.

The three topics provide insight into the TREC document assessment process. Although the assessment guidelines give some freedom to the human assessors, many examples above can hardly be attributed to individual differences. We believe these errors may be due to (i) differences between the traditional TREC tracks that only code for topic relevance, and the 2019 Decision Track, which also required assessors to code for credibility and efficacy, and (ii) the time limitations on the judges who were expected to spend around 30 seconds per document. In fact, it unexpectedly took them longer to evaluate each document, and Topic 14 was not assessed due to these difficulties. Nevertheless, these are inevitable and cannot be entirely eliminated, which leads us to the following section.

Summary of Assessment Errors

The above examples show that errors may occur at different levels, either during the judgment process or experiment design. Some of these errors may not have a noticeable impact on the scores, while others may result in significant deviance. Nevertheless, these error types demand deeper analysis. For the first type, we have presented examples where the misjudged document explicitly supports the efficacy and has no information to support otherwise. We also presented some examples of online shopping websites that were judged as credible; and pages that are irrelevant to the topic but labelled as correct.

These error types give some insight into the difficulties when developing successful search engines. One problem is that there is an error rate in the judgments, which prevents us from perceiving the correct judgment. This error may be random, but can also be more systematic. For example, according to the TREC assessment guidelines, a document is inconclusive if it discusses both options but not sharply supports one. Finding pieces of information that present both viewpoints may be harder and erroneous judgments are more likely in such cases. Moreover, Cochrane reviews of some topics (e.g. “acupuncture for epilepsy”) are not clear enough and can be categorized as inconclusive or unhelpful. In such cases, correct and credible documents, which read very similar to the medical reports, can easily fall into the efficacy category that is different from the ground truth's and thus mapped to “incorrect”.

There is another systemic problem in the design of the experiment. According to the TREC assessment guidelines, a document is inconclusive if it “mentions the medical intervention but does not provide any information on its efficacy, benefits, or disadvantages”. Forums where users share their experiences, websites with pages of internal search results with the relevant titles and online shopping pages may fall into this category. However, this type of “inconclusive” has no relationship with the ground truth that there is no scientific consensus. In such cases, many noncredible documents will be mapped to “correct”.

In summary, the three topics we analyzed were challenging the hypothesis that credibility correlates with the accuracy of the information. But the above examination shows that there are errors in judgment processes and experimental design that may be weakening the measured relationship. Some of these errors are idiosyncratic and are inevitable, but the rest are systematical, and a better design would overcome these difficulties. Also note that if the assessment were flawless, these three topics would not have large number of credible and correct documents relative to the relevant number of documents. Nevertheless, the analysis shows some limitations on our study.

Chapter 5

Discussion & Conclusion

5.1 Discussion of Results

In our study, we proposed algorithms to improve the correctness and credibility of information presented in SERPs. We hypothesized that this could be done by targeting the documents of higher credibility. Our runs improved the credibility and correctness of SERPs, but the improvements in the task to retrieve only correct documents were relatively smaller. We then analyzed the annotated collection to measure the relationship.

The performance results in Section 4.1 and 4.3 show that we can improve the credibility of information presented in search engines by employing a logistic regression model. When we used the credibility classifier not to filter noncredible documents but to reorder the ranks by transforming retrieval scores, we improved the baseline nDCG@10 scores by 19%. These scores are higher than manual runs by Abualsaud et al. (2019) who used HiCAL for this task and improved the same baseline nDCG@10 scores by 6.06%. The gap is unexpected because HiCAL puts a human in the loop who manually evaluates each document and is more likely to find documents that satisfy the criteria. Nevertheless, the improvements in overall 50 topics were insignificant with some exceptional cases.

There are three main reasons why it is hard to improve credibility in the TREC Decision Track. First, due to the design of the track, a document is considered as credible if it is also relevant to the topic. Therefore, filtering out (or reordering) the noncredible documents from the top BM25 scores even with a perfect credibility classifier will have a partial impact since we do not have a perfect algorithm to retrieve all documents relevant to the topic. In other words, the classifier will leave many credible but nonrelevant documents. Second, it is not possible to improve the search engine when there is no relevant document to the query or no credible document.

The third reason is that the assessment guidelines allow for a degree of subjectivity, which is personal by nature, and NIST assessors' perceived credibility might not overlap with the researchers'. The manual runs' credibility performances by Abualsaud et al. (2019) present a good example of why such a discrepancy matters in this task.

With this in mind, we also analyzed the credibility classifier's performance and the extent to which our perceived credibility overlaps with NIST assessors'. Although we spent effort to make sure that Decision Track's topics and our training corpus are not related, we found 77 documents that were in both our corpus and the qrels file. Although the set is too small, it gives insight into the differences in perceived credibility judgments, where we were more conservative in labelling a document as credible. This was also reflected on the performance of our classifier in predicting credibility of the track's documents, where the accuracy is 0.5976. This small score can be partially attributed to our classifier's architecture as the 10-fold validation accuracy is 0.6727 when the train/validation sets were split based on the topics. However, when we calculated 10-fold validation on qrels shared by NIST by splitting 50 topics into 10 groups, the accuracy is 0.9358. In other words, when the classifier was trained on NIST judgments and predicting the credibility of documents that are out of the training topics, the score is quite high. This is also reflected in the runs' performances as they improved the baseline very significantly in credibility related tasks.

On the other hand, we improved correctness by employing classifiers trained on NIST assessors' judgments. Note that the improvement seems partly due to the spaminess of the pages. When the documents having spaminess scores less than 30 were removed, the scores improved 11.59% relative to the baseline, which is quite close to scores of runs using credibility filters (14.84%) and runs that use credibility for reordering (15.96%). Nevertheless, the contributions are less than the manual runs' reported by Abualsaud et al. (2019) (23.76%).

Here the correctness scores are generally low, and the improvements over the baseline are insignificant for the similar reasons we mentioned for credibility. First, a document cannot be correct unless it is relevant to the topic, and the performance principally relies on the retrieval scores by the BM25 algorithm. Different than credibility, we are using credibility by hoping it will filter out the incorrect documents or reduce their positions in the SERPs. However, as the analysis of the dataset indicates, 42.0% of the credible subset of the collection is correct. This adds only 4.2% to an overall 37.8% correct proportion in the whole dataset. Therefore credibility is unlikely to have a marginal impact on the runs' performances in the correctness task.

It is also important to note that credibility, correctness and relevance are correlated with each other. In our runs, although not significant, credibility classifier improved relevance by 5.59%, which can be attributed to a similar effect of spam filtering. Similarly, the manual runs by Abualsaud et al. (2019) achieved higher scores in the correctness task (23.76%) than credibility (6.06%) even though the researchers labelled documents based on their relevance and credibility only.

As we mentioned above, one reason why the runs did not perform well in correctness task is that, for some topics, there were no correct and credible documents found in the pool of documents that the NIST assessors labelled, or the number of correct documents was very small. For both cases, all runs performed very poorly (mostly scored 0) in credibility and correctness tasks, which makes it harder to reject the hypothesis that the runs perform differently than the baseline. Removing these topics improved the p-values, and gradually we achieved some significant results at 0.01 level.

Another problem we mentioned is the subjectivity of the credibility judgments. When we trained our classifiers using NIST assessors' credibility judgments, the runs achieved higher scores than the submitted runs. In the correctness task, most runs that were submitted to track performed worse than the baseline, but when trained the classifier with NIST's judgments, they improved the baseline by 14% to 19%. In the credibility task, the runs' performances improved dramatically; over the subset of 25 topics (with more than 10 correct documents), three runs performed better than the benchmark at 0.01 significance level. In almost all cases, the best performing run was the UWatMDS_BM25_ZS that modifies the BM25 rankings by first removing the most spamy documents, then reordering the remaining based on their credibility classifier scores.

The main goal of this thesis is to promote both correct and credible information in the SERPs. Our runs performed well in this task. The runs that reorder the documents based on their credibility scores improved the baseline from 20% to 41%, and the runs using NIST judgments were generally better than the submitted runs performing better than the baseline at 0.05 level. When we removed the 11 topics for which NIST assessors did not find any credible and correct documents, the best performing run became UWatMDS_BMF_C95, which is filtering noncredible documents out using the classifier trained with our judgments.

It is worth asking how the classifiers can predict the documents' credibility with a accuracy. Our literature review implies some parallelisms between the nature of human credibility judgment and how the classifiers work. Overall, the literature suggests that users evaluate credibility in two steps, first by finding peripheral cues such as page design, layout, document type (e.g. blog, hospital page), availability of author credentials and references, and then with an in-depth examination of these signals, such as the use of language and reputation of the author or the organizer. Metzger (2007) also suggest that many users do not go to the second step due to cognitive limitation unless they have appropriate skills and make their decisions based only on peripheral cues.

The nature of our classifier is analogical to these theories to some extent. We trained our classifier on the raw HTML codes, which involve the cues mentioned above, including the page design, layout, colouring, the author credentials. The supervised classifier can also extract some details about the source reputation (e.g. MayoClinic, Oxford University, Center for Disease Control), some keywords about the use of language (e.g. holistic solutions, herbal remedies). Moreover, it can also extract features that the user may never notice, such as some outlinks to credible pages, or health on the net code (HONcode) issued by a health credibility tracking organization.

5.2 Weakness and Limitations

There are a couple of limitations and weaknesses of our study that deserve to be mentioned here. Our study is based on TREC Decision Track that is designed in a collaboration between University of Queensland, University of Copenhagen and University of Waterloo. The author actively contributed when selecting the topics and determining treatment efficacy based on the medical reports by Cochrane Review.

When selecting the topics, we were primarily inspired by previous studies without being aware of the extent to which they are represented in the ClueWeb12-B13 collection. As a result, in the TREC Decision Track collection, for 14 topics, there was no correct and credible document, and for 11 others, there were less than 10 correct documents. After removing these topics, 2, 7 and 16 topics remained for which the treatment is unhelpful, inconclusive, and helpful, respectively. Therefore the topic distribution itself was inherently biased towards "helpful". When we regard the types of noncredible documents such as ads to promote a product, this biased distribution will put credibility classifiers at a disadvantage, because many noncredible documents will naturally be correct when the topic is discussing a helpful treatment. A better research design must ensure there are enough correct documents for this type of research.

Moreover, when training the credibility classifier, we constructed our training and test corpora by determining 30 topics. We retrieved 1000 documents per topic and evaluated the credibility of 3533. Although we ensured there is no topic jointly included in our topic list and in the track's, there were 77 evaluated documents in our training corpus that were also in the track's qrels file. Therefore, our training corpus was not entirely distinct from the documents that were evaluated in the track. Nevertheless, our analysis also includes runs using a credibility classifier trained on NIST judgments, which performed better overall. During its training, we ensured that no document was shared in both training and test sets.

Our literature review discusses that the perceived credibility assessments are subjected to biases and this includes us and NIST assessors. Briefly, when the user is displayed too much information, to reduce the effort and come up with a decision, they focus more on peripheral cues that are insufficient for their purpose. Recall that the assessors were not exposed to the mentioned biases in the same way the searchers were exposed to as the collection is old and the vast majority of the pages were not displayed properly. However, one can argue that they were exposed to different type of bias as they had limited time to label the documents.

Another weakness is regarding the difficulties in assessment and the size of the dataset. 22842 documents were labelled by the NIST assessors; however, when a document was irrelevant to the topic, its credibility was not evaluated to reduce the labour. In return, 4065 documents could be used for training. The size is relatively small for today's technologies, such as DNNs, which could be useful to improve credibility. However, it would be infeasible to evaluate all documents' credibility regardless of their topic relevance and might not be appropriate for the previously mentioned connection between credibility and relevance.

Lastly, our study only used a character-based logistic regression to detect credibility, which might not be sophisticated enough. The choice of character 4-grams was arbitrary, and we did not optimize among other values of n using k-fold validation. Moreover, the alternative methods mentioned in the literature review, such as biased PageRank algorithms and more complex models could be employed in this research. We did not go further due to the computational difficulties and time constraints.

5.3 Conclusions

In this thesis, we addressed credibility and correctness in health-related searches, and proposed algorithms to promote correct and credible information presented in SERPs. We also analyzed the TREC Decision Track collection and discussed the limitations of the improvement.

In our literature review, we first addressed the notion of credibility. The studies on perceived credibility show that users often do not evaluate the cues recommended by experts. Instead, they focus more on superficial cues, but depending on the users' involvement and research skills, they may also evaluate the organizer behind the website, author credentials and other such details. Moreover, perceived credibility models suggest that users evaluate credibility in multiple steps, first by noticing some cues and later evaluating them to make a final decision. We argue that is model can be modelled by machine learning models.

Studies in Human-Computer Interaction indicate that human cognitive heuristics are malfunctioning during health searches so that users tend to put more trust on the pages presented at a higher position in the SERPs. Nevertheless, this tendency can be used to help users to make better decisions; in other words, if the users are presented with the correct information, their decisions can be improved.

Motivated by this fact, we proposed algorithms that improved the ranks of correct and credible documents in the SERPs. Assuming that the correctness of the information presented in the document correlates decently with the credibility of the information, our algorithms targetted credible documents. To this end, we trained a character-based logistic regression classifier on raw HTML files to evaluate all cues in the document, including design and layout, colour, author credentials, content. Our runs improved the performance of baseline BM25 run in finding correct & credible documents around 40% to 50%. The most consistent results were achieved by reordering documents, using a credibility classifier's scores to adjust retrieval scores after soft spam filtering. This method was also shown to be more successful in credibility and correctness tasks separately. Moreover, these algorithms did not harm the relevance performances. However, in the most strict case to find correct & credible documents, the best performance was achieved by filtering documents based solely on their credibility scores.

These improvements were not significant overall. The reason can be (i) the discrepancy between our perceived credibility vs. NIST assessors', (ii) the design of the task, which gives priority to relevance, and (iii) the fact that there were not enough correct and/or credible documents found for some topics. To further investigate our algorithms' effectiveness, we excluded such topics from our analysis and trained our algorithms with NIST judgments rather than our credibility judgments. Thanks to the classifier that can accurately predict the credibility of a document, these changes gradually improved the performances and their significance in improving the baseline.

Our analysis of the annotated collection measured the relationship between credibility and correctness. We found that the correctness of credible documents distinguishes from the noncredible documents when the underlying treatment is unhelpful according to the scientific consensus. However, when the treatment efficacy is helpful, or the scientific consensus is inconclusive, the proportion of correct documents in the credible subset is not significantly different from the whole collection.

Our analysis also showed that this type of research is subject to systematic errors. One challenge is the assessment errors, which can be deceptive; an in-depth analysis of three topics shows that they can cause large measurement errors, and eventually removing some topics from the analysis. Another challenge arises from experiment design. We found that the definition of inconclusiveness eventually caused some documents to be regarded as correct, even though they did not reflect the scientific consensus. These errors challenge engineers when developing decision support systems.

5.4 Future Directions

Our analysis indicates that there is a degree to which credibility classifier can improve the search engines. In our study, we used a character-based logistic regression model to extract both peripheral and content cues and improved the performances of the runs. One can use more complicated methods, such as character-based DNNs, to increase classification accuracy. Such an approach will be useful to detect credibility since bag of word based methods, such as logistic regressions, cannot capture the contextual dependencies and patterns.

The models on perceived credibility, e.g. Dual Processing Model, indicate that users' decisions are more influenced by superficial cues such as page design and layout than central cues. One can train a Convolutional Neural Network (CNN) model on screenshots of the web documents to better extract these features to imitate how users perceive credibility, with each convolution focusing on different aspects of the page. This can be favourable because visual cues may be too hard to be captured from raw HTML, CSS and JS codes. Also, the web content can also be inputted to CNN to extract central cues, such as authors affiliation and organization behind the web page. However, training neural networks demands a larger training set, which is not possible for our study regarding the scale of our experiment. Also, the documents that we used in our study are old and many cannot be displayed properly. Nevertheless, if one can construct a larger training corpus, it is possible to improve classification scores to a greater extent.

Another improvement can be made by separating content cues and peripheral cues and training a classifier accordingly. Literature indicates three types of credibility, source, medium and message, which can be examined separately with proper cleaning methods that will separate different types of features. One can also extract document content using an HTML cleaner to tokenize it into words, tokenize the rest of the code into characters, and to train separate classifiers. These scores can be combined either with, e.g. linear transformations, or in layers of a deep neural network.

The above can help determine reputation and source credibility to a certain degree, but an analysis of the document's network properties has also proven to be effective for this task. Castillo et al. (2007) discusses that content and link-based filtering are orthogonal to each other, and it is worth using TrustRank (Gyöngyi et al., 2004), SpamRank (Benczur et al., 2005) and other biased PageRank algorithms in the context of credibility. This type of scoring is particularly suitable for source and reputation credibility and has the potential to improve the scores substantially.

One problem in calculating the scores to rank the documents was to determine how to combine them. We decided on the combining rules manually by comparing the performance of runs on a test corpus. A better method would be using learning to rank algorithms to come up with a better rule.

Lastly, our analysis also shows that credibility may not be the ultimate solution to present the correct information to the users. One way to solve the problem could be stance detection/opinion mining to detect whether or not the document claims the treatment is effective. Accordingly, a document will remain in the SERP if its position towards treatment efficacy is aligned with medical consensus. One challenge of using such an approach is the presumption that the search engine is effectively retrieving highly relevant documents; in fact, one document may discuss more than one treatment for the health issue and mining its opinion may be challenging. Besides, this approach assumes that the search engine can automatically detect searchers' intentions and match the query with scientific consensus effectively. Although this can be achieved for some popular topics, it may not be generalizable with today's information systems.

Bibliography

Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M. D., Cormack, G. V., and Grossman, M. R. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '18, pages 1317–1320, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210176. URL http://doi.acm.org/10.1145/3209978.3210176.

Abualsaud, M., Beylunioğlu, F. C., Smucker, M. D., and Duimering, R. P. UWaterlooMDS at the TREC 2019 Decision Track. 2019.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194, 2008.

Allen, R. W., Schwartzman, E., Baker, W. L., Coleman, C. I., and Phung, O. J. Cinnamon use in type 2 diabetes: an updated systematic review and meta-analysis. *The Annals of Family Medicine*, 11(5):452–459, 2013.

Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, pages 295–303, 1951.

Barthes, R. *Image-music-text.* Macmillan, 1977.

Becchetti, L., Castillo, C., Donato, D., Baeza-Yates, R., and Leonardi, S. Link analysis for web spam detection. *ACM Transactions on the Web (TWEB)*, 2(1):2, 2008.

Benczur, A. A., Csalogany, K., Sarlos, T., and Uher, M. Spamrank–fully automatic link spam detection work in progress. In *Proceedings of the first international workshop on adversarial information retrieval on the web*, pages 1–14, 2005.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

Bessière, K., Pressman, S., Kiesler, S., and Kraut, R. Effects of internet use on health and depression: a longitudinal study. *Journal of Medical Internet Research*, 12(1):e6, 2010.

Castillo, C., Donato, D., Gionis, A., Murdock, V., and Silvestri, F. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

Castillo Ocaranza, C., Mendoza, M., and Poblete Labra, B. Predicting information credibility in time-sensitive social media. 2013.

Chaiken, S. The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39. Hillsdale, NJ: Lawrence Erlbaum, 1987.

Conroy, N. J., Rubin, V. L., and Chen, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

Cormack, G. V. University of waterloo participation in the trec 2007 spam track. In *TREC*, 2007.

Cormack, G. V., Smucker, M. D., and Clarke, C. L. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

Davison, B. Propagating trust and distrust to demote web spam. 2006.

Davison, B. D. Recognizing nepotistic links on the web. *Artificial Intelligence for Web Search*, pages 23–28, 2000.

De Choudhury, M., Morris, M. R., and White, R. W. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1365–1376. ACM, 2014.

Dori-Hacohen, S. Controversy analysis and detection. 2017.

Dori-Hacohen, S., Yom-Tov, E., and Allan, J. Navigating controversy as a complex search task. In *SCST@ ECIR*. Citeseer, 2015.

Eiron, N., McCurley, K. S., and Tomlin, J. A. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, 2004.

Fetterly, D., Manasse, M., and Najork, M. Detecting phrase-level duplication on the world wide web. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177, 2005.

Fischer, P. and Greitemeyer, T. A new look at selective-exposure effects: An integrative model. *Current Directions in Psychological Science*, 19(6):384–389, 2010.

Fischer, P., Jonas, E., Frey, D., and Schulz-Hardt, S. Selective exposure to information : The impact of information limits. *European Journal of Social Psychology*, 35(4):469–492, 2005.

Fischer, P., Schulz-Hardt, S., and Frey, D. Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information. *Journal of personality and social psychology*, 94(2):231, 2008.

Fogg, B., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., et al. What makes web sites credible? a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68, 2001.

Fogg, B. J. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723, 2003.

Fogg, B. J. and Tseng, H. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87, 1999.

Fox, S. and Duggan, M. Majority of adults look online for health information, 2013. URL https://www.pewinternet.org/2013/01/15/health-online-2013/.

Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P., Castillo, C., Mendoza, M., and Poblete, B. Predicting information credibility in time-sensitive social media. *Internet Research*, 2013.

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, K. P. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*, pages 61–70, 2012.

Gigerenzer, G. and Todd, P. M. *Simple heuristics that make us smart.* Oxford University Press, USA, 1999.

Ginsca, A. L., Popescu, A., Lupu, M., et al. Credibility in information retrieval. *Foundations and Trends® in Information Retrieval*, 9(5):355–475, 2015.

Grier, C., Thomas, K., Paxson, V., and Zhang, M. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37, 2010.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587. VLDB Endowment, 2004.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555, 2009.

Helft, P. R., Eckles, R. E., Johnson-Calley, C. S., and Daugherty, C. K. Use of the internet to obtain cancer information among cancer patients at an urban county hospital. *Journal of Clinical Oncology*, 23(22):4954–4962, 2005.

Hovland, C. I. and Weiss, W. The influence of source credibility on communication effectiveness. *Public opinion quarterly*, 15(4):635–650, 1951.

Hovland, C. I., Janis, I. L., and Kelley, H. H. Communication and persuasion. 1953.

Hu, Y. and Sundar, S. Effects of online health sources on credibility and behavioral intentions. *Communication research*, 37(1):105–132, 2010.

Kahneman, D. and Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

Kalbfleisch, P. J. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. In *Communication yearbook 27*, pages 307–350. Routledge, 2003.

Kastenmüller, A., Fischer, P., Jonas, E., Greitemeyer, T., Frey, D., Köppl, J., and Aydin, N. Selective exposure: The impact of framing information search instructions as gains and losses. *European Journal of Social Psychology*, 40(5):837–846, 2009.

Kizilaslan, N. and Erdem, N. Z. The effect of different amounts of cinnamon consumption on blood glucose in healthy adult individuals. *International Journal of Food Science*, 2019, 2019.

Koh, Y. J. and Sundar, S. S. Effects of specialization in computers, web sites, and web agents on e-commerce trust. *International journal of human-computer studies*, 68(12): 899–912, 2010.

Kruger, J., Wirtz, D., Van Boven, L., and Altermatt, T. W. The effort heuristic. *Journal of Experimental Social Psychology*, 40(1):91–98, 2004.

Lau, A. Y. and Coiera, E. W. Do people experience cognitive biases while searching for information? *Journal of the American Medical Informatics Association*, 14(5):599–608, 2007.

Lauckner, C. and Hsieh, G. The presentation of health-related search results and its impact on negative emotional outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 333–342, 2013.

Leach, M. J. and Kumar, S. Cinnamon for diabetes mellitus. *Cochrane database of systematic reviews*, (9), 2012.

Lee, K., Caverlee, J., and Webb, S. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442, 2010.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

Liao, Q. V. and Fu, W.-T. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2359–2368. ACM, 2013.

Liao, Q. V. Effects of cognitive aging on credibility assessment of online health information. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 4321–4326. 2010.

Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology*, 25(2):115–126, 2006.

Lucassen, T., Muilwijk, R., Noordzij, M. L., and Schraagen, J. M. Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology*, 64(2):254–264, 2013.

Ludolph, R., Allam, A., and Schulz, P. J. Manipulating google's knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *Journal of medical Internet research*, 18(6), 2016.

Maierean, S. M., Serban, M.-C., Sahebkar, A., Ursoniu, S., Serban, A., Penson, P., Banach, M., Lipid, analysis Collaboration, B. P. M., et al. The effects of cinnamon supplementation on blood lipid concentrations: a systematic review and meta-analysis. *Journal of clinical lipidology*, 11(6):1393–1406, 2017.

Mejova, Y., Zhang, A. X., Diakopoulos, N., and Castillo, C. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*, 2014.

Metzger, M. J. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology*, 58(13):2078–2091, 2007.

Metzger, M. J. and Flanagin, A. J. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics*, 59:210–220, 2013.

Metzger, M. J., Flanagin, A. J., and Medders, R. B. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 60(3):413–439, 2010.

Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450, 2012.

Novin, A. and Meyers, E. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 175–184, 2017.

Pogacar, F. A., Ghenai, A., Smucker, M. D., and Clarke, C. L. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 209–216. ACM, 2017.

Rafehi, H., Ververis, K., and Karagiannis, T. Controversies surrounding the clinical potential of cinnamon for the management of diabetes. *Diabetes, Obesity and Metabolism*, 14(6):493–499, 2012.

Rieh, S. Y. Judgment of information quality and cognitive authority in the web. *Journal of the American society for information science and technology*, 53(2):145–161, 2002.

Rieh, S. Y. and Danielson, D. R. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364, 2007.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Ross, L., Greene, D., and House, P. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301, 1977.

Samal, L., Saha, S., Chander, G., Korthuis, P. T., Sharma, R. K., Sharp, V., Cohn, J., Moore, R. D., and Beach, M. C. Internet health information seeking behavior and antiretroviral adherence in persons living with hiv/aids. *AIDS Patient Care and STDs*, 25(7):445–449, 2011.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

Sillence, E., Briggs, P., Harris, P. R., and Fishwick, L. How do patients evaluate and make use of online health information? *Social science & medicine*, 64(9):1853–1862, 2007.

Simon, H. A. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

Simon, H. A. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

Stanford, J., Tauber, E. R., Fogg, B., and Marable, L. *Experts vs. online consumers: A comparative credibility study of health and finance Web sites.* Consumer Web Watch, 2002.

Sundar, S. S. *The MAIN model: A heuristic approach to understanding technology effects on credibility.* MacArthur Foundation Digital Media and Learning Initiative, 2008.

Thon, F. M. and Jucks, R. Believing in expertise: How authors' credentials and language use influence the credibility of online health information. *Health communication*, 32(7): 828–836, 2017.

Todd, P. M. and Gigerenzer, G. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(5):727–741, 2000.

Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International journal of human-computer studies*, 64(11):1071–1083, 2006.

Trotman, A., Jia, X.-F., and Crane, M. Towards an efficient and effective search engine. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 40–47, 2012.

Tversky, A. and Kahneman, D. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

Voorhees, E. M. and Harman, D. Overview of trec 2002. In *Trec*, 2002.

Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science*, 359 (6380):1146–1151, 2018.

Warnick, B. Online ethos: Source credibility in an "authorless" environment. *American Behavioral Scientist*, 48(2):256–265, 2004.

Wathen, C. N. and Burkell, J. Believe it or not: Factors influencing credibility on the web. *Journal of the American society for information science and technology*, 53(2):134–144, 2002.

WebWatch, C. A matter of trust: what users want from web sites. *Consumer WebWatch, Yonkers, NY, http://www. consumerwebwatch. org/news/report1. pdf accessed*, 17, 2002.

White, R. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. ACM, 2013.

White, R. W. Belief dynamics in web search. *Journal of the Association for Information Science and Technology*, 65(11):2165–2178, 2014.

White, R. W. and Hassan, A. Content bias in online health search. *ACM Transactions on the Web (TWEB)*, 8(4):25, 2014.

White, R. W. and Horvitz, E. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)*, 27(4):1–37, 2009.

White, R. W. and Horvitz, E. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)*, 33(4):18, 2015.

Wierzbicki, A. *Web Content Credibility.* 2018.

Wu, B., Goel, V., and Davison, B. D. Topical trustrank: Using topicality to combat web spam. In *Proceedings of the 15th international conference on World Wide Web*, pages 63–72, 2006.

Zhou, X. and Zafarani, R. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.