SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di Informatica Sistemistica e Comunicazione

Dottorato di Ricerca in Informatica
Ciclo XXXVI

Addressing the Challenge of Online Health Misinformation:Detection, Retrieval, and Explainability

Author: Upadhyay Rishabh Gyanendra

Registration number: 865291

Supervisor: Prof. Marco Viviani

Co-Supervisor: Prof. Gabriella Pasi

Coordinatore / Coordinator: Prof. Leonardo Mariani

ACADEMIC YEAR 2022/23

Abstract

In today's digital age, online platforms serve as a primary conduit for individuals seeking health-related information. While the web provides a vast repository of health knowledge, it has simultaneously birthed a daunting challenge: the proliferation of online health misinformation. This malady, when unchecked, poses serious repercussions for public health, as individuals, often untrained in medical nuances, make health decisions based on misleading or outright false information. Addressing this pressing concern, my thesis delves deep into understanding and mitigating the challenge of *Online Health Misinformation*, exploring avenues of detection, retrieval, and explainability.

Our research journey began with a focus on the *detection* of health misinformation, by utilizing *structural*, *content*, and *context-aware* strategies. This new model was uniquely poised to assess the truthfulness of online health content. By exploiting a specialized medical lexicon, the model crafted embedded representations of web pages, thereby comprehending subtle nuances associated with health misinformation. The innovation lay in the model's capability to also consider URLs embedded within these pages, which proved instrumental in the classification effort. Comparative evaluations across diverse datasets underscored the superiority of our model against traditional machine learning techniques, which predominantly hinge on handcrafted features. Moreover, the strategic inclusion of a domain-specific pre-trained representation considerably amplified the model's efficiency. In the subsequent phase, we built upon these foundational findings to birth the *Vec4Cred* model - an advanced approach tailored explicitly for detecting health misinformation online. Vec4Cred was underscored by a multi-layered framework, focusing on embedding representations of various web page attributes. The model's prowess lay in its capacity to seamlessly integrate embedding representations from parts-of-speech tags and keywords from linked pages. Experimental outcomes affirmed the model's aptitude in combating online health misinformation, underscoring its adaptability and efficiency. Forward-looking, the model beckons enhancement through advanced contextual embedding methodologies, thereby continually refining its accuracy in misinformation detection.

Yet, merely detecting misinformation is not the panacea; the *retrieval* of truthful health information is equally paramount and constitutes the next step of our work. We explored various methodologies to address this, leading to an unsupervised retrieval strategy. This technique distinctively juxtaposed online health narratives with scholarly articles, ensuring the retrieved information was not only contextually relevant but also firmly anchored in scientific validation. Our contributions in the realm of *Consumer Health Search* (CHS) further extended the boundaries of relevance assessment. By integrating multidimensional relevance, we ensured that retrieval outputs were not only topically aligned but also truthful. In light of challenges

observed in existing IR literature, our research also proposed a Transformer-based re-ranking model that exploited Passage Retrieval techniques. The central tenet was to extract the most pertinent passage of a document, thus ensuring topical relevance and information truthfulness. Empirical results resonated with our hypothesis, establishing the model's supremacy over conventional re-ranking solutions.

The challenge of misinformation, however, is not just about detection and retrieval. In an era where trust in online information is eroding, *explainability* becomes a cornerstone. Here, our research made strides in ensuring that search results, especially within the CHS context, were not only accurate but also explainable. By weaving together advanced textual retrieval, representation techniques, and Named Entity Recognition, our models presented health information with a layer of clarity. Importantly, we subjected our approaches to rigorous evaluations, leveraging user-centric studies to glean feedback and refine our methodologies.

In conclusion, this thesis represents a confluence of innovative methodologies and empirical insights aimed at fostering a safer and more informative *Online Health Information* (OHI) ecosystem. Contributions span across proposing novel models for health misinformation detection, the formulation of a novel multi-dimensional retrieval methodology, and the development of explainability measures for CHS tasks. As the web burgeons with health narratives, the tools and techniques espoused herein offer a beacon of hope, ensuring that truthfulness remains paramount.

Acknowledgements

As I finish my Ph.D. journey at the University of Milano-Bicocca, I feel an immense sense of gratitude towards all those who have supported and guided me along this path.

I extend my deepest thanks to my supervisors, *Prof. Marco Viviani* and *Prof. Gabriella Pasi*. Their constant support, wise counsel, and patient guidance have been vital in shaping my research and personal growth. I am truly honored to have worked under their mentorship.

My heartfelt appreciation goes out to *Prof. Allan Hanbury* from the Technical University of Wien, *Dr. Petr Knoth* from The Open University, and *Prof. Arkaitz Zubiaga* from Queen Mary University of London. The opportunity to collaborate and exchange ideas with such distinguished scholars has greatly enriched my research experience and broadened my academic perspectives.

The love, understanding, and support of my family have been my constant source of strength and motivation. I am forever grateful for their sacrifices and unwavering belief in me.

I also want to thank my friends and colleagues in Italy, who have been like a second family to me. They have been like a second family to me. Their friendship, collaborative spirit, and encouragement have not only made my academic journey smoother but also filled it with cherished memories and laughter. The sense of community we shared, the brainstorming sessions, the late-night discussions, and the shared meals have all contributed to making my time here both productive and enjoyable. They have indeed made my stay in Italy truly memorable.

I acknowledge the financial support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860721. This support has been essential for the completion of my research.

As I embark on a new chapter in my life, I carry with me the memories and lessons learned during my Ph.D. I am deeply thankful to everyone who has been a part of this rewarding journey.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Rishabh Upadhyay

Contents

Figures and Tables

Figures

Tables

Nomenclature

AIDS Acquired Immunodeficiency Syndrome
AP Average Precision
AUC Area Under the Receiver Operating Characteristic Curve
BERT Bidirectional Encoder Representations from Transformers
BiLSTM Bidirectional Long Short-Term Memory
BioBERT Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
BM25 Best Matching 25
CAM Convex Aggregating Measure
CHS Consumer Health Search
CLEF Conference and Labs of the Evaluation Forum
CMU Carnegie Mellon University
CNNs Convolutional Neural Networks
COVID Coronavirus Disease
DARPA Defense Advanced Research Projects Agency
DL Deep Learning
DNNs Deep Neural Networks
DOM Document Object Model
F1 F1 Score - The harmonic mean of precision and recall
GM Geometric Mean Score - Commonly used for imbalanced datasets
GUI Graphical User Interface
HON Health on the Net
HTML HyperText Markup Language
IR Information Retrieval
IRS Information Retrieval System
KGs Knowledge graphs
KNN K-Nearest Neighbors
MAP Mean Average Precision
MedCERTAIN/MedCIRCLE Collaboration for Internet Rating, Certification, Labelling and Evaluation of Health Information
MRT Mean Response Time
NB Naive Bayes
NDCG Normalized Discounted Cumulative Gain
NER Named Entity Recognition
NLTK Natural Language Toolkit
OHI Online Health Information

POS Part-Of-Speech
QA Question Answering
RNNs Recurrent Neural Networks
RSV Retrieval Status Value
RTA Retrospective Talk-Aloud
SARS Severe Acute Respiratory Syndrome
SciBERT A BERT model trained on scientific text
SGD Stochastic Gradient Descent
SVM Support Vector Machine
T5 Text-to-text transformer
TF-IDF Term Frequency-Inverse Document Frequency
TREC Text REtrieval Conference
URL Uniform Resource Locator
XAI eXplainable Artificial Intelligence
XIR eXplainable Information Retrieval

Publications

Publications included in this thesis

1. *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on web2vec."*** In GoodIT 2021: Proceedings of the ACM Conference on Information Technology for Social Good, pp. 19-24. 2021.
2. *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"Vec4Cred: a model for health misinformation detection in web pages."*** *Multimedia Tools and Applications*, 82(4), 5271-5290, 2023.
3. *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"An unsupervised approach to genuine health information retrieval based on scientific evidence."*** In WISE 2022: Proceedings of the International Conference on Web Information Systems Engineering, pp. 119-135. Cham: Springer International Publishing, 2022.
4. Somnath Banerjee, *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"Summary in Action: A Trade-off between Effectiveness and Efficiency in Multidimensional Relevance Estimation."*** In IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 12-18, 2023.
5. *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"A Passage Retrieval Transformer-Based Re-Ranking Model for Truthful Consumer Health Search."*** In ECML-PKDD 2023: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 355-371, 2023.
6. *Rishabh Upadhyay*, Petr Knoth, Gabriella Pasi, and Marco Viviani. ***"Explainable online health information truthfulness in Consumer Health Search."*** *Frontiers in Artificial Intelligence*, 6, 1184851, 2023.

Other publications during doctoral studies

1. Suominen, Hanna, et al. ***"Overview of the CLEF eHealth evaluation lab 2021."*** In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, pp. 308-323, 2021.
2. *Rishabh Upadhyay*, Gabriella Pasi, and Marco Viviani. ***"Leveraging Socio-contextual Information in BERT for Fake Health News Detection in Social Media."*** In Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks, pp. 38-46. 2023.

Extra Information

Financial Support

My doctoral research journey has been funded by the European Union's Horizon 2020 Research and Innovation program, specifically through Marie Sklodowska-Curie Grant Agreement (No. 860721). This grant aligns with the objectives and visions of the *Domain Specific Systems for Information Extraction and Retrieval* (DoSSIER). As an initiative of the EU Horizon 2020 ITN/ETN, DoSSIER is dedicated to understanding, modeling, and addressing the distinct information requirements of professional users in specialized domains.

Keywords

Health Misinformation, Information Retrieval, Consumer Health Search.

PART I

The Online Misinformation Context

Chapter 1

Introduction

In recent years, thanks to the possibilities that Web 2.0 technologies have provided us with to generate and disseminate content without the control provided by traditional communication media – through the so-called *disintermediation* (Eysenbach, 2008) process – we have increasingly had to face problems related to the risk of coming into contact with misinformation of various kinds. In a matter of few seconds, a message can spread among tens of millions of people, at little to no cost (Berners-Lee, Cailliau, Groff, and Pollermann, 2010), disregarding its genuineness. In this context, the problem has been studied for some years now and solutions have been sought to limit the spread of misinformation in specific domains. In particular, several works of literature addressing the problems of *fake news* and *opinion spam* have been proposed, detailed, and summarized in specific surveys (Ferrara, 2019; Yadollahi, Shahraki, and Zaiane, 2017; Zhou and Zafarani, 2020).

However, one domain has only recently been investigated with respect to the spread of misinformation, and that is the domain of health, which, instead, is particularly critical with respect to the damage to one's well-being that one might suffer if guided by incorrect or otherwise distorted information. The damage caused by health misinformation can also come to affect society as a whole, think, for example, of the consequences of the set of unverified news stories that have been spread in recent years about Covid-19 (Barua, Barua, Aktar, Kabir, and Li, 2020; Love, Blumenberg, and Horowitz, 2020). This is also due to the increasing use of technology to access *Online Health Information* (OHI) and people's reliance on that information (Thapa, Visentin, Kornhaber, West, and Cleary, 2021). According to the Pew Research Center,[1] already in 2013 one in three adults in the United States went online to try to identify a diagnosis to their symptoms, even going so far as to exclude the figure of the doctor in terms of making decisions with respect to their own health (Fox S, 2013). Similarly, by means of a recent Eurostat survey,[2] it was shown that also in Europe online health information seeking has been steadily increasing over the years, especially among young people. Similar studies have been

1. According to `https://www.pewresearch.org/`, the "Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes, and trends shaping the world. We conduct public opinion polling, demographic research, content analysis, and other data-driven social science research. We do not take policy positions".
2. `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1` (accessed on May 25, 2022).

carried out in other geographic areas such as the Arab world, where it was found that 85.9% of the subjects surveyed use online health information, from diet to the side effects of smoking (Bahkali, Almaiman, El-Awad, Almohanna, Al-Surimi, and Househ, 2016); China, where 87.44 percent of online users search for health information (Wong and Cheung, 2018); India, where 92 percent of people use the Web as a starting point for health information search (Akerkar, Kanitkar, Bichile, et al., 2005); and Australia, where the use of the Internet is well entrenched, with about 17 million Australians actively online,[3] and almost 80% of them seeking out health information on the Web (Chen, Conroy, and Rubin, 2015).

As is evident from the studies and statistics just cited, searching for health information online can affect healthcare decisions and outcomes (Thapa et al., 2021), depending in particular on the aim for which the health search is performed. Sometimes, such search is carried out for *instructional purposes*, and is in any case supported by a medical expert (Powell, Inglis, Ronnie, and Large, 2011). This is a search activity that often consists of several stages, including, for example, *before visiting their doctor*: discover the possible meaning of symptoms; *during investigations*: be reassured that the doctor is performing the right tests, prepare for the results, etc.; *after diagnosis*: contact online support groups to seek second opinions; *when choosing treatments*: search information about treatment options and side effects, experimental treatments and alternative; *before treatment*: find out what to take to hospital, what will happen, and what it will be like. In other cases, the search for health information takes place for self-diagnosis purposes, when traditional medical consultations are inaccessible, or when the individual prefers to take a more self-guided approach to healthcare. This can also include searches aimed at holistic or preventive health care, where the focus is on lifestyle adjustments and wellness strategies rather than specific disease treatments. In both cases, we can speak of so-called *Consumer Health Search* (CHS), i.e., searching for health information conducted by persons who are not experts in the field.

1.1 Consumer Health Search

Consumer Health Search (CHS), as previously mentioned, refers to non-experts seeking health-related information online. Previous literature has equivalently referred to this activity as “consumer health information seeking” (Keselman, Browne, and Kaufman, 2008; McCray, Ide, Loane, and Tse, 2004) and “health online” (Fox and Duggan, 2013). An investigation (White and Horvitz, 2009) of commercial web search query logs shows that approximately 2% of queries in their log are related to health. The wide range of knowledge made possible by the internet has allowed individuals unprecedented access to health information. However, this phenomenon brings its own set of challenges and open issues.

3. https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/online-landscape-review-may-2014.pdf

- *Quality Variability*: Research reveals that most of the health information online is of subpar quality (Berland, Elliott, Morales, Algazy, Kravitz, Broder, Kanouse, Muñoz, Puyol, Lara, et al., 2001). Unlike traditional sources that were closely linked to field experts, web-based data undergoes repeated dissemination by diverse individuals, amplifying misinformation risks (Diviani, van den Putte, Giani, and van Weert, 2015). Inadequate or misleading online medical details can escalate anxiety (Singh and Brown, 2016), fear (Baumgartner and Hartmann, 2011), and even disease susceptibility (Norr, Capron, and Schmidt, 2014). The ambiguity and complexity of such information can diminish its value, especially for those with limited health literacy (Lee, 2008).
- *Over-reliance on Web Diagnostics*: Today's patients/consumers frequently consult online medical resources prior to physician visits, often broaching the veracity of such information during consultations (Hu, Bell, Kravitz, and Orrange, 2012; Moreland, French, Cumming, et al., 2015). This behavior varies based on factors like age, gender, and education (Li, Theng, and Foo, 2016). The digital age ensures that physicians increasingly engage with patients informed by online resources (Wong and Cheung, 2019). While this can occasionally be beneficial, over-reliance on the Internet often strains the physician-patient rapport (Bianco, Zucco, Nobile, Pileggi, Pavia, et al., 2013; Mota, Ferreira, Costa Neto, Falbo, and Lorena, 2018). Furthermore, online health searches frequently correlate with more doctor visits (Baumgartner and Hartmann, 2011), suggesting the web's medical content may be lacking information or quality. Enhancing education, *refining retrieved documents*, and fostering eHealth literacy can empower users to better leverage the Internet's vast, invaluable health resources (Wong and Cheung, 2019).

To navigate these challenges, a multi-pronged approach is essential. While there is a need for platforms to ensure the quality of the health information they host, there's also a crucial role for digital health literacy initiatives to empower consumers to discern and utilize online health information appropriately.

1.2 A Clarification on Terminology

Before diving into the main open challenges and solutions associated with health misinformation, it is crucial to lay the groundwork by clarifying the terminology that's been used up to this point. This will cover various facets of *information authenticity*, even when they pertain to fields not closely linked to health.

1.2.1 Information Authenticity

When discussing the property of information to be *truthful*, a word we employ broadly in this review, several concepts connected to or resembling it have been addressed in the literature, sometimes with overlapping or slightly varying meanings:

- *Authenticity*: In general, the term refers to "being true to the self in terms of an individual's thoughts, feelings, and behaviors reflecting their true identity" (Van Leeuwen, 2001). Philosophically, 'authentic' often resonates with "of undisputed origin or authorship", or in a nuanced context, being "faithful to an original". When referred to information, it indicates a genuine representation of facts, without adulteration or misrepresentation.
- *Credibility* (*Believability*): often defined as believability, credibility can be referred to as "the degree of belief that may be attributed to a chunk of information (a message) or its source" (Fogg and Tseng, 1999; Tseng and Fogg, 1999). When referring to source credibility, it mainly relies on two notions, i.e., *trustworthiness* and *expertise* (Hovland, Janis, and Kelley, 1953); In (Kim and Brown, 2015; Wathen and Burkell, 2002), five factors influencing believability are described: characteristics of the source, of the receiver, of the content, of the communication medium, and of the context of the content receiver. In fact, it is important to emphasize the aspect of *subjectivity* related to the concept of credibility; in (Freeman and Spyridakis, 2004; Schwarz and Morris, 2011b) it is defined as a perceived quality of the information receiver, based on the extent to which they are willing to trust the information (source). Despite credibility can sometimes be confused with the term *trustworthiness* (illustrated in detail below), while credibility is "the level of belief that is perceived about" (how credible is), trustworthiness is related to the "level of positive belief about the perceived confidence in" (reliability in) a person, an object, or a process (Al-Khalifa and Binsultan, 2011);
- *Quality*: being quality information has been defined in several ways: as information that is "fit for use" by information consumers (Huang, Lee, and Wang, 1998); information that "meets specifications or requirements and also meets or exceeds customer expectations" (Kahn, 1998); "information to be of high value to its users" (Lesca, Lesca, Lesca, and Caron-Fasan, 2010); and "information to meet the functional, technical, cognitive, and aesthetic requirements of information producers, administrators, consumers, and experts" (Eppler, 1999). It is based on various dimensions such as accuracy, consistency, timeliness, completeness, conciseness, amount of data, reputation, relevance, reliability, etc. (Knight and Burn, 2005; Lopez, Blobel, and Gonzalez, 2016; Weitzman, Cole, Kaci, and Mandl, 2011)
- *Reliability*: this term has been used in relation to distinct objects, such as Internet-based application components, expected end-user behavior, and information content (Adams, 2006). Regarding the latter, it was used as a synonym for information quality (Sih, 1992). It has also been used as a synonym for information accuracy, although reliability, quality, and accuracy are different aspects of information (Templeton and Franklin, 1992). While

the definition of quality has been elucidated earlier, accuracy specifically refers to the closeness of a piece of information to the actual, true, or factual value it represents. Accuracy relates to "the correctness of the output information" (Bailey and Pearson, 1983) and It is one of the elements of data quality (Wang and Strong, 1996).

- *Trustworthiness*: it can be defined as the perceived likelihood that information will preserve a user's trust in it but also have characteristics such as the competence of the information source (Kelton, Fleischmann, and Wallace, 2008). Some researchers have related the concept of trustworthiness to that of quality and provenance (Bertino and Lim, 2010).
- *Truthfulness*: The term 'truthfulness' is described by the Cambridge Dictionary as the attribute of being honest without encompassing or conveying falsehoods. In a more academic context, (Soprano, Roitero, La Barbera, Ceolin, Spina, Mizzaro, and Demartini, 2021) elaborates on this by defining seven distinct facets of truthfulness, including correctness, neutrality, comprehensibility, precision, completeness, speaker trustworthiness, and informativeness. Meanwhile, (Rubin and Lukoianova, 2013) emphasizes deception detection as a technique designed to discern the authenticity of verbal expressions, discerning whether they are rooted in truth or otherwise. Moreover, within the vast expanse of Big Data, the concepts of 'deception' and 'truthfulness' often find themselves utilized interchangeably. In terms of AI (Evans, Cotton-Barratt, Finnveden, Bales, Balwit, Wills, Righetti, and Saunders, 2021), a system can be truthful if it: Avoid lying, Avoid using true statements to mislead or misdirect, Be clear, informative, and (mostly) cooperative in conversation and Be well-calibrated, self-aware, and open about the limits of their knowledge. In addition, authors (Evans et al., 2021) also mentioned Truthfulness is a more demanding standard than honesty: "*a fully truthful system is almost guaranteed to be honest*".
- *Veracity*: this concept became widely used among computer scientists around 2012, when it was introduced as the fourth characteristic of Big Data, identified by the four *V*s, i.e., *Volume*, *Variety*, *Velocity* and *Veracity*.[4] According to distinct English dictionaries, veracity can be defined as: the quality of being true, honest, accurate (*Cambridge*); conformity with truth or fact; devotion to the truth; the power of conveying or perceiving truth (*Merriam-Webster*); habitual observance of truth in speech or statement; truthfulness (*Dictionary.com*). According to IBM's reports,[5] veracity can

4. https://www.informationweek.com/pdf_whitepapers/approved/1372892704_analytics_the_real_world_use_of_big_data.pdf
5. https://www.informationweek.com/pdf_whitepapers/approved/1372892704_analytics_the_real_world_use_of_big_data.pdf, http://docplayer.net/40836703-Solutions-big-data-ibm.html

be intended as managing “data uncertainty” and managing “data in doubt”. Some other literature states that veracity “focuses on information quality” (Ramachandramurthy, Subramaniam, and Ramasamy, 2015), and that the main dimensions of veracity are “Objectivity, Truthfulness, Credibility (OTC)” (Rubin and Lukoianova, 2013).

Table 1.1, taken from (Zhou and Zafarani, 2020), illustrates some characteristics that may be related to the above definitions.

Table 1.1: Web Factor Affecting Trustworthiness and Credibility

Factors	Trustworthiness	Credibility
Author Authority	✓	✓
Familiarity	✓	✓
Currency	✓	✓
Usefulness	✓	
Credentials		✓
References	✓	✓
Accuracy	✓	✓
Understandable	✓	
Motivation		✓
Beliefs		✓
Relevance		✓
Easy to Use	✓	
Recommended	✓	
Easy to Access	✓	
Contact details	✓	✓
Brand/Logo	✓	
Privacy Policy		✓
Personalisation	✓	
Affiliations		✓
FAQ Section	✓	
Slow	✓	
Textual deficits		✓
Sponsors		✓
Agreement (corroboration)	✓	
User Expertise	✓	
Storage of Resources	✓	
Recency	✓	
Age	✓	
Reputation		✓
Endorsement		✓
Intent (Author)		✓
Expectation		✓

It seems that while there's considerable overlap between the terms, there is a concerted effort in academia to distinguish between them. Based on the tables and information provided, there is a spectrum of misinformation, ranging from unintentional errors to deliberate deception. It is essential to recognize these nuances, especially in an age of digital information where false narratives can be rapidly amplified. The idea is to adopt a more encompassing term that captures the essence of all these variations.

As can be seen from this non-exhaustive list of concepts, many of them are closely interrelated, others capture objective aspects, and others more subjective aspects related to information and its perception.

1.2.2 Information Disorder

First of all, let us start from a recent document published by the Council of Europe that defines, regardless of the domain taken into consideration, *information disorder* as constituted by three different components, i.e., *mis-*, *dis-* and *mal-information* (Wardle and Derakhshan, 2017). The differences between these three types of information with respect to their lack of genuineness can be described using the dimensions of harm and falseness (Wardle et al., 2018):

- *Misinformation* is information that is wrong or incorrect, but not intended to cause harm. This includes the case of people who in good faith, wanting to help, spread false content online without being aware of it.
- *Disinformation* is false information that is deliberately created or disseminated with the express purpose of causing harm. This includes disinformation campaigns that are often linked to obtaining financial, political, and social benefits;
- *Malinformation* is genuine information that is shared to cause harm. This includes private or revealing information that is spread to harm a person or reputation (e.g., the deplorable revenge porn phenomenon).

Above table 1.2, provides a concise mapping of various definition concepts related to the dissemination of information to three terms for information disorder: misinformation, disinformation, and malinformation, as defined in a survey (El Mikati et al., 2023). For each definition concept, the table denotes the frequency associated with each of the three types of information disorder. For instance, the concept of "False/inaccurate/incorrect" was linked 15 times to misinformation and 13 times to disinformation. Some concepts, such as "Unintentional," are solely attributed to misinformation, whereas concepts like "Intentional" span across all three terms. Interestingly, while misinformation and disinformation have a broader spectrum of associated concepts, malinformation has specific and fewer associations. This mapping serves as a valuable reference for understanding the nuances and overlaps between these often misunderstood terms in the realm of information quality.

Table 1.2: Mapping definition concepts to terms of misinformation, disinformation, and malinformation (El Mikati et al., 2023).

Definition Concept	Misinformation	Disinformation	Malinformation
False/inaccurate/incorrect	15	13	-
Fabricated	-	2	-
Accurate	-	-	2
Clearly verifiably false	3	-	-
Misleading	5	8	-
Unintentional	7	-	-
Intentional	5	15	2
Based on expert opinion	2	-	-
Used in the wrong context	-	-	1
Political reasons	-	5	-
Purpose to instill doubt	-	2	-
Purpose to manipulate	-	4	-

Table 1.3: Comparison of different types of misinformation

Type	Characteristics	Objectiveness	Severity	Integrity
Rumours	Ambiguous	Not Sure	Low	Not Sure
False Information	Deception	Yes	High	False
Fake News	Misguided	Yes	Medium	False
Disinformation	Mislead/deceive	Yes	Medium	False
Spam	Confused	Yes	Low	Not Sure

The one just provided is only one of the possible distinctions between the three concepts. In particular, there is debate in the literature about the concept of misinformation; in fact, other definitions have been provided in which it is defined as created either deliberately emerged when people share their opinions and comments or due to honest reporting mistakes or incorrect interpretations (Hernon, 1995; Wu, Morstatter, Carley, and Liu, 2019). Misinformation can also be defined as misrepresentation of information by an actor, due to lack of understanding, attention, or even cognitive biases (Fallis, 2009).

Table 1.4, taken from (Islam, Liu, Wang, and Xu, 2020a), illustrates five fake news-related terms, i.e., Rumours, False Information, Fake News, Disinformation, and Spam.

Table 1.4: A Comparison between Concepts related to Fake News

Concept	Authenticity	Intention	News
Deceptive News	Non-Factual	Mislead	Yes
False News	Non-Factual	Undefined	Yes
Satire News	Non-Unified	Entertain	Yes
Disinformation	Non-Factual	Mislead	Undefined
Misinformation	Non-Factual	Undefined	Undefined
Cherry-picking	Commonly Factual	Mislead	Undefined
Clickbait	Undefined	Mislead	Undefined
Rumors	Undefined	Undefined	Undefined

1.3 Types of Misinformation

In the context of our exploration of health misinformation, it is pertinent to discuss and define the various forms of misinformation. Misinformation types include, but are not limited to, the following:

- *Bot*: Bots, shorthand for software robots, are computer programs that operate either fully automatically or with human involvement. They can be used to disseminate both malicious and benign information. For instance, Twitter accounts like @big_ben_clock, which tweets the time every hour, or Botivist, which recruits volunteers and donations, are examples of bots. According to CMU researchers, of the top 50 influential retweeters about COVID or coronavirus, 82% are bots [6]. Bots are particularly prevalent on social media platforms such as Reddit, Twitter, YouTube, and Facebook.
- *Deepfakes*: Deepfakes refer to content, such as a video or audio recording, where someone's face or voice is artificially replaced with someone else's using AI. For instance, fake videos or photos claiming that alcohol, extreme heat, or cold can kill the coronavirus have been circulated[7].
- *Doxing*: The term 'doxing' originated from the phrase 'dropping documents' or 'dropping dox' on someone[8]. According to the Oxford British [9], it refers to the act of searching for or publishing private or identifying information about a particular individual on the internet without their permission. Doxing can affect anyone, including celebrities, ordinary individuals, and even children and adolescents.

---

6. https://www.cmu.edu/news/stories/archives/2020/may/twitter-bot-campaign.html
7. https://www.reuters.com/article/uk-factcheck-hospital-consume-alcohol-co-idUSKBN21J6W2
8. https://www.wired.com/2014/03/doxing/
9. https://www.oxfordreference.com/display/10.1093/acref/9780191803093.001.0001/acref-9780191803093-e-405

- *Fake news*: The definition provided by a study (Allcott and Gentzkow, 2017) is: "the news articles that are intentionally and verifiably false, and could mislead readers". One of the article (Zhang and Ghorbani, 2020) refers to fake news as "all kinds of false stories or news that are mainly published and distributed on the Internet, in order to purposely mislead, befool or lure readers for financial, political or other gains". It was also termed as a political weapon (after its transmission in the 2016 US election) (Meel and Vishwakarma, 2020). (Cui, Wang, and Lee, 2019) defined it as a modified version of an original news story that is spread intentionally and very difficult to identify.
- *Hoax*: A hoax is defined as a humorous or malicious deception. Hoaxes are often associated with urban legends and rumors. These are stories that contain false or inaccurate facts presented as legitimate truths (Kumar, West, and Leskovec, 2016). False death reports of celebrities are common examples of hoaxes[10].
- *Propaganda*: Propaganda refers to stories or news that aim to harm the interests of a specific context or party. These types of false information have serious consequences as they can significantly influence human history by swaying political elections or causing instability in a country.
- *Rumours*: Rumours refer to news whose truthfulness is ambiguous or never confirmed. (Walker and Blaine, 1991) refers to it as a proposition of belief in circulation within a community without proof or evidence of its authenticity. Rumours are widely circulated on online social networks therefore several studies have analysed it. Some rumors include stories quercetin, essential oils, and other supplements can protect from COVID, antiperspirant deodorants cause breast cancer, and drinking cold water after meals can lead to cancer, etc. It was also perceived as doubtful truth that is easy to spread widely online e.g., AIDS rumor in the 90s [11](Zubiaga, Liakata, and Procter, 2017).
- *Satire*: According to an article(Burfoot and Baldwin, 2009), satire news is news that contains a lot of irony and humor. Wikipedia defines it as a type of parody presented in a format typical of mainstream journalism, and called a satire (i.e. a way of criticizing people or ideas in a humorous way) because of its content.
- *Trolling*: Cambridge Dictionary defines trolling as a message that someone leaves on the internet that is intended to annoy people. In the context of information disorder, the term trolling refers to a troll user who posts controversial information that aims to do things to annoy, disrupt, and provoke other users. Traditionally, trolls use fringe communities such as Reddit, YouTube, Facebook, Twitter, etc (Zannettou, Caulfield, De Cristofaro, Kourtelris, Leontiadis, Sirivianos, Stringhini, and Blackburn, 2017).
- *Fabrication*: Fabrication is something made up, like a lie. Wikipedia refers to a lie as "an assertion that is believed to be false, typically used with the purpose of deceiving someone". It is the most common and widely spread false information.

---

10. https://www.snopes.com/fact-check/adam-sandlerdeath-hoax-2/
11. https://www.documentcloud.org/documents/4780336-85-11-30-Lehrman-Amsterdam-News.html

- *Biased or one-sided*: Defined as news or stories that are biased to a side. Hyperpartisan news (Potthast, Kiesel, Reinartz, Bevendorff, and Stein, 2017) is the biased or one-sided news in the political context. Those are news that are extremely biased towards a person/party/situation/event.
- *Clickbait (Chen et al., 2015)*: Refers to a headline of misleading or sensationalist content created with the sole purpose of sharing misleading content or increasing page views. This type of false information is not new and it has appeared for decades, starting from the "newspaper era," and this phenomenon is known as yellow journalism. But yellow journalism is the least severe of false information since some yellow journalism just uses clickbait that catches the audience's attention. In addition, it may exaggerate the facts or promote the spread of rumors.

1.4 Health Misinformation and Open Issues

The topic of *health misinformation* is becoming increasingly prevalent in our societies due to its dynamic dissemination across a myriad of sources such as the web and social media, along with its broad applicability to a vast range of health topics (Wang, McKee, Torbica, and Stuckler, 2019). Health misinformation takes many forms, including hoaxes, rumors, fake news, fake reviews, and false facts (Vyas and El-Gayar, 2020). The proliferation of such categories, as exemplified by the term "Infodemic" introduced by (Rothkopf, 2003) during the Severe Acute Respiratory Syndrome (SARS) epidemic, is continual. According to the World Health Organization, an infodemic refers to an excessive spread of both correct and incorrect health information, which can subsequently result in the propagation of misinformation, disinformation, malinformation, and rumors during a health crisis (Organization et al., 2020).

Recently, during the times of COVID-19, Misinformation has caused confusion and led people to decline COVID-19 vaccines, reject public health measures such as masking and physical distancing, and use unproven treatments (Roozenbeek, Schneider, Dryhurst, Kerr, Freeman, Recchia, Van Der Bles, and Van Der Linden, 2020). For example, a recent study (Loomba, de Figueiredo, Piatek, de Graaf, and Larson, 2021) showed that even brief exposure to COVID-19 vaccine misinformation made people less likely to want a COVID-19 vaccine. Misinformation has also led to harassment of and violence against public health workers, health professionals, airline staff, and other front-line workers tasked with communicating evolving public health measures (Mello, Greene, and Sharfstein, 2020).

Health misinformation is not a recent phenomenon. In the late 1990s, a poorly designed study later retracted, falsely claimed that the Measles, Mumps, and Rubella (MMR) vaccine causes autism (Rao and Andrade, 2011). Health misinformation is also a global problem. In South Africa, for example, "AIDS denialism"—a false belief denying that HIV causes AIDS—was adopted at the highest levels of the national government, reducing access to effective treatment

and contributing to more than 330,000 deaths between 2000 and 2005 (Chigwedere, Seage III, Gruskin, Lee, and Essex, 2008). Health misinformation has also reduced the willingness of people to seek effective treatment for cancer, heart disease, and other conditions (Swire-Thompson, Lazer, et al., 2020).

Through a comprehensive review of academic literature and governmental publications, I investigated the various interpretations of Health Misinformation. Defining "*health misinformation*" is a challenging task, and every definition has some limitations. Some researchers state that for information to be considered misinformation, it has to go against "*scientific consensus*" (Sylvia Chou, Gaysynsky, and Cappella, 2020). Others consider misinformation to be information that is contrary to the "*best available evidence*" (Sell, Hosangadi, Smith, Trotochaud, Vasudevan, Gronvall, et al., 2022). Several other definitions, such as "*a health-related claim of fact that is currently false due to a lack of scientific evidence*" (Chou, Oh, and Klein, 2018; Shah, Surian, Dyda, Coiera, Mandl, and Dunn, 2019; Zhang, Pian, Ma, Ni, Liu, et al., 2021), "*information that contradicts the widely accepted scientific understanding of a subject*" (Swire-Thompson and Lazer, 2019), and "*beliefs about factual matters unsupported by expert opinions*" (Kim, Ahn, Atkinson, and Kahlor, 2020; Nyhan and Reifler, 2010; Yang, Sangalang, Rooney, Maloney, Emery, and Cappella, 2018). Most of the approaches recognize that what counts as misinformation can change over time with new evidence and scientific consensus.

1.4.1 Terminology for this Research

Within the vast domain of health information, misinformation occupies a particularly contentious space. For the purposes of clarity in this work, I adhere to the definition of health misinformation as given in (Sell et al., 2022; Sylvia Chou et al., 2020). It describes health misinformation as: "*a health-related claim that is based on anecdotal evidence, false, or misleading owing to the lack of existing scientific knowledge / best available scientific evidence at that time*".

It is vital to understand the nuances of this definition. Firstly, the definition underscores the significance of differentiating between anecdotal narratives and claims grounded in verifiable scientific evidence. Secondly, it implies that the veracity of a claim is not always static but is contingent on the prevailing scientific understanding at a given time. This definition does not account for the intent behind the creation of the misinformation, i.e., whether it was disseminated with the intention of causing harm or not. Additionally, the realm of scientific consensus is fraught with challenges; discerning who qualifies as an expert, determining what level of agreement is necessary, and outlining what constitutes the best and most relevant evidence, remains debatable.

Having elucidated the concept of health misinformation, the intricacies that surround its identification and evaluation become evident. At the core of our assessment is the determination of the "truthfulness" of health information, which refers to its "*factual accuracy of the claim in relation to established medical and scientific knowledge best available scientific evidence at that time.*" Yet, as we navigate the vast landscape of online health information, another term often emerges: i.e., *credibility*. Credibility pertains to the trustworthiness and reliability of the source as defined before, which may not always align perfectly with sheer factual accuracy.

In the context of this study, our primary aim remains to provide access to *truthful* information. However, due to the nuances of evaluating online information and the constraints posed by available labeled datasets, we often encounter labels and metrics of "credibility" as evaluated by renowned evaluation initiatives like TREC and CLEF in the context of health misinformation detection. It is worth noting that for the purposes of our research, we approximate the concept of truthfulness with that of credibility considered by such initiatives, especially in the absence of other specific datasets addressing truthfulness directly in association with topical relevance.

1.5 Organization of the Work and Research Questions

This thesis aims to address the critical issue of ensuring truthful document detection and retrieval, with a particular emphasis on the realm of health information. With the proliferation of information, especially in the digital age, it is become imperative to design systems that sieve through vast data to retrieve trustworthy and accurate health documents. This work is structured around a series of questions and methodologies aiming to develop such systems and understand the depth and breadth of truthful information retrieval.

Specifically, this thesis work is divided into four parts. Part I is devoted to providing an introduction to the problem and discussing the main literature solutions that have been proposed so far, as illustrated in detail in Chapter 2. Each of its subsections focuses on a particular forthcoming part of this thesis. Further, we highlight here the two main contributions of this thesis which we investigate, respectively, in Parts II (Misinformation Detection), and III (Retrieval and Explainability). Finally, Part IV is devoted to drawing conclusions and illustrating further research directions. Below we detail the research aspects treated in Parts II and III in particular, which form the heart of the work, and highlight the research questions.

Part II: Misinformation Detection

Integration of Structural and Context-aware Approaches for Misinformation Detection: This part of the thesis delves into the integration of structural and context-aware methodologies for misinformation detection. The objective here is to discern whether a symbiotic relationship between these approaches can enhance the accuracy of detecting misinformation in health documents. This approach aims to utilize the best of both structural and contextual data, enhancing the robustness and precision of misinformation detection. The detailed methodologies and findings related to this approach are explored in Chapter 3, and an improved version of the approach in Chapter 4.

Part III: Retrieval and Explainability

Engineering Unsupervised Models for Genuineness Evaluation: Recognizing that not all research environments provide the labeled datasets for training, especially in this topic, the thesis proceeds to discuss the potential of unsupervised models. The primary research question in this segment is about the capability of such models to assess the truthfulness of information in health documents without the need for fine-tuning or training a machine learning model. This avenue is crucial for scaling the system to large datasets where manual labeling becomes impractical. The methodologies and outcomes of these unsupervised models are covered extensively in Chapter 5.

Document Passages Vs. Full-text Retrieval for Online Health Information: As the thesis progresses, it probes into the efficiency of using relevant document passages over traditional full-text retrieval methods. The premise is straightforward: can cherry-picking query-relevant passages from documents helpful in providing topical relevant as well as truthful information than utilizing the entire document? This approach can not only save time but also ensure that users get the most truthful information for the specific query without being overwhelmed. The comprehensive exploration and results of this hypothesis span Chapter 6 and Chapter 7.

Enhancing Interpretability and Explainability: The final dimension this thesis touches upon is the explainability of automated systems. With an increasing reliance on Language Models for information retrieval, it is paramount that these systems not only provide accurate results but also offer a rationale behind their choices. Users, more than ever, are keen on understanding why a particular piece of information is deemed truthful. Hence, methods to augment the explainability of these systems are vital. The strategies and outcomes related to this question are explained in Chapter 8.

Research Questions

This thesis is motivated by the need for for improved access to truthful health information, which therefore involves both the problem of identifying misinformation and retrieval of truthful information, especially in the context of Consumer Health Search. Based on that, the overall research question is: *How can we tackle the health misinformation problem by designing algorithms and search engines to ensure access to both relevant and truthful health information? Additionally, how can we make users understand the truthfulness of the retrieved results?*

This research question can be decomposed into research sub-questions that are addressed in the different chapters of the thesis:

R1 **Chapter 3 and Chapter 4:** *How can we effectively amalgamate structural and context-aware methodologies to boost the accuracy of misinformation detection in health-related documents?*

R2 **Chapter 5:** *Can we develop an unsupervised model that accurately evaluates the truthfulness of information in health-related documents?*

R3 **Chapter 6 and Chapter 7:** *Can we enhance the effectiveness of retrieval of truthful health information by focusing on* document summaries (Chapter 6) *and* query-relevant document passages (Chapter 7) *rather than employing full-text?*

R4 **Chapter 8:** *What methodologies can be employed to increase the explainability of automated systems, ensuring they provide a clear rationale for the truthfulness of health-related content?*

Each of these chapters, detailed in Parts II and III of this Thesis, provides both a theoretical and empirical analysis to offer robust answers and solutions.

Chapter 2

Literature Review

The study of health misinformation demands a multi-faceted approach due to its complexity and wide-reaching implications. In order to have a well-rounded understanding of the field and to identify key gaps in knowledge, an exhaustive literature review was undertaken. This review can be divided into three key areas: behavioral approaches, algorithmic approaches for detection, and retrieval approaches.

Behavioral approaches primarily focus on the role of individuals or groups in the creation, spread, and reception of health misinformation. These are often referred to as *interactive approaches*, as they usually involve interacting with the participants through interviews or surveys to gather insights. A variety of factors are considered in these approaches, such as demographic information, psychological traits, social networks, and more. Understanding these variables provides valuable context and aids in developing more effective strategies to counteract the impact of health misinformation.

Algorithmic approaches, on the other hand, involve the use of computational methods to identify, classify, and mitigate health misinformation. These can include *machine learning techniques*, *semantic web technologies*, and *knowledge graph-based* methods. Machine learning, for instance, can be used to predict whether a piece of information is likely to be genuine or misinformation based on patterns found in known examples. Semantic web and knowledge graph approaches aim to harness the interconnectedness of information on the web.

In this context, it is imperative to mention the advent of sophisticated language models like GPT-4 (Bubeck, Chandrasekaran, Eldan, Gehrke, Horvitz, Kamar, Lee, Lee, Li, Lundberg, et al., 2023), which have significantly influenced the field. ChatGPT, built upon OpenAI's GPT architecture, demonstrates an advanced understanding of natural language, making it a valuable tool for both generating and analyzing textual content in the health domain.

However, it's crucial to acknowledge the inherent limitations of such language models. Despite their sophistication, these models can sometimes "hallucinate" information - generating content that is convincingly articulated yet factually incorrect or misleading (Chen and Shu, 2023). This is particularly problematic in health-related contexts where accuracy and evidence-based information are critical. The ability of these models to cross-reference and validate

the generated content against credible sources or evidence remains limited. Thus, while they offer remarkable capabilities in language understanding and content generation, their outputs, especially in sensitive areas like health, must be carefully reviewed and corroborated with established scientific evidence and data.

Finally, retrieval approaches focus on the problem of fetching truthful health information from a sea of data. This involves ensuring that trustworthy, scientifically sound information is presented to users, thereby reducing the exposure to and impact of health misinformation. These approaches are of particular importance in an era where an abundance of information, both correct and incorrect, is readily accessible.

In this realm, Retrieval Augmented Generation (RAG) (Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, Lewis, Yih, Rocktäschel, et al., 2020) emerges as a groundbreaking approach. RAG combines the power of language models with the efficiency of information retrieval systems. By fetching relevant documents or data snippets in response to queries, RAG enhances the capability of language models like ChatGPT to provide more accurate and contextually relevant information. This hybrid approach is particularly promising for health information retrieval, where accuracy and reliability are paramount.

Overall, this literature review aims to synthesize the knowledge from these different areas and provide a comprehensive picture of the state-of-the-art in health misinformation research.

2.1 Interactive Approaches

In the realm of truthfulness assessment for online health content, 31 studies, employing interactive approaches, have been identified. These studies aim to elucidate the criteria users employ in their evaluation of health information truthfulness, gathered primarily through user interactions or interviews. The insights gained from these studies could inform the design of automatic or semi-automatic systems to assess truthfulness. The participants of these studies comprise diverse groups, encompassing the general public, patients, students, and older adults. While some studies provide predefined criteria for credibility evaluation, others allow participants to employ their own standards. Roughly half the studies focus on specific health issues, with the remaining addressing general health or unspecified topics. Notably, three studies concentrate on the perspectives of older adults.

In terms of data, 11 studies (roughly 35%) employ predefined datasets, while 20 studies (about 65%) use search engines for data sourcing. The demographic breakdown shows a significant focus on adults, with 26 studies specifically aimed at this group, and three studies targeting adolescents. Nine studies specifically recruited participants with particular conditions, while others enlisted the participation of patient's relatives (3 studies), the general public (10 studies), and university or school students (5 studies).

Kerr et al. (Kerr, Murray, Stevenson, Gore, and Nazareth, 2006) zeroed in on identifying specific criteria pertaining to information content, presentation, and trustworthiness for the quality evaluation of web pages concerning chronic conditions. In this study, 40 participants, aged between 30-79, were selected to analyze predefined web pages. Similarly, *Marshall et al.* (Marshall and Williams, 2006) investigated 15 criteria for the quality evaluation of web pages, which included authority, language, contacts, appropriateness, accessibility, graphics, currency, balance, layout, font, and comparison with sources and previous websites. For this investigation, 32 participants were selected and provided with predefined websites and booklets for evaluation.

Further, two studies conducted by *Sillence et al.* (Sillence, Briggs, Harris, and Fishwick, 2007a,0) focused on the quality assessment of web pages related to Menopause and Hormone Replacement Therapy (HRT) and Hypertension, respectively. In the first study, 15 participants experiencing menopause and undergoing HRT were asked to search for web pages related to HRT and menopause and document information about the quality factors. In the second study, 13 participants diagnosed with hypertension were enlisted to search for and document information about the quality factors of web pages related to hypertension.
In another study, *Sillence et al.* (Sillence and Briggs, 2007) sought to identify trust factors important in selecting documents from search results. This study enrolled 42 participants and included discussions on health-related topics and potential trust factors of web pages.
Hoffman-Goetz et al. (Hoffman-Goetz and Friedman, 2007) aimed to investigate the influence of Aboriginal women's beliefs on the selection of credibility factors for health-related web pages, specifically concerning breast cancer. They selected 25 participants to evaluate two web pages related to breast cancer and then conducted interviews to discuss their thoughts on credibility factors.
Freeman et al. (Freeman and Spyridakis, 2009) examined the impact of publisher's contact information on assessing the credibility of health-related content. They recruited 188 university students to evaluate a set of web articles on diabetes. After this, the participants completed a questionnaire on the credibility of the web article and the influence of contact information.
MacKert et al. (MacKert, Kahlor, Tyler, and Gustafson, 2009) conducted a study with the aim of evaluating the credibility of online articles related to child and adolescent obesity. They recruited 43 parents with low health literacy and asked them to use search engines to find relevant articles. The study concluded with discussions about their opinions on the credibility of the articles.
Lastly, *Liao et al.* (Liao, 2010) explored the effect of aging on the assessment of the credibility of health-related web pages. They selected 24 participants aged between 19 and 78, who were tasked to rate the credibility of eight health-related web pages (sourced from revolutonhealth[1]) on a 7-point scale.

1. `https://www.revolutionhealth.org/`

The study (Kim, Park, and Bozeman, 2011) examined the evaluation behavior of college students and health experts. Eleven college students and three health experts were interviewed. Participants were asked to select the best website related to preconception via a search engine, and subsequently articulate their reasons for their choice, alongside their evaluation process.
(Feufel and Stahl, 2012) aimed to identify qualitative differences between skilled and less-skilled web users in terms of their approach to online health information. The attitudes, cognitive strategies, and technical skills of both groups were compared. Ten participants were selected for the skilled group and twelve for the less-skilled group. Information was gathered through verbal interviews.
A distinct study (Colombo, Mosconi, Confalonieri, Baroni, Traversa, Hill, Synnot, Oprandi, and Filippini, 2014) investigated the factors that come into play when evaluating health-related information. The study recruited 60 participants, including 40 Multiple Sclerosis (MS) patients and 20 of their relatives. Participants were asked to search for information online and participate in an audio-recorded discussion, which was transcribed for subsequent analysis.
McPherson et al. (McPherson, Gofine, and Stinson, 2014) set out to evaluate the reliability of online articles pertaining to chronic conditions in children and young people. Six participants, ranging in age from 11 to 23, were asked to assess 100 websites related to chronic conditions.
A study (Fay, Lynette, and Kiwanuka-Tondo, 2014) studied perceptions and cultural relevance of online articles about HIV/AIDS among black female college students. Forty participants, aged between 18 and 24, were asked to evaluate websites about HIV/AIDS from the National Institutes of Health (NIH).
In (Briones, 2015), the focus was on young people's assessment of health information. Fifty participants, aged between 18 and 25, were included in the study. Participants were queried about the quality of health information available on the internet and social media.
A user study (Santer, Muller, Yardley, Burgess, Ersser, Lewis-Jones, and Little, 2015) examined the experiences of parents seeking information about childhood eczema on the internet. A total of 31 parents were interviewed for periods ranging between 30 and 60 minutes. The interviews explored their beliefs and understanding of eczema and their experiences seeking information online from both formal and informal sources.
Subramaniam et. al. (Subramaniam, St Jean, Taylor, Kodama, Follman, and Casciotti, 2015) researched adolescents' information-seeking process and information assessment. The study was conducted as part of the HackHealth program across three schools, with 30 students participating. The researcher used participant observation, surveys, interviews, and web browser activity analysis for assessment.
A study (Diviani, van den Putte, Meppelink, and van Weert, 2016) sought to gain insight into the relationship between health literacy and online health information assessment. The study used a mixed-methods approach, employing forty-four interviews followed by short questionnaires. Both qualitative and quantitative analyses of the data were conducted.
In a user study (Sillence, Hardy, Medeiros, and LeJeune, 2016), trust factors in online risk

information about "raw" or "unpasteurized" milk were examined. Two studies were conducted: one using eye-tracking data from thirty-three consumers, and another involving interviews with forty-one consumers. The studies aimed to explore the trust factors of milk consumers. Preselected websites were used for the research.

In the research conducted by *Alesem et al.* (Alsem, Ausems, Verhoef, Jongmans, Meily-Visser, and Ketelaar, 2017), the evaluation criteria used by parents of children with physical disabilities when searching for health information online were explored. Interviews with 15 parents focused on their interpretation of information and their information needs.

In a study by *Champlin et al.* (Champlin, Mackert, Glowacki, and Donovan, 2017), the authors endeavored to understand the health literacy of patients and their methods for seeking and evaluating online health information. The study recruited 40 participants of diverse health literacy levels with a mean age of 39. Through a semi-structured interview, participants were asked to narrate their experiences with online health information seeking and evaluation.

Cusack et al. (Cusack, Desha, Del Mar, and Hoffmann, 2017) targeted understanding the health information assessment processes of high school students. Their study explored student attitudes and perceptions toward health information evaluation. This qualitative research utilized semi-structured interviews with 27 Australian high school students aged 12–15 years, aiming to gain insights into their behaviors and comprehension of health interventions.

A study by *Peddie et al.* (Peddie and Kelly-Campbell, 2017) explored the online information-seeking process for people with hearing impairment in New Zealand. The research involved 11 participants and collected data via questionnaires about internet usage. The questionnaire was constructed to scrutinize participants' decision-making and opinions about various websites.

Research by Klawitter and Hargittai (Klawitter and Hargittai, 2018) analyzed the online health information-seeking process of American adults. The researchers recruited 76 adults who used the Internet for various health tasks, followed by post-observation interviews.

Zhang and Kaufman (Zhang and Song, 2020) conducted an exploratory study to understand how older adults evaluate the quality of online health information. The research involved four older participants who evaluated preselected web pages. Data collection included recording eye and mouse movements along with interviews.

Choi et al. (Choi, 2020) conducted another exploratory study focused on older adults, examining their credibility assessment factors. The study analyzed data from 21 older adults from the US with a mean age of 70.3. Data were collected via face-to-face interviews.

Lastly, a lab-based experiment by Chang and Hsieh (Chang, Zhang, and Gwizdka, 2021) evaluated the relationship between online health information and consumers' eHealth literacy. The aim was to investigate the impact of eHealth literacy on the utilization of credibility indicators and criteria. The study involved 25 participants who evaluated 15 web pages from government,

commercial, and online forum sources using a Gaze-and-mouse-movement-cued retrospective talk-aloud (RTA) method. In this technique, participants verbalized their thoughts after evaluating the web pages. The results from the RTA method were subsequently recorded and analyzed by the researchers.

Outcome

The task of evaluating the quality of online health information is marked by challenges such as the scarcity of human assessors, the voluminous quantity of web articles requiring evaluation, and the absence of a universally accepted "gold standard." Consequently, the criteria used to evaluate credibility often differ among users and studies. The most frequently applied criteria are Authorship, Currency, and Language, while Argument Quality and Balance are the least commonly employed.

Three categories of indicators - Source, Content, and Design - are typically employed to evaluate the credibility of online health information. Their application varies depending on the user and the nature of the information in question (Fogg, 2003), and many researchers have adopted these indicators for their evaluations (Choi and Stvilia, 2015; Sun, Zhang, Gwizdka, and Trace, 2019).

A review of the literature indicates that Content and Source indicators are generally perceived to positively influence credibility, while Design indicators are often seen as negatively impactful. Moreover, the influence of each indicator appears to be dependent on the source of the information. For instance, Source and Content indicators are more commonly used for government websites, while Content indicators are employed more frequently for commercial websites (Chang et al., 2021). Older adults have been observed to concentrate more on Content and Design indicators (Choi, 2020). In a study focused on hearing impairment, it was discovered that Design and Source were deemed the most significant quality indicators (Peddie and Kelly-Campbell, 2017).

Several studies have also explored the relationship between health literacy and credibility assessments (Briones, 2015; Champlin et al., 2017; Diviani et al., 2016; Song, Zhao, Song, and Zhu, 2019; Subramaniam et al., 2015).

The indicators of source credibility, particularly when it comes to providing high-quality health information online, are diverse and multifaceted. These indicators can impact the perceived ability and willingness of the source to provide credible information. Critical indicators include the domain type - *.org, .gov, or .edu*, for example - as websites operated by government agencies or educational institutions are often deemed more credible (McPherson et al., 2014; Subramaniam et al., 2015). The identity of the website owner, such as a parent organization or educational institution, also impacts credibility (Alsem et al., 2017; Kerr et al., 2006; MacKert et al., 2009; Peddie and Kelly-Campbell, 2017; Santer et al., 2015; Sillence and Briggs, 2007;

Sillence et al., 2007a,0). Other critical factors include the type of site (*chatrooms, forums, online discussions, Wikipedia*) (Colombo et al., 2014; Feufel and Stahl, 2012; Kerr et al., 2006; McPherson et al., 2014), and the presence of references (*recommendations, links*) that can enhance the perceived credibility of a source (Chang et al., 2021; Cusack et al., 2017; Diviani et al., 2016; Feufel and Stahl, 2012; Santer et al., 2015; Sillence et al., 2007b; Subramaniam et al., 2015).

Content indicators refer to the perceived quality of the presented health information and are assessed based on cues and heuristics. These indicators allow users to infer the accuracy, semantic and structural completeness, and recency of the information. There are various content-related indicators that affect the credibility of the information, including *content types such as factual and personal information* (Kerr et al., 2006; Marton, 2010; Sillence et al., 2007b), *content attributes such as quantity and balance* (Diviani et al., 2016; Kerr et al., 2006; Marshall and Williams, 2006; Sillence et al., 2007a,1), *writing and language, including factors such as grammar, simplicity of terms, and conciseness* (Choi, 2020; Freeman and Spyridakis, 2009; Kerr et al., 2006; Liu, Song, and Zhang, 2021a; MacKert et al., 2009; Sillence and Briggs, 2007; Sillence et al., 2007a,0) and *Authorship* (Champlin et al., 2017; Choi, 2020; Cunningham and Johnson, 2016; Cusack et al., 2017; Diviani et al., 2016; Sillence et al., 2016) and *currency, or the frequency of updates* (Briones, 2015; Cusack et al., 2017; Diviani et al., 2016; Kerr et al., 2006; Marton, 2010).

Design indicators are frequently used to make decisions about the credibility of a page and go beyond just the aesthetic quality. These indicators include *interface design, such as font and graphics* (Briones, 2015; Chang et al., 2021; Cunningham and Johnson, 2016; Kerr et al., 2006; Sillence et al., 2007b), *interaction design, such as links and logins* (Cunningham and Johnson, 2016; Kerr et al., 2006; Sillence and Briggs, 2007; Sillence et al., 2007a; Subramaniam et al., 2015) and *navigation design such as easier-to-use interface have higher web credibility* (Fay et al., 2014; Kerr et al., 2006; Peddie and Kelly-Campbell, 2017).

Aside from the indicators mentioned above, the evaluation of the quality of online health information is also impacted by individual factors like the individual's personal situation, knowledge, and beliefs. One of the most commonly recognized factors is the individual's prior knowledge and experience of a source, with users trusting sites that they have had positive experiences within the past (Diviani et al., 2016; Sillence et al., 2007b). Health literacy has been noted as another important factor in evaluating the credibility of content (Briones, 2015; Champlin et al., 2017; Diviani et al., 2016; Song et al., 2019; Subramaniam et al., 2015). Some studies have also highlighted the impact of society on the evaluation of credible information, such as in the case of Aboriginal women (Hoffman-Goetz and Friedman, 2007) and African women (Fay et al., 2014). Research has also been conducted on adolescent users and the difficulties they face in evaluating online health information, as seen in studies by (Subramaniam et al., 2015) and (Cusack et al., 2017).

2.2 Algorithmic Approaches

Our review has uncovered articles that focus on the use of different algorithmic methods or models for automatically assessing the credibility of online health content. Certain research concentrates on specific medical conditions such as cancer (Kinkead, Allam, and Krauthammer, 2019; Xie and Burstein, 2011), diabetes (Belen Salam and Temizel, 2015), vaccination (Meppelink, Hendriks, Trilling, van Weert, Shao, and Smit, 2020a; Shah et al., 2019), and side effects (Hoang, Liu, Pratt, Zheng, Chang, Roughead, and Li, 2018; Mukherjee, Weikum, and Danescu-Niculescu-Mizil, 2014).

The methods employed for credibility assessment are modeled to tackle either the ranking or classification problems related to health content. Over half of the articles focus on website content, with most of this research conducted before 2016. In more recent years, machine learning has found extensive application in analyzing and creating models for different data types such as audio, text, and images. Over 90% of the reviewed articles rely on statistical machine learning models, which will be further explored in Subsection 2.2.1. Meanwhile, a few articles deal with earlier generation classifiers such as rule-based models (Wang and Liu, 2007; Zhang, Burkell, Cui, and Mercer, 2018) or content labeling (Mayer, Karampiperis, Kukurikos, Karkaletsis, Stamatakis, Villarroel, and Leis, 2011).

2.2.1 Machine Learning Models

Machine learning models, specifically Support Vector Machine (SVM), Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN), are commonly employed for the classification of misinformation. Recently, Deep-learning approaches have also gained traction in this field.

Support Vector Machine (SVM): The SVM algorithm, often utilized for classification and regression tasks, constructs hyperplanes to segregate data points by maximizing the margin between classes (Cristianini, Shawe-Taylor, et al., 2000). SVM has been effectively used in various studies dealing with the credibility of health information. For instance, (Gaudinat, Grabar, and Boyer, 2007) leveraged SVM for the development of a model aimed at classifying web content based on nine distinct Health on the Net (HON) principles. Other research efforts have employed SVM for tasks such as classifying websites into reliable and non-reliable categories (Al-Jefri, Evans, Ghezzi, and Uchyigit, 2017; Liu, 2014; Sondhi, Vinod Vydiswaran, and Zhai, 2012a), distinguishing between true and fake websites (Abbasi, Zahedi, and Kaza, 2012), and assessing website trustworthiness utilizing content features (Park, Sampathkumar, Luo, and Chen, 2013). In addition, SVM has been utilized for classifying statements related to medicine side effects (Mukherjee et al., 2014).

Naive Bayes Classifier: Bayesian classifiers are probabilistic classification methods, where the naive Bayes model makes the assumption that the features are independent, given a class (Friedman, Hastie, and Tibshirani, 2001). The Naive Bayes (NB) algorithm has seen application by numerous authors for the classification of various health-related content, such as websites, and news articles. For instance, Boyer (Boyer and Dolamic, 2015) utilized it for website classification and reported a concordance between the manual (HON) and automatic (NB) systems ranging from 79% to 95%. This classifier has been applied to specific health issues like breast cancer (Xie and Burstein, 2011) and Zika (Ghenai and Mejova, 2017), in addition to more generic medical content (Abbasi et al., 2012; Afsana, Kabir, Hassan, and Paul, 2020; Gaudinat et al., 2007). Moreover, it has been used for the classification of vaccine-related websites (Meppelink et al., 2020a; Shah et al., 2019).

Random Forest: The Random Forest is an ensemble classifier and serves as a modification/improvement of bagging, wherein it builds a large collection of decorrelated trees before averaging them (Friedman et al., 2001). Known for their resistance to variance and ability to handle over-fitting, Random Forests are also frequently used for feature selection (Afsana et al., 2020; Dhoju, Kabir, Rony, and Hassan, 2019; Ghenai and Mejova, 2017; Kinkead et al., 2019; Liu, Yu, Wu, Qing, and Peng, 2019b; Shah et al., 2019; Zhao, Da, and Yan, 2021). A study by (Ghenai and Mejova, 2017) employed a Random Forest for classifying Zika-related posts, achieving an F-measure score of 94.5%. Similarly, (Zhao et al., 2021) compared the Random Forest to four other classifiers for health-related misinformation on forums, with the Random Forest yielding the best results—an F-measure of 84%.

K-Nearest Neighbour: The k-Nearest-Neighbours (kNN) algorithm is a non-parametric classification method (Hand, Mannila, and Smyth, 2001). As a memory-based method, it leverages the distance between data points for classification purposes. Selecting an appropriate value for 'k' is crucial for the effectiveness of kNN, which also lends it the moniker 'lazy learning'. Some authors have employed kNN alongside other classifiers for comparative purposes, as seen in (Gaudinat et al., 2007) for website content classification and (Al-Jefri et al., 2017) for the classification of specific medical (Shingles, Flu, Migraine) website content.

Decision Tree: This algorithm seeks to generate classification rules by training data to facilitate decision-making in the test set. The process involves splitting attributes into different branches based on their values. J48, as discussed in (Xie and Burstein, 2011), is an example of a decision tree algorithm employed for assessing the quality of online health content.

Logistic Regression: Over the past decade, logistic regression has become a popular method for analysis and classification. Each independent feature is multiplied by a specific weight and then summed. This sum feeds into a sigmoid function, yielding an outcome within the continuous range between 0 and 1. By applying an activation function rule, these values can be converted to discrete form. Al-Jefri et al. (Al-Jefri et al., 2017) employed logistic regression,

along with SVM and Stochastic Gradient Descent (SGD), for website content classification. In this context, logistic regression delivered superior F1 measures compared to other algorithms. Other studies also employed logistic regression for training web content (Meppelink et al., 2020a; Oroszlányová, Teixeira Lopes, Nunes, and Ribeiro, 2018).

Deep Neural Networks: First introduced to the machine learning community in 1986 (Dechter, 1986), Deep Learning (DL) represents a significant evolution in machine learning. Deep Neural Networks (DNNs) utilize a layered architecture, adding complexity and depth to traditional single-layer neural nets (Schmidhuber, 2015). Numerous types of DNNs exist, each with their own specific applications. For instance, a Convolutional Neural Network yielded an F1 score of 61% for the prediction of prescribed drug side effects (Nguyen, Sugiyama, Kan, and Halder, 2020).

Transformers: Emerging from the foundational work of Vaswani et al. in 2017, transformer-based architectures have redefined the boundaries of state-of-the-art results in various natural language processing tasks. These architectures, fundamentally different from conventional RNNs and CNNs, allow the model to pay varying attention to different parts of the input data throughout training (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2017). One of the transformative outcomes from this domain is BERT (Devlin, Chang, Lee, and Toutanova, 2018), pre-trained on vast textual collections and subsequently fine-tuned for tasks ranging from misinformation detection to quality estimation. Variants like RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov, 2019a) further optimized BERT's capabilities. In specialized applications, domain-specific transformers like SciBERT (Beltagy, Lo, and Cohan, 2019) have been utilized, and novel techniques such as domain adaptation on transformer embeddings have been introduced (Dharawat, Lourentzou, Morales, and Zhai, 2020a; Hossain, Logan IV, Ugarte, Matsubara, Young, and Singh, 2020). (Mattern, Qiao, Kerz, Wiechmann, and Strohmaier, 2021) supplemented BERT with user and post interaction features, while studies delving into Covid-19 misconceptions employed semantic similarity measures using sentence transformers (Reimers and Gurevych, 2019b) alongside BERTScore (Zhang and Song, 2020). The versatility of transformers is also evident in multilingual tasks, with models like XLM-R (Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzmán, Grave, Ott, Zettlemoyer, and Stoyanov, 2019) and mBERT (Devlin et al., 2018) being pivotal. The applicability of transformers is diverse, ranging from rumor detection in Arabic using MARBERT (Abdul-Mageed, Elmadany, and Nagoudi, 2020) to misinformation detection in Chinese by leveraging translated BERT embeddings (Du, Dou, Xia, Cui, Ma, and Philip, 2021). The text-to-text transformer (T5) (Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu, 2020) further showcases the potential breadth of application in veracity detection and beyond.

Other ML Models: Various machine learning models have been employed by researchers in the field of health misinformation detection. These models include Rule Based Classifiers, ZeroR, SPA, Linear Regressions, Neural Networks, Ensemble Methods, and Attention Based Models (Kinkead et al., 2019).

- *Rule-Based Model*: This model, which utilizes IF-THEN rules for classification (Tung, 2009), has been employed for specific health conditions such as skin disease (Wang and Liu, 2007) and depression treatment (Zhang et al., 2018). It has demonstrated precision and accuracy rates of 98.07% and 46% respectively in these applications.
- *ZeroR*: A frequency-based model that predicts the majority class, ZeroR has been used for the quality assessment of breast cancer-related content (Xie and Burstein, 2011).
- *Linear Regression*: Linear Regression, a simple and linear supervised machine learning algorithm, has been employed for bias detection in the quality ranking of diabetes-related websites (Belen Salam and Temizel, 2015).
- *Neural Networks*: A multi-layer perceptron was trained to classify misinformation for specific conditions such as shingles, flu, and migraines. This model yielded the highest F measure of 84.756%, outperforming other algorithms.
- *Ensemble Method*: While the Random Forest model has been previously discussed, other ensemble models used in the field include Gradient Boosting (Liu et al., 2019b), Voting Classifier (Afsana et al., 2020), and XGBoost (Zhao et al., 2021).

2.2.2 Other Types of Approaches

During the literature search, a limited number of studies were identified that utilized non-machine learning methods such as Semantic Web and Knowledge Graph approaches.

Semantic Framework: The Semantic Web concept involves defining and linking web data in a manner that enables machine utilization for various applications, including quality assessment (Eysenbach, 2005). Essentially, the Semantic Web is an extension of the conventional web, in which assessors or users provide additional machine-readable data (markers) along with human-readable content. This approach can be particularly useful for detecting health misinformation, as it can aid in verifying that the information, along with its associated markers, is accurate, trustworthy, and relevant. However, relatively few studies have explored the use of the Semantic Web for quality or misinformation detection.

One such project, MedCERTAIN/MedCIRCLE (Collaboration for Internet Rating, Certification, Labelling and Evaluation of Health Information) (Eysenbach, 2005), involves a medical community collaboration to assess health information by providing standardized, machine-readable statements about specific health websites. Information is presented using agreed-upon vocabularies, facilitating machine processing. The core concept behind this project was to rank results by integrating topical relevance with quality factors. Similarly, the POWDER framework has been employed for content and quality labeling of medical websites (Mayer

et al., 2011). This two-step framework includes Quality Labeling (MedIEQ) and User Content Labeling (QUATRO plus). The main goal of this framework is to assist certification and filtering organizations in adding, updating, and maintaining machine-readable labels for health-related web content, while also empowering users to contribute. Furthermore, a framework was introduced to capture credible treatment options and disease-related information from online sources using third-party applications (Lai, Vong, and Then, 2012). This framework includes two processes: one focused on mining disease-related information from PubMed, and the other centered on identifying treatments by parsing the top 20 web pages that passed the HON evaluation. Topical analysis is also performed to improve the framework and better retrieve user query results.

Another framework (Konstantinidis, Kummervold, Luque, and Vognild, 2015) aimed at enriching Norwegian Electronic Health Records using the Semantic Web, provides both patients and physicians with tips. This framework allows the physician to select trusted information from the matches obtained by keyword searching from a list of medical resources. The selected information is then transferred to the patient tip for further information.

In conclusion, the Semantic Web offers a promising approach to improving the quality and accuracy of online health information by providing machine-readable data in addition to human-readable content. By using standardized vocabularies and labeling frameworks, the quality of health information can be effectively assessed and filtered, reducing the risk of misinformation.

2.2.3 Knowledge-based Approach

Knowledge graphs (KGs) have emerged as powerful tools for representing complex, unstructured documents in a structured graph format. This format consists of a collection of relational knowledge facts. Research leveraging knowledge graphs for quality or fact-checking purposes has proliferated in various fields such as politics and news media.

A notable study in healthcare misinformation detection, DETERRENT (Cui, Seo, Tabar, Ma, Wang, and Lee, 2020), utilized a medical knowledge graph. This algorithm comprises three components: Information Propagation, Knowledge-Aware Attention, and Prediction Layer. The Information Propagation component integrates the article-entity bipartite graphs and medical knowledge graphs into a unified relational graph, facilitating the propagation of knowledge between articles and nodes. R-GCN (Schlichtkrull, Kipf, Bloem, van den Berg, Titov, and Welling, 2017), a relational graph convolutional network, models the data with node-level attention, as not all relations are crucial for misinformation detection. After obtaining neighboring node embeddings from the information propagation step, a knowledge-guided embedding layer is employed to secure the article embeddings.

DETERRENT surpasses other content-based and graph-based models in performance. The model was trained and evaluated on two different types of medical conditions, namely Diabetes and Cancer, achieving F-measures of 84.74% and 93.09% respectively.

Discussion: The choice of model in health misinformation detection hinges on a trade-off between complexity, interpretability, and performance. Traditional models like Linear Regression and Rule-Based Classifiers offer transparency and ease of interpretation, crucial for understanding and justifying classifications in sensitive health contexts. However, these models may not capture the nuanced and non-linear patterns inherent in health misinformation as effectively as more complex models mentioned above. Others', while powerful in detecting subtle misinformation cues, demand substantial training data and computational resources, and their opaque nature poses challenges. This trade-off underscores the importance of selecting models not only based on performance metrics but also considering the context of application, the availability of data, and the need for transparency in model decisions.

2.2.4 Data Source

Patients or their relatives often utilize search engines to gather more information related to a disease or its treatment before consulting health experts or doctors (Hesse, Nelson, Kreps, Croyle, Arora, Rimer, and Viswanath, 2005). In the early years of the current decade, research predominantly focused on evaluating the quality of web pages and the ranking of quality content on search engines. With over 1.8 billion websites presently in circulation [2], credibility is a paramount factor when searching for information.

Research by Kinkead et al. (Kinkead et al., 2019) utilized Google Trends to identify the three most commonly searched diseases. From this, 269 web pages relating to treatment options were selected from Google and Yahoo, which were then rated using the DISCERN instrument by two manual annotators. Meppelink et al. (Meppelink et al., 2020a) utilized Google.nl to search for information using 13 terms (e.g., "vaccinations safe" and "vaccinations unsafe") related to childhood vaccinations. The study retrieved textual content from 476 web pages, and after preprocessing and duplicate removal, the final experiment was conducted on 468 unique web pages.

A different study (Shah et al., 2019) focused on vaccine-related web pages, monitoring links from tweets to gather data, ultimately analyzing 144,878 web pages for quality assessment. Al-Jefri et al. (Al-Jefri et al., 2017) focused on specific medical conditions such as shingles, flu, and migraines. Their study used datasets for flu prevention (Maki, Evans, and Ghezzi, 2015) and migraines (Yaqub and Ghezzi, 2015). For creating a dataset related to "shingles", the Google search engine was used to retrieve web pages related to "shingles treatment". Finally, 111 web pages were retrieved, annotated, and used for further experimentation.

2. https://www.internetlivestats.com/

Various studies used web pages to study a range of medical conditions such as diabetes (Belen Salam and Temizel, 2015), skin disease (Wang and Liu, 2007), depression (Zhang et al., 2018), and generic medical studies (Gaudinat et al., 2007), (Gaudinat, Cruchet, Chrawdhry, and Boyer, 2010), (Mayer et al., 2011), (Sondhi et al., 2012a), (Abbasi et al., 2012), (Liu, 2014), (Boyer and Dolamic, 2015), and (Oroszlányová et al., 2018). Kaicker et al. (Kaicker, Debono, Dang, Buckley, and Thabane, 2010) used keywords such as “pain”, “chronic pain”, “back pain”, “arthritis”, and “fibromyalgia” to search Google, Yahoo, and MSN for web pages, selecting the top 20 pages from each search engine. Xie et al. (Xie and Burstein, 2011) focused on breast cancer and used the Breast Cancer Knowledge Online (BCKOnline) portal for data collection.

A significant proportion of researchers (over 90%) employing interactive approaches used web pages, focusing on the indicators and criteria users apply while evaluating credibility. In most research, participants were allowed to use the search engine for specific or generic medical terms, and then asked to share their experiences in the form of an interview. These topics included menopause and hormone replacement therapy (Sillence et al., 2007b), hypertension (Sillence et al., 2007a), multiple sclerosis (Colombo et al., 2014), childhood eczema (Santer et al., 2015), obesity (Subramaniam et al., 2015), physical disabilities (Alsem et al., 2017), hearing (Peddie and Kelly-Campbell, 2017), and generic medical terms (Ye, 2011), (Feufel and Stahl, 2012), (Briones, 2015), (Diviani et al., 2016), (Champlin et al., 2017), (Cusack et al., 2017), and (Klawitter and Hargittai, 2018).

2.2.5 Motivation for the Proposed Detection Model

As our exploration of the existing literature reveals, a wide array of models and methodologies have been employed to tackle the critical challenge of assessing the credibility/misinformation of online health content. Methods have spanned from conventional machine learning techniques such as Support Vector Machines, Naive Bayes, and ensemble techniques like Random Forests, to more contemporary deep learning paradigms including Transformers and Deep Neural Networks. Yet, there remain gaps and limitations in these existing methodologies, necessitating further exploration and the introduction of more comprehensive models.

While many of the aforementioned models excel in specific scenarios, they often have limitations in holistic contexts. For instance, purely structural approaches may overlook nuanced contextual elements within health content. On the other hand, solely context-aware models might fail to account for the broader structure and interconnected elements of the content, which often hold the key to misinformation detection.

Given the importance of both the structural and contextual aspects, there is a growing realization that an integrated approach could potentially harness the strengths of both worlds. This insight forms the cornerstone of our proposed model: the integration of structural and context-aware methodologies for misinformation detection. This integrated methodology does not just juxtapose structural and contextual elements; rather, it strategically intertwines them, ensuring that each component informs and refines the other. The approach aims to construct a holistic understanding of content, wherein the structure offers a scaffold to the content.

The in-depth exploration of this integrated approach is laid out in the ensuing chapters of this thesis. Chapter 3 provides an overview of the foundational version of the proposed methodology, and Chapter 4 delves into its refined version. These chapters demonstrate not just the theoretical underpinnings of the approach but also its empirical effectiveness when benchmarked against the prevalent models discussed in this literature review.

2.3 Retrieval Approaches

In this section, recent aspects that attempt to combine the classification or ranking aspects of health-related information and misinformation in Information Retrieval should be discussed. Most of the works were submitted as *runs* to the TREC (2020, 2021, and 2022) Health Misinformation Track and CLEF eHealth (2020 and 2021) for the ad-hoc retrieval task.

Most of the works submitted to TREC focus on considering the *correctness* and *credibility* dimensions in the re-ranking phase. In this strategy, a ranking is produced by a given Information Retrieval model; the obtained ranking is then re-ranked by considering the additional relevance dimensions. For Topical assessment, most of the retrieval approaches submitted by distinct research groups at the 2020 Health Misinformation Track employ the classical BM25 as the baseline ranking model. Among them, the CiTUS (Fernández-Pichel, Losada, Pichel, and Elsweiler, 2020a), H2oloo (Pradeep, Ma, Zhang, Cui, Xu, Nogueira, Lin, and Cheriton, 2020), KU (Lima, Wright, Augenstein, and Maistro, 2020), NLM (Mrabet, Sarrouti, Abacha, Gayen, Travis, Goodwin, Rae, Rogers, and Demner-Fushman, 2020), VOHCoLAB (Gonçalves and Martins, 2020), and RealSakaiLab (Tao and Sakai, 2020) groups. ChatNoir, a distinct BM25F-based model (Bevendorff, Stein, Hagen, and Potthast, 2018), was used as a baseline by the Webis group (Bevendor, Bondarenko, Fröbe, Günther, Völske, Stein, and Hagen, 2020). However, additional IR models have been proposed by different groups in implementing their retrieval approaches. Among them, many variations were also employed, such as BM25 with a language model combined with pseudo-relevance feedback RM3 (Lima et al., 2020) and BM25 with pointwise re-ranker T5 (Mrabet et al., 2020; Pradeep et al., 2020).
In 2021, the TREC track continued its focus on *ad-hoc retrieval*, with a significant shift towards the importance of health-related searches on the web. As a divergence from past years, the 2021 document collection utilized web crawls, consistent with the approach in 2019. These

health-based web search challenges highlight the significance of eliminating or demoting incorrect information, which could be harmful, from the search results. The aim was not only to find relevant documents but to prioritize those with correct and credible information. The topics for this year were framed as questions, like "*Should I apply ice to a burn?*", paired with a more concise query version, say "*put ice on a burn*". Each topic was geared towards evaluating the effectiveness of treatments for health concerns, and a stance was provided for each topic, supported by evidence from a credible source.

Regarding the development of models for assessing correctness and credibility, several solutions have been proposed. In particular, in 2020, referring to correctness, the CiTUS group (Fernández-Pichel et al., 2020a) submitted a run in which descriptions and answers were manually combined. For example, given the description: "*Can vitamin D does cure Covid-19?*" and the answer: "*No*", they were converted into the expression: "*vitamin D does not cure Covid-19*". In this case, correctness scores were obtained by computing the maximum sentence similarity via cosine similarity, between the manually generated expression and all the sentences in each document. A similar hand-crafted expression-related model was also used by the H2oloo group (Pradeep et al., 2020). The KU team (Lima et al., 2020) treated the correctness assessment task as a misinformation detection task, and tackled it with a stance detection model. This model produces probabilities for agreement, disagreement, and neutrality between topics and documents. Similarly, the NLM group (Mrabet et al., 2020) considered converting the topic description to an affirmative sentence, and a Natural Language Inference model was used to infer "whether the most relevant sentence from the documents had an entailment/neutral/contradiction relation to the affirmative sentence". For 2021, understanding the *supportiveness* of the content is paramount. Many models have been employed in recent endeavors to ensure that the retrieved documents or information pieces robustly align with the core query. DigiLab (Zhang, Naderi, Jaume-Santero, and Teodoro, 2022), for instance, embarked on a dual-phase ranking approach. After an initial retrieval using the BM25 model, the team employed BERT-based models specifically fine-tuned on both scientific corpora and Wikipedia to judge the *supportiveness* of the retrieved documents. Similarly, CiTIUS group (Fernández-Pichel, Losada, Pichel, and Elsweiler, 2020b) leaned on RoBERTa to represent sentences and compute the similarity between passages and the underlying topic, aiming to gauge how well the passages buttress the core topic. H2oloo group's (Pradeep et al., 2020) approach, building upon Pyserini's default BM25, took advantage of various T5 models and Vera (Pradeep, Ma, Nogueira, and Lin, 2021) in its re-ranking phase, aligning documents with the user's core intent and ensuring they provide relevant supportive evidence.

Regarding credibility assessment, two of the models considered this issue as a *binary classification* problem, with handcrafted features such as link-based, content-based, commercial features, etc. (Fernández-Pichel et al., 2020a; Lima et al., 2020). The first, focused on training a Support Vector Machine (SVM) classifier (Fernández-Pichel et al., 2020a), while the second used a voting classifier (Lima et al., 2020), an ensemble of various models and predicts an

output based on their highest probability of chosen class as the output. Two other research groups considered the credibility assessment task as *claim verification* (Mrabet et al., 2020) and *fact verification* (Pradeep et al., 2020) of the most topically relevant sentence from each document. The RealSakaiLab (Tao and Sakai, 2020) group focused on the following hypothesis: "The more similar a document is to others, the more likely the document is credible". Under this hypothesis, they computed a so-called *majority score*. Cosine similarity is the measure used to compute the similarity among documents represented as TF-IDF vectors. Finally, the VOHCoLAB (Gonçalves and Martins, 2020) group focused on the usage of Kullback-Leibler divergence among documents and some evidence texts, under the hypothesis that "correct information paraphrases the actual evidence; therefore vocabulary distribution will be similar". In 2021, DigiLab (Zhang et al., 2022) made notable strides here by employing a random forest model trained on the Microsoft Credibility dataset. By further amalgamating this with a catalog of credible sites, they managed to foster a degree of trustworthiness in their retrieval. UPV (Schlicht, de Paula, and Rosso, 2021) added another layer of sophistication by harnessing the RoBERTa model to assess credibility. This was achieved by comparing the similarity between documents and a reference standard, which was governed by strict credibility criteria. UWaterlooMDS (Abualsaud, Chen, Ghajar, Minh, Smucker, Tahami, and Zhang, 2021), in one of its manual run approaches, focused on re-ranking using RoBERTa, fine-tuned on the BoolQ dataset (Clark, Lee, Chang, Kwiatkowski, Collins, and Toutanova, 2019), further bolstering the credibility of the retrieved content. Webis (Bondarenko, Fröbe, Gohsen, Günther, Kiesel, Schwerter, Syed, Völske, Potthast, Stein, et al., 2021), through their use of Anserini's BM25, followed by re-sequencing with argumentative axioms, ensured that the credibility of content was held in high regard.

2.3.1 Motivation for Proposed Retrieval Approaches

The extensive survey of the literature, especially within the realm of *Retrieval Approaches*, has uncovered a multitude of methodologies that attempt to merge classification or ranking aspects of health-related information and misinformation. One observed trend in the reviewed literature, especially in the works submitted to TREC, has been the emphasis on re-ranking based on the *correctness* and *credibility* dimensions. While these approaches have shown some degree of success, they primarily lean on supervised models that require labeled datasets. The dependency on labeled data makes it a challenging proposition to scale these systems for broader applications, especially in contexts where labeled data is scarce or expensive to obtain. This gap in the existing literature presents a potent motivation to explore unsupervised models, as proposed in Chapter 5. The unique proposition of these unsupervised models is their capacity to assess the genuineness of information without relying on fine-tuning or training, offering a solution that can potentially scale to vast datasets.

Additionally, while the majority of the literature has been focused on the full-text retrieval of health documents, the efficiency and precision of this approach, especially in an online setting, remain questionable. The above-mentioned challenges provide a clear rationale for the exploration of the approach, in Chapter 6 and Chapter 7. The fundamental inquiry here is whether selecting specific, query-relevant passages can be more efficient and effective. Not only could this approach ensure users are met with concise, relevant, and truthful information, but it also addresses the pitfalls associated with full documents like redundancy and information overload.

2.4 Explainability in Artificial Intelligence

The concept of *eXplainable Artificial Intelligence* (XAI) refers to the ability of an AI system to provide clear and understandable explanations for its decision-making processes and outcomes (Adadi and Berrada, 2018; Guidotti, Monreale, Ruggieri, Turini, Giannotti, and Pedreschi, 2018). Similarly, according to (Bansal, Wu, Zhou, Fok, Nushi, Kamar, Ribeiro, and Weld, 2021), XAI addresses the challenge of understanding and interpreting the recommendations made by an AI model by generating explanations for its predictions. The *Defense Advanced Research Projects Agency* (DARPA) provides a wider definition of the purpose of XAI as to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)", and "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning, Stefik, Choi, Miller, Stumpf, and Yang, 2019). From this latter definition, it emerges that explainability may increase *trust* in AI systems (Inam, Terra, Mujumdar, Fersman, and Vulgarakis, 2021; Miller, 2019). In particular, according to (Bjerring and Busch, 2021), the ability to provide explanations for an AI's predictions increases the likelihood of people trusting the AI system and following its predictions. One of the prevalent approaches to obtain XAI is to enhance system *transparency*. Transparency can be defined as the capability of a system to expose the reasoning processes behind its applications to the user (Gedikli, Jannach, and Ge, 2014). More specifically, transparency helps users understand systems' intentions, capabilities, and decision-making processes, which enhances the mutual understanding and awareness between users and the model (Bhaskara, Skinner, and Loft, 2020; Shin, 2021). For this reason, researchers and practitioners have been working to develop new techniques and methods to improve the transparency and, therefore, the explainability of AI systems. These efforts have included the development of *interpretable machine learning models*, as well as the use of *visualization tools* and other similar solutions, e.g., automatically generated *Natural Language Processing explanations*, to make the inner workings of AI systems more transparent (Bach, Binder, Montavon, Klauschen, Müller, and Samek, 2015).

2.4.1 Explainability in Tackling Online Misinformation

A variety of approaches have been proposed in the last years to address the problem of the spread of misinformation online. Most formulate the problem as a *binary classification* task, distinguishing truthful information from misinformation, thereby possibly incurring the automatic information filtering problem discussed in the Introduction, and through techniques that often do not allow the user to fully understand how such a classification was generated (Islam, Liu, Wang, and Xu, 2020b; Viviani and Pasi, 2017; Zhou and Zafarani, 2020). Recently, some approaches have been developed to provide explanations for misinformation detection results. For example, in the fake news detection context, (Shu, Cui, Wang, Lee, and Liu, 2019) have developed an explainable fake news detection system that utilizes a co-attention mechanism in deep neural networks to capture explainable content in news articles and user comments. (Lu and Li, 2020) have proposed a graph-aware co-attention neural network scheme to generate explanations for fake news detection by analyzing user comments and retweet patterns on social media. (Kou, Zhang, Shang, and Wang, 2020) have designed a graph neural network approach to detect and explain multi-modal fauxtography posts on social media. With regard to some approaches developed in the field of health, (Ayoub, Yang, and Zhou, 2021) have proposed an explainable COVID-19 misinformation detection method that learns semantic representations of COVID-19 posts based on deep Natural Language Processing models, but only uses some words extracted from the posts for explanations. (Kou, Shang, Zhang, and Wang, 2022) have designed a duo hierarchy attention-based approach, namely HC-COVID, that uses specific and generalized knowledge facts in a hierarchical crowd-source knowledge graph to explain COVID-19 misinformation effectively. However, these approaches to the explainability of results still apply to solutions that make a binary classification between information and misinformation.

For this reason, too, efforts have been made in recent years to address the problem of online misinformation by developing *Information Retrieval Systems* (IRS) that produce a ranked list of results that meet a user's information need while trying to uprank truthful results (Clarke, Rizvi, Smucker, Maistro, and Zuccon, 2020; Pradeep et al., 2021; Suominen, Goeuriot, Kelly, Alemany, Bassani, Brew-Sam, Cotik, Filippo, González-Sáez, Luque, et al., 2021; Upadhyay, Pasi, and Viviani, 2022). Such systems are relevant to our work as they do not produce a strict truthfulness judgment (used for binary classification), leaving the final decision to the user based on their investigation of the ranked list. Furthermore, this decision-making process can be complemented and supported by XAI solutions. Indeed, in the last bunch of years, there has been a growing interest in the field of *eXplainable Information Retrieval* (XIR), to improve the transparency of IR systems. While there are similarities between XIR and the broader field of XAI, there are also some notable differences due to the specific tasks, inputs, and output types involved in Information Retrieval, according to the classification provided by (Anand, Lyu, Idahl, Wang, Wallat, and Zhang, 2022). Three types of XIR solutions can be

detailed: *post-hoc interpretability*, *interpretability by design*, and *grounding to IR principles*. Post-hoc interpretability involves providing explanations for decisions made by pre-trained machine learning models (Lundberg and Lee, 2017; Ribeiro, Singh, and Guestrin, 2016). Several methods fall under this category, such as *feature attribution*, *free-text explanation*, and *adversarial example* methods. Feature attribution methods, also known as feature importance or saliency methods, generate explanations for an individual token by attributing the model output to input features (Polley, Janki, Thiel, Hoebel-Mueller, and Nuernberger, 2021; Qiao, Xiong, Liu, and Liu, 2019; Singh and Anand, 2019). On the other hand, free-text explanation methods provide explanations using natural language. Some methods are constituted by *point-wise explanations*, which use transformer-based models to generate free text explanations for individual query-document pairs (Rahimi, Kim, Zamani, and Allan, 2021); others are constituted by *list-wise explanations*, which use encoder-decoder transformers to generate text to explain all documents contained in a ranked result list for a given query (Yu, Rahimi, and Allan, 2022). Lastly, adversarial example methods are commonly used to demonstrate the fragility or robustness of machine learning models and are typically used in classification tasks. However, in a retrieval task, the adversarial perturbation can be used to make a document rank higher or lower in the search results than it would. In the model proposed by (Raval and Verma, 2020), adversarial examples for black-box retrieval models are generated to lower the position of a top-ranked document using a stochastic evolutionary algorithm with a one-token-at-a-time replacement strategy. However, a challenge with post hoc interpretability methods is the difficulty in determining the extent to which the model behavior is understood. (Rudin, 2019) argued that interpretable-by-design models should be used as much as possible, especially for high-stakes decision-making situations. One way to increase the transparency of data-driven machine learning models is to determine if the trained models follow well-established IR principles. There are currently two research directions, (i) trying to align the predictions of ranking models with certain axioms; and (ii) examining the models to see if they incorporate known relevance factors such as matching, term proximity, and semantic similarity (Anand et al., 2022).

2.5 Concluding Remarks

This chapter provided a comprehensive overview of existing approaches in addressing health misinformation, spanning behavioral, algorithmic, and retrieval methodologies. Through this exploration, several key insights and challenges emerged, underscoring the complexity and urgency of the issue.

Firstly, behavioral approaches highlighted the intricate interplay of individual and social factors in the creation and spread of health misinformation. While insightful, these approaches often lack the scalability needed to address misinformation on a web-wide scale. Algorithmic approaches, including various machine learning models, have shown promise in automatically detecting

misinformation. However, challenges remain in terms of accuracy, adaptability to evolving misinformation tactics, and the need for large, annotated datasets. Retrieval approaches focus on promoting access to truthful information, but often grapple with balancing relevance and truthfulness, especially in consumer health search scenarios.

These findings from the literature review underscore the multifaceted nature of the health misinformation problem and the need for innovative solutions that can tackle these challenges effectively. This sets the stage for the research questions that this thesis aims to address, each targeting specific aspects of the health misinformation dilemma:

R1 **Misinformation Detection:** How can we effectively amalgamate structural and context-aware methodologies to boost the accuracy of misinformation detection in health-related documents?

R2 **Unsupervised Evaluation:** Can we develop an unsupervised model that accurately evaluates the truthfulness of information in health-related documents?

R3 **Effective Retrieval:** Can we enhance the effectiveness of retrieval of truthful health information by focusing on *document summaries (Chapter 6)* and *query-relevant document passages (Chapter 7)* rather than employing full-text?

R4 **Explainability in Automated Systems:** What methodologies can be employed to increase the explainability of automated systems, ensuring they provide a clear rationale for the truthfulness of health-related content?

Each of these questions represents a critical component in the broader effort to combat health misinformation and ensure access to reliable and accurate health information for consumers. The subsequent chapters of this thesis are dedicated to exploring these questions, presenting innovative approaches, and contributing to the body of knowledge in this domain.

PART II

Health Misinformation Detection

Chapter 3

A Structural and Context-Aware Approach to Health Misinformation Detection

The explosive growth of user-generated online content, unrestrained by external control, has facilitated the spread of misinformation. The pressing need for effective solutions has opened up numerous application areas, from opinion spam to fake news detection. The online dissemination of health information has recently gained particular attention due to the potential serious risks associated with misinformation. Early research efforts in this field largely revolved around user-based studies applied to Web page content. However, with the recent advent of the COVID-19 pandemic, the focus has shifted towards automated methods for both Web pages and social media content.

These automated methodologies primarily rely on handcrafted features derived from online content coupled with machine learning techniques. With the primary focus being on Web page content, the opportunity remains to explore features related to structural, content, and context information for assessing the credibility of these pages. Thus, this chapter primarily focuses on studying the effectiveness of such features in association with a deep learning model. We delve into an embedded representation of Web pages recently proposed in the context of phishing Web page detection known as *Web2Vec*. Our aim is to unearth the potential of this approach in addressing the challenge of health misinformation.

3.1 Methodology

The proposed methodology is an adapted version of the original *Web2Vec* model (Feng et al., 2020), developed initially for phishing web page detection. *Web2Vec* is predicated on the embedded representation of the URL, content, and Document Object Model (DOM) structure of a given web page. These representations are used by a hybrid Convolutional Neural Network

(CNN) - Bidirectional Long Short-Term Memory (BiLSTM) network to extract both local and global features. An attention mechanism then combines these features, emphasizing the most significant ones. Multichannel output vectors are concatenated and supplied to a classifier to determine the category of the tested web page.

In the proposed approach, specific characteristics related to the problem of assessing the credibility of health information are considered. First, a specific vocabulary related to the medical field is used when generating an embedded representation of web pages, as it is crucial in detecting health misinformation. In addition, rather than focusing on the features related to the URL of the web page under evaluation (as in the original *Web2Vec* model), the proposed approach considers features related to the URLs present within the page itself. These URLs can provide a better indication of whether the links refer to reliable or unreliable external sources (for instance, the presence of commercial links).

The solution thus proposed can be divided into the following phases:

1. **Data Parsing:** Each HTML page in the dataset is parsed to extract the page links, content, and Document Object Model (DOM) structure. The data obtained from this phase are then used in the subsequent stage.
2. **Data Representation:** Word-level and sentence-level embedding representations are generated for the web page content, while the DOM structure and links are represented through HTML tag and URL embeddings.
3. **Feature Extraction:** A CNN-BiLSTM network is used to extract features from the generated representations.
4. **Web Page Classification:** Using densely connected layers, health-related web pages are classified as credible or not credible.

3.1.1 Data Parsing

The data parsing operation closely follows the approach utilized in *Web2Vec*, with a unique exception in link parsing, which in our approach is applied to the content of the HTML page instead.

DOM Corpus

HTML files embody a standard semi-structured data format. This hierarchical structure is represented using HTML tags, which are organized according to the Document Object Model (DOM) structure. With a focus on such structure, an ordered list of tags is extracted, starting from high-level tags until the “children” tags, namely *HTML*, *HEAD*, *META*, *LINK*, *TITLE*, *SCRIPT*, *BODY*, *DIV*, *TABLE*, *TR*, *TD*, and *IMG*. These HTML tags are treated as words and constitute the word-level corpus for the DOM structure to be used in the subsequent data representation phase.

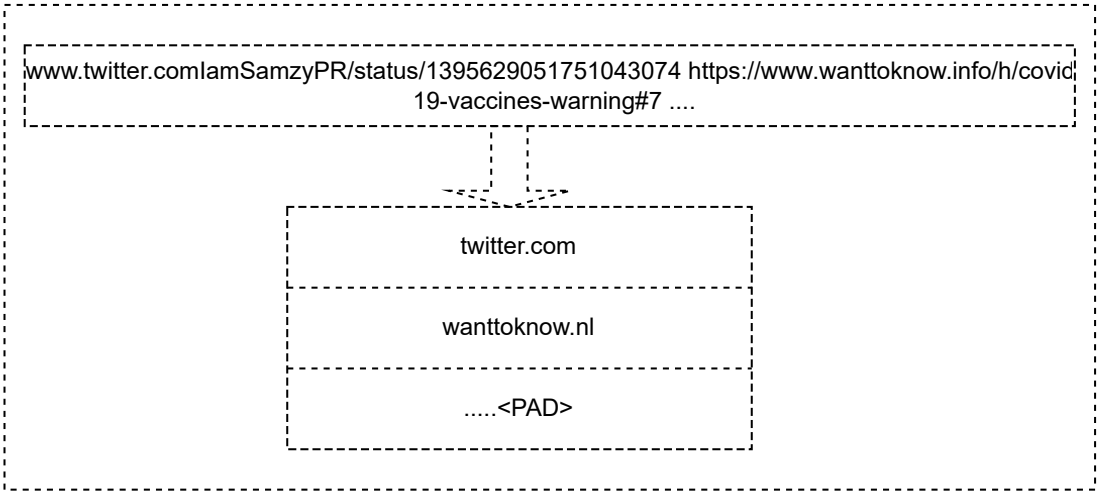

Figure 3.1: Construction of the word-level corpus for links.

Content Corpus

Each web page is parsed, with only the textual content considered (links and tags are excluded). Both word-level and sentence-level corpuses are constructed. The word-level corpus comprises each distinct word present on the page, while the sentence-level corpus identifies sequences of words separated by the '.' character. Notably, we consider a fixed-length dimension for each word sequence.

Link Corpus

The link corpus is created considering the links present in the HTML page. We focus particularly on the domain names extracted from the URL of the websites referenced within the HTML page. These domain names, illustrated in Figure 3.1, constitute the word-level corpus to be employed in the data representation phase.

3.1.2 Data Representation

In this phase, the word- and sentence-level corpora corresponding to the web page's DOM structure, content, and links, generated in the previous phase, are formally represented. The aim is to capture their semantic relationships through word embedding. Specifically, a Keras embedding layer is utilized, which is based on a supervised method. This method enhances the semantic representation while training the model using back-propagation. It is important to note that a separate embedding layer is defined for the DOM corpus, the word-level content and link corpora, and the sentence-level content corpus.

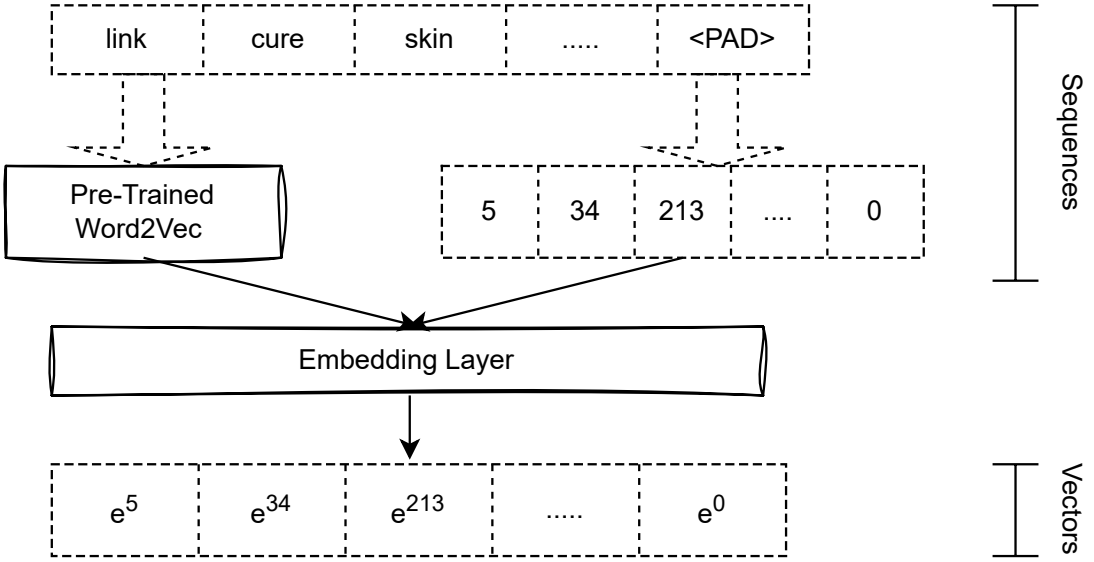

Figure 3.2: The word-level embedding phase for the content.

Differing from Web2Vec, this work includes domain-specific information related to the medical field by adding a word2vec layer pre-trained on PubMed as a weight initializer in the Keras embedding layer. This is specifically considered for the content word-level embedding. In this manner, the word2vec weights serve as weight initializers for the embedding layer, as illustrated in Figure 3.2.

3.1.3 Feature Extraction

The extraction of features, akin to the original Web2Vec methodology, utilizes a CNN-BiLSTM network supplemented with an attention mechanism. This network is applied to the embedding representations derived from the previous phase.

Convolutional Neural Networks (CNNs), are routinely employed in modern times for local feature extraction from data. Bidirectional Long Short-Term Memory (Bi-LSTM) networks augment this ability to learn features from sequences by effectively associating a word with its context (Fan, Gongshen, Kui, and Zhaoying, 2018). The attention mechanism is then employed to boost the model's predictive capacity.

The CNN utilized here shares its architectural design with the one used in Web2Vec, characterized by a feed-forward network model structure. The hidden layer is bifurcated into a convolution layer and a pooling layer. Each fully connected layer is succeeded by a dropout layer (with a dropout ratio of 0.05) to prevent over-fitting (Feng et al., 2020).

The output from the CNN layer is used as input for the BiLSTM layer. This layer utilizes Long Short-Term Memory in both directions, i.e., forward and backward, thereby preserving the sequential order of the data. It also allows for the detection of relationships between previous inputs and the output. The BiLSTM, being a sequential and memory-based model, can effectively learn the long-term dependencies present in the web page, as well as enhance the feature extraction using local features from the CNN. To counter possible over-fitting, dropout learning and L2 regularization are incorporated to bolster model training.

The inclusion of an attention layer becomes vital when assessing the credibility of health-related information. Within a single document, certain segments may be more credible or less credible than others. The presence of even a small quantity of non-credible features within an otherwise credible page (or vice versa) can negatively impact its final evaluation. The purpose of the attention layer is, therefore, to pay extra attention to the most discriminative features concerning the problem at hand. In this work, we particularly refer to the concept of additive attention (Bahdanau, Cho, and Bengio, 2014).

3.1.4 Web Page Classification

The process of web page classification entails categorizing web pages into two categories: credible and not credible. This is done using a binary classifier, which comprises a fully connected layer with a sigmoid function in the final layer. This layer amalgamates the features extracted from the previous layers related to the four corpora considered (i.e., the DOM corpus, the word-level content and link corpora, and the sentence-level content corpus).

To calculate the classification loss, the cross-entropy loss function and L2 regularization are employed to prevent over-fitting. Formally, the error between the target label (t) and the predicted label (y) is calculated as:

$$\text{Error}(t-y) = -\frac{1}{N}\sum_{n=1}^{N}[t_n \log y_n + (1-t_n)\log(1-y_n)]$$

The total loss is then given by:

$$\text{Loss} = \text{Error}(t-y) + \lambda \sum_{n=1}^{N} w_n^2$$

where w is the weight matrix of the layer and λ is the so-called L2 penalty parameter.

3.2 Experimental Setup

In this section, we outline the key components of our experimental setup for evaluating the performance of our proposed model. We describe the datasets used, the baseline models, and the evaluation metrics applied in our experiments. This section serves as the foundation for understanding the methodology and results presented in the following sections.

3.2.1 Datasets

The performance of our proposed model was evaluated using three datasets:

1. **Microsoft Credibility Dataset** (Schwarz and Morris, 2011a): This dataset comprises 1,000 Web pages across various domains such as Health, Finance, Politics, and more. Each Web page comes with a credibility rating provided on a five-point Likert scale, where 1 represents “very non-credible” and 5 represents “very credible". Following the methodology outlined in (Fernández-Pichel, Losada, Pichel, and Elsweiler, 2021), we pre-processed the labels by removing the middle value 3, and mapping 4-5 rating values to credible Web pages and 1-2 rating values to non-credible Web pages. For our study, we focused on the 130 available health-related Web pages, with 104 being credible and 26 being non-credible. Due to the high data imbalance, we employed the SMOTE (Chawla, Bowyer, Hall, and Kegelmeyer, 2002) oversampling method on the minority class.
2. **Medical Web Reliability Corpus** (Sondhi, Vydiswaran, and Zhai, 2012b): This manually generated dataset is balanced, with binary labels associated with Web pages indicating reliability. Reliable websites were randomly selected from HON-accredited websites[1], while unreliable websites were found on the Web using queries, structured as the disease name + “*miracle cure*”. The dataset consists of 360 Web pages, 180 reliable and 180 unreliable. After a cleaning phase to remove blank and no-longer accessible pages, we worked with 170 reliable Web pages and 176 unreliable Web pages.
3. **CLEF eHealth 2020 Task-2 Dataset** (Goeuriot, Suominen, Kelly, Miranda-Escalada, Krallinger, Liu, Pasi, Gonzalez Saez, Viviani, and Xu, 2020): This dataset, consisting of a larger number of documents compared to the previously discussed datasets, was built specifically to assess the topical relevance, readability, and credibility of Web pages consisting of medical content, as part of the *Consumer Health Search* (CHS) task[2]. Credibility ratings are expressed on a four-point scale, from 0 to 3. These ratings were converted to binary values by considering 0-1 values as non-credible and 1-2 values as credible. Ultimately, we worked with 5,509 credible and 6,736 non-credible Web pages.

1. https://www.hon.ch/en/
2. https://clefehealth.imag.fr/?page_id=610

3.2.2 Baselines and Evaluation Metrics

We evaluated the effectiveness of our proposed approach against various baselines. These baselines consist of solutions developed for assessing the credibility of both general and health-related information, which utilize both textual and other types of handcrafted features in association with Machine Learning. Specifically, we considered the textual-feature-based model proposed in (Meppelink et al., 2020b), the multi-feature-based model suggested in (Fernández-Pichel et al., 2021) that encompasses another multi-feature-based model discussed in (Sondhi et al., 2012b), and a BioBERT-SVM model developed for evaluation purposes within this work. This decision was based on the satisfactory results that BERT embeddings have achieved in association with SVM in fake news and misinformation detection problems (Dharawat et al., 2020b; Glazkova et al., 2020; Karande, Walambe, Benjamin, Kotecha, and Raghu, 2021). Specifically, for this baseline, we considered BERT embeddings pre-trained on PubMed articles to adapt to the biomedical domain (Lee, Yoon, Kim, Kim, Kim, So, and Kang, 2019).

In relation to the above-mentioned baselines, we used the following evaluation metrics: F1 measure, accuracy, and AUC. These metrics have often been utilized in various literature works related to misinformation detection and credibility assessment (Cui et al., 2020; Meppelink et al., 2020b). The `scikit-learn` library (Buitinck, Louppe, Blondel, Pedregosa, Mueller, Grisel, Niculae, Prettenhofer, Gramfort, Grobler, Layton, VanderPlas, Joly, Holt, and Varoquaux, 2013) was used for training the Machine Learning models employed as baselines.[3] We applied 5-fold stratified cross-validation to evaluate the results.

3.2.3 Results and Discussion

In this section, we present and discuss the results of our proposed solution in relation to each dataset and baseline, as described in the previous sections. We also discuss these results in terms of the evaluation metrics introduced earlier. For reference, the baselines are abbreviated as follows: NB-CountVec and LR-TF-IDF represent the most effective approaches proposed in (Meppelink et al., 2020b), based on the application of a Naive Bayes and Logistic Regression classifier to textual features rendered as count vectors and TF-IDF vectors, respectively; MFB-SVM refers to the multi-feature model outlined in (Fernández-Pichel et al., 2021); and BioBERT-SVM, as elaborated in Section 3.2.2:

Moreover, we also considered two variants of the Web2Vec model:

- Web2Vec(C): This refers to the Web2Vec model trained solely on content embeddings with default weight initializers.
- Web2Vec(C-D): This denotes the Web2Vec model trained on both content and DOM embeddings with default weight initializers.

3. `https://scikit-learn.org/`

These additional baselines were compared with distinct implementations of the proposed model based on Web2Vec for assessing credibility, denoted as Cred-W2V. In particular:

- Cred-W2V(C): This refers to the proposed model trained on content embeddings, using the PubMed word2vec layer as a weight initializer.
- Cred-W2V(C-D): This represents the proposed model trained on content embeddings with the PubMed word2vec layer, and on DOM embeddings with default weights.
- Cred-W2V(C-D-L): This denotes the proposed model trained on content, DOM, and link embeddings with default weight initializers.
- Cred-W2V(C-D-L)*: This refers to the proposed model trained on content, DOM, and link embeddings, using the PubMed word2vec layer as a weight initializer.

Table 3.1: Evaluation results.

	Metrics	D1	D2	D3	D3(BI)
NB-CountVec	Accuracy	74.55	94.43	64.89	64.9 ± 3.00
	F1	83.22	94.71	67.84	67.2 ± 3.00
	AUC	67.02	93.98	64.12	64.6 ± 2.93
LR-TF-IDF	Accuracy	75.35	94.29	68.6	67.9 ± 2.55
	F1	85.82	94.37	71.3	70.9 ± 2.80
	AUC	47.18	93.21	67.6	67.8 ± 2.55
MFB-SVM	Accuracy	70.03	94.73	66.15	63.8 ± 3.50
	F1	75.97	93.52	46.03	46.7 ± 2.50
	AUC	57.44	93.98	47.78	50.2 ± 0.10
BioBERT-SVM	Accuracy	72.1	94.1	70.74	69.8 ± 2.00
	F1	44.67	94.2	65.34	65.3 ± 4.00
	AUC	63.2	94.1	69.56	67.0 ± 3.00
Web2Vec(C)	Accuracy	*78.34*	*94.81*	*70.34*	*69.5* ± 2.50
	F1	*85.67*	*94.49*	*71.56*	*68.9* ± 2.75
	AUC	*65.34*	*94.54*	*70.18*	*68.9* ± 2.10
Cred-W2V(C)	Accuracy	*78.34*	***96.1***	***71.38***	***71.5*** ± 1.75
	F1	***86.34***	***95.21***	***72.35***	***71.8*** ± 2.25
	AUC	***68.13***	***95.98***	***71.59***	***70.9*** ± 2.10
Web2Vec(C-D)	Accuracy	*80.7*	*96.4*	*72.12*	*71.9* ± 2.22
	F1	*88.28*	*96.12*	*73.69*	*72.5* ± 1.70
	AUC	*74.34*	*96.32*	*71.71*	*71.1* ± 1.75
Cred-W2V(C-D)	Accuracy	***86.9***	***97.57***	***73.58***	***72.5*** ± 2.20
	F1	***91.62***	***97.69***	***77.98***	***75.5*** ± 2.15
	AUC	***80.07***	***97.42***	***73.59***	***72.8*** ± 1.40
Cred-W2V(C-D-L)	Accuracy	***84.12***	***96.23***	***73.98***	***73.4*** ± 1.70
	F1	***90.45***	***96.24***	***75.74***	***75.1*** ± 1.97
	AUC	***78.17***	***96.26***	***73.85***	***73.4*** ± 2.10
Cred-W2V(C-D-L)*	Accuracy	***89.89***	***98.32***	***74.12***	***72.7*** ± 2.0
	F1	***93.78***	***97.01***	***76.61***	***75.5*** ± 2.1
	AUC	***85.69***	***97.71***	***74.56***	***75.4*** ± 1.00

As depicted in Table 1, the proposed model for health misinformation detection outperforms all the baseline models, which are reliant on handcrafted features and Machine Learning techniques, across all datasets and evaluation metrics considered. Furthermore, in comparison to the application of the original Web2Vec model to the problem considered in this paper, our proposed model yields superior results (values in bold). This is true both when the word2vec layer trained on PubMed is incorporated into the original architecture, and when we account for the embeddings of the links present within the webpages to be evaluated. Specifically, by comparing the results of the Cred-W2V(C-D), Cred-W2V(C-D-L), and Cred-W2V(C-D-L)* models, it can be inferred that the impact of incorporating a pre-trained embedded representation on a domain-specific lexicon is predominant in enhancing the effectiveness of the proposed approach.

3.3 Summary and Outlook

This chapter delved deep into the realm of health misinformation detection by harnessing the power of structural-, content-, and context-aware strategies. A model was introduced, founded on the enhancements made to the existing Web2Vec model. This model's objective was to appraise the credibility of health content available on the web. A salient feature of this model is its aptitude to comprehend the peculiar nuances associated with the credibility of health information. It achieves this by leveraging a dedicated medical vocabulary to craft an embedded representation of web pages. Additionally, the model factored in the URLs embedded within these pages, a strategy that showcased significant merit during the classification endeavor.

Our experimental evaluation, conducted across multiple datasets, demonstrated that the proposed model performs better than other machine learning techniques that rely solely on handcrafted features. The results also revealed that the inclusion of a pre-trained embedded representation based on a domain-specific lexicon significantly enhances the model's effectiveness.

The next chapter will extend the discussion by proposing an advanced version of the current model. The focus will be on addressing some of the limitations observed in this study, such as the information from external links, and integrating additional features that can potentially enhance the model's performance. We will introduce further modifications to the model architecture, more robust training strategies, and consider a more comprehensive and diverse range of feature sets. This extension aims to further improve the model's capability to accurately detect health misinformation.

Chapter 4

Vec4Cred: Improving the Web2Vec Approach for Health Misinformation Detection

Building upon the challenges observed in previous approaches, this chapter introduces Vec4Cred, a new model designed to overcome some of the key limitations of existing methods. Inspired by the Web2Vec model, originally developed for phishing Web page detection, Vec4Cred represents an innovative extension tailored to the unique characteristics of health-related content.

This chapter provides a detailed examination of Vec4Cred, outlining its structure, methodology, and the underlying principles that guide its design. The development of Vec4Cred represents a continuation of our previous work explained in Chapter 3, but with substantial enhancements and refinements.

By exploring a data-driven approach that considers automatically learned embedding features, Vec4Cred opens new horizons in the assessment of health-related content genuineness. Its design and evaluation contribute valuable insights to the ongoing discourse on how technology can be leveraged to discern reliable information in the rapidly evolving landscape of online health content.

4.1 Methodology

The proposed solution, Vec4Cred, aims to evaluate health misinformation by leveraging the embedded representations of health-related Web page characteristics. The methodology emphasizes the automatic extraction of features and the deployment of a multi-layer architecture designed to address the unique challenges of health information genuineness. Below, we describe each phase of the Vec4Cred model in detail, following the four-step process outlined in Fig. 4.1.

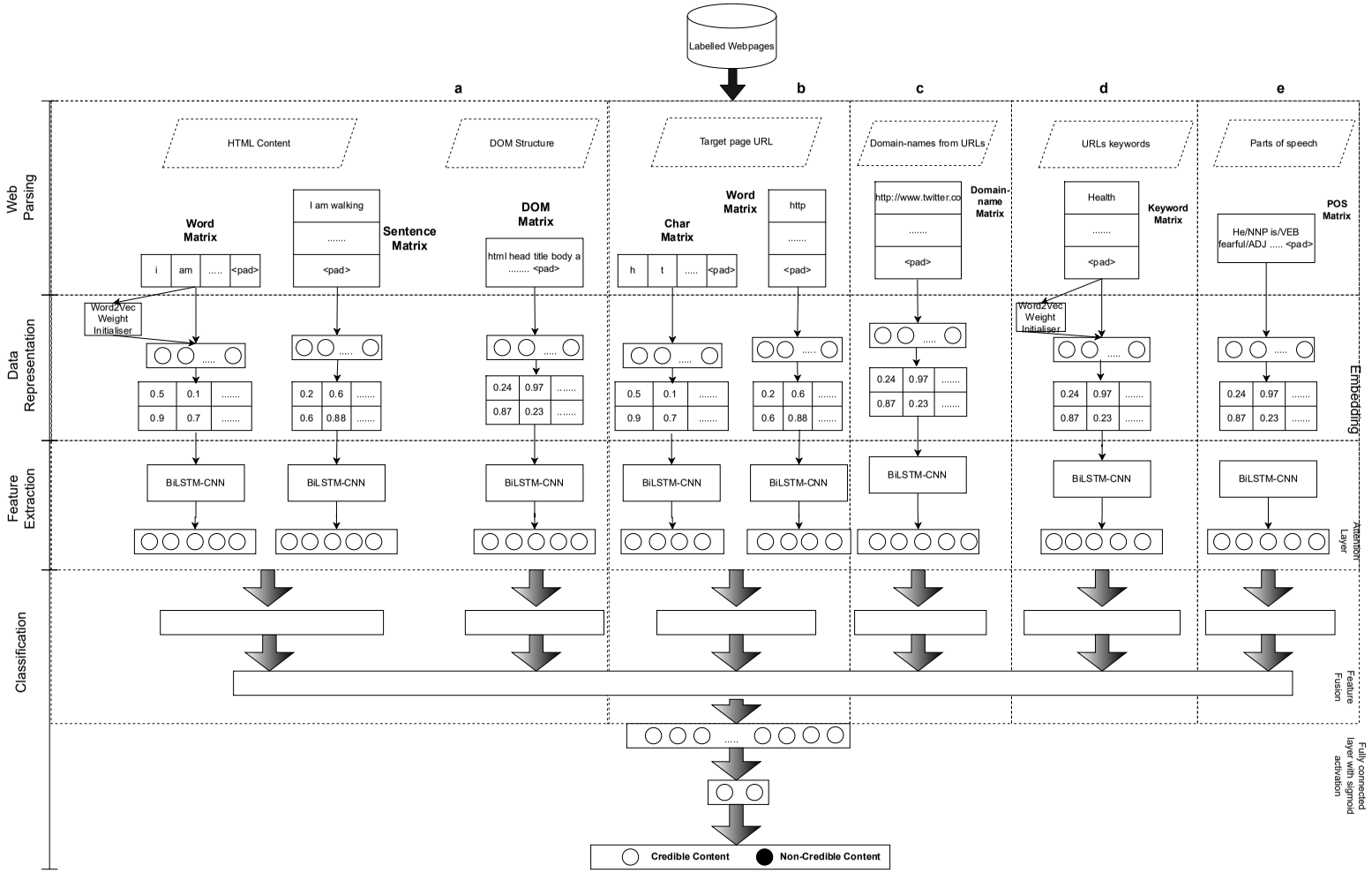

Figure 4.1: The multi-layer architecture of Vec4Cred. In particular, several configurations of the model are illustrated. In (a), only the Web page content and its DOM structure are considered; such information is employed in all the model configurations; (a) + (b) represents the model configuration in which the URL of the Web Page is also considered, as in the Web2Vec model (Feng et al., 2020); (a) + (c) represents the model configuration proposed in (Upadhyay et al., 2021), considering the links present in the content of the target Web page; (a) + (b) + (c) is the model configuration in which we add the URLs in the form of domain-names present in the target Web page; with the addition of (d), we indicate the model configuration considering also the keyword extracted from the pages referred by the links present in the target Web Page; finally, the last configuration of the model, represented by the addition of (e), considers parts of speech from the target Web Page content

4.1.1 Data Parsing

The first step in the Vec4Cred model involves parsing various elements from each Web page in the dataset. This includes:

- The Web page content, including textual and structural information.
- The Document Object Model (DOM) structure, which represents the organization of the content.
- The target URL and any URLs present within the page content.
- The content of the pages linked from the target Web page, focusing on keywords and indicators of health information genuineness.
- Parts Of Speech (POS) extracted from the content of the target Web page, capturing grammatical aspects.

The extracted data form word-level and sentence-level corpora as explained in Chapter 3, encompassing keywords and parts of speech, which serve as essential components for the following phases.

DOM Structure Parsing

HTML files, characterized by a typical semi-structured data format, maintain a hierarchical arrangement represented through HTML tags. These tags follow the Document Object Model (DOM) structure. By focusing on this structure, an ordered list of tags was extracted, commencing with high-level tags and descending to "children" tags, i.e., `HTML`, `HEAD`, `META`, `LINK`, `TITLE`, `SCRIPT`, `BODY`, `DIV`, `TABLE`, `TR`, `TD`, `IMG`. These HTML tags were treated as words, composing a word-level corpus for the DOM structure to be used in the subsequent data representation phase.

Web Page Content Parsing

The content of each Web page is parsed to extract only unstructured textual content, excluding links and tags. Both word-level and sentence-level corpora are formed. The word-level corpus consists of individual words found on the page, while the sentence-level corpus identifies word sequences separated by the '.' character. Specifically, a fixed-length dimension (around 500 characters, mirroring the average size of word sequences in the dataset) is maintained for each word sequence after experimenting with various dimensions.

Additionally, parts of speech are extracted from the Web page. This inclusion of POS tags acknowledges existing research using text analysis for fake news detection and related tasks, focusing on mining linguistic information (Choudhary and Arora, 2021; Horne and Adali, 2017; Markowitz and Hancock, 2014; Pérez-Rosas, Kleinberg, Lefevre, and Mihalcea, 2017). Notable

findings reveal that fake news often contains a significant number of personal pronouns and other distinct grammatical features (Gupta, Kumaraguru, Castillo, and Meier, 2014). The extracted parts of speech create the POS-level corpus for subsequent phases, as illustrated in Fig. 4.2.

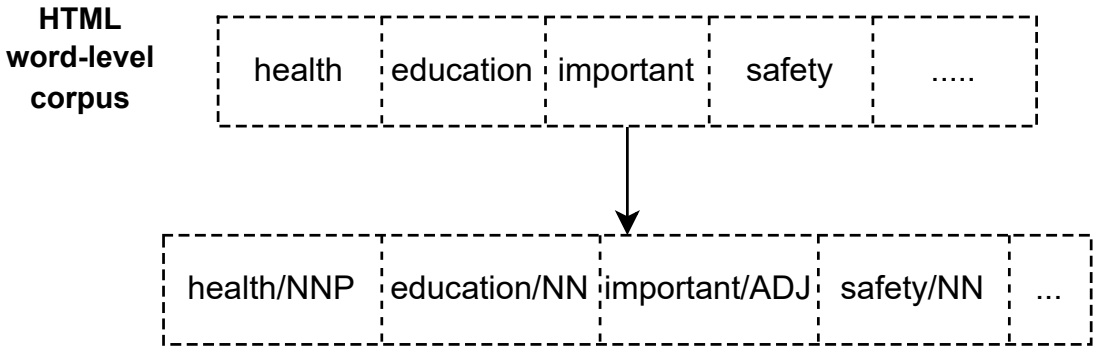

Figure 4.2: Example of the construction of the POS-level corpus.

URL Parsing

The URL parsing phase concerns the extraction of the target Web page's URL and the URLs within the page's content. Here, domain names are culled from the URLs, a tactic identified as beneficial for misinformation detection (Choi and Stvilia, 2015; Hong et al., 2006; Rieh and Belkin, 2000). When considering the target page, the appearance and order of the domain names may enable the model to discern associations with genuine or non-genuine information. Thus, the sequence of domain names—exemplified in Fig. 4.3—forms the word-level corpus for the next phase.

Furthermore, keywords are automatically extracted from the content of pages linked within the target Web page. This effort is inspired by literature findings suggesting that referencing external sources can serve as an indicator of information genuineness [38]. Accordingly, this work emphasizes the content of externally referenced pages. Two methods were trialed for keyword extraction: TextRank [36], a graph-based text summarization technique, and YAKE [6], a statistics-based approach for keyword extraction. The latter proved quicker and more effective on a sample containing an average of 100 referenced links in the target Web pages. Consequently, the top-20 keywords were extracted using YAKE for each linked page, creating a word-level corpus for subsequent phases. An example of this keyword extraction phase is depicted in Fig. 4.4.

In summary, the data parsing section elucidates the comprehensive process of interpreting the Web page's structural and content elements. By capturing the details of the DOM structure, textual content, and URLs, this phase sets the groundwork for the subsequent stages of the Vec4Cred model.

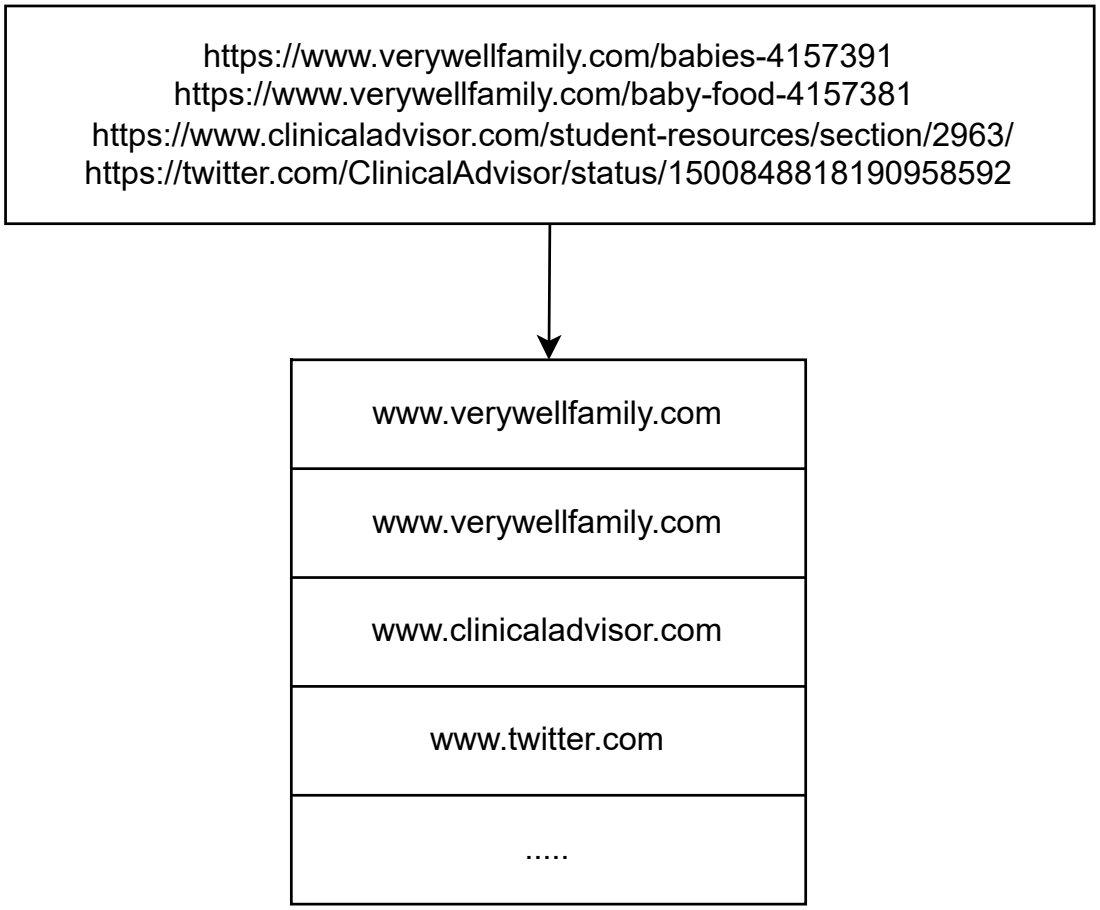

Figure 4.3: Example of the construction of the word-level corpus from URLs in the target page.

4.1.2 Data Representation

In the data representation phase of the Vec4Cred model, a formal structure is developed to encapsulate the semantic relationships found within the word-level and sentence-level corpora. These corpora are derived from the various components of the Web page, including the DOM structure, content, URLs, keywords extracted from linked pages, and parts of speech extracted from the content of the target Web page. The data representation phase builds upon the data parsing phase detailed in Chapter 3. While the parsing phase focuses on extracting various components and elements from the Web page, the representation phase transforms these raw data into formalized semantic relationships. An illustration of this phase can be found in Fig. 3.2.

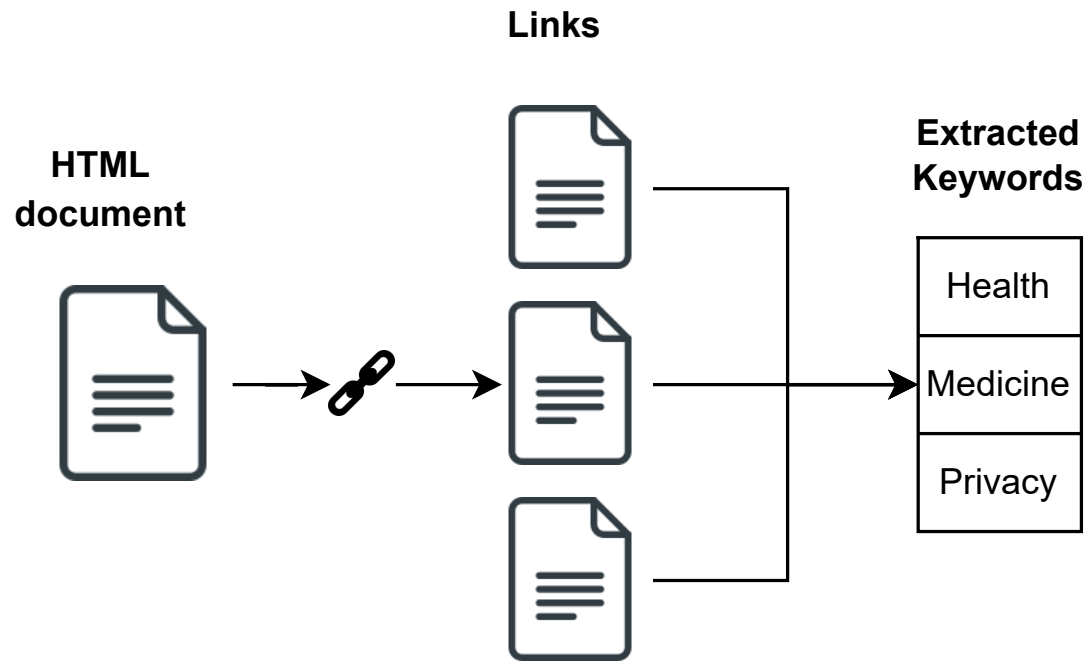

Figure 4.4: Example of the construction of the word-level corpus for keywords extracted from the linked page content in the target Web page.

Embedding Representation

To translate these diverse elements into a coherent representation, a Keras embedding layer is employed. This embedding layer operates on a supervised method that enhances the representations while training the model using backpropagation (Ketkar, 2017). Each data corpus obtained through parsing the Web pages—such as the word-level DOM corpus, word-level and sentence-level corpora from Web page content, word-level URL corpora, word-level keyword corpus, and POS-level corpus—has a designated and independent embedding layer. This separation ensures that each aspect of the Web page is distinctly modeled.

Domain-Specific Embedding

Distinct from generic embedding models, Vec4Cred incorporates domain-specific information pertaining to the medical field. This is achieved by employing a word2vec layer (Mikolov, Chen, Corrado, and Dean, 2013) pre-trained on PubMed, specifically applied to the Web page content and the word-level keyword corpus. The word2vec weights serve as initializers for the embedding layer, much like the approach used in [58]. This unique methodology allows the embedding layer to encapsulate both the general semantics of the words and the specialized terminology prevalent within the medical field.

4.1.3 Feature Extraction

Building on the foundational representations constructed in the previous phase, we proceed to extract pertinent features through a sophisticated blend of CNN-BiLSTM networks supplemented by an attention mechanism. This approach is particularly pertinent for the nature of health information where both local patterns and sequential context are paramount.

CNN Feature Extraction

Adopting the model structure from (Upadhyay et al., 2021), our Convolutional Neural Network (CNN) is structured as a feed-forward model. The hidden layers comprise both a convolution layer and a pooling layer. Borrowing from the preventive mechanisms of over-fitting, a dropout layer with a ratio of 0.05 is included after each fully connected layer. As described in detail in (Feng et al., 2020), the convolution and pooling operations are employed to ensure the extraction of only the most salient features.

BiLSTM for Contextual Understanding

The outputs generated from the CNN layer serve as the input foundation for the BiLSTM layer. As a sophisticated evolution of the LSTM, the Bidirectional Long Short-Term Memory (BiLSTM) encapsulates both forward and backward sequences, ensuring data integrity in its sequential order. This duality not only empowers the detection of relationships between precedent inputs and resultant outputs but also enables the absorption of long-term dependencies within the Web page. Furthermore, it refines the feature quality by integrating localized features harnessed from the CNN.

Attention Layer: Honing in on Details

The addition of the attention layer, in the case of assessing the genuineness of health information, is dictated by the fact that in the same document there may be parts characterized by “more genuine" and “less genuine" information. In this situation, even the presence of a small amount of “non-genuine" features characterizing a genuine page (or vice versa), can negatively affect its final evaluation. The purpose of the attention layer is, therefore, to pay particular attention with respect to the most discriminant features with respect to the considered problem; in this work, we have referred in particular to the concept of *additive attention* (Bahdanau et al., 2014).

Concluding Remarks on Feature Extraction

By judiciously combining the strengths of CNNs for localized feature extraction, BiLSTMs for contextual understanding, and the attention mechanism for nuanced discernment, this phase paves the way for a robust evaluation of health information genuineness. The techniques employed here draw parallels with the model in Chapter 3.

4.1.4 Web Classification

The approach to webpage classification discussed in this section parallels the methodology delineated in the preceding chapter, specifically in Section 3.1.4.

4.2 Experimental setup

This section provides an empirical foundation upon which the theoretical constructs of our model, Vec4Cred, are tested, and validated. As the results of experimentation unfold, they offer insights into the model's strengths, potential areas of improvement, and its comparative performance relative to established benchmarks.

It is important to note that the experimental setup outlined in this chapter is grounded in the same datasets as utilized in 3.2.1 exploration of the Web2Vec model. This choice facilitates a consistent comparative analysis, ensuring that any disparities in performance can be attributed more to the nuances of the models themselves rather than variations in the data. Utilizing identical datasets also promotes transparency and replicability in our experimental procedures, offering future researchers a stable platform from which to base further investigations or adaptations.

In the subsequent sections, we delve deeper into the specifics of our experimentation process, detailing the dataset characteristics, the experimental setup, the metrics used for evaluation, and the results obtained.

4.2.1 Baselines and Evaluation Metrics

To maintain a robust analytical framework, it's crucial to draw comparative inferences from the performance of our model relative to existing ones. Therefore, we have chosen specific baseline models against which the efficacy of the proposed Vec4Cred model will be evaluated.

Baseline Models

Traditional methods, often grounded in textual and metadata features paired with machine learning, have previously been scrutinized for their capability to evaluate the authenticity of health information. As reported in (Upadhyay et al., 2021), these traditional solutions were surpassed by approaches pivoting on the embedding of web page attributes. Consequently, the following models have been selected as our baselines:

- Web2Vec (Baseline): it refers to the Web2Vec model applied to the health misinformation domain trained on Web page content, DOM structure and the URL of each Web page with default weight initialization;
- Web2Vec+L (Baseline): it corresponds to the previous baseline trained, in addition, by considering the word-level corpus constituted by the domain-names extracted from links present in the target Web page;
- GoodIT (Baseline): it refers to the model proposed in (Upadhyay et al., 2021) and presented at the GoodIT 2021 Conference,[1] trained on Web page content, DOM structure and domain-names extracted from the links present in the target Web page, with a word2vec layer trained on PubMed acting as weight initializer (i.e., the model is constituted by components (a) and (c) of the architecture illustrated in Figure 4.1);
- Vec4Cred (a-c-d): it refers to the first configuration of the Vec4Cred model tested in this paper, which constitutes an improvement w.r.t. the GoodIT (Baseline) model, by exploiting the keywords extracted from the content of the Web pages referred from links in the target Web page. This model employs a word2vec layer trained on PubMed acting as weight initializer (i.e., the model is constituted by components (a), (c) and (d) of the architecture illustrated in Figure 4.1);
- Vec4Cred (a-c-e): it refers to the second configuration of the Vec4Cred model tested in this article, which is constituted by the GoodIT (Baseline) model to which are added POS tags extracted from the target Web page content. Also this model employs a word2vec layer trained on PubMed acting as weight initializer (i.e., the model is constituted by components (a), (c) and (e) of the architecture illustrated in Figure 4.1);
- Vec4Cred (a-c-d-e): it refers to the last configuration of the Vec4Cred model tested in this article, which combines the two above-mentioned configurations. Specifically, this model is trained on Web page content, DOM structure, the word-level corpus constituted by domain-names of the links present in the target Web page, the keywords extracted from the pages referred from such links, and the POS tags extracted from the target Web page content, with a word2vec layer trained on PubMed acting as weight initializer (i.e., the model is constituted by components (a), (c), (d) and (e) of the architecture illustrated in Figure 4.1).

1. http://www.grc.upv.es/goodit2021/

4.2.2 Evaluation Metrics

To assess the effectiveness of both the considered baselines and the different configurations of the proposed Vec4Cred model, the following evaluation metrics have been taken into account: $f1$-*measure*, *accuracy*, and *Area Under the ROC Curve* (AUC). Such metrics have often been used in various literature works related to misinformation detection and credibility assessment (Cui et al., 2020; Meppelink et al., 2020a). 5-fold stratified cross-validation has been applied in the evaluation process.

4.2.3 Results and Discussion

Three datasets were utilized for evaluation, as delineated in Section 3.2.1. These datasets are referred to as D1, representing the Microsoft Credibility Dataset; D2, signifying the Medical Web Reliability Corpus; and D3, denoting the CLEF eHealth 2020 Task-2 Dataset. Notably, due to the extensive labeled data in D3, it was feasible to compute the Binomial Proportion Confidence Intervals at 95% confidence, as elaborated in (Blyth and Still, 1983). These intervals, termed as Binomial Intervals (BI), are represented as D3(BI).

Table 4.1: Evaluation results.

		D1	D2	D3	D3(BI)
Web2Vec (Baseline)	Accuracy	80.34	-	72.31	71.32 ± 2.0
	F1	86.80	-	73.16	71.88 ± 1.7
	AUC	69.84	-	71.34	70.77 ± 1.0
Web2Vec+L (Baseline)	Accuracy	81.11	-	72.56	72.23 ± 1.0
	F1	86.78	-	73.10	71.98 ± 1.5
	AUC	78.44	-	71.11	72.00 ± 1.0
GoodIT (Baseline)	Accuracy	89.89	98.32	74.12	72.70 ± 2.0
	F1	93.78	97.01	76.61	75.00 ± 2.1
	AUC	85.69	97.71	74.56	$75.40 + 1.0$
Vec4Cred (a-c-d)	Accuracy	**90.03**	**99.05**	**80.18**	79.33 ± 2.0
	F1	**93.99**	**99.21**	**79.87**	79.01 ± 1.0
	AUC	**86.89**	**98.89**	**79.17**	78.19 ± 1.0
Vec4Cred (a-c-e)	Accuracy	**90.88**	**99.70**	**82.34**	81.00 ± 1.0
	F1	**94.01**	**99.41**	**82.98**	81.00 ± 1.8
	AUC	**88.27**	**99.40**	**81.01**	79.00 ± 3.4
Vec4Cred (a-c-d-e)	Accuracy	**90.47**	**99.71**	**82.56**	82.00 ± 1.3
	F1	**94.21**	**99.71**	**83.11**	82.00 ± 1.2
	AUC	**88.25**	**99.70**	**81.11**	81.00 ± 1.0

Table 4.1 presents the outcomes of the experimental assessments for the chosen baselines and various Vec4Cred configurations in relation to the three aforementioned datasets and the selected evaluation metrics. The evaluation of the D2 dataset using the first two baselines was not feasible. This limitation arises because the URLs of the target web pages in the D2 dataset are not provided.

In Chapter 33, the model, referred to as the enhanced Web2Vec model, was trained using links present within the target web pages, rather than the URL of the target pages themselves. This approach was chosen to capture the context and content more effectively from within the page, rather than relying on the URL structure alone.

In contrast, Chapter 4 revisits the original Web2Vec model, where the focus shifts back to including the URL of the target web pages alongside other features. The original Web2Vec model integrates the URL as a key feature in its training process, utilizing the structure and semantics of the URL for a more comprehensive understanding of the web page's content and context.

Regarding the evaluation of the D2 dataset, it's important to clarify that the inability to use the first two baselines for this dataset was due to the lack of URLs of the target web pages in the dataset. As the original Web2Vec model, in relies on URL features, the absence of these URLs in the D2 dataset rendered the evaluation of this particular dataset with the first two baselines unfeasible.

First and foremost, every configuration of the Vec4Cred model outperformed all three baseline models across all three datasets and evaluation metrics, with the top results highlighted in bold. This underscores the efficacy of utilizing embedded representations of various web page attributes, as expounded in prior studies. Coupling this with domain-specific pre-training on health-centric data (e.g., from PubMed) and integrating genuineness-associated features—like those extracted from related web page content and POS tags in the target web page—proves highly promising in addressing health misinformation detection on web pages.

A granular analysis of results from distinct Vec4Cred configurations reveals that emphasizing the grammatical nuances of the target web page, through the incorporation of POS tags, as seen in the Vec4Cred (a-c-e) model, yields slightly superior outcomes compared to merely focusing on the content of linked pages in the target web page, as shown in Vec4Cred (a-c-d). This observation is further cemented by the results from the Vec4Cred (a-c-d-e) model, which do not deviate significantly (though marginally better) from the Vec4Cred(a-c-e) model. This suggests that future enhancements might benefit from not just examining keywords from reference pages in the target web page, but also their grammatical construct, thereby refining the Vec4Cred model further.

What in our opinion makes interesting these results related to the proposed model, which proves to be effective, is that it acts only by taking into account information directly extractable from the Web page, without referring to external information that could be difficult to find, such as those of the authors of the page and their role, nor additional handcrafted features as used in other approaches in the literature that have proven to be however inferior to the approach proposed already in (Upadhyay et al., 2021). Nevertheless, it would be interesting to be able to use some domain knowledge (where available) to test if this could further increase the effectiveness of the proposed model.

4.3 Summary and Outlook

In this chapter, we introduced Vec4Cred, a sophisticated model geared towards health misinformation detection on web pages. This endeavor builds upon and refines the preliminary model presented in (Upadhyay et al., 2021), which itself drew inspiration from the Web2Vec model (Feng et al., 2020). The model, tailored specifically for health misinformation detection, hinges on a multi-layered framework that capitalizes on embedding representations of web page attributes, meticulously pre-trained using health-centric data. Contrasting with the earlier version presented in Chapter 3, this model integrates enhancements that emphasize the nuances of genuine information within the health sector, which are pivotal in distinguishing misinformation. Key enhancements spotlight the grammatical constructs of the target web page and the embedded content within its referenced pages. To be specific, the model adeptly incorporates embedding representations from POS tags found in the primary web page and keywords identified within linked pages.

The outcomes from our experiments substantiate the efficiency and adaptability of the Vec4Cred model in mitigating health misinformation spread on the web. Its potency stems from its ability to discern domain-specific semantic nuances by exclusively leveraging attributes that can be autonomously extracted from web pages. Charting the course ahead, there's ample scope to augment the model's capabilities. Potential avenues include delving into advanced contextual embedding methodologies, such as BERT, or exploring deeper linguistic intricacies associated with web pages linked within the primary content. This evolution will further fine-tune the model's accuracy in detecting health misinformation.

PART III

Truthful and Explainable Consumer Health Search

Chapter 5

An Unsupervised Model for Truthful Health Document Retrieval

Identifying credible health information is crucial. With the increasing trust individuals place in online sources, ensuring that the retrieved information is not only relevant but also genuine becomes imperative. There's a pressing need for mechanisms that can effectively differentiate genuine health information from misleading or false claims.

The primary focus of this chapter is to introduce and delve deep into a novel unsupervised approach that seeks to address this challenge. This method, contrasting many in the literature, operates without human intervention, emphasizing the retrieval of health information that is backed by scientific evidence. By cross-referencing claims made in online health content with scientific articles, our approach aims to elevate the credibility of the retrieved information, ensuring users access content that is both relevant and genuine. This chapter will present the methodology, its underpinning principles, the experiments conducted using the *TREC 2020 Health Misinformation Track* dataset, and the outcomes of these experiments.

In the succeeding sections, we will explore the intricacies of our approach, providing readers with a comprehensive understanding of its mechanics, its significance, and its potential implications for future health information retrieval systems.

5.1 Methodology

The proposed solution, is based on the development of a retrieval model capable of considering both topical relevance and information truthfulness in providing access to health-related content. The model focuses, in particular, on the idea of calculating the second criterion on the basis of comparing health documents and medical journal articles, which are considered reliable sources of scientific evidence for a given query. In this way, we obtain two query-dependent relevance scores related to each distinct criterion, which are combined through a suitable

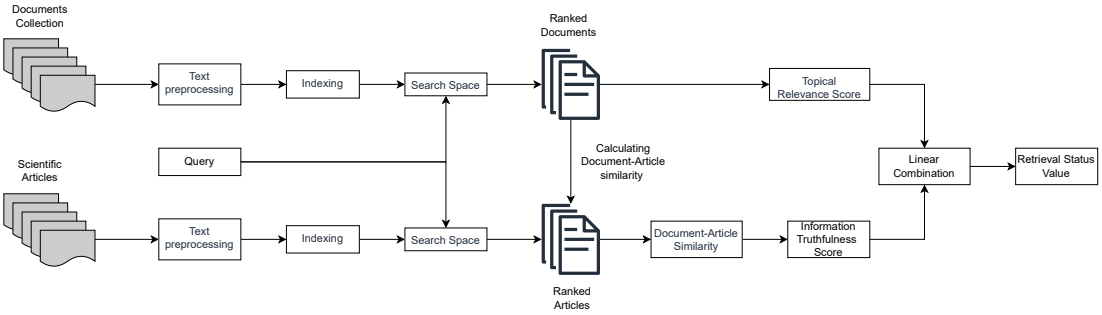

Figure 5.1: The proposed retrieval model, considering both topical relevance and information truthfulness (based on scientific evidence in the form of medical journal articles).

aggregation strategy for obtaining the final Retrieval Status Value (RSV), based on which the estimated relevant documents are ranked. Neither human intervention, nor complex knowledge bases, nor labeled datasets are needed for this purpose. The architecture of the proposed model is illustrated in Fig. 5.1.

5.1.1 Computing Topical Relevance

Topical relevance constitutes the core relevance dimension in any Information Retrieval System (IRS), and assesses how well the content of a document topically meets the information needs of users, which are usually expressed by means of a query (Croft, Metzler, and Strohman, 2010). There are several approaches in literature to estimate topical relevance, one of the most effective is still Okapi BM25 (Robertson, Walker, Beaulieu, Gatford, and Payne, 1996), which is a lexical-based unsupervised model, a baseline for distinct IR tasks (Rosa, Rodrigues, Lotufo, and Nogueira, 2021), based on a probabilistic interpretation of how terms contribute to the relevance of a document and uses easily computed statistical properties such as functions of term frequencies, document frequencies, and document lengths. Using BM25, the topical relevance score of a document d with respect to a query q, denoted as $trs(d,q)$, is calculated as follows:

$$trs(d,q) = \sum_{t \in q,d} \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right) \cdot \frac{tf(t,d) \cdot (k1+1)}{tf(t,d) + k1 \cdot (1 - b + b \cdot \frac{ld}{L})} \tag{5.1}$$

The left part of the equation allows to compute the inverse document frequency of a term with respect to the entire document collection; specifically, N denotes the total number of documents in the collection, and $df(t)$ refers to the document frequency for the term t, i.e., the number of documents in which t appears. In the second part, $tf(t,d)$ denotes the term frequency, i.e., the number of times the term t appears in the document d. Since document collections usually are constituted by documents with different lengths, length normalization is

performed in the denominator; specifically, ld refers to length of the document d, L refers to the average document length, while $k1$ (a positive tuning parameter that calibrates the document term frequency scaling) and b (determines the document length scaling) are internal BM25 parameters.

5.1.2 Computing Information Truthfulness

Various approaches have been proposed in the literature to evaluate information truthfulness,[1] whether health-related or not, whether applied to IR or not.

In our approach, we commence by indexing open-source articles culled from distinguished medical journals,[2] notably the Journal of the American Medical Association (JAMA),[3] and eLife,[4], considered as sources of trustworthy scientific evidence. Utilizing the BM25 algorithm, we sought topically relevant articles, employing queries derived from the dataset earmarked for evaluation in this study, as detailed in Section 5.2.1. A cosine similarity measure facilitated the comparison of each retrieved journal article against every retrieved document corresponding to the designated query. In order to represent the documents and journal articles, we incorporated two BERT-based textual representation models: one pre-trained on MSMarco,[5] and the other on the Pubmed and PubMed Central (PMC) datasets.[6] This process yielded dense vector representations constructed from chunks of 512 tokens, complemented by a sliding-window encompassing 450 words to retain contextual continuity across the document. For the top-n retrieved documents and the top-k retrieved journal articles,[7] we engineered an $n \times k$ similarity matrix. Within this matrix, rows symbolize documents, columns epitomize journal articles, and each matrix cell captures the similarity score juxtaposing the document and the journal article, as depicted in Fig. 5.2.

To compute the information truthfulness score for each document d relative to a query q, represented as $its(d,q)$, we linearly amalgamated the similarity scores between d and the top-k journal articles ji deemed pertinent to the identical query for which d was procured. This was accomplished by weighing scores in proportion to the ranked positions of the retrieved journal articles. Formally:

$$its(d,q) = w_1 \cdot cos(d, j_1) + w_2 \cdot cos(d, j_2) + \cdots + w_k \cdot cos(d, j_k) \tag{5.2}$$

---

1. Although there are numerous terms that have been used in the literature, to refer to this dimension of relevance (e.g., credibility, veracity, genuineness, etc.), in this and other works we prefer to use the concept of truthfulness as an abstract term that can grasp various aspects of the above concepts.
2. `https://openmd.com/guide/finding-credible-medical-sources`
3. `https://jamanetwork.com/`
4. `https://elifesciences.org/`
5. `https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4`
6. `https://github.com/dmis-lab/biobert`
7. It's worth noting that $k \ll n$, to keep the focus on document retrieval and consider only the most relevant journal articles.

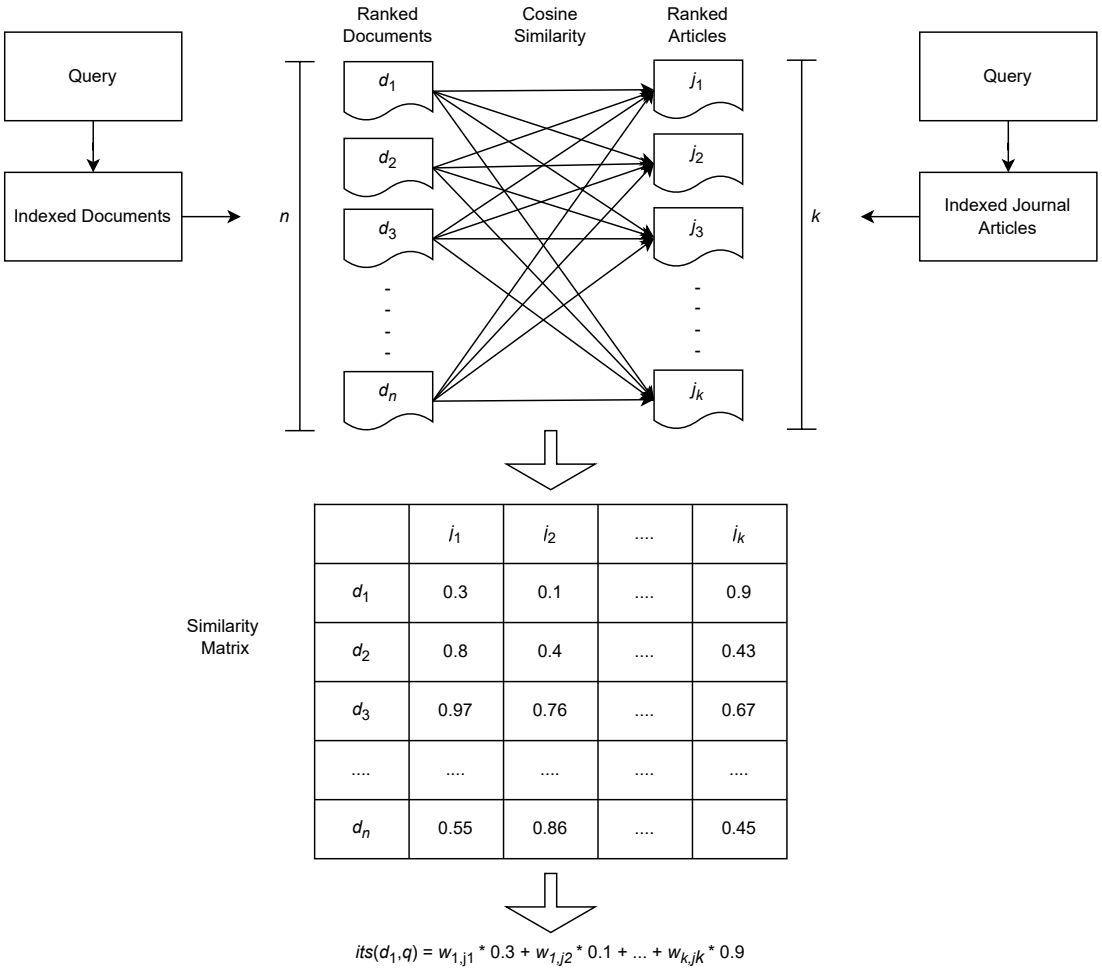

Figure 5.2: Information truthfulness score calculation. q denotes the query that is used to retrieve both documents and journal articles..

In Eq. 2, $w_1, w_2, \ldots, w_k$ denote the weights assigned to each similarity score, such that $\sum w_i = 1$ and $w_i \geq w_{i+1} (1 \leq i \leq k-1)$. This second condition serves to consider the position in the rank in which the journal articles were positioned with respect to the similarity to the documents retrieved (i.e., the higher the position, the higher the weight). The way in which the w_i weights are actually assigned, for evaluation purposes, is illustrated in detail in Sect. 5.2.3.

5.1.3 Computing the Retrieval Status Value

Upon computing the two relevance dimension scores, which are contingent on the user-formulated query, we recognized the imperative to amalgamate these scores to discern the Retrieval Status Value, symbolized as $RSV(d,q)$. This value epitomizes the ultimate relevance score of a document concerning a specific query, taking into account both its topical relevance and information truthfulness. We elected to employ a linear combination for the scores. Formally expressed as:

$$RSV(d,q) = w_{trs} \cdot trs(d,q) + w_{its} \cdot its(d,q) \tag{5.3}$$

In Eq. 3, w_{trs} denotes the weight assigned to the topical relevance score, and w_{its} denotes the weight assigned to the information truthfulness score. Also in this case, each weight w_{**s} is actually assigned, for evaluation purposes, as illustrated in Section 5.2.3. In the same section, the solution adopted to normalize the two relevance dimension scores in the same numerical range is also explained, since they are calculated in different ranges.

5.2 Experimental Setup

This section describes the experimental evaluation framework that was set up to assess the effectiveness of the retrieval model presented in this article. A BM25 baseline and several model configurations are evaluated on a public dataset and by means of suitable evaluation metrics. The purpose of this experimental evaluation is to punctually assess the effectiveness of such configurations of the proposed approach in using external reputed sources (medical journal articles) to consider information truthfulness as a query-dependent dimension of relevance, compared to the simple baseline chosen that uses topical relevance alone.

5.2.1 The TREC Health Misinformation Track Dataset

The TREC Health Misinformation Track fosters research on retrieval methods that promote reliable and correct information over misinformation for health-related decision-making tasks.[8] In this work, we used a subset of the dataset provided by the Track in its 2020 edition (Clarke et al., 2020). The original dataset is constituted by CommonCrawl news,[9] sampled from January, 1st 2020 to April 30th, 2020, which contains health-related news articles from all over the world. For our experiments, given the large volume of the original dataset, we selected 219,245 English news related to COVID-19. The dataset has a fixed structure, organized into topics. Each topic includes a title, a description, which reformulates the title as a question, a yes/no answer, which is the actual answer to the description field based on the provided evidence, and a narrative, which describes helpful and harmful documents in relation to the given topic. For example, for the topic title field: 'ibuprofen COVID-19', the value of the other attributes in the dataset are, for the description: '*Can ibuprofen worsen COVID-19?*', for the yes/no answer: 'no', and for the narrative: '*Ibuprofen is an anti-inflammatory drug used to reduce fever and treat pain or inflammation*'.

8. https://trec-health-misinfo.github.io/
9. https://commoncrawl.org/2016/10/news-dataset-available/

The considered dataset also consists of an evaluation set of 5,340 labeled data. The data is labeled with respect to usefulness, answer, and credibility. Usefulness corresponds to topical relevance, answer indicates if the document provides an answer to the query contained in the description field, and credibility is the concept that, in the document collection, is used to indicate information truthfulness. In this work, we just considered as labels usefulness and credibility. Both of them are provided on a binary scale, i.e., useful or non-useful, and credible or non-credible.

5.2.2 Evaluation Metrics

The TREC Health Misinformation Track not only furnishes publicly available data but also offers a robust evaluation tool implementing standard Information Retrieval (IR) metrics. In particular, our evaluation employs multiple dimensions of relevance, using Average Precision (AP), Normalized Discounted Cumulative Gain for the first 10 results (NDCG@10), and two variants of the Convex Aggregating Measure (CAM).

The Multidimensional Metric (MM) framework allows for the inclusion of diverse relevance criteria in the evaluation of an Information Retrieval System (IRS) besides topical relevance. Initially, the evaluation results for each relevance dimension are computed independently using distinct metrics. Inspired by the measures employed in TREC Decision Track 2019, we considered both AP and NDCG@10. These scores are then fused into a singular metric using a weighted harmonic mean, an approach sensitive to lower-than-average values, thereby rewarding systems that exhibit consistent performance across all relevance dimensions.

The Convex Aggregating Measure (CAM) combines distinct evaluation results according to the following formula:

$$CAM(r) = \lambda_{\text{rel}} M_{\text{rel}}(r) + \lambda_{\text{cred}} M_{\text{cred}}(r)$$

where M_{rel} and M_{cred} represent topical relevance and credibility measures, respectively. In our study, we experimented with both Mean Average Precision (MAP) and NDCG@n for different numbers of n retrieved results. We set $\lambda_{\text{rel}} + \lambda_{\text{cred}} = 1$, and for our evaluation, both were assigned a weight of 0.5, consistent with TREC 2020 Health Misinformation Track guidelines.

5.2.3 Implementation Technical Details

Basic IR Operations

To index documents and compute topical relevance, we used the BM25 algorithm as implemented in PyTerrier with default parameters. Document retrieval employed topic descriptions from the TREC 2020 dataset as queries. This same procedure was extended for journal article retrieval, serving as the basis for calculating the information truthfulness score.

Assignment of Weights

We evaluated different methods for weight assignment, including heuristic, greedy strategies, and ad-hoc models. For the scope of this article, we followed the weight assignment methodology outlined in [37]. Using this approach, we performed grid search on ten randomly selected queries to optimize weights for both topical relevance score $\text{trs}(d,q)$ and information truthfulness score $\text{igs}(d,q)$ in terms of CAM with MAP. Weight parameters for relevance dimensions were also heuristically tested with configurations that either balanced or skewed importance toward topical relevance or information truthfulness.

Normalization of Relevance Dimension Scores

Due to the different numerical ranges of topical relevance and information truthfulness scores, we employed min-max normalization to bring them to a common scale. Specifically, the normalized topical relevance score $\text{trs}'(d,q)$ is computed as:

$$\text{trs}'(d,q) = \frac{\text{trs}(d,q) - \min_{\text{trs}}(q)}{\max_{\text{trs}}(q) - \min_{\text{trs}}(q)}$$

where $\min_{\text{trs}}(q)$ and $\max_{\text{trs}}(q)$ represent the minimum and maximum topical relevance scores for all documents associated with query q, respectively.

5.2.4 Results and Discussion

This section presents and discusses the outcomes of our experiments, focusing on the comparative performance of different retrieval models as outlined in this research. We employed the simple BM25 retrieval model as our baseline and compared its performance against other configurations, utilizing the metrics previously specified.

Table 5.1 highlights the results in terms of Average Precision (AP) and NDCG@10. The results in this table were subjected to a t-test for statistical significance ($p < 0.05$) (Smucker et al., 2007), with significant results denoted by an asterisk.

Observing the results, it becomes evident that Models (4), (5), and (6), all of which employ BioBERT embeddings, show significant improvement over the baseline BM25 model. These models not only outperform the baseline but also yield better results than the configurations utilizing BERT embeddings. This suggests that BioBERT's specialized vocabulary and training corpus offer an advantage in retrieving more relevant and genuine documents in the healthcare domain.

Additionally, Models (5) and (6) indicate that varying the weight attributed to the different dimensions of relevance—topical relevance and information truthfulness—can further fine-tune the performance, as seen from the statistically significant improvements in AP and NDCG@10 scores.

Table 5.1: Comparison of model performances using MM evaluation framework. Metrics include Average Precision (AP) and NDCG@10. All evaluations consider the same number of top-k journal articles, specifically $k = 10$. Significant results are marked with $*$, indicating $p < 0.05$ according to the t-test (Smucker et al., 2007).

Model	w_{trs}	w_{igs}	AP	NDCG@10	Embeddings
BM25	-	-	0.461	0.8601	-
Model (1)	0.5	0.5	0.469	0.8676	BERT
Model (2)	0.6	0.4	0.474	0.8701	BERT
Model (3)	0.4	0.6	0.476	0.8747	BERT
Model (4)	0.5	0.5	0.479	0.8785*	BioBERT
Model (5)	0.6	0.4	0.481*	0.8813*	BioBERT
Model (6)	0.4	0.6	0.493*	0.8951*	BioBERT

Also from Table 5.2 we observe that BioBERT-based models performs better than BERT-based ones, almost under each model configuration. To test the effectiveness of both textual representations as the number of articles taken as scientific evidence increased, i.e., for $k = 5$, $k = 10$, and $k = 15$, we kept fixed the number of retrieved documents on which the assessments were made (specifically, $n = 20$), and employed both Model (3) and Model (6), which are the ones who provided the best results in Table 5.2 for the BERT and BioBERT representations.

In Table 5.3, we observe that increasing the number of journal articles taken into account as scientific evidence actually contributes positively to the improved results obtained. The superiority of the model based on the BioBERT representation is confirmed, regardless of the number of articles considered.

5.3 Summary and Outlook

In an era marked by the pervasive spread of health misinformation online, our research has holistically addressed the exigency of provisioning online users with information that's not only topically relevant but also truthful. We have championed a unique retrieval model that inherently integrates scientific evidence, sourced from acclaimed international medical journals, to ascertain what we term as "information truthfulness".

What sets our approach apart from existing methodologies in literature is the elimination of dependencies on subject matter experts, manually curated knowledge bases, or the harnessing of labeled datasets in tandem with supervised algorithms to gauge information truthfulness. Instead, our novel unsupervised method undertakes a direct comparative analysis of online health narratives with scholarly articles.

Table 5.2: Experimental results in terms of Convex Aggregating Measure (CAM), w.r.t. both Mean Average Precision (MAP) and NDCG@n, for the top-n documents (# n docs) considered in different runs. The number of top-k journal articles considered as scientific evidence (# k j.arts) is fixed, i.e., $k = 10$. Statistically significant results.

Model	# n docs	w_{trs}	w_{igs}	# k j.arts	CAM_{MAP}	$\text{CAM}_{\text{NDCG@}n}$	Embeddings
BM25	5	1	-	-	0.0631	0.1435	-
Model (1)		0.5	0.5	10	0.0641	0.1434	BERT
Model (2)		0.6	0.4	10	0.0685	0.1475	BERT
Model (3)		0.4	0.6	10	0.0697	0.1495	BERT
Model (4)		0.5	0.5	10	0.0701	0.1487	BioBERT
Model (5)		0.6	0.4	10	0.0721	0.1500	BioBERT
Model (6)		0.4	0.6	10	**0.0894**	**0.1688**	BioBERT
BM25	10	1	-	-	0.1047	0.2052	-
Model (1)		0.5	0.5	10	0.1073	0.2057	BERT
Model (2)		0.6	0.4	10	0.1085	0.2084	BERT
Model (3)		0.6	0.4	10	0.1145	0.2151	BERT
Model (4)		0.5	0.5	10	0.1124	0.2112	BioBERT
Model (5)		0.6	0.4	10	0.1177	0.2161	BioBERT
Model (6)		0.4	0.6	10	**0.1249**	**0.2299**	BioBERT
BM25	15	1	-	-	0.0631	0.1435	-
Model (1)		0.5	0.5	10	0.1399	0.249	BERT
Model (2)		0.6	0.4	10	0.1435	0.2535	BERT
Model (3)		0.4	0.6	10	0.1485	0.2552	BERT
Model (4)		0.5	0.5	10	0.1489	0.2541	BioBERT
Model (5)		0.6	0.4	10	0.1507	0.259	BioBERT
Model (6)		0.4	0.6	10	**0.1597**	**0.2702**	BioBERT
BM25	20	1	-	-	0.1676	0.285	-
Model (1)		0.5	0.5	10	0.1649	0.2845	BERT
Model (2)		0.6	0.4	10	0.1726	0.2905	BERT
Model (3)		0.4	0.6	10	0.1797	0.2945	BERT
Model (4)		0.5	0.5	10	0.1753	0.2902	BioBERT
Model (5)		0.6	0.4	10	0.1783	0.2948	BioBERT
Model (6)		0.4	0.6	10	**0.1978**	**0.3102**	BioBERT

Table 5.3: Comparison of Model (3) and Model (6) by considering the same number, i.e., $n = 20$, of retrieved documents and a different number of top-k journal articles (# k j.arts), as scientific evidence. Statistically significant results.

Model	# k j.arts	CAM_{MAP}	$\text{CAM}_{\text{NDCG@20}}$	Embedding
Model (3)	5	0.1698	0.285	BERT
Model (6)		**0.1787**	**0.2953**	BioBERT
Model (3)	10	0.1797	0.2945	BERT
Model (6)		**0.1978**	**0.3102**	BioBERT
Model (3)	15	0.1810	0.2912	BERT
Model (6)		**0.1975**	**0.3109**	BioBERT

While our approach offers promising avenues for truthful health information retrieval, it is crucial to acknowledge the limitations inherent in our study to provide a complete and balanced understanding of our research. Notably, our investigations were conducted using a single dataset focused exclusively on COVID-19 related documents, and this dataset was in one language only. This constraint limits the generalizability of our findings across other health conditions, languages, and cultural contexts. Moreover, the number of experiments conducted, varying only a limited number of parameters (k and n), does not fully explore the potential search space, and the evaluation was primarily technical, lacking the direct involvement of real users to assess the outputs of different systems. These limitations indicate potential areas for future expansion and refinement.

As we continue our exploration in the realm of truthful health information retrieval, several promising research directions beckon. Key among these is the need to juxtapose our retrieval model with established baselines prevalent in the Information Retrieval (IR) domain. This will provide a comprehensive understanding of our model's performance, both in terms of retrieval effectiveness and computational efficiency.

In the subsequent chapter, we turn our attention to an evolved IR model by zeroing it on summaries in Chapter 6 or specific text passages in 7, this model offers a departure from traditional full-text retrieval. Such a nuanced method not only curtails the negative influences of irrelevant content in lengthy documents but also amplifies the document's query-relevant truthfulness assessment. Our preliminary ventures, especially within the CHS paradigm, vindicate the superiority of this passage-centric re-ranking approach over traditional full-text retrieval frameworks. This underlines its immense potential and paves the way for its adoption as the focal point of our future research undertakings.

Chapter 6

Leveraging Document Summarization for Enhanced Relevance Dimensions in Consumer Health Search

In the quest for improving the performance of domain-specific search systems, this chapter explores the possibility of leveraging document summarization to enhance the estimation of relevance dimensions beyond topicality. Our focus is on the health domain, where the Ad-Hoc Retrieval task from the TREC 2020 Health Misinformation Track poses distinct challenges due to the involvement of multiple relevance dimensions. We aim to strike a balance between effectiveness and efficiency by introducing a re-ranking approach that utilizes document summaries to estimate domain-specific relevance scores.

Taking a deeper dive into the health domain, we navigate the contours of the TREC Health Misinformation Track, which accentuates the Ad-Hoc Retrieval task (Clarke et al., 2020). At its core, the task implores the development of advanced retrieval models capable of returning information that is not just topically relevant but also stands up to the scrutiny of correctness and credibility. Given the pivotal nature of health-related information, ensuring the authenticity and accuracy of the returned results becomes paramount. In our exploration, we underscore the importance of tying credibility to the query's theme and hence, opting for real-time estimation.

To drill down further, we present a re-ranking strategy that initiates with full documents ranked in alignment with their topical relevance. These documents then undergo a re-ranking transformation, steered by an overall relevance score—a linear amalgamation of scores from each relevance dimension, encompassing *topicality*, *correctness*, and *credibility*.

6.1 Methodology

Our proposed solution embraces a re-ranking architecture that leverages document summarization for multidimensional relevance estimation. The overarching architecture of this solution is meticulously represented in Figure 7.1.

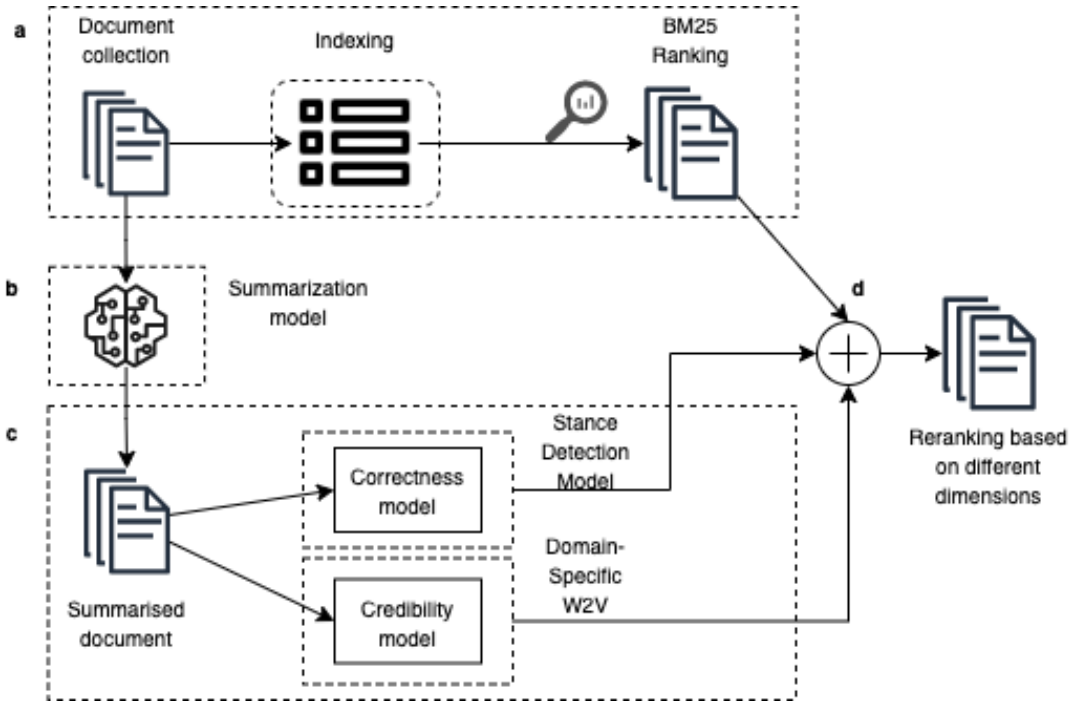

Figure 6.1: Architecture of the proposed summarization-based approach for multidimensional relevance estimation.

The methodology can be segmented into distinct yet interrelated phases:

1. **Topicality Estimation** (refer to Figure 7.1a): This initial phase hinges on the prominent BM25 model (Robertson et al., 1996) to compute the topical relevance of documents. The result of this phase is a ranked list of top-k documents based predominantly on their topical relevance.
2. **Document Summarization** (refer to Figure 7.1b): Before diving into the subsequent dimensions of relevance, namely *correctness* and *credibility*, documents undergo a summarization process.
3. **Estimation of Additional Relevance Dimensions** (refer to Figure 7.1c): Building on the summarized documents, we proceed to estimate the aforementioned dimensions of relevance - *correctness* and *credibility*.

4. **Re-ranking based on Multidimensional Relevance** (refer to Figure 7.1d): The culmination of our methodology is the re-ranking phase. Here, the initial list produced during the topicality estimation phase is restructured. This restructuring is influenced by the scores of the additional relevance dimensions ascertained from the summarized documents.

For a more granular breakdown of the Ad-Hoc Retrieval Task, the nuances of the estimation process for relevance dimensions excluding topicality, and their amalgamation into a comprehensive estimate, the forthcoming Sections 6.1.1–6.1.4 provide a deep dive.

6.1.1 Topicality Estimation

Topical relevance constitutes the core relevance dimension in any IRS, and assesses how well the content of a document topically meets the information needs of users, which are usually expressed by means of a query (Croft et al., 2010). There are several approaches in literature to estimate topical relevance, one of the most effective is still Okapi BM25 (Robertson et al., 1996), which is a lexical-based unsupervised model, a strong baseline for distinct IR tasks (Rosa et al., 2021), based on a probabilistic interpretation of how terms contribute to the relevance of a document and uses easily computed statistical properties such as functions of term frequencies, document frequencies, and document lengths. Using BM25, the topical relevance score of a document d with respect to a query q, denoted as $trs(d,q)$, is calculated as follows:

$$trs(d,q) = \sum_{t \in q,d} \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right) \cdot \frac{tf(t,d) \cdot (k1+1)}{tf(t,d) + k1 \cdot (1 - b + b \cdot \frac{ld}{L})} \tag{6.1}$$

The left part of the equation allows to compute the inverse document frequency of a term with respect to the entire document collection; specifically, N denotes the total number of documents in the collection, and $df(t)$ refers to the document frequency for the term t, i.e., the number of documents in which t appears. In the second part, $tf(t,d)$ denotes the term frequency, i.e., the number of times the term t appears in the document d. Since document collections usually are constituted by documents with different lengths, length normalization is performed in the denominator; specifically, ld refers to length of the document d, L refers to the average document length, while $k1$ (a positive tuning parameter that calibrates the document term frequency scaling) and b (determines the document length scaling) are internal BM25 parameters.

6.1.2 Document Summarization

In the realm of text summarization, two major approaches dominate: extractive and abstractive summarization. The former involves extracting whole sentences or phrases from the original text to form the summary, while the latter involves generating new sentences, often leading to more coherent and concise summaries. For the needs of our approach, we predominantly leaned towards *extractive summarization*.

Our chosen method, the *TextRank* algorithm, is a long-established technique in the field of extractive summarization (Mihalcea and Tarau, 2004). TextRank is an unsupervised, graph-based approach that ranks sentences in a document based on their relative significance. It relies on the idea that sentences with more and stronger connections to other sentences in the document are likely to be more informative. Despite being nearly two decades old, TextRank's simplicity, transparency, and lack of requirement for training data continue to make it a viable choice, particularly in applications where explainability is paramount.

To capture the semantic essence of sentences and to compute their similarity, we harnessed pre-trained *GloVe* embeddings (Pennington, Socher, and Manning, 2014). These embeddings transform sentences into dense vector representations in a high-dimensional space, making it feasible to gauge the semantic closeness of different sentences.

The culmination of the TextRank algorithm's processing results in a set of sentences ranked by their relevance. In line with our methodology, we cherry-picked only the top ten sentences deemed most salient by the algorithm to form the extractive summary of each document.

As we move forward, it's worth exploring how the integration of more recent advancements in NLP and deep learning could further enhance the performance and efficiency of our summarization component, possibly integrating or comparing with methods like BERT. Nevertheless, the decision to utilize TextRank in our current model was driven by a strategic choice to prioritize simplicity, efficiency, and explainability in the context of our specific application.

6.1.3 Relevance Dimensions Estimation

The primary goal of the Ad-Hoc Retrieval task, as described in the TREC 2020 Health Misinformation Track Overview Paper (Clarke et al., 2020), is to develop a ranking model that places emphasis on credible and accurate information. The task can yield various outcomes in the context of documents, ranging from them being useful, correct, and credible to being incorrect and non-useful.

Given the importance of understanding these outcomes, we delve into the details of the data provided, particularly concerning the dimension of *correctness*. Every topic is accompanied by a *treatment–disease* pair where the disease invariably is COVID-19. The pair is further supplemented by a *description* framed as a question pertaining to the treatment's efficacy

against COVID-19. An associated *answer* field, limited to binary values “yes” or “no,” is provided as a ground truth for model development. As highlighted in (Clarke et al., 2020), a document’s correctness is ascertained if its contents match the provided answer, making correctness a query-dependent relevance dimension.

The dimension of *credibility* can be derived from diverse methodologies, considering the text or any metadata associated with the documents. Although not inherently query-dependent, our approach opts to assess credibility during run-time on summarized documents, seeking to align this relevance dimension with the query that mirrors users’ informational needs.

To gauge *correctness*, we took inspiration from (Fernández-Pichel et al., 2020a), devising a model that formulates a *topical expression* by amalgamating the topic’s *description* and the provided *answer*. We then determine the similarity between this expression and each document under scrutiny for correctness. Using Sentence-BERT (Reimers and Gurevych, 2019a), a pre-trained BERT-based model, each sentence within a document is transformed into embeddings. The final correctness score for each document is derived by calculating the *cosine similarity* between its sentences and the topical expression, with the apex cosine-similarity score being chosen.

Addressing *credibility*, increasingly spotlighted in recent IR research (Goeuriot, Pasi, Suominen, Bassani, Brew-Sam, González-Sáez, Upadhyay, Kelly, Mulhem, Seneviratne, et al., 2021; Putri, Viviani, and Pasi, 2021), our approach utilizes a credibility assessment model. This model interprets the task as a binary classification challenge, akin to various literary proposals (Fernández-Pichel et al., 2021; Lima, Wright, Augenstein, and Maistro, 2021). Our model employs a supervised learning method that focuses on linguistic features. These features, labeled as *textual representation* (Di Sotto and Viviani, 2022), are sourced from *Word2Vec* pretrained on PubMed (Pyysalo, Ginter, Moen, Salakoski, and Ananiadou, 2013). Utilizing these domain-specific features, a *Logistic Regression* classifier is trained for the task.

6.1.4 Overall Relevance Estimation

In the re-ranking phase, we amalgamated the diverse relevance scores employing a *linear combination* method (Wu, Bi, Zeng, and Han, 2009). During this process, we experimented with various *aggregation schemes*.

Aggregation Schemes

Formally, given the score and weight pairs for topicality (s_t, α), correctness (s_{co}, β), and credibility (s_{cr}, γ), we explored four aggregation schemes represented as $agr_{\#}(s_t, s_{co}, s_{cr})$:

- $agr_1(s_t, s_{co}, s_{cr}) = lc_p(lc_p(s_t, s_{co}), s_{cr}) = (1-\gamma) * (\alpha * s_t + \beta * s_{co}) + \gamma * s_{cr}$
- $agr_2(s_t, s_{cr}, s_{co}) = lc_p(lc_p(s_t, s_{cr}), s_{co}) = (1-\beta) * (\alpha * s_t + \gamma * s_{cr}) + \beta * s_{co}$
- $agr_3(s_t, s_{co}, s_{cr}) = lc_p(s_t, lc_p(s_{co}, s_{cr})) = \alpha * s_t + (1-\alpha) * (\beta * s_{co} + \gamma * s_{cr})$
- $agr_4(s_t, s_{co}, s_{cr}) = lc_s(s_t, s_{co}, s_{cr}) = \alpha * s_t + \beta * s_{co} + \gamma * s_{cr}$

The terms lc_s and lc_p define two linear combination techniques used by the aforementioned schemes:

- lc_s: linear combination *simple*, fusing the three dimensions simultaneously,
- lc_p: linear combination *paired*, merging a maximum of two dimensions concurrently.

To ensure each relevance score for the different dimensions remains within the [0,1] range, we implemented the *min-max normalization* technique (Lee, 1997). We trialed multiple *weighting schemes* to pinpoint the best weights for different relevance dimensions.

Weighting Schemes

An initial strategy was to allocate identical weights to each relevance dimension, a method advocated in previous works (Adhikari and Agrawal, 2014; Wu, 2012; Wu et al., 2009). However, this approach proved less effective than the straightforward BM25-based IR model focusing solely on topicality. Consequently, we adopted an alternative methodology, inspired by (Wu et al., 2009).

We selected 10 queries at random and carried out a *grid search* on the four aggregation schemes with different weights, evaluating them using the CAM_{map} metric— the TREC 2020 Health Misinformation Track's official metric (Clarke et al., 2020). This assessment helped identify the most effective *aggregation-weighting scheme* combination. Consequently, the optimal weight configurations linked with the four considered aggregation schemes (denoted as ws_i for the aggregation scheme agr_i, where $i = 1, \ldots, 4$) are:

- $ws_1 : \alpha = 0.8, \beta = 0.2, \gamma = 0.5$
- $ws_2 : \alpha = 0.8, \beta = 0.5, \gamma = 0.2$
- $ws_3 : \alpha = 0.8, \beta = 0.5, \gamma = 0.5$
- $ws_4 : \alpha = 0.8, \beta = 0.1, \gamma = 0.1$

The determined optimal weights underline the anticipated supremacy of topicality in relevance estimation (Saracevic, 2007). Additionally, the relevance measures of credibility and correctness might be influenced by the potential information loss from document summarization.

6.2 Experimental Setup

This section initially discusses the implementation details of the solutions used within this paper (Section 6.2.1), and then presents and discusses the results of the evaluation of the proposed approach in terms of both effectiveness and efficiency (Section 6.2.2).

6.2.1 Implementation Details

For evaluation purposes, we considered the dataset provided in the *AdHoc Retrieval Task* of the TREC-2020 Health Misinformation Track. Based on available computational resources, we considered a subset of the entire dataset, consisting of around 1 million documents related to 46 assessed topics; for a complete description of the original dataset, please refer to (Clarke et al., 2020).

For the development of the proposed approach, *PyTerrier* (Macdonald and Tonellotto, 2020) has been used for indexing and producing the initial ranking based on the BM25 model. For the estimation of correctness, in the sentence representation phase, we used the `sentence-transformers` framework and the best performing (in terms of effectiveness) pre-trained Sentence-BERT model named `all-mpnet-base-v2`(Reimers and Gurevych, 2019a). For credibility assessment, we used the `scikit-learn` (Buitinck et al., 2013) implementation of Logistic Regression and the `gensim` implementation (Řehůřek and Sojka, 2010) of the *Word2Vec* model pre-trained on PubMed.

The implementation of extractive summarization leveraged the Python `NetworkX` library (Hagberg, Swart, and S Chult, 2008). To perform word embedding in summarization, we employed *GloVe* pre-trained on 100-dimensional vectors on 6B token corpus (Wikipedia 2014 + Gigaword 5) with 400K word vocabulary. Summarization was performed offline, not impacting the run-time estimation of relevance dimensions.

6.2.2 Results and Discussions

This section illustrates and discusses the results produced by the proposed approach, in terms of both effectiveness and efficiency. In particular, the performed evaluations aim to reply to the following research questions (related to the considered Ad-Hoc Retrieval task):

- *R1: Does the proposed re-ranking solution, which estimates domain-specific relevance dimensions based on document summaries instead of full documents, maintain the retrieval effectiveness?*
- *R2: Does the proposed solution actually improve efficiency over baselines?*
- *R3: Can the proposed solution strike a satisfactory balance between effectiveness and efficiency?*

R1

To reply to this question, we consider: (i) a *baseline* consisting of a retrieval model that takes into account only topicality on full-length documents (i.e., BM25); (ii) the *four aggregation schemes* as formulated in Section 6.1.4, considering all the *three relevance dimensions* (i.e., topicality, correctness, and credibility) in the re-ranking phase, with their respective *optimal weights* computed as described in Section 6.1.4. Experiments are performed both on *full-length documents* (FULL) and *summarized documents* with distinct lengths of 150, 100, and 50 words (SUM-150, SUM-100, and SUM-50).

To compute the results over the baseline and aggregation schemes, we employ the evaluation scripts and metrics provided by the TREC-2020 Health Misinformation Track (Clarke et al., 2020). Specifically, the evaluation metrics used are the *Convex Aggregating Measure* (CAM$_{map}$) (Lioma, Simonsen, and Larsen, 2017), CAM$_{n\mathrm{DCG}}$, and *compatibility* (Clarke, Vtyurina, and Smucker, 2021). In particular, the results are illustrated by considering two and three relevance dimensions at a time when computing evaluation metrics, as done in TREC. When considering two dimensions, CAM$_{map}$ focuses on dimension pairs (correctness, credibility) and (topicality, credibility), whereas CAM$_{n\mathrm{DCG}}$ considers (topicality, correctness) and (topicality, credibility). Both CAM$_{map}$ and CAM$_{n\mathrm{DCG}}$ also consider the three dimensions together. Compatibility scores, detailed in (Clarke et al., 2021), are computed for *harmful* and *helpful* results, as defined in (Clarke et al., 2020). To guarantee statistical significance in comparing results, we used the *paired t-test* as it is the most commonly used in IR evaluations (Urbano, Lima, and Hanjalic, 2019). We considered p-values less than 0.05 as significant.

Table 6.1 reports the effectiveness results. From the table, it emerges that the proposed multidimensional relevance estimation approach produces significantly better results for agr_2, agr_3, and agr_4 compared to the baseline (BM25),[1] also when we employ a summarized version of documents to compute correctness and credibility (for agr_3 and agr_4). Upon inspecting SUM-150, SUM-100, and SUM-50, it becomes evident that our method maintains its performance level even with reduced document length, albeit with a slight drop-off compared to the FULL document results. Notably, SUM-150, under the agr_3 scheme, achieved the best performance, scoring 0.5664 in the CAM$_{map}$ metric, which is marginally lower than the FULL agr_3, but significantly better than the BM25 baseline. This result indicates that our method can indeed maintain good effectiveness, even when working with significantly shortened document summaries. The *compatibility* measure also showed improvements over the BM25 baseline for both harmful and helpful queries, demonstrating the adaptability of our method in different contexts. This is particularly evident for the FULL scheme, but SUM also maintains reasonable performance.

1. According to statistical significance computed using paired t-test with p-value < 0.05.

Table 6.1: Comparative evaluation of effectiveness (on 1,000 retrieved documents)

	2 Aspects (CAM_{map})		$\text{CAM}_{n\text{DCG}}$		3 Aspects		Compatibility		
Model	**Scheme**	**Cor-Cre**	**Top-Cre**	**Top-Cor**	**Top-Cre**	**CAM_{map}**	**$\text{CAM}_{n\text{DCG}}$**	**harmful**	**helpful**
BM25		0.3824	0.6478	0.6930	0.8069	0.5201	0.6385	0.2101	0.5245
FULL	agr_1	0.3237	0.4939	0.6876	0.7443	0.4236	0.6209	0.1556	0.4376
SUM-150		0.3046	0.4635	0.6807	0.7435	0.3971	$0.6141_{\dagger}$	0.1855	0.4455
SUM-100		0.2910	0.4599	0.6736	0.7455	0.3908	0.6066	0.1707	0.4357
SUM-50		0.2889	0.4566	0.6708	0.7398	0.3883	0.6049	0.168	0.4313
FULL	agr_2	0.4128	0.6844	0.7233	0.8220	0.5693^{*}	0.6767^{*}	0.1927	0.5903
SUM-150		0.3694	0.6186	0.7140	0.8046	0.5189	0.6524	0.1823	0.5138
SUM-100		0.3657	0.6177	0.7104	0.8037	0.5174	0.6487	0.1819	0.5129
SUM-50		0.3654	0.6146	0.7067	0.8032	0.5147	0.6469	0.1807	0.5116
FULL	agr_3	0.4412	0.7019	0.7514	0.8438	0.5844^{*}	0.6869^{*}	0.1976	0.5515
SUM-150		0.4362	0.6869	0.7347	0.8367	$\underline{0.5664}^{*}$	$\underline{0.6707}^{*}_{\dagger}$	0.1945	0.5313
SUM-100		0.4052	0.6663	0.7241	0.8261	0.5601^{*}	$0.6667^{*}_{\dagger}$	0.2032	0.5309
SUM-50		0.4043	0.6645	0.7238	0.8258	0.5596^{*}	$0.6659^{*}_{\dagger}$	0.1995	0.5297
FULL	agr_4	0.4047	0.6560	0.7261	0.8203	0.5523^{*}	0.6713^{*}	0.2009	0.5302
SUM-150		0.4009	0.6523	0.7243	0.8200	$0.5493^{*}_{\dagger}$	$0.6648^{*}_{\dagger}$	0.1994	0.5206
SUM-100		0.3965	0.6519	0.7204	0.8192	$0.5477^{*}_{\dagger}$	$0.6603^{*}_{\dagger}$	0.1962	0.5202
SUM-50		0.3951	0.6475	0.7185	0.8169	$0.5452^{*}_{\dagger}$	$0.6589^{*}_{\dagger}$	0.1959	0.5188

(Note: Top: topicality, Cre: credibility, Cor: correctness. Topicality is always estimated on full-length documents.)

If we compare the two best-performing aggregation schemes, i.e., agr_3 and agr_4, we found that agr_3 achieves better performance, however, the drop in effectiveness for summarized documents is lower than that for full documents for agr_4.

We recall that topicality is estimated, in every aggregation scheme, on the full content of documents, and this can have a positive impact on the results but, as it emerges in particular from the aggregation schemes agr_3 and agr_4, particular configurations can be found in which the multidimensional aspect is effective with respect to the overall assessment of relevance even using summarized documents. The high performance of agr_3 can be attributed to several factors. Firstly, it emphasizes the topicality dimension by factoring it separately into the aggregation, which aligns well with the concept of document relevance where topicality is often the most crucial criterion. Secondly, by initially combining the correctness and credibility dimensions, agr_3 computes what might be called a *reliability score* (if a piece of information is correct and credible, it can reasonably be considered reliable (Egala, Liang, and Boateng, 2022)). In many domain-specific search contexts, such as *health search* where correctness and credibility have particular importance (given the possible harm that would be done in incurring false or incorrect information), this score can positively impact the document's overall relevance. Furthermore, the agr_3 aggregation scheme is particularly effective when used with document summaries (SUM-150, SUM-100, and SUM-50). It maintains high performance even with reduced document length, indicating its robustness and adaptability. This is critical in large-scale or time-sensitive search tasks, where efficiency and effectiveness are both paramount.

R2

The reply to the second research question deals with the evaluation of the efficiency of the proposed solution. In this case, we assessed *computational time performances*, which were measured on a server with AMD Ryzen 7 5800h, and GeForce RTX 3070 Mobile. The results presented in Table 6.2 show that using summarized documents can actually lead to a gain in terms of computational time. In particular, we observed the *relative gain* in terms of *Mean Response Time* (MRT) when considering:

- The average time required to respond to each query for which 1000 documents were retrieved, denoted as $mrt(Q)$;
- The average time required to estimate the individual relevance dimensions of a single document, denoted as $mrt(d)$

In a noteworthy detail, the relative gain of $mrt(Q)$ for summarized documents of 150, 100, and 50 words were 23.92%, 38.12%, and 44.93%, respectively. Concerning the relative efficiency gain in estimating $mrt(d)$, with respect to correctness, it improved by 23.72%, 37.77%, and 44.73%, respectively for summarized documents of 150, 100, and 50 words, and by 60%,

88%, and 93.6% respectively for credibility, for summarized documents with respect to the same above-mentioned number of words. This analysis suggests that the use of summarized documents can result in significant efficiency gains, but we still need to verify the trade-off between the effectiveness and efficiency of our solution in order to fully comment on the results.

Table 6.2: Comparative evaluation of efficiency

	Avg. words/d	$mrt(d)$ (sec)			$mrt(Q)$ (sec)
Data		Top	Cre	Cor	
FULL	1024.60	4.016×10^{-2}	0.005	8.002×10^{-2}	805.6848
SUM-150	148.02	4.016×10^{-2}	0.002	6.102×10^{-2}	612.6813
SUM-100	99.10	4.016×10^{-2}	6×10^{-4}	4.981×10^{-2}	499.1034
SUM-50	49.67	4.016×10^{-2}	3.2×10^{-4}	4.428×10^{-2}	443.6153

Note: Top: topicality, Cre: credibility, Cor: correctness. Topicality is always estimated on full-length documents, hence $mrt(d)$ does not change for Top.

R3

With a view to answering the last research question, we decided to use a *data visualization* solution by plotting *effectiveness-efficiency graphs* that consider all the four aggregation schemes and different types of documents (i.e., both full and summarized documents, at distinct summarization lengths). On these graphs, illustrated in Figures 6.2 and 6.3, the reader can observe *effectiveness* as measured by the CAM_{MAP} metric (Figure 6.2) and by the $\text{CAM}_{n\text{DCG}}$ metric (Figure 6.3). In both figures, *efficiency* is assessed in terms of $mrt(Q)$. These are the same metrics used to assess effectiveness and efficiency respectively in Tables 6.1 and 6.2.

The graphs vividly portray the inherent trade-off between efficiency and effectiveness associated with each applied aggregation scheme. Specifically, each curve refers to an aggregation scheme, and the graphical symbols on the curves (i.e., circle, square, triangle, and diamond) represent the best trade-offs between effectiveness and efficiency for each aggregation scheme. In particular, in both graphs, we can observe that the utilization of the $aggr3$ aggregation scheme across different document types (FULL, SUM-150, SUM-100, and SUM-50) consistently outperforms the alternatives.

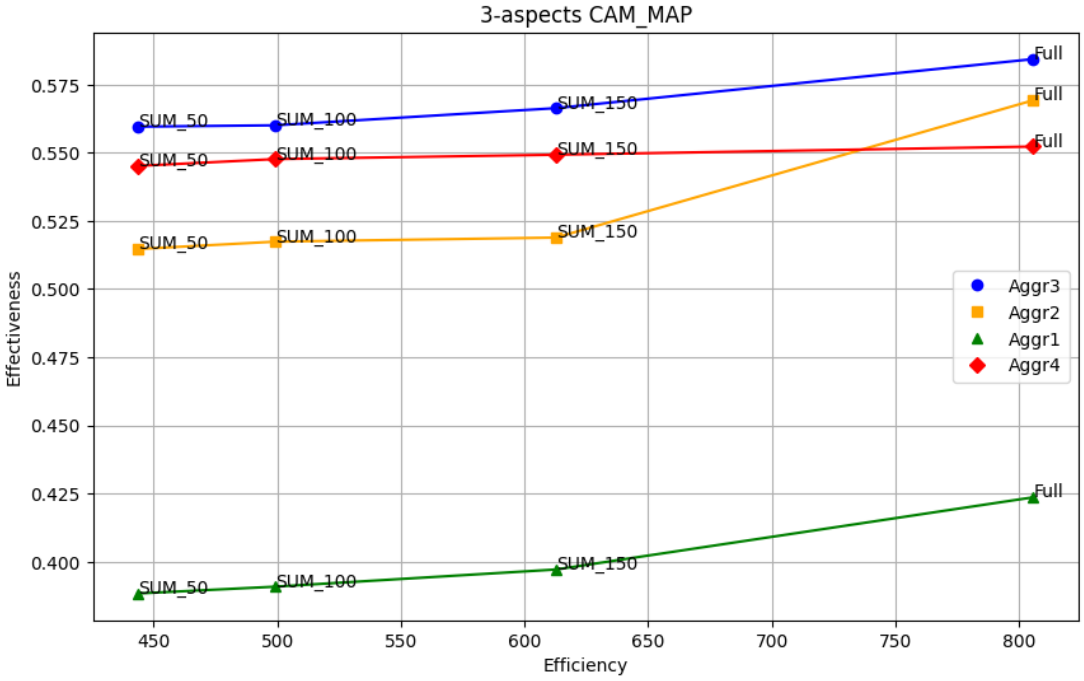

Figure 6.2: The x-axis represents the efficiency score in terms of $mrt(Q)$, while the y-axis represents the effectiveness score in terms of CAM$_{map}$.

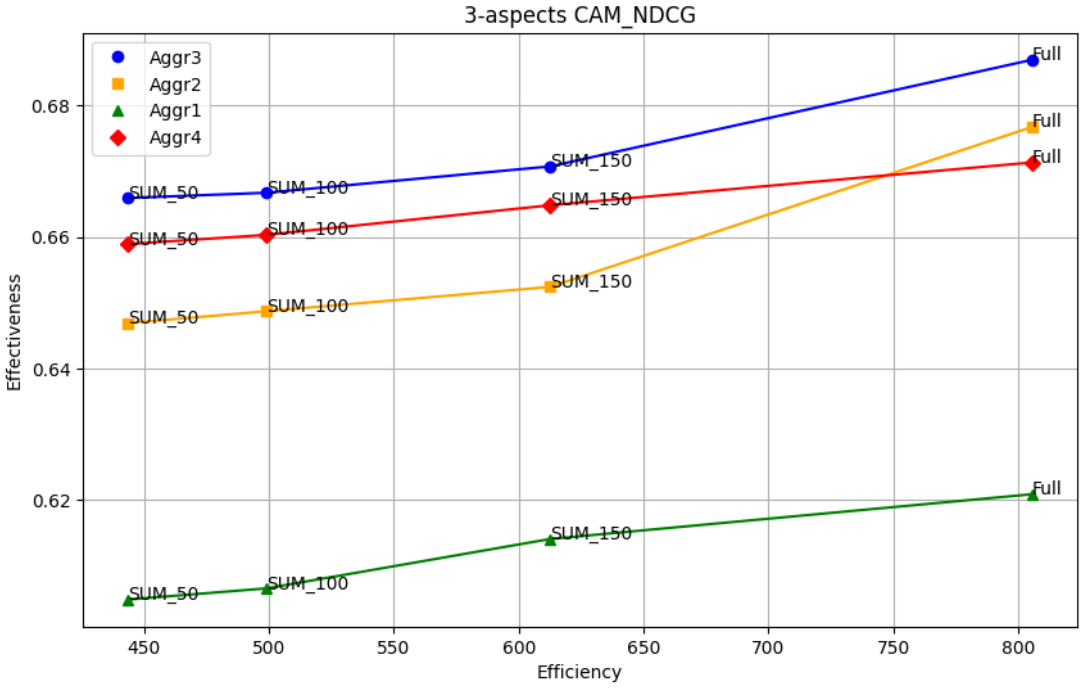

Figure 6.3: The x-axis represents the efficiency score in terms of $mrt(Q)$, while the y-axis represents the effectiveness score in terms of CAM$_{n\text{DCG}}$.

6.3 Summary and Outlook

While the progress made in this chapter is notable, a few avenues remain open for exploration. First, there is potential in deepening our understanding of how each relevance dimension individually contributes to the comprehensive estimation. This opens up prospects to experiment with novel aggregation schemes and re-ranking methodologies, particularly those leveraging the power of Transformer-based models, which, as of now, are predominantly employed in single-dimensional relevance scenarios.

In the next chapter, we pivot our attention towards refining the concept of re-ranking in Information Retrieval (IR). Although prevalent re-rankers extensively analyze entire document texts to arrive at a relevance score, this exhaustive scrutiny can sometimes lead to less-than-ideal retrieval outcomes.

Chapter 7

A Passage Retrieval, Transformer-based Re-ranking Model for Consumer Health Search

Chapter 6 primarily revolved around multidimensional relevance assessment, emphasizing the need of balance between effectiveness and efficiency, especially in specialized domains like Consumer Health Search. It highlighted the potential of using summarized documents for estimating relevance dimensions.

Conventional re-rankers typically analyze the full text of documents to compute an aggregate relevance score, a practice that often culminates in sub-optimal retrieval outcomes due to the noise introduced by query-unrelated content. Even some of the state-of-the-art Transformer-based re-rankers, while focusing on specific text passages rather than the entire document, limit their analysis to topical relevance. Bridging this gap, this chapter introduces an advanced IR model that deploys re-ranking techniques with a focus on carefully selected text passages within documents. This nuanced approach serves the dual purpose of reducing the noise attributed to query-irrelevant content and enabling a more accurate assessment of a document's truthfulness, thereby achieving a more effective retrieval process.

The remainder of this chapter is organized as follows: Section 1 presents the methodology; Section 2 outlines the experimental setup and evaluation metrics; and Section 3 discusses the conclusions drawn from this research.

7.1 Methodology

This chapter proposes a new model for Information Retrieval, focusing on passage-based re-ranking for multi-dimensional relevance. Specifically, the Passage Retrieval Transformer-based re-ranking model is introduced. This model consists of four primary stages: (i) first-stage retrieval using BM25, (ii) passage segmentation, (iii) Passage Retrieval, and (iv) Transformer-based re-ranking of documents. These stages are described in detail in the following sections and illustrated in Figure 7.1.

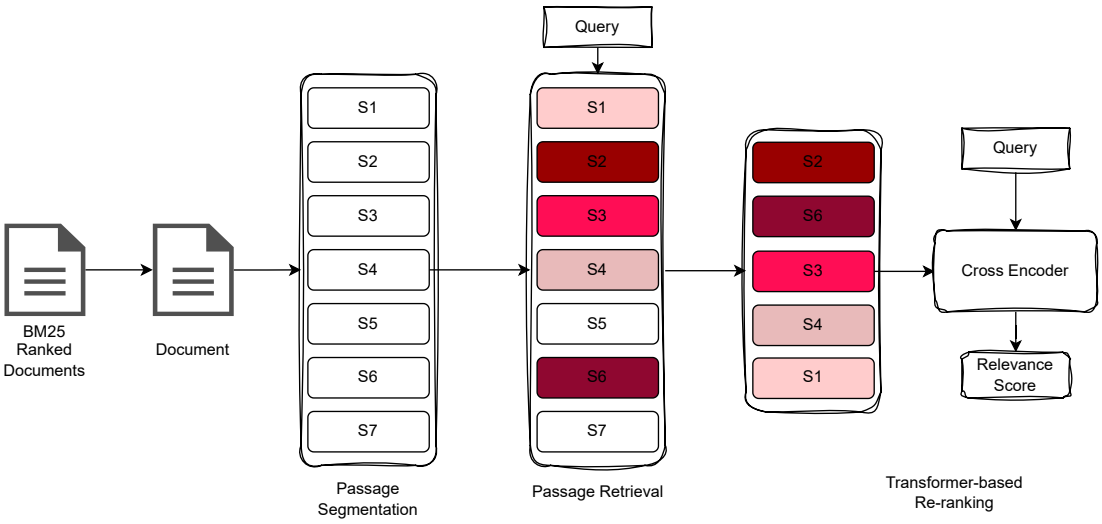

Figure 7.1: The four stages of the Passage Retrieval Transformer-based re-ranking model.

7.1.1 First-stage Retrieval: BM25

The model's *first-stage retrieval* employs the BM25 retrieval model (Robertson, Zaragoza, et al., 2009). This stage calculates a *topicality score*, denoted as BM25(q,d), based on word frequency and distribution for both the query q and the document d. The resulting score is used to rank a list of potentially relevant documents. The mathematical formulation for BM25 is given by Equation 7.1:

$$\text{BM25}(q,d) = \sum_{t \in q,d} \log \frac{N - df(t) + 0.5}{df(t) + 0.5} \cdot \frac{tf(t,d) \cdot (k_1 + 1)}{tf(t,d) + k_1 \cdot (1 - b + b\frac{l_d}{L})} \tag{7.1}$$

In this equation, N represents the total number of documents in the collection, $df(t)$ stands for the document frequency of term t, $tf(t,d)$ refers to the term frequency in document d, l_d signifies the length of document d, and L represents the average length of documents in the corpus. Parameters k_1 and b are tunable and used to scale term frequency and document length, respectively.

7.1.2 Passage Segmentation

The second stage, *passage segmentation*, breaks down the documents retrieved from the first stage into smaller text units, referred to as passages. Unlike previous works such as KeyBLD (Li and Gaussier, 2021), which use blocks, our approach employs the NLTK sentence tokenizer to segment documents into individual sentences, as shown in Figure 7.2.

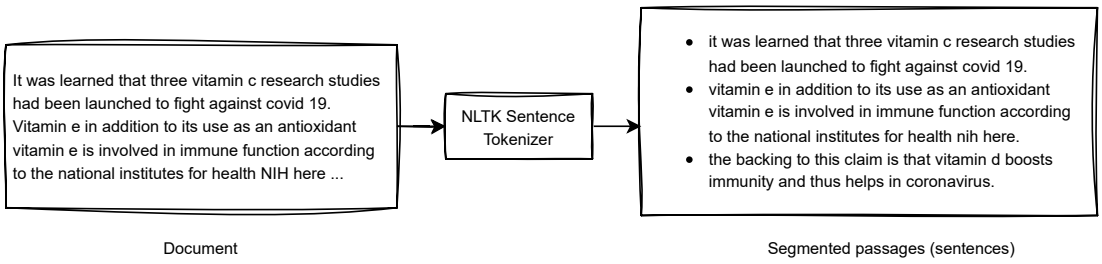

Figure 7.2: Example of a document segmented into sentences in the passage segmentation stage.

We hypothesize that sentences can offer more granular information for both topical relevance and truthfulness evaluation. A block of text might encompass multiple ideas, making it less suited for single-query evaluations and for assessing the truthfulness of individual claims. To validate this hypothesis, a preliminary evaluation was conducted on publicly available datasets, specifically targeting the truthful health IR task. The results of this evaluation are discussed in Section 7.2.3.

7.1.3 Passage Retrieval

After the document sentences are extracted in the passage segmentation stage, the *Passage Retrieval* phase aims to select the most (topically) relevant sentences for a given query from each document. This retrieval is conducted on the top-k documents selected by the first-stage retrieval, discussed in Section 7.1.1. These top-k documents are considered to be “globally" relevant to the query.

We employ BioBERT (Lee, Yoon, Kim, Kim, Kim, So, and Kang, 2020) to encode both the query and the sentences into vector representations. BioBERT is a state-of-the-art language model particularly suited for biomedical text applications, and it has been effective in various Natural Language Processing (NLP) tasks like Named Entity Recognition (NER) (Bhatia, Celikkaya, Khalilia, and Senthivel, 2019; Liu, Hu, Xu, Xu, and Chen, 2021b).

NER helps identify *Named Entities*, such as disease and medication names, within the text. By incorporating NER into our model, we enhance the context of sentences extracted during the passage segmentation, providing a more accurate basis for comparison with the query. Specifically, we use NER to identify entities related to *disease* and *medication* in the sentences. This ensures that sentences closely align with the specific entities mentioned in the query, as illustrated in Figure 7.3.

Formally, we compute the similarity score $\sigma(q,s)$ between a query q and a sentence s as follows:

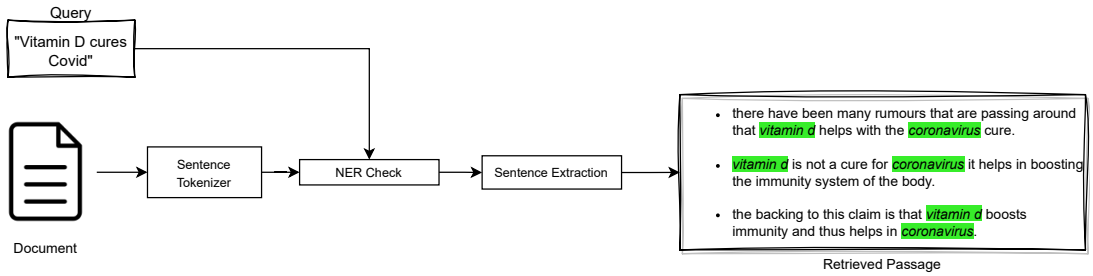

Figure 7.3: Query-relevant Passage Retrieval enhanced with NER.

$$\sigma(q,s) = \begin{cases} \cos(q,s), & \text{if } \mathrm{NER}_q(\mu,\delta) = \mathrm{NER}_s(\mu,\delta) \\ w_d \cdot \cos(q,s), & \text{otherwise} \end{cases} \tag{7.2}$$

Here, μ and δ represent the medication and disease entities, respectively. The function $\cos(q,s)$ denotes the cosine similarity between the vector representations of q and s. $\mathrm{NER}_x(\mu,\delta)$ represents the Named Entities identified in x, where $x \in \{q,s\}$. Finally, $w_d \in [0,1]$ is a discount weight used to penalize the similarity score when the Named Entities in the query and the sentence do not correspond. To find the optimal value for w_d, we conducted a grid search using the NDCG metric, as further discussed in Section 7.2.

7.1.4 Transformer-based Re-ranking

After calculating the similarity values for each query-sentence pair, the next step is to select the top-h most relevant sentences. These sentences form a *sentence-based document*, which serves as the basis for the re-ranking process. As will be further elaborated, the selection of an optimal number of sentences (h) is critical for the overall effectiveness of the model.

Sentence-based Documents

Formally, a sentence-based document is denoted as $\tilde{d}$ and is defined as:

$$\tilde{d} = s_1 \oplus s_2 \oplus \cdots \oplus s_h \tag{7.3}$$

Here, $\oplus$ signifies the concatenation of sentences, and $s_1, s_2, \ldots, s_h$ are the sentences that are ranked in the top-h positions based on their $\sigma(q,s)$ values. Each sentence is individually scored against user query. Only those sentences with the highest relevance scores are chosen to represent the document in its sentence-based form. This ensures that each sentence-based document is a distilled, focused representation of the original, containing only the most query-relevant information.

Cross-encoder Re-ranking

We employ a *cross-encoder* re-ranker (Devlin et al., 2018) for this stage of the process. In Information Retrieval (IR), a cross-encoder combines a query (q) and a candidate document (d) into a single input sequence for a Transformer-based model like BERT. Using Transformer attention mechanisms, the model computes a *Retrieval Status Value* ($\mathrm{RSV}(q,d)$) that quantifies the relevance between q and d. Formally:

$$\mathrm{RSV}(q,d) = \mathrm{CE}([\mathrm{CLS}]\ q\ [\mathrm{SEP}]\ d\ [\mathrm{SEP}]) \cdot W \tag{7.4}$$

In this equation, CE represents the cross-encoder, and '[CLS]' and '[SEP]' are special tokens that signify the beginning and separation of sequences, respectively. W is a learned weight matrix.

For our model, the cross-encoder operates on the sentence-based documents $\tilde{d}$ rather than the complete original documents. This is expressed as:

$$\mathrm{RSV}(q,\tilde{d}) = \mathrm{CE}([\mathrm{CLS}]\ q\ [\mathrm{SEP}]\ \tilde{d}\ [\mathrm{SEP}]) \cdot W \tag{7.5}$$

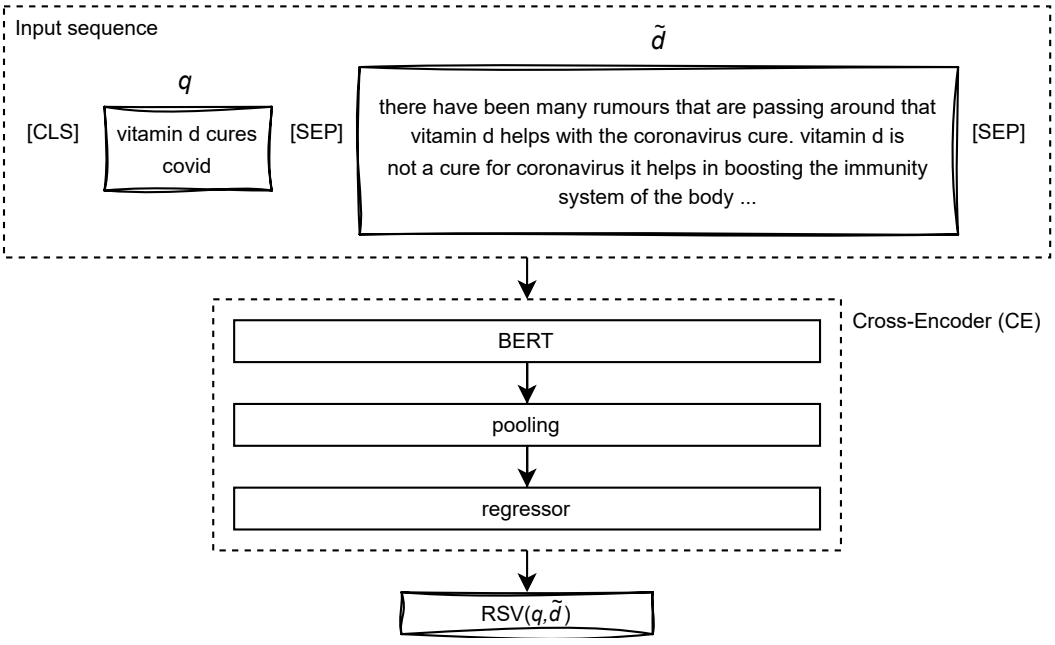

Figure 7.4: Overview of the Cross-Encoder architecture used in the proposed model.

Importantly, the labels used for fine-tuning the BERT model are based on both topicality and truthfulness, enabling our model to consider these two dimensions during re-ranking. By working with sentence-based documents, the cross-encoder receives a more compact and focused representation, potentially improving the retrieval scoring. This approach also reduces the computational complexity and time required for the re-ranking process.

7.2 Experimental Setup

We focused on the *ad-hoc retrieval* task within the context of the TREC-2020 Health Misinformation Track (Clarke et al., 2020) and the CLEF-2020 eHealth Track (Goeuriot et al., 2020) for our evaluation. Both tracks pertain to *Consumer Health Search* (CHS) and give weight to *credibility* as an essential factor of relevance, in addition to topicality.[1] We utilized a subset of 1 million documents from each track, with the TREC-2020 Track encompassing 46 topics linked to Coronavirus and the CLEF-2020 Track including 50 medical conditions. The TREC-2020 Health Misinformation Track categorizes documents into binary labels, with those that meet the criteria of being “topically relevant and credible" labeled as “1", and the remaining labeled as “0". The same binary labeling procedure applies to both topicality and credibility for the CLEF-2020 eHealth Track.

7.2.1 Implementation Details

We used *PyTerrier* (Macdonald, Tonellotto, MacAvaney, and Ounis, 2021) for indexing and initial BM25-based retrieval. Separate indexes were created for the TREC-2020 and CLEF-2020 datasets. The top 500 documents from the first-stage retrieval served as the input for the re-ranking phase. We employed BioBERT (Lee et al., 2020), specifically the `dmis-lab/biobert-v1.1` variant,[2] for re-ranking due to its suitability for health-related content.

For training and testing, we used an 80-20 split on one dataset (e.g., TREC-2020), and all queries and documents from the other dataset (e.g., CLEF-2020) were used for validation, and vice versa. The BioBERT model was fine-tuned for 10 epochs using the Adam optimizer and a learning rate of 2×10^{-5}. We utilized a batch size of 4 and a maximum sequence length of 512 tokens. For implementation, the *HuggingFace* library (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz, et al., 2019), `cross-encoder` package from the *Sentence-Transformers* library (Reimers and Gurevych, 2019a), and *PyTorch* (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, Desmaison, Köpf, Yang, DeVito, Raison, Tejani, Chilamkurthy, Steiner, Fang, Bai, and Chintala, 2019) were used.

---

1. Given the current lack of datasets in Consumer Health Search that are labeled with respect to both topicality and truthfulness (understood as the factuality of the information, as previously introduced), in the experiments we approximate this concept with that of credibility used in datasets.
2. `https://huggingface.co/dmis-lab/biobert-v1.1`

7.2.2 Baselines and Evaluation Metrics

For performance comparison, we evaluated the proposed model against the following baselines:

- BM25: the BM25 model implemented by PyTerrier;
- WAM: the aggregation-based multidimensional relevance model presented in (Upadhyay et al., 2022), based on a simple weighted average of distinct relevance scores. Specifically, weights associated with topicality and credibility are set as in the best model described in (Upadhyay et al., 2022). This model is tested with different percentages of relevant sentences, i.e., 5%, 10%, 15%, 20%, 25%, and Full Document;
- KeyBLD: the model for key-block detection that selects the most informative blocks from a document based on their topical relevance to the query;
- PARADE:[3] the Passage Retrieval model for document ranking that uses aggregation techniques to combine relevance signals from a document's passages;
- CE_{full} (512 tokens relevant passages): the cross-encoder model for re-ranking as proposed in (Devlin et al., 2018), based on Equation (7.4), and with the maximum length obtainable for a BERT document, i.e., 512 tokens.

In the proposed solution, the last cross-encoder for re-ranking is employed in association with different percentages of relevant sentences constituting the sentence-based document, according to Equation (7.5). In this case, the model is denoted as CE_p, where p indicates a given percentage of sentences, i.e., 5%, 10%, 15%, 20%, 25%. For instance, $\text{CE}5\%$ would mean that the top 5% of sentences (ranked by relevance) are concatenated to form the document representation used in the cross-encoder for re-ranking. The evaluation metrics considered for experiments are *Normalized Discounted Cumulative Gain* at 10 and 20 (NDCG@10, NDCG@20), *Precision* at 10 and 20 (P@10, P@20), *Mean Reciprocal Rank* at 10 (MRR@10), and *Mean Average Precision* (MAP). All results are statistically significant according to a paired t-test $(p < 0.05)$ with Bonferroni correction for multiple testing, as described in (Weisstein, 2004).

7.2.3 Results and Discussion

In this subsection, we address two pivotal research questions to evaluate the performance and implications of our proposed solution. Specifically, we examine:

- R1. *What is the impact of using sentence-level representations instead of block-level representations for document re-ranking based on topicality and truthfulness?*
- R2. *Is the utilization of Passage Retrieval and Transformer-based re-rankers more effective than the approaches currently documented in the literature?*

3. https://github.com/canjiali/PARADE/

R1: Sentence-level vs. Block-level Representations

Our initial investigation sought to ascertain the optimal text passage length for re-ranking. Specifically, we aimed to determine whether using a single sentence, two sentences, or blocks would yield the best results. Specifically, the 2-sentence representation approach, a method that differs from individual sentence consideration by dividing the document into consecutive pairs of sentences, each treated as a single unit for relevance scoring. After scoring these sentence pairs for their relevance to the query, we ranked them and selected the top pairs and concatenate them to form document. For this comparison, we employed the CE_{full} re-ranking model, currently regarded as the most effective in existing literature, to fill the 512-token limit of BERT documents with top-h passages. The outcomes of this preliminary evaluation are captured in Table 7.1.

Table 7.1: Performance comparison of the CE_{full} cross-encoder re-ranker using different textual passage lengths to populate the 512-token limit for BERT documents on both CLEF and TREC datasets. Metrics in bold denote the best performance across the different configurations.

		CLEF					
	Passage Type	NDCG@10	NDCG@20	P@10	P@20	MRR@10	MAP
CE_{full}	1 sentence	**0.2843**	**0.2848**	**0.2811**	**0.2818**	**0.4801**	**0.1474**
	2 sentences	0.2531	0.2511	0.2503	0.2495	0.4221	0.1023
	blocks	0.2632	0.2612	0.2661	0.2615	0.4434	0.1231
		TREC					
	Passage Type	NDCG@10	NDCG@20	P@10	P@20	MRR@10	MAP
CE_{full}	1 sentence	**0.6055**	**0.6023**	**0.6059**	**0.6011**	**0.6997**	**0.2986**
	2 sentences	0.5601	0.5578	0.5545	0.5396	0.6311	0.2589
	blocks	0.5691	0.5671	0.5631	0.5403	0.6324	0.2677

The data in Table 7.1 demonstrates that using a single sentence as the text passage yields the best performance across all measured metrics for both the CLEF and TREC datasets. This is most probably because sentences are more succinct compared to blocks or two-sentence passages, which can contain irrelevant or conflicting information. The results imply that employing one sentence as the passage type can enhance the cross-encoder model's effectiveness in document retrieval tasks. Nevertheless, determining the best approach may be contingent on the datasets and the task considered, and additional experimentation may be required.

R2: Efficacy of CE$_p$ Model in Comparison with Literature Approaches

To ascertain the effectiveness of the proposed CE$_p$ re-ranking model, we conducted comprehensive experiments and compared its performance against well-established baseline models in the literature. The models were evaluated across two standard benchmarks—CLEF and TREC datasets. The comparative performance is delineated in Table 7.2.

Table 7.2: Comparison of the performance of different models on CLEF and TREC datasets, with various percentages of relevant passages and Full Document (512 tokens in the cross-encoder model) as input. In bold the best results.

		CLEF					
Model	Rel. Passage	NDCG@10	NDCG@20	P@10	P@20	MRR@10	MAP
BM25		0.1054	0.1578	0.1081	0.1954	0.1578	0.0764
WAM	Full Document	0.0865	0.1591	0.1002	0.2034	0.1232	0.0632
	5%	0.0912	0.1699	0.1096	0.2156	0.1503	0.0694
	10%	0.0993	0.1643	0.1195	0.2213	0.1596	0.0701
	15%	0.1031	0.1694	0.1254	0.2284	0.1612	0.0744
	20%	0.1342	0.1864	0.1495	0.2443	0.1965	0.0985
	25%	0.1032	0.1703	0.1295	0.2294	0.1664	0.0792
KeyBLD		0.2635	0.261	0.2645	0.2645	0.4431	0.1233
PARADE		0.2512	0.2534	0.2551	0.2593	0.4342	0.1213
CE_{full}	512 tokens	0.2843	0.2848	0.2811	0.2818	0.4801	0.1474
CE_5	5%	0.2956	0.2958	0.2899	0.2931	0.5083	0.1499
CE_{10}	10%	0.3145	0.3058	0.3002	0.3012	0.5293	0.1552
CE_{15}	15%	0.3215	0.3198	0.3112	0.3098	0.5453	0.1659
CE_{20}	20%	**0.3475**	**0.3446**	**0.3423**	**0.3445**	**0.5923**	**0.1878**
CE_{25}	25%	0.3398	0.3223	0.3301	0.3311	0.5545	0.1599

		TREC					
Model	Relevant Passage	NDCG@10	NDCG@20	P@10	P@20	MRR@10	MAP
BM25		0.4166	0.4231	0.4177	0.4266	0.5107	0.2142
WAM	Full Document	0.5065	0.5164	0.4976	0.5001	0.5546	0.2453
	5%	0.5112	0.5199	0.4999	0.5051	0.6012	0.2579
	10%	0.5231	0.5221	0.5034	0.5093	0.6231	0.2734
	15%	0.5225	0.5223	0.5087	0.5102	0.6333	0.2788
	20%	0.5546	0.5533	0.5234	0.5212	0.6443	0.2945
	25%	0.5264	0.5288	0.5097	0.5143	0.6332	0.2834
KeyBLD		0.5432	0.5443	0.5342	0.5403	0.6324	0.2677
PARADE		0.5693	0.5664	0.5634	0.5669	0.6589	0.2785
CE_{full}	512 tokens	0.6055	0.6023	0.6059	0.6011	0.6997	0.2986
CE_5	5%	0.6194	0.6156	0.6012	0.6001	0.7211	0.3223
CE_{10}	10%	0.6534	0.6429	0.6267	0.6144	0.7345	0.3414
CE_{15}	15%	0.6623	0.6602	0.6322	0.6234	0.7541	0.3568
CE_{20}	20%	**0.6934**	**0.6801**	**0.6511**	**0.6311**	**0.7834**	**0.3784**
CE_{25}	25%	0.6634	0.6597	0.6374	0.6232	0.7431	0.3493

A number of observations can be made based on the results presented:

- The BM25 model, often used as a baseline in Information Retrieval (IR) tasks, lags considerably behind other deep learning-based models in performance metrics.
- The WAM model demonstrates respectable performance when it utilizes query-relevant passages to compress the document to 20% of its original size. Despite this, its efficacy is still surpassed by the CE model in various configurations.
- The CE model, particularly in its proposed configuration as CE_p, outperforms all baseline models in both CLEF and TREC datasets. Notably, the CE_p model exhibits peak performance when the document is truncated to 20% of its original size using query-relevant passages.
- Interestingly, the performance of the CE_p model begins to decline when more than 20% of the document is considered. This indicates a diminishing return on effectiveness beyond this point, likely due to the incorporation of less relevant or noisy data.

The results affirm the central thesis of this research, validating the efficacy of the CE_p model, especially when the document is reduced to a specific fraction (20%) of its original size using query-relevant passages. This outcome suggests that over-inclusion of sentences beyond this threshold compromises the quality of results, reiterating the value of selective passage retrieval.

7.3 Summary and Outlook

This study has addressed gaps in the existing Information Retrieval (IR) literature pertaining to multidimensional relevance, with a particular focus on the Consumer Health Search task within the health domain. While extant models often employ a two-phase re-ranking approach that first applies a standard IR model, followed by a more nuanced re-ranking phase, these models have shown limitations in both effectiveness and efficiency. This is because they generally consider the entire document for re-ranking, and they predominantly focus on only one dimension of relevance—topical relevance.

In response to these limitations, we introduced a novel Transformer-based re-ranking model that leverages Passage Retrieval techniques. The purpose is to extract the most contextually and factually relevant portions of a document, thereby capturing both topical relevance and information truthfulness. Empirical results demonstrate that our proposed model significantly outperforms existing re-ranking solutions, including those based on Transformers like BERT, particularly in scenarios requiring multidimensional relevance considerations.

Several promising avenues of research emerge from this study:

- **Explainability Layer:** We aim to further enhance our model by incorporating an explainability layer. This layer would elucidate how each passage contributes to the dimensions of relevance under consideration, thereby making the re-ranking process more transparent and interpretable.

- **Domain Adaptation:** Future work also involves fine-tuning and applying the proposed model in various other domains. Nonetheless, it is imperative to note that the availability of high-quality datasets with both topicality and truthfulness labels remains a challenge, not just in the health domain but across the IR landscape.
- **Evaluation Framework:** Given the paucity of comprehensive datasets, there is a need for concerted efforts to develop evaluation frameworks and initiatives that can offer such resources (Fernández-Pichel, Meyer, Bink, Frummet, Losada, and Elsweiler, 2023; Hofstätter, Althammer, Schröder, Sertkan, and Hanbury, 2020; Petrocchi and Viviani, 2023).

This study thus lays the foundation for developing more effective and efficient re-ranking algorithms that address both topical and truthful relevance, an essential requirement in our current age of information overload and misinformation.

Chapter 8

Considering the Explainability of Information Truthfulness in Consumer Health Search

The landscape of health information retrieval has primarily been dominated by methods that focus on binary classification, labeling information as either "*correct*" or "*misinformation*". However, such an approach, despite its utility, presents certain limitations. Predominantly, it confines users to accept one of two predefined outcomes on the truthfulness of information, often without offering the necessary context. This black-box nature of information assessment hampers users' ability to make informed decisions and poses potential risks.

This chapter offers a fresh perspective on this challenge. Moving beyond the traditional binary categorization, it embraces an *ad-hoc* retrieval paradigm, closely aligned with the *Consumer Health Search (CHS)* task. Building on the foundation laid in Chapter 5, which introduced an unsupervised approach to genuine health information retrieval, this chapter extends the model by weaving in an element of explainability. By intertwining the ranked retrieval of relevant and genuine documents with explanations rooted in scientific evidence, this chapter presents a more comprehensive, transparent, and trustworthy health information retrieval framework. This knowledge base-driven approach, grounded in medical journal articles, ensures that users not only receive relevant and truthful information but also understand the rationale behind its ranking.

To substantiate the efficacy and user-acceptance of this method, this chapter also delves into rigorous evaluation measures, both *quantitative* and *qualitative*. Through a standard classification benchmark and a comprehensive user study, we illuminate the efficacy and potential of our proposed approach in enhancing the landscape of online health information retrieval.

8.1 Methodology

The primary objective of this research is to introduce a methodology that seamlessly incorporates explainability into the process of retrieving genuine health information introduced in Chapter 5. This section meticulously details the underpinnings of the approach, building upon established foundational concepts and integrating them with the novel elements introduced in this chapter.

8.1.1 Adding Explainability for Information Truthfulness

The proposed solution aims at providing users with *scientific evidence* for truthfulness related to the ranking of documents produced by the model described in Chapter 5. To extract this evidence, as illustrated in Figure 8.1, we first retrieve *query-relevant passages* from retrieved documents, i.e., we identify portions of text in each document that are topically relevant with respect to the query. Then, we use these passages to extract *passage-based evidence* from journal articles that are topically relevant with respect to the query. Both query-relevant passages and passage-level evidence are then shown to users by means of a *Graphical User Interface* (GUI), which will be illustrated in detail in Section 8.2.4. This should help to increase the user's understanding of the obtained ranking and provide insight into the reasoning behind the truthfulness of each document in the ranked list.

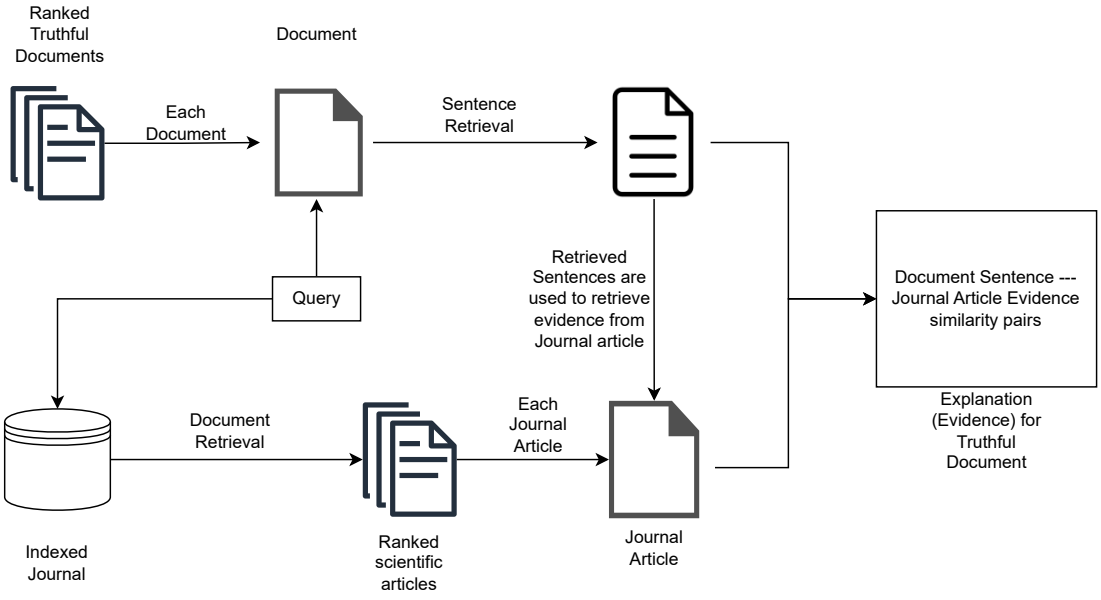

Figure 8.1: High-level outline of the scientific evidence extraction process to be provided to users.

Extracting Query-relevant Passages from Documents

This section details the process of extracting the most important passages from a document in relation to a given query. In fact, the IR model described in Chapter 5 returns a ranked list of documents, which were estimated to be “globally" relevant to a query. In our explainability model, we want to extract from such documents only those text passages that are topically relevant with respect to that query. For the purpose of this paper, one *sentence* was chosen as the size of a textual passage within a document. The high-level overview of this approach is illustrated in Figure 8.2.

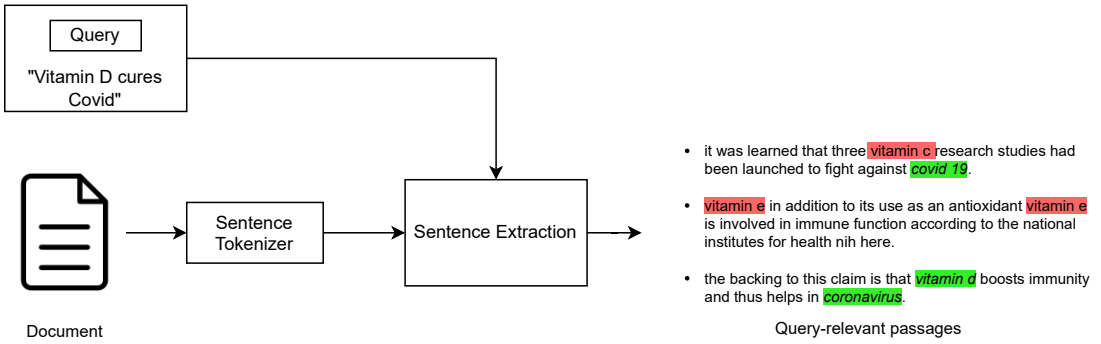

Figure 8.2: Query-relevant passage extraction.

Specifically, to extract *query-relevant passages* (sentences) we considered several strategies. The first strategy is based on representing queries and sentences as TF-IDF vectors, whose similarity is calculated by means of *cosine similarity*, according to which the sentences are ranked. The second strategy is based on using the BM25 model to obtain a ranked list of sentences relevant to the query. The last strategy involves the use of BioBERT to represent queries and sentences and again the use of cosine similarity to obtain a ranked list of sentences. In particular, BioBERT is a leading-edge language model in the biomedical field (Lee et al., 2020). It has proven to be particularly effective in various Natural Language Processing tasks related to medical texts, including *Question-Answering* (QA) (Das and Nirmala, 2022; Poerner, Waltinger, and Schütze, 2020) and *Named Entity Recognition* (NER) (Bhatia et al., 2019; Liu et al., 2021b).

In particular, NER is a process of identifying *Named Entities*, i.e., real-world entities, such as people, organizations, places, dates, and more, in unstructured text. It can improve the sentence extraction process by providing context and additional information about such entities mentioned in medical sentences. Indeed, in medical texts, Named Entities play a crucial role in answering a query exactly; e.g., if the query is about “vitamin C" it would be incorrect to return a sentence that contained “vitamin D," no matter how similar the two vector representations may be. For this reason, it was decided to incorporate NER in the three query-relevant passage extraction models considered (i.e., TF-IDF, BM25, and BioBERT), as high-level illustrated in Figure 8.3.

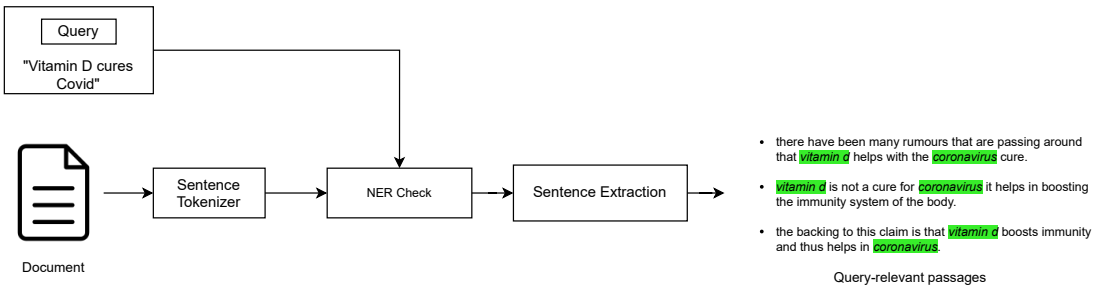

Figure 8.3: Query-relevant passage extraction with NER.

In particular, we compared the two Named Entities *medication*, denoted as μ, and *disease*, denoted as δ, present in both the query and the considered sentences. In this way, the similarity score between a query and a sentence (obtained either by means of cosine similarity or the BM25 similarity) has been modified so as to decrease it in the absence of correspondence between Named Entities. Formally:

$$\sigma(q,s) = \begin{cases} sim(q,s), & \text{if } \mathrm{NER}_q(\mu,\delta) = \mathrm{NER}_s(\mu,\delta) \\ w_d \cdot sim(q,s), & \text{otherwise} \end{cases} \tag{8.1}$$

where $\sigma(q,s)$ indicates the similarity score between the query q and a sentence s, $sim(q,s)$ indicates the similarity function employed to compute $\sigma(q,s)$, which can be either $\cos(q,s)$ or $\mathrm{BM25}(q,s)$ depending on the employed model, $\mathrm{NER}_x(\mu,\delta)$ indicates the Named Entities extracted from x ($x \in \{q,s\}$), and w_d ($w_d \in [0,1]$) is a discount weight,[1] employed to decrease the value of $\sigma(q,s)$ in the case of non-corresponding Named Entities in q and s.

Extracting Passage-based Evidence from Journal Articles

After extracting the query-relevant passages from the documents, the step described in this section involves identifying, within scientific journal articles, pieces of *passage-based evidence* that support the query-relevant passages. This operation is performed on the scientific articles that had been identified as “globally" relevant to the query by the IR model shown in Chapter 5.

This can be achieved by using the same models illustrated in Section 8.1.1 for query-related passages extraction, i.e., models based on TF-IDF, BM25, or BioBERT, in association with Named Entity Recognition, and as high-level illustrated in Figure 8.4.

---

1. For finding the optimal w_d value, we performed a grid search using 5 queries (randomly selected) and document related to those queries. The grid search involved systematically testing different values of w_d within a predefined range, and evaluating the performance of the system for each value of w_d using a predefined set of metrics (F1). The aim of this process was to identify the value of w_d that yielded the best performance in terms of the selected metrics, and therefore the best overall performance for the system.

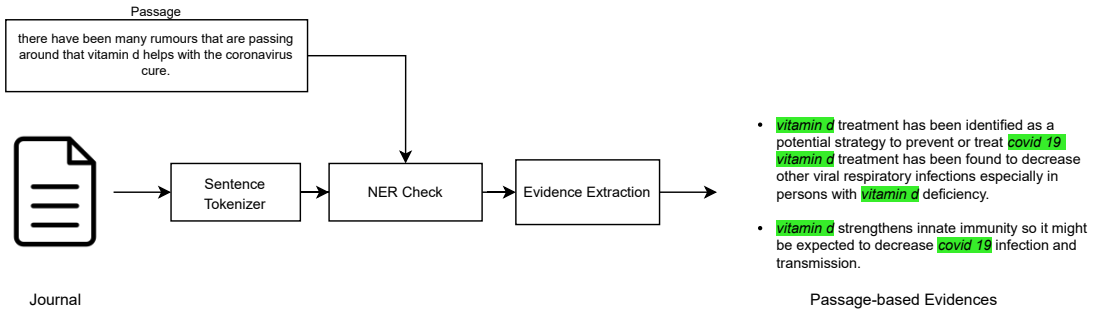

Figure 8.4: Evidence Extraction using NER

8.2 Experimental Setup

In the experimental evaluations conducted, two critical facets were examined. Firstly, a *quantitative* assessment was performed to discern whether the integration of query-relevant passages and passage-based evidence truly enhances the identification of truthful documents. Secondly, a *qualitative* evaluation was undertaken to gauge the efficacy of the proposed approach concerning the explainability of results. This was accomplished through a meticulous user study involving human assessors.

8.2.1 Dataset: TREC “Health Misinformation Track”

The primary dataset deployed for the implementation and validation of the proposed model is derived from the TREC 2020 “Health Misinformation Track” (Clarke et al., 2020). This track emphasizes promoting reliable health information over potential misinformation in medical decision-making scenarios. The original data, extracted from *CommonCrawl news*[2] spanning January to April 2020, encompasses global health-related news articles. From this voluminous set, we strategically chose a subset of 219,245 COVID-19-associated English news articles. However, this subset has an inherent skew, with a larger representation of negative samples compared to the positive ones.

Structured systematically, the dataset comprises *topics*. Each topic possesses a *title*, a descriptive *question*, a *yes/no answer* reflecting the veracity of the description, and a *narrative* detailing the significance of relevant documents concerning the topic. An exemplar is the topic “ibuprofen COVID-19”, elucidated further with “Can ibuprofen exacerbate COVID-19?" yielding the answer “no", and the narrative explicating the role and nature of ibuprofen.

2. https://commoncrawl.org/2016/10/news-dataset-available/

Additionally, an evaluation subset of 5,340 labeled data points is available. Labels categorize documents based on *usefulness*, *answer*, and *credibility*. Here, usefulness mirrors topical relevance, the answer indicates the document's response to the posed query, and credibility functions as a proxy for truthfulness.[3] This study prioritizes the usefulness and credibility labels, distinctly classified as useful/not useful and credible/not credible.

8.2.2 Implementation Details

Document indexing and the application of BM25-based retrieval models were facilitated using the BM25 functionality within *PyTerrier* (version 0.7.0).[4] Default parameters were retained. Document retrieval utilized the topic description from the TREC 2020 dataset as the primary query. This modality was consistently applied to source journal articles pertinent to the query, subsequently utilized to derive evidence (explanations). The BioBERT model `dmis-lab/biobert-v1.1`[5], equipped with sentence-transformers and trained on biomedical data (Lee et al., 2020), was harnessed for this task. The `sent_tokenise` functionality from *NLTK* (version 3.8) was our choice for passage tokenization.[6] The Graphical User Interface (GUI) was crafted using *anvil*[7], offering user-friendly deployment. All experimental setups were orchestrated in *Python* (version 3.7).

8.2.3 Quantitative Evaluation of Effectiveness

The objective of quantitative model evaluation is to determine whether the similarity scores between query-relevant passages and passage-based evidence pieces are effective in identifying the truthfulness of retrieved documents with respect to a query. Indeed, the purpose of the article is to provide explanations to users based on such similarity, so they must also prove effective with respect to the task of identifying truthful information as a whole.

Hence, the similarity scores between a query-relevant passage and pieces of passage-based evidence were computed, and this score was employed to classify documents as truthful or non-truthful. In particular, we considered two solutions to calculate the final similarity score between the query-relevant passage and pieces of passage-based evidence. The first solution considers the similarity scores between the query-relevant passage and different pieces of passage-based evidence and calculates their *mean*. The second solution considers the *maximum* similarity score between the query-relevant passage and the different pieces of

---

3. For further contextualization, the TREC "Health Misinformation Track" (Clarke et al., 2020) employed human assessors for document labeling, evaluating the perceived believability or *credibility* of content (McKnight and Kacmar, 2007). Though closely aligned, credibility and truthfulness are not synonymous. Assessors were instructed to utilize objective criteria where feasible.
4. `https://github.com/terrier-org/pyterrier`
5. `https://huggingface.co/dmis-lab/biobert-v1.1`
6. `https://www.nltk.org/`
7. `https://anvil.works/`

passage-based evidence as the similarity score. With respect to both models, we tested query-relevant passage and passage-based evidence extraction models based on TF-IDF, BM25, and BioBERT, with and without the application of NER. We performed the experiments considering a variable number of retrieved documents (i.e., 10, 20, 50, and 100), a variable number of retrieved scientific journal articles (i.e., 1, 5, and 10), a variable number of query-relevant passages extracted from the documents (i.e., 5 and 10). In all cases, the number of pieces of passage-based evidence taken into account was equal to 5.

In performing experiments with respect to each of these configurations, we applied *five-fold cross-validation* in the following way. Query and document pairs were randomly and independently divided into five folds, with each fold containing a subset of the total data. In each iteration of the cross-validation process, one fold was used as the test set, while the other four folds were used as the training set to train the model. The model was then used to calculate the similarity score between the query-relevant passages extracted from the documents and the pieces of passage-based evidence from the journals, based on the considered queries. This approach allowed us to evaluate model performance using all available data, while ensuring that the evaluation was not biased by using different subsets of data for training and testing in each iteration of the cross-validation process. Performance was evaluated in terms of F1 score (F1), *Geometric Mean* score (GM), commonly used for imbalanced datasets (Davagdorj, Lee, Pham, and Ryu, 2020), and *Area Under the* ROC *Curve* (AUC).

Tables 8.1 and 8.2 show the model performance using the BioBERT model for both query-relevant passage and passage-based evidence extraction, without and with the application of NER. Standard deviation values for all the results presented in the table are between ± 0.01 to ± 0.03. The presented results are averaged results for each fold under each parameter configuration. In the tables, the column "#docs" indicates the number of considered retrieved documents, "#journals" the number of considered journal articles, and "#doc-passages" the number of retrieved passages per document. The section indicated by "mean-similarity" shows the results obtained by computing the mean similarity among the retrieved query-relevant passages and pieces of passage-based evidence, while the "max-similarity" section presents the result obtained by considering the highest similarity score among them.

As mentioned earlier, these results are higher than those using the TF-IDF and BM25 models in extracting query-relevant passages and passage-based evidence. For the sake of conciseness, in Tables 8.3 and 8.4 We illustrate the results for these other two models compared to BioBERT only with respect to the best parameter configuration.

Overall, we can observe that the BioBERT model, both with and without the application of NER, outperforms all other models in terms of F1 score, GM, and AUC. Furthermore, incorporating NER generally improves the performance of the models across the board. In addition, the BioBERT model with NER achieves the highest F1 score, GM, and AUC, indicating better

Table 8.1: Quantitative evaluations of the BioBERT model without NER.

Description	#docs	#journals	F1	GM	AUC	F1	GM	AUC
			#doc-passages = 10			#doc-passages = 5		
					mean-similarity			
	100	10	0.61	0.55	0.52	0.6054	0.534	0.507
	50	10	0.613	0.54	0.53	0.613	0.543	0.5871
	20	10	0.601	0.53	0.587	0.6401	0.532	0.581
	10	10	0.64	0.54	0.579	0.6398	0.534	0.596
	100	5	0.604	0.5503	0.521	0.631	0.557	0.512
	50	5	0.601	0.549	0.567	0.634	0.534	0.533
	20	5	0.619	0.543	0.594	0.667	0.545	0.601
	10	5	0.625	0.521	0.59	0.6465	0.556	0.578
	100	1	0.634	0.567	0.556	0.645	0.567	0.534
	50	1	0.654	0.538	0.567	0.634	0.534	0.564
	20	1	0.698	0.534	0.6013	0.688	**0.584**	0.607
	10	1	0.7	0.546	0.581	0.6742	0.534	0.587
					max-similarity			
BioBERT w/o NER	100	10	0.623	0.54	0.54	0.612	0.56	0.53
	50	10	0.63	0.546	0.544	0.632	0.546	0.546
	20	10	0.678	0.571	**0.613**	0.687	0.546	**0.624**
	10	10	0.672	0.578	0.608	0.647	0.567	0.617
	100	5	0.634	0.5467	0.534	0.612	0.545	0.53
	50	5	0.64	0.567	0.545	0.638	0.565	0.546
	20	5	0.678	0.557	0.624	0.645	0.53	0.617
	10	5	0.657	0.566	0.614	0.641	0.534	0.601
	100	1	0.6533	0.566	0.546	0.564	0.567	0.536
	50	1	0.655	0.546	0.534	0.645	0.567	0.567
	20	1	**0.703**	**0.589**	0.607	**0.698**	0.557	0.614
	10	1	0.695	0.59	0.6	0.687	0.587	0.598

performance in misinformation detection with respect to other baselines.[8] We also note that the "max-similarity" model performs better than the "mean-similarity" model. It is also clear from the tables that the application of NER leads to a significant increase in performance, enabling more accurate identification and retrieval of topically relevant sentences that contain important entities or concepts related to the query and evidence from the journal articles. In general, with respect to effectiveness in classifying health misinformation, using only some of the passages

8. We are aware that these results are still far from optimal, especially in the sensitive context of identifying misinformation in the medical field. However, this is not the main purpose of the article, which is focused on illustrating the explainability of these results to the user so that they can make their own decisions.

Table 8.2: Quantitative evaluations of the BioBERT model with NER.

Description	#docs	#journals	F1	GM	AUC	F1	GM	AUC
			#doc-passages = 10			#doc-passages = 5		
					mean-similarity			
	100	10	0.66	0.581	0.56	0.6565	0.5801	0.566
	50	10	0.66	0.574	0.58	0.655	0.573	0.5871
	20	10	0.66	0.57	0.63	0.6901	0.583	0.633
	10	10	0.67	0.58	0.62	0.6701	0.581	0.623
	100	5	0.66	0.5801	0.564	0.664	0.582	0.559
	50	5	0.665	0.574	0.581	0.6547	0.573	0.581
	20	5	0.6852	0.585	0.624	0.694	0.585	0.631
	10	5	0.6816	0.585	0.616	0.6858	0.586	0.614
	100	1	0.684	0.588	0.576	0.6733	0.584	0.573
	50	1	0.684	0.582	0.597	0.667	0.576	0.588
	20	1	0.7107	0.593	0.6313	0.728	**0.603**	0.627
	10	1	0.7129	0.597	0.607	0.7042	0.594	0.602
					max-similarity			
BioBERT w NER	100	10	0.6568	0.58	0.56	0.655	0.58	0.558
	50	10	0.66	0.575	0.584	0.661	0.574	0.584
	20	10	0.707	0.591	**0.643**	0.696	0.586	**0.644**
	10	10	0.702	0.592	0.639	0.673	0.582	0.638
	100	5	0.6524	0.5798	0.552	0.652	0.579	0.55
	50	5	0.67	0.577	0.579	0.659	0.574	0.579
	20	5	0.698	0.587	0.639	0.679	0.58	0.639
	10	5	0.677	0.583	0.633	0.695	0.588	0.628
	100	1	0.6733	0.584	0.569	0.674	0.584	0.566
	50	1	0.679	0.579	0.593	0.675	0.578	0.591
	20	1	**0.7334**	**0.606**	0.637	**0.722**	0.599	0.636
	10	1	0.74	0.619	0.622	0.711	0.597	0.619

Table 8.3: Quantitative evaluations of the TF_IDF, BM25, and BioBERT models without NER.

Description	#docs	#journals	F1	GM	AUC	F1	GM	AUC
			#doc-passages = 10			#doc-passages = 5		
TF_IDF	20	1	0.6335	0.4823	0.4684	0.6337	0.4697	0.4433
BM25	20	1	0.6745	0.5131	0.5337	0.6701	0.5019	0.5195
BioBERT w/o NER	20	1	**0.703**	**0.589**	**0.607**	**0.698**	**0.557**	**0.614**

and not the whole document, we can see that classification performance, especially in terms of F1 score, can be considered quite satisfactory as a preliminary result although not exceptional,

Table 8.4: Quantitative evaluations of the TF_IDF, BM25, and BioBERT models with NER.

Description	#docs	#journals	F1	GM	AUC	F1	GM	AUC
			#doc-passages = 10			#doc-passages = 5		
TF_IDF	20	1	0.6545	0.4998	0.4777	0.6443	0.4923	0.4663
BM25	20	1	0.6985	0.5432	0.5542	0.6881	0.5213	0.5305
BioBERT w NER	20	1	**0.7334**	**0.606**	**0.637**	**0.722**	**0.599**	**0.636**

considering that the classification of misinformation is not the purpose of the article. We must also remember that there may be a potential decoupling between the concept of truthfulness used in this article and the concept of credibility that was used as a classification label in the dataset under consideration, in the absence of other datasets useful for the purpose in the health domain.

8.2.4 Qualitative Evaluation of Effectiveness

The objective of the qualitative model evaluation is to understand the effectiveness of the proposed explainability strategy by assessing the usefulness of the information and scientific evidence provided to users by means of a *user study*. This can help improve the proposed model and guide the development of additional tools or techniques to improve the explainability of the results obtained by means of the model.

The user study was conducted with 18 human assessors, all doctoral and master's students experienced in NLP and IR, respecting the age and gender balance criteria. The study was performed by means of a specifically-designed *Graphical User Interface* (GUI). Assessors were given clear guidance on the domain under consideration, how to use the GUI, and what aspects to evaluate.

In the following, the GUI is detailed in Section 8.2.4. By means of this GUI, the users were required to perform some *tasks*, illustrated in Section 8.2.4. Later, users were required to answer a *questionnaire*, detailed in Section 8.2.4. Based on this questionnaire, it was possible to assess the *outcome* of user satisfaction with respect to the explainability of the results obtained, as discussed in Section 8.2.4.

The Graphical User Interface

The appearance of the developed GUI is illustrated in Figure 8.5. Here are visible some key components of the interface, which can be summarized into five main panels, as follows.

(a) *The Query Panel*: it presents the set of 12 randomly-chosen queries from those available in the dataset (i.e., 48) from which human assessors can choose;
(b) *The Ranking Panel*: it presents the ranked list of the top-5 documents retrieved w.r.t. a query, by using the IR model detailed in Chapter 5. In particular,

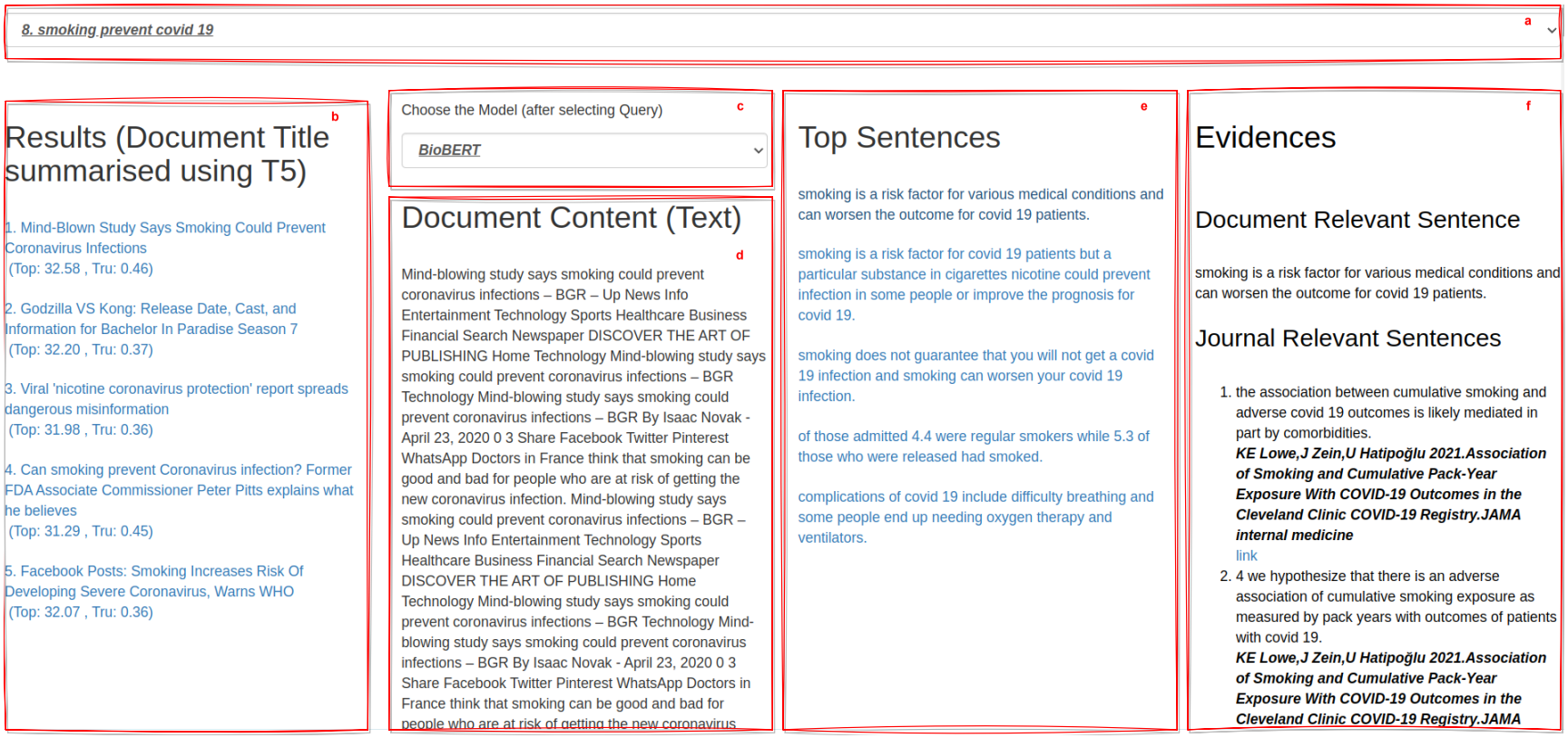

Figure 8.5: The Graphical User Interface.

in this panel, the title associated with these documents are presented. Since in the original dataset, no titles described the content of documents, we employed the T5 model over documents to produce significant titles.[9] This ranked list further associates, to each document title, the document's topicality score (*Top*) and truthfulness score (*Tru*);
(c) *The Sentence Extraction Model Panel*: it allows human assessors to choose among the list of the three distinct models for extracting query-relevant passages and pieces of passage-based evidence in the form of sentences. As illustrated in Section 8.1.1, the three models are based on TF-IDF, BM25, and BioBERT;
(d) *The Document Content Panel*: by selecting a document from panel (b), this panel shows its content and highlights the query-relevant passages identified by using the sentence extraction model selected in the panel (c);
(e) *The Top Sentences Panel*: this panel illustrates the list of the query-relevant passages extracted for the query selected in panel (a), in the document selected in panel (b), for the extraction model selected in panel (c);
(f) *The Evidence Panel*: it shows human assessors pieces of passage-based evidence from journal articles for the selected sentence in panel (e).

The Tasks

The tasks were designed to have human assessors test different query-relevant passage and passage-based evidence extraction models to determine the best way to explain the truthfulness of a document. In particular, the human assessors were required to perform the following tasks.

(i) *Evaluate the Ranking*: in this task, assessors were asked to select a query from panel (a), analyze the documents in the obtained ranking against that query and the associated topicality and truthfulness scores in panel (b), and evaluate, based on these and the content of the retrieved documents, which dimension of relevance they believed had the greatest impact on the final ranking;
(ii) *Evaluate the Query-relevant Passages*: in this task, assessors were asked to evaluate, for each document returned in the ranking based on the query chosen in panel (a), what was the best query-relevant passages extraction model. To do this, each assessor had to first choose a document from the ranking in panel (b), choose a model from panel (c), and analyze the highlighted sentences in panels (d) and (e);
(iii) *Evaluate the Passage-based Evidence*: in this task, assessors were asked to evaluate, for each query-relevant passage in panel (e) extracted from each document found in panel (b) against the query in panel (a) and the model chosen

9. The T5 model is a pre-trained language model developed by Google that uses a Transformer-based architecture to generate text (Raffel et al., 2020). It was trained on a large corpus of text using a task-specific approach called "text-to-text" learning, where the model learns to perform a specific Natural Language Processing task by mapping input text to output text. For our title generation task, we used the T5 implementation provided at the following address: `https://huggingface.co/fabiochiu/t5-small-medium-title-generation`

in panel (c), the usefulness and reliability of the scientific evidence associated with each step and illustrated in panel (f), to determine whether the supporting scientific evidence was sufficient and clear to determine the truthfulness of the document.

The Questionnaire

The questionnaire was used to collect information on the perceived quality of both query-relevant passage and passage-based evidence extraction models and to understand the users' level of satisfaction with the explainability of the truthfulness of the retrieved documents. In particular, the questionnaire contains a set of questions related to panel (b): for assessing the clarity and influence of topicality and truthfulness in the ranking of the document; to panels (c), (d), and (e): for finding the best method to retrieve sentences and to understand the effectiveness of this choice; and to panels (c), (e) and (f): for assessing the usefulness and quality of evidence provided.

Questions related to ranking – panel (b) – are as follows:

- *Are topicality and truthfulness scores useful to understand the ranking?*
- *Do you think this ranking is more influenced by topicality or truthfulness?*

Questions related to sentence extraction (query-relevant passages) – panels (c), (d), and (e) – are as follows:

- *Are the highlighted sentences topically related to the query by using either TF-IDF, BM25 or BioBERT?*
- *Which of the three models best captures the previous aspect?*
- *Do the highlighted sentences (with the best model between TF-IDF, BM25, or BioBERT) provide sufficient information to determine the truthfulness of the document?*
- *Do you think highlighting a single sentence is enough to capture both the topicality and truthfulness of the document?*

Questions related to sentence extraction (passage-based evidence) – panels (c), (e) and (f) are as follows:

- *Are the top sentences (with the best model between TF-IDF, BM25, or BioBERT) correctly supported by scientific evidence (scientific journal articles)?*
- *Does the scientific evidence provide sufficient information to assess the truthfulness of the document?*
- *Do you think the information sources (the scientific journal articles) associated with each highlighted sentence are trustworthy?*

Some questions in the questionnaire were designed in a way that allows participants to answer *yes*, *no*, *don't know*, or *other*. Some questions in the questionnaire include multiple-choice questions that allow participants to choose specific methods or ways for document ranks, extracted sentences, and evidence.[10]

Outcome of the Questionnaire

The responses to the questionnaire were collected and analyzed to gain insights into the users' perspectives on the proposed model. Given our 18 human assessors, a total of 36 responses were gathered, with three responses per question. In particular, to evaluate the *inter-rater reliability* of the study, we computed *Fleiss' kappa measure* (Fleiss, 1971) for each question as rated by three raters. Fleiss' kappa quantifies the level of agreement among multiple raters, with values closer to 1 indicating stronger agreement. Table 8.5 displays the mean Fleiss' kappa values for each question across all questions.

Overall, the table shows fairly high Fleiss' kappa scores, ranging from a low of 0.64 to a high of 0.91, indicating a satisfactory to high level of agreement among assessors for each question.

Q1–Q2. *Ranking* From Figure 8.6, when considering question Q1, it is interesting to note that the majority of the respondents answered *yes*, indicating that they consider the visualization of both topicality and truthfulness scores to be useful in understanding the obtained ranking. However, there is also a non-negligible number of respondents who answered *partly*, suggesting that some respondents may not fully understand the concepts or how they are related to the ranking.

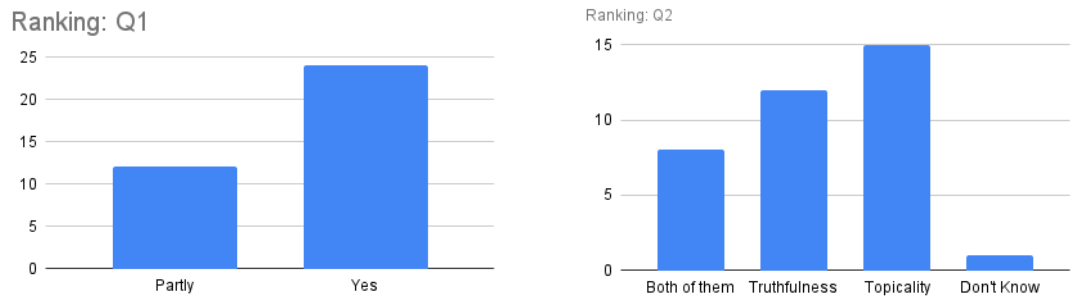

Figure 8.6: Outcome of the questions related to ranking.

10. The questionnaire template provided to human assessors is available at the following address: `https://rb.gy/ziqa3o`

Table 8.5: Mean Fleiss' kappa score for each question for 3 raters

ID & Category	Question	Fleiss' kappa
Q1. *Ranking*	*Are topicality and truthfulness scores useful to understand the ranking?*	0.85
Q2. *Ranking*	*Do you think this ranking is more influenced by topicality or truthfulness?*	0.65
Q3. *Passage extraction*	*Are the highlighted sentences topically related to the query by using either TF-IDF, BM25 or BioBERT?*	0.88
Q4. *Passage extraction*	*Which of the three models best captures the previous aspect?*	0.89
Q5. *Passage extraction*	*Do the highlighted sentences (with the best model between TF-IDF, BM25, or BioBERT) provide sufficient information to determine the truthfulness of the document?*	0.70
Q6. *Passage extraction*	*Do you think highlighting a single sentence is enough to capture both the topicality and truthfulness of the document?*	0.64
Q7. *Evidence extraction*	*Are the top sentences (with the best model between TF-IDF, BM25, or BioBERT) correctly supported by scientific evidence (scientific journal articles)?*	0.85
Q8. *Evidence extraction*	*Does the scientific evidence provide sufficient information to assess the truthfulness of the document?*	0.78
Q9. *Evidence extraction*	*Do you think the information sources (the scientific journal articles) associated with each highlighted sentence are trustworthy?*	0.91

Regarding question Q2, there is more variability in the responses, with approximately equal numbers of respondents choosing *topicality* and *truthfulness* as the factors having the greatest impact on ranking. This suggests a different perception of the respondents with respect to the importance of these factors in the process of ranking the results, which needs to be investigated more in the future also considering the psychological aspects of assessors. However, the limited number of responses *don't know* indicates that only a few respondents fail to get an idea of which dimension of relevance is actually most important with respect to the results obtained.

Q3–Q6. *Passage Extraction* The results for question Q3, as illustrated in Figure 8.7, show that most respondents answered with *yes*, indicating that the highlighted sentences were mostly considered topically related to the query using the TF-IDF, BM25 or BioBERT models. However, it is worth noting that the responses were not entirely unanimous, with some users responding with *partly*, suggesting that there may be room for improvement in accurately identifying and extracting the most relevant passages. Ultimately, user responses offer valuable insights that can guide future improvements to the proposed model, in particular when analyzing the replies to the next questions.

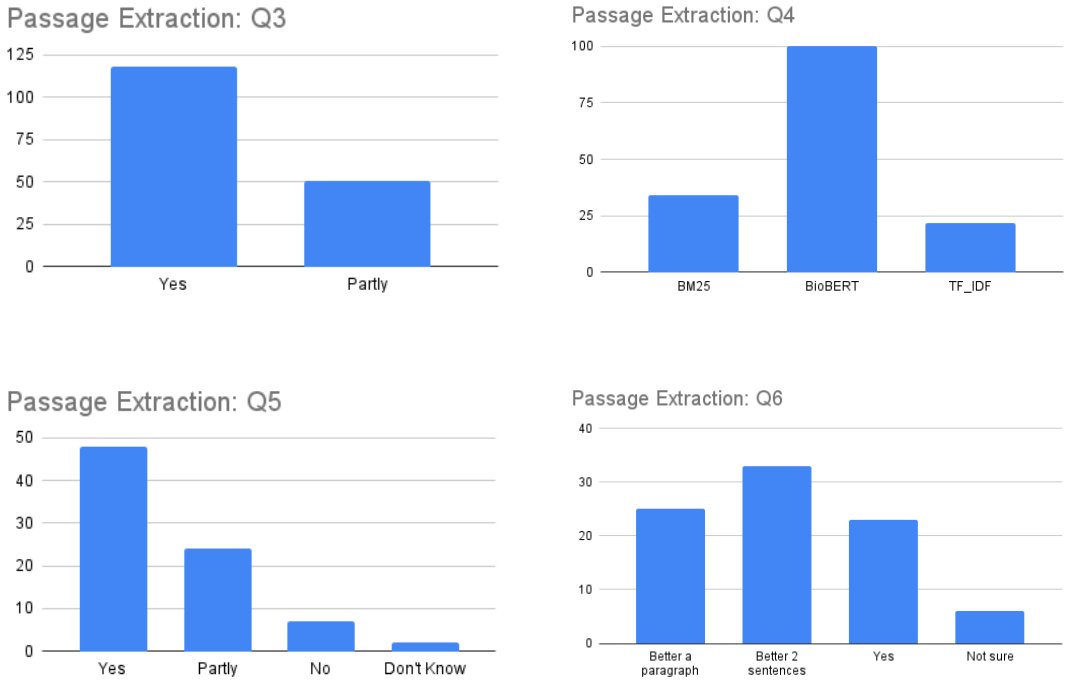

Figure 8.7: Outcome of the questions related to query-relevant passage extraction.

In response to question Q4, which asks users to identify the best algorithm for topicality-based passage extraction between TF-IDF, BM25, and BioBERT, the majority of users choose BioBERT. This is illustrated in Figure 8.7. However, there is a non-negligible number of users who chose the other algorithms, suggesting also in this case some perception differences among assessors, maybe due to specific queries and/or documents.

Regarding question Q5, responses were mixed, with the maximum number of respondents answering affirmatively (*yes* or *partly*) and some answering negatively (*no* or *don't know*). This suggests that while the majority of users found the highlighted passages informative, others still did not consider them sufficient as meaningful sentences from which to determine the document's truthfulness. It is important to note that this question is complex, as it involves not only the (topical) relevance of the highlighted passages to the query, but also their ability to provide starting points for identifying evidence for or against the document's truthfulness.

Finally, when considering question Q6, which asks whether singling out single sentences as passages are sufficient to capture both topicality and truthfulness aspects of the document, the answers are quite varied. Users who believe that a single sentence is sufficient to capture both aspects are a minority. In general, most believe that a better approach would be to consider a passage consisting of more text, such as *two sentences* or *a paragraph*. This highlights the importance of considering such feedback in order to take into account a different granularity of text passages presented to users in the future.

Q7–Q9. *Evidence Extraction* The replies associated with question Q7, summarized in Figure 8.8, generally indicate agreement among participants. However, while most respondents answered *yes*, indicating that the highlighted passages were supported by scientific evidence, a still significant number of respondents answered *partly*. This may suggest that not all highlighted passages are indeed fully supported by scientific evidence and/or that there may be a mix of both fully and partly supported passages.

Similarly, for question Q8, the majority of respondents answered *yes* indicating that the scientific evidence provided sufficient information to understand the document's truthfulness, while a significant number of respondents answered *partly*, suggesting that not all of the extracted evidence was fully helpful in determining the document's truthfulness. Some respondents also answered *don't know*. The responses suggest that scientific evidence can globally play a crucial role in supporting the explainability of the truthfulness of a document, even if for a small number of respondents this is not fully sufficient.

Finally, for question Q9, respondents expressed different opinions and uncertainties about the reliability of the sources associated with each piece of evidence. While a large percentage of respondents answered *yes*, indicating that they thought the sources were reliable, there were also some *partly* or *don't know* responses, suggesting uncertainty or lack of information on the part of the respondent. This may be due to the respondents' lack of health literacy to confidently assess the reliability of sources or the complexity or ambiguity of the question.

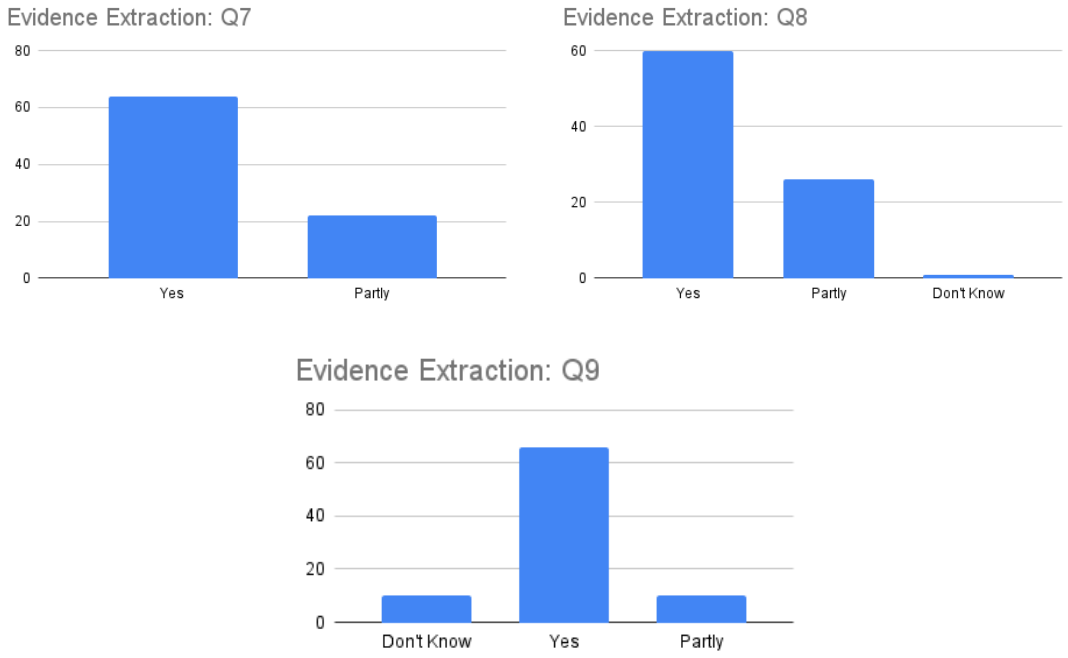

Figure 8.8: Outcome of the questions related to passage-based evidence extraction.

8.3 Summary and Outlook

In this chapter, we have presented a new approach to add explainability to the search results in the context of CHS, particularly regarding the truthfulness of the information. In particular, to provide truthful information with explanations (evidences) for those retrieved results.

To carry out the extraction of (topically) relevant passages from documents and corresponding scientific evidence from scientific articles, we used various textual retrieval and representation techniques, with and without the aid of Named Entity Recognition (NER) techniques to consider the specificity of certain entities in the health domain. The proposed solution was evaluated both from a quantitative and qualitative point of view. The latter evaluation took place, in particular, by means of a user study, in which users were asked to perform tasks and answer a questionnaire.

Through this questionnaire, we were able to obtain valuable information from the assessors regarding their perception of the explainability of the results obtained. In particular, with respect to the ranking obtained and the effectiveness of the relevance dimensions, the extraction of textual passages from documents and scientific evidence from scientific articles and their usefulness in explaining why a document found was actually judged as satisfactory by the majority of respondents. We analyzed responses using the *Fleiss' kappa score* to assess the inter-rater reliability of the questionnaire and found that the level of agreement among raters was generally high.

However, the results of our user survey also revealed some limitations and room for improvement with respect to the proposed solution. For example, it appears that identifying textual passages in the form of single sentence explanations may not always be sufficient to provide a good starting point for assessing both the topical relevance and truthfulness of a document; some psychological factors of the users or other factors related to the dataset should be further investigated for a better understanding of the actual impact of the different dimensions of relevance; however, some assessors found it difficult to estimate the reliability of the information sources (and this is a problem closely related to health literacy); moreover, the quantitative evaluation has given encouraging albeit not excellent results.

In addressing the limitations highlighted in our user survey, it's important to consider the methodological aspects of the study that may influence the outcomes. Notably, the tasks were not rotated among the users, which could have contributed to a bias in the results. Task rotation helps ensure that any learning effects are evenly distributed across the different conditions of the study, thus providing a more balanced and reliable insight into user performance and preferences. Moreover, the fact that model names were revealed to users introduces a significant bias, particularly given the subject cohort's potential preconceived notions or preferences towards certain models. This transparency might inadvertently influence their perceptions and interactions with the system, skewing the results. As we consider these methodological improvements, it is also essential to delve deeper into understanding the psychological factors and dataset-related characteristics that may affect user interaction and evaluation of the systems, thereby refining our approach to assessing the impact of different dimensions of relevance and the overall system effectiveness.

PART IV

A Final Overview

Chapter 9

Discussion and Conclusions

9.1 Discussion

This thesis has embarked on an expedition to detect, retrieve, and explain truthful online health information. The driving force behind this investigation has been the immense necessity for users to have improved access to truthful health information, especially within the purview of Consumer Health Search. It is an undeniable fact that in today's digital age, misinformation runs rampant, and the health domain is not exempt from this malaise. The ramifications of this problem are not just academic in nature but have real-world consequences that affect lives. As such, the principal research question this work seeks to address is: *How can we tackle the health misinformation problem by designing algorithms and search engines to ensure access to both relevant and truthful health information? Additionally, how can we make users understand the veracity of the results they receive?*

R1: How can we effectively amalgamate structural and context-aware methodologies to boost the accuracy of misinformation detection in health-related documents?: Reflecting upon our research trajectory, Chapter 3 and Chapter 4 delved deep into the first sub-question (Chapter 1, R1), probing the potential of amalgamating structural and context-aware techniques to detect misinformation in health-related documents. The findings suggested that a nuanced blend of these methodologies augmented the accuracy of detection. Relying solely on textual content or on structural elements might lead to oversights. However, when combined, they create a robust mechanism that comprehensively processes documents, ensuring that even subtle cues of misinformation are identified.

R2: Can we develop an unsupervised model that accurately evaluates the truthfulness of information in health-related documents?: Navigating to the second sub-question (Chapter 1, R2), Chapter 5 embarked on a quest to discern the feasibility of an unsupervised model that could evaluate the truthfulness of health-related content. Traditional supervised approaches often demand vast labeled datasets, not always available in specialized domains like health. The results from this chapter illustrated that unsupervised models, when designed meticulously, can indeed be powerful tools to gauge truthfulness without explicit human annotation. The success of this approach underlines the importance of leveraging inherent patterns within the data to make informed decisions about its truthfulness.

Can we enhance the effectiveness of retrieval of truthful health information by focusing on document summaries and query-relevant document passages rather than employing full-text?: The journey then transitioned to Chapter 6 and Chapter 7, answering the third sub-question (Chapter 1, R3). The traditional paradigm of relying on full-text retrieval was challenged, positing that more concise, contextually apt representations—like document summaries or specific passages—could offer a more effective retrieval of truthful health information. By homing in on key passages or summaries of documents, the chances of retrieving misleading or tangential information were reduced. It's an approach that respects the user's time while ensuring they obtain topically relevant as well as truthful content.

What methodologies can be employed to increase the explainability of automated systems, ensuring they provide a clear rationale for the truthfulness of health-related content?: Finally, Chapter 8 addressed the pressing issue encapsulated in the fourth sub-question (Chapter 1, R4): How do we ensure that our automated systems do not operate as enigmatic black boxes? The quest for truthfulness in health-related content is not just about retrieval; it's equally pivotal that users understand why certain content is deemed truthful. The methodologies developed in this chapter focus on increasing the explainability of our systems. By providing clear rationales and context for their decisions, these systems not only gain user trust but also facilitate an environment where users can critically engage with the content and understand its truthfulness.

9.2 Conclusion

The digitized age has ushered in a huge amount of information, available at the fingertips of countless individuals. This rise in digital content, particularly in the health domain, while empowering, has also paved the way for a rampant proliferation of misinformation. Such misleading content, when concerned with health, can bear dire consequences, from exacerbating conditions to risking lives. With this in mind, my doctoral dissertation embarked on a comprehensive exploration into effective models and methodologies to detect and counter health misinformation online and finally also try to add explanations for them.

In Chapter 3, our venture commenced with a focused examination of health misinformation detection. By leveraging an enhanced Web2Vec model, the novel approach blended structural-, content-, and context-aware strategies. The model's design aimed to unravel and comprehend the unique nuances linked to the truthfulness of health data on the web. The findings were unequivocal—our proposed model displayed superior effectiveness over other contemporary techniques. However, like all models, it was not devoid of its constraints, prompting the need for an advanced iteration.

Our research progression led to Vec4Cred in Chapter 4. Stemming from its predecessor's foundation and inspired by the earlier works, Vec4Cred incorporates the zenith of health misinformation detection. By incorporating multi-layered web page attribute representations and focusing on grammatical constructs and embedded content, Vec4Cred showcased a heightened ability to decipher domain-specific semantic nuances. The implications of this advanced model were profound, suggesting future potential in harnessing advanced contextual embedding methods and deeper linguistic analysis.

Chapter 5 pivoted the discourse towards "information truthfulness," a concept underscoring the need to combine topical relevance with truthfulness. Our proposed retrieval model was pioneering—eschewing dependencies on expert intervention, and instead, paralleling online narratives with scholarly articles in an unsupervised manner. But our quest did not halt here; it foreshadowed a move towards an evolved IR model that would amplify truthfulness assessment in specific document summaries or passages.

The theme of multidimensional relevance took center stage in Chapter 6. This chapter proposed an innovative method to maintain a fine balance between retrieval effectiveness and efficiency. A re-ranking solution emerged as a beacon in this quest, with strategic aggregation schemes ensuring the integration of varied relevance scores. The relationship observed between efficiency and effectiveness highlighted the intricate dynamics in play, hinting at vast prospects for future research.

Chapter 7 refined this concept further. Traditional IR methodologies, albeit powerful, showcased limitations, especially in multi-dimensional relevance assessment. Responding to these challenges, our innovative Transformer-based re-ranking model was introduced. This model seamlessly merged Passage Retrieval techniques with the essence of traditional IR, extracting salient passages of documents that mirrored both relevance and truthfulness. Empirical data stood a testament to the model's prowess, as it outstripped many of its contemporaries.

Finally, Chapter 8 ventured into a domain of utmost importance—explainability. In the complex realm of CHS, it is not just about providing truthful information; it's about furnishing this information with clear explanations. Our novel solution integrated diverse textual retrieval techniques, enhanced with NER, to extract pertinent passages from documents and present corroborating evidence from the scientific literature. Through comprehensive evaluation, particularly user studies, this approach received favorable feedback. However, challenges remained, particularly in fully addressing the different dimensions of relevance, understanding underlying user psychology, and enhancing information source reliability.

To conclude, this dissertation has been a manifestation of the journey traversed in the vast domain of health misinformation detection, information truthfulness, and explainability. From pioneering models to intricate re-ranking solutions, from information retrieval to explainability, every chapter, and every study has been a step towards ensuring that users are furnished with health information that's not just relevant, but also truthful. While significant strides have been made, the dynamic nature of the digital domain suggests that this is but the dawn. Misinformation remains a formidable adversary, ever-evolving, and adapting. Our endeavors have laid a robust foundation, and it's upon this bedrock that future researchers and scientists can build, innovate, and fortify the bastion of truth in the digital health realm.

Bibliography

Ahmed Abbasi, Fatemeh Mariam Zahedi, and Siddharth Kaza. Detecting fake medical web sites using recursive trust labeling. *ACM Transactions on Information Systems*, 30(4), 2012. ISSN 10468188. 10.1145/2382438.2382441.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*, 2020.

Mustafa Abualsaud, Irene Xiangyi Chen, Kamyar Ghajar, Linh Nhi Phan Minh, Mark D Smucker, Amir Vakili Tahami, and Dake Zhang. Uwaterloomds at the trec 2021 health misinformation track. In *Proceedings of the Thirtieth REtrieval Conference Proceedings (TREC 2021). National Institute of Standards and Technology (NIST), Special Publication*, pages 1–18, 2021.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

SA Adams. *Under Construction: Reviewing and Producing Information Reliability on the Web*. PhD thesis, Erasmus School of Health Policy & Management (ESHPM), 2006.

Ratnadip Adhikari and RK Agrawal. Performance evaluation of weights selection schemes for linear combination of multiple forecasts. *Artificial Intelligence Review*, 42(4):529–548, 2014.

Fariha Afsana, Muhammad Ashad Kabir, Naeemul Hassan, and Manoranjan Paul. Automatically Assessing Quality of Online Health Articles. *arXiv*, (August):1–12, 2020. ISSN 2168-2194. 10.1109/jbhi.2020.3032479.

Shashank M Akerkar, M Kanitkar, LS Bichile, et al. Use of the internet as a resource of health information by patients: a clinic-based study in the indian population. *Journal of Postgraduate Medicine*, 51(2):116, 2005.

Majed M. Al-Jefri, Roger Evans, Pietro Ghezzi, and Gulden Uchyigit. Using machine learning for automatic identification of evidence-based health information on the Web. In *ACM International Conference Proceeding Series*, volume Part F1286, pages 167–174, 2017. ISBN 9781450352499. 10.1145/3079452.3079470.

Hend Al-Khalifa and Mohammed Binsultan. An experimental system for measuring the credibility of news content in Twitter. *IJWIS*, 7:130–151, 2011. 10.1108/17440081111141772.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

M. W. Alsem, F. Ausems, M. Verhoef, M. J. Jongmans, J. M.A. Meily-Visser, and M. Ketelaar. Information seeking by parents of children with physical disabilities: An exploratory qualitative study. *Research in Developmental Disabilities*, 60:125–134, 2017. ISSN 18733379. 10.1016/j.ridd.2016.11.015. URL `http://dx.doi.org/10.1016/j.ridd.2016.11.015`.

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*, 2022.

Jackie Ayoub, X Jessie Yang, and Feng Zhou. Combat covid-19 infodemic using explainable natural language processing models. *Information processing & management*, 58(4): 102569, 2021.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

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

Salwa Bahkali, Reem Almaiman, Mamoun El-Awad, Huda Almohanna, Khaled Al-Surimi, and Mowafa Househ. Exploring the Impact of Information Seeking Behaviors of Online Health Consumers in the Arab World. *Studies in health technology and informatics*, 226: 279–282, 2016.

James E Bailey and Sammy W Pearson. Development of a tool for measuring and analyzing computer user satisfaction. *Management science*, 29(5):530–545, 1983.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, and Mingze Li. Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119, 2020.

Susanne E Baumgartner and Tilo Hartmann. The role of health anxiety in online health information search. *Cyberpsychology, behavior, and social networking*, 14(10):613–618, 2011.

Rahime Belen Salam and Tugba Taskaya Temizel. A framework for automatic information quality ranking of diabetes websites. *Informatics for Health and Social Care*, 40(1):45–66, 2015. ISSN 17538165. 10.3109/17538157.2013.872109.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Gretchen K Berland, Marc N Elliott, Leo S Morales, Jeffrey I Algazy, Richard L Kravitz, Michael S Broder, David E Kanouse, Jorge A Muñoz, Juan-Antonio Puyol, Marielena Lara, et al. Health information on the internet: accessibility, quality, and readability in english and spanish. *jama*, 285(20):2612–2621, 2001.

Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. World-Wide Web: The Information Universe. *Internet Research*, 20:461–471, 2010. 10.1108/10662241011059471.

Elisa Bertino and Hyo-Sang Lim. Assuring data trustworthiness-concepts and research challenges. In *Workshop on Secure Data Management*, pages 1–12. Springer, 2010.

Janek Bevendor, Alexander Bondarenko, Maik Fröbe, S. Günther, Michael Völske, Benno Stein, and Matthias Hagen. Webis at trec 2020: Health misinformation track extended abstract. In *TREC*, 2020.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic chatnoir: Search engine for the clueweb and the common crawl. In *ECIR*, 2018.

Adella Bhaskara, Michael Skinner, and Shayne Loft. Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, 50(3):215–224, 2020.

Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. Comprehend medical: a named entity recognition and relationship extraction web service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE, 2019.

Aida Bianco, Rossella Zucco, Carmelo Giuseppe A Nobile, Claudia Pileggi, Maria Pavia, et al. Parents seeking health-related information on the internet: cross-sectional study. *Journal of medical Internet research*, 15(9):e2752, 2013.

Jens Christian Bjerring and Jacob Busch. Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*, 34:349–371, 2021.

Colin R Blyth and Harold A Still. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.

Alexander Bondarenko, Maik Fröbe, Marcel Gohsen, Sebastian Günther, Johannes Kiesel, Jakob Schwerter, Shahbaz Syed, Michael Völske, Martin Potthast, Benno Stein, et al. Webis at trec 2021: deep learning, health misinformation, and podcasts tracks. In *The Thirtieth REtrieval Conference Proceedings (TREC 2021). National Institute of Standards and Technology (NIST), Special Publication*, pages 500–335, 2021.

Célia Boyer and Ljiljana Dolamic. Automated detection of HONcode website conformity compared to manual detection: An evaluation. *Journal of Medical Internet Research*, 17 (6):e135, 2015. ISSN 14388871. 10.2196/jmir.3831.

Rowena Briones. Harnessing the Web: How E-Health and E-Health Literacy Impact Young Adults' Perceptions of Online Health Information. *Medicine 2.0*, 4(2):e5, 2015. ISSN 1923-2195. 10.2196/med20.4327.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164, 2009.

Sara Champlin, Michael Mackert, Elizabeth M. Glowacki, and Erin E. Donovan. Toward a Better Understanding of Patient Health Literacy: A Focus on the Skills Patients Need to Find Health Information. *Qualitative Health Research*, 27(8):1160–1176, 2017. ISSN 15527557. 10.1177/1049732316646355.

Yung Sheng Chang, Yan Zhang, and Jacek Gwizdka. The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior. *Computers in Human Behavior*, 115(July 2020):106629, 2021. ISSN 07475632. 10.1016/j.chb.2020.106629. URL `https://doi.org/10.1016/j.chb.2020.106629`.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

Pride Chigwedere, George R Seage III, Sofia Gruskin, Tun-Hou Lee, and Max Essex. Estimating the lost benefits of antiretroviral drug use in south africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 49(4):410–415, 2008.

Wonchan Choi. Older adults' credibility assessment of online health information: An exploratory study using an extended typology of web credibility. *Journal of the Association for Information Science and Technology*, 71(11):1295–1307, 2020. ISSN 23301643. 10 .1002/asi.24341.

Wonchan Choi and Besiki Stvilia. Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology*, 66(12):2399–2414, 2015.

Wen-Ying Sylvia Chou, April Oh, and William MP Klein. Addressing health-related misinformation on social media. *Jama*, 320(23):2417–2418, 2018.

Anshika Choudhary and Anuja Arora. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171, 2021.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Charles LA Clarke, Saira Rizvi, Mark D Smucker, Maria Maistro, and Guido Zuccon. Overview of the trec 2020 health misinformation track. In *TREC*, 2020.

Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. Assessing top-preferences. *ACM Transactions on Information Systems (TOIS)*, 39(3):1–21, 2021.

Cinzia Colombo, Paola Mosconi, Paolo Confalonieri, Isabella Baroni, Silvia Traversa, Sophie J. Hill, Anneliese J. Synnot, Nadia Oprandi, and Graziella Filippini. Web search behavior and information needs of people with multiple sclerosis: Focus group study and analysis of online postings. *Journal of Medical Internet Research*, 16(7), 2014. ISSN 14388871. 10.2196/ijmr.3034.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.

Limeng Cui, Suhang Wang, and Dongwon Lee. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 41–48, 2019.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 492–502, 2020. 10.1145/3394486.3403092.

Anna Cunningham and Frances Johnson. Exploring trust in online health information: a study of user experiences of patients.co.uk. *Health Information and Libraries Journal*, 33(4): 323–328, 2016. ISSN 14711842. 10.1111/hir.12163.

Leila Cusack, Laura N. Desha, Chris B. Del Mar, and Tammy C. Hoffmann. A qualitative study exploring high school students' understanding of, and attitudes towards, health information and claims. *Health Expectations*, 20(5):1163–1171, 2017. ISSN 13697625. 10.1111/hex.12562.

Baivab Das and S Jaya Nirmala. Improving healthcare question answering system by identifying suitable answers. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, pages 1–6. IEEE, 2022.

Khishigsuren Davagdorj, Jong Seol Lee, Van Huy Pham, and Keun Ho Ryu. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. *Applied Sciences*, 10(9):3307, 2020.

Rina Dechter. Learning while searching in constraint-satisfaction problems. 1986.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation. In *arXiv*, 2020a. URL `http://arxiv.org/abs/2010.08743`.

Arkin Dharawat et al. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID-19 misinformation. *arXiv preprint arXiv:2010.08743*, 2020b.

Sameer Dhoju, Muhammad Ashad Kabir, Md Main Uddin Rony, and Naeemul Hassan. Differences in health news from reliable and unreliable media. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, (May):981–987, 2019. 10.1145/3308560.3316741.

Stefano Di Sotto and Marco Viviani. Health misinformation detection in the social web: An overview and a data science approach. *International Journal of Environmental Research and Public Health*, 19(4):1–20, 2022. ISSN 1660-4601. 10.3390/ijerph19042173. URL `https://www.mdpi.com/1660-4601/19/4/2173`.

Nicola Diviani, Bas van den Putte, Stefano Giani, and Julia CM van Weert. Low health literacy and evaluation of online health information: a systematic review of the literature. *Journal of medical Internet research*, 17(5):e112, 2015.

Nicola Diviani, Bas van den Putte, Corine S. Meppelink, and Julia C.M. van Weert. Exploring the role of health literacy in the evaluation of online health information: Insights from a mixed-methods study. *Patient Education and Counseling*, 99(6):1017–1025, 2016. ISSN 18735134. 10.1016/j.pec.2016.01.007. URL `http://dx.doi.org/10.1016/j.pec.2016.01.007`.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and S Yu Philip. Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE, 2021.

Sulemana Bankuoru Egala, Decui Liang, and Dorcas Boateng. Social media health-related information credibility and reliability: An integrated user perceived quality assessment. *IEEE Transactions on Engineering Management*, 2022.

Ibrahim K El Mikati, Reem Hoteit, Tarek Harb, Ola El Zein, Thomas Piggott, Jad Melki, Reem A Mustafa, and Elie A Akl. Defining misinformation and related terms in health-related literature: Scoping review. *Journal of Medical Internet Research*, 25:e45731, 2023.

Martin J Eppler. Qualitätsstandards—ein instrument zur sicherung der informationsqualität in multimedia-produktionen. In *Qualitätssicherung bei Multimedia-Projekten*, pages 129–149. Springer, 1999.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

Gunther Eysenbach. MedCERTAIN/MedCIRCLE: Using Semantic Web Technologies for Quality Management of Health Information on the Web. pages 217–226, 2005. 10.1007/0-387-27652-1_18.

Gunther Eysenbach. *Credibility of health information and digital media: New perspectives and implications for youth*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

Don Fallis. A conceptual analysis of disinformation. 2009.

Yang Fan, Liu Gongshen, Meng Kui, and Sun Zhaoying. Neural feedback text clustering with bilstm-cnn-kmeans. *IEEE Access*, 6:57460–57469, 2018. 10.1109/ACCESS.2018.2873327.

Cobb Payton Fay, Kvasny Lynette, and James Kiwanuka-Tondo. Online HIV prevention information. *Managerial Auditing Journal*, 28(2):2–3, 2014.

Jian Feng, Lianyang Zou, Ou Ye, and Jingzhou Han. Web2vec: Phishing webpage detection method based on multidimensional features driven by deep learning. *IEEE Access*, 8:221214–221224, 2020. 10.1109/ACCESS.2020.3043188.

Marcos Fernández-Pichel, D. Losada, J. C. Pichel, and David Elsweiler. Citius at the trec 2020 health misinformation track. In *TREC*, 2020a.

Marcos Fernández-Pichel, David E Losada, Juan Carlos Pichel, and David Elsweiler. Citius at the trec 2020 health misinformation track. In *TREC*, 2020b.

Marcos Fernández-Pichel, David Losada, Juan C Pichel, and David Elsweiler. Reliability prediction for health-related content: A replicability study. In *European Conference on Information Retrieval, Lucca, Tuscany, Italy*, 2021.

Marcos Fernández-Pichel, Selina Meyer, Markus Bink, Alexander Frummet, David E Losada, and David Elsweiler. Improving the reliability of health information credibility assessments. In *Proceedings of ROMCIR 2023, European Conference on Information Retrieval*, 2023.

Emilio Ferrara. The history of digital spam. *Communications of the ACM*, 62(8):82–91, 2019.

Markus A. Feufel and S. Frederica Stahl. What do web-use skill differences imply for online health information searches? *Journal of Medical Internet Research*, 14(3):1–11, 2012. ISSN 14388871. 10.2196/jmir.2051.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

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

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

Susannah Fox and Maeve Duggan. Health online 2013. *Health*, 2013:1–55, 2013.

Duggan M. Pew Research Center Fox S. Health online 2013 url, 2013. URL `http://www.pewinternet.org/2013/01/15/health-online-2013/`.

Kris S. Freeman and Jan H. Spyridakis. Effect of contact information on the credibility of online health information. *IEEE Transactions on Professional Communication*, 52(2):152–166, 2009. ISSN 03611434. 10.1109/TPC.2009.2017992.

Krisandra S Freeman and Jan H Spyridakis. An examination of factors that affect the credibility of online health information. *Technical communication*, 51(2):239–263, 2004.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Arnaud Gaudinat, Natalia Grabar, and Célia Boyer. Machine learning approach for automatic quality criteria detection of health web pages. *Studies in Health Technology and Informatics*, 129:705–709, 2007. ISSN 18798365.

Arnaud Gaudinat, Sarah Cruchet, Pravir Chrawdhry, and Celia Boyer. Enriching trustworthiness of Health Web Pages through the Semantic Web. 2010.

Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

Amira Ghenai and Yelena Mejova. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, page 518, 2017. ISBN 9781509048816. 10.1109/ICHI.2017.58.

Anna Glazkova et al. g2tmn at Constraint@ AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. *arXiv preprint arXiv:2012.11967*, 2020.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Gonzalez Saez, Marco Viviani, and Chenchen Xu. Overview of the clef ehealth evaluation lab 2020. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–271. Springer, 2020.

Lorraine Goeuriot, G Pasi, H Suominen, Elias Bassani, Nicola Brew-Sam, Gabriela González-Sáez, RG Upadhyay, L Kelly, P Mulhem, S Seneviratne, et al. Consumer health search at clef ehealth 2021. In *CLEF 2021 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS*, Romania, 2021. Springer.

Simão N. Gonçalves and Flávio Martins. Voh.colab at trec 2020 health misinformation track. In *TREC*, 2020.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer, 2014.

Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining (adaptive computation and machine learning)*. MIT Press, 2001.

Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, 12(2):133–139, 1995.

Bradford W. Hesse, David E. Nelson, Gary L. Kreps, Robert T. Croyle, Neeraj K. Arora, Barbara K. Rimer, and Kasisomayajula Viswanath. Trust and sources of health information. The impact of the internet and its implications for health care providers: Findings from the first health information national trends survey. *Archives of Internal Medicine*, 165(22):2618–2624, 2005. ISSN 00039926. 10.1001/archinte.165.22.2618.

Tao Hoang, Jixue Liu, Nicole Pratt, Vincent W. Zheng, Kevin C. Chang, Elizabeth Roughead, and Jiuyong Li. Authenticity and credibility aware detection of adverse drug events from social media. *International Journal of Medical Informatics*, 120(January 2017):157–171, 2018. ISSN 18728243. 10.1016/j.ijmedinf.2018.10.003. URL `https://doi.org/10.1016/j.ijmedinf.2018.10.003`.

Laurie Hoffman-Goetz and Daniela B. Friedman. A qualitative study of Canadian aboriginal women's beliefs about "credible" cancer information on the internet. *Journal of Cancer Education*, 22(2):124–128, 2007. ISSN 08858195. 10.1007/BF03174361.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

Traci Hong et al. The influence of structural and message features on web site credibility. *Journal of the American Society for Information Science and Technology*, 57(1):114–127, 2006.

Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. Covidlies: Detecting covid-19 misinformation on social media. In *Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. Communication and persuasion. 1953.

Xinyi Hu, Robert A Bell, Richard L Kravitz, and Sharon Orrange. The prepared patient: information seeking of online support group members before their medical appointments. *Journal of health communication*, 17(8):960–978, 2012.

Kuan-Tse Huang, Yang W Lee, and Richard Y Wang. *Quality information and knowledge*. Prentice Hall PTR, 1998.

Rafia Inam, Ahmad Terra, Anusha Mujumdar, Elena Fersman, and Aneta Vulgarakis. Explainable ai – how humans can trust ai. 2021.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10:1–20, 2020a. ISSN 18695469. 10.1007/s13278-020-00696-x. URL `https://doi.org/10.1007/s13278-020-00696-x`.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20, 2020b.

Beverly K Kahn. Product and service performance model for information quality: an update. In *Proc. 1998 International Conference on Information Quality*. MIT, 1998.

Jatin Kaicker, Victoria B. Debono, Wilfred Dang, Norman Buckley, and Lehana Thabane. Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. *BMC Medicine*, 8, 2010. ISSN 17417015. 10.1186/1741-7015-8-59.

Hema Karande, Rahee Walambe, Victor Benjamin, Ketan Kotecha, and TS Raghu. Stance detection with bert embeddings for credibility analysis of information on social media. *PeerJ Computer Science*, 7:e467, 2021.

Kari Kelton, Kenneth R Fleischmann, and William A Wallace. Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3):363–374, 2008.

Cicely Kerr, Elizabeth Murray, Fiona Stevenson, Charles Gore, and Irwin Nazareth. Internet interventions for long-term conditions: Patient and caregiver quality criteria. *Journal of Medical Internet Research*, 8(3):1–12, 2006. ISSN 14388871. 10.2196/jmir.8.3.e13.

Alla Keselman, Allen C Browne, and David R Kaufman. Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association*, 15(4): 484–495, 2008.

Nikhil Ketkar. Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer, 2017.

Carolyn Mae Kim and William J Brown. Conceptualizing credibility in social media spaces of public relations. *Public Relations Journal*, 9(4):1–17, 2015.

Hye Kyung Kim, Jisoo Ahn, Lucy Atkinson, and Lee Ann Kahlor. Effects of covid-19 misinformation on information seeking, avoidance, and processing: A multicountry comparative study. *Science Communication*, 42(5):586–615, 2020.

Hyojin Kim, Sun Young Park, and Ingrid Bozeman. Online health information search and evaluation: Observations and semi-structured interviews with college students and maternal health experts. *Health Information and Libraries Journal*, 28(3):188–199, 2011. ISSN 14711834. 10.1111/j.1471-1842.2011.00948.x.

Laura Kinkead, Ahmed Allam, and Michael Krauthammer. Autodiscern: Rating the quality of online health information with hierarchical encoder attention-based neural networks. *arXiv*, pages 1–13, 2019.

Erin Klawitter and Eszter Hargittai. Shortcuts to Well Being? Evaluating the Credibility of Online Health Information through Multiple Complementary Heuristics. *Journal of Broadcasting and Electronic Media*, 62(2):251–268, 2018. ISSN 15506878. 10.1080/08838151.2018.1451863. URL `https://doi.org/10.1080/08838151.2018.1451863`.

Shirlee-ann Knight and Janice Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8, 2005.

Stathis Th Konstantinidis, Per Egil Kummervold, Luis Fernandez Luque, and Lars Kristian Vognild. A proposed framework to enrich norwegian EHR system with health-trusted information for patients and professionals. *Studies in Health Technology and Informatics*, 213(July):149–152, 2015. ISSN 18798365. 10.3233/978-1-61499-538-8-149.

Ziyi Kou, Daniel Yue Zhang, Lanyu Shang, and Dong Wang. Exfaux: A weakly-supervised approach to explainable fauxtography detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 631–636. IEEE, 2020.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25, 2022.

Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.

Chee Ping Lai, Wan Tze Vong, and Patrick H.H. Then. A patient-centric framework for multisourced actionable health solution. *2012 International Conference on Biomedical Engineering, ICoBE 2012*, (February):573–578, 2012. 10.1109/ICoBE.2012.6178982.

Chul-Joo Lee. Does the internet displace health professionals? *Journal of health communication*, 13(5):450–464, 2008.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019. ISSN 1367-4803. 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Joon Ho Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–276, USA, 1997. ACM.

Humbert Lesca, Elisabeth Lesca, Nicolas Lesca, and Marie-Laurence Caron-Fasan. *Gestion de l'information: Qualité de l'information et performances de l'entreprise*. EMS Editions, 2010.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Jinhui Li, Yin-Leng Theng, and Schubert Foo. Predictors of online health information seeking behavior: Changes between 2002 and 2012. *Health informatics journal*, 22(4):804–814, 2016.

Minghan Li and Eric Gaussier. Keybld: Selecting key blocks with local pre-ranking for long document information retrieval. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2207–2211, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379.

Qingzi Vera Liao. Effects of cognitive aging on credibility assessment of online health information. *Conference on Human Factors in Computing Systems - Proceedings*, pages 4321–4326, 2010. 10.1145/1753846.1754147.

Lucas Chaves Lima, Dustin Wright, Isabelle Augenstein, and Maria Maistro. University of copenhagen participation in trec health misinformation track 2020. *ArXiv*, abs/2103.02462, 2020.

Lucas Chaves Lima, Dustin Brandon Wright, Isabelle Augenstein, and Maria Maistro. University of copenhagen participation in trec health misinformation track 2020. In *TREC*, Maryland, USA, 2021. NIST.

Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 91–98, Japan, 2017. ACM.

Jiaying Liu, Shijie Song, and Yan Zhang. Linguistic features and consumer credibility judgment of online health information. *University of Illinois*, 2021a.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. Med-bert: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, 18(8):5600–5608, 2021b.

Rey Long Liu. Automatic quality measurement for health information on the internet. *International Journal of Intelligent Information and Database Systems*, 8(4):340–358, 2014. ISSN 17515866. 10.1504/IJIIDS.2014.068340.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019a.

Yue Liu, Ke Yu, Xiaofei Wu, Linbo Qing, and Yonghong Peng. Analysis and detection of health-related misinformation on Chinese social media. *IEEE Access*, 7:154480–154489, 2019b. ISSN 21693536. 10.1109/ACCESS.2019.2946624.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

DM Lopez, Bernd Blobel, and C Gonzalez. Information quality in healthcare social media–an architectural approach. *Health and Technology*, 6(1):17–25, 2016.

Jennifer S Love, Adam Blumenberg, and Zane Horowitz. The parallel pandemic: Medical misinformation and covid-19. *Journal of General Internal Medicine*, 35(8):2435–2436, 2020.

Yi-Ju Lu and Cheng-Te Li. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Craig Macdonald and Nicola Tonellotto. Declarative experimentation in information retrieval using pyterrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 161–168, China, 2020. ACM.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In *Proceedings of the 30th ACM International Conference on Information Knowledge Management*, CIKM '21, page 4526–4533, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. 10.1145/3459637.3482013. URL `https://doi.org/10.1145/3459637.3482013`.

Michael MacKert, Leeann Kahlor, Diane Tyler, and Jamie Gustafson. Designing e-health interventions for low-health-literate culturally diverse parents: Addressing the obesity epidemic. *Telemedicine and e-Health*, 15(7):672–677, 2009. ISSN 15563669. 10.1089/tmj.2009.0012.

Ali Maki, Roger Evans, and Pietro Ghezzi. Bad news: analysis of the quality of information on influenza prevention returned by google in english and italian. *Frontiers in immunology*, 6:616, 2015.

David M Markowitz and Jeffrey T Hancock. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937, 2014.

Lyndsay A. Marshall and Dorothy Williams. Health information: Does quality count for the consumer?: How consumers evaluate the quality of health information materials across a variety of media. *Journal of Librarianship and Information Science*, 38(3):141–156, 2006. ISSN 09610006. 10.1177/0961000606066575.

Christine Marton. How women with mental health conditions evaluate the quality of information on mental health web sites: A qualitative approach. *Journal of Hospital Librarianship*, 10 (3):235–250, 2010. ISSN 15323277. 10.1080/15323269.2010.491422.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. Fang-covid: A new large-scale benchmark dataset for fake news detection in german. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91, 2021.

Miguel A. Mayer, Pythagoras Karampiperis, Antonis Kukurikos, Vangelis Karkaletsis, Kostas Stamatakis, Dagmar Villarroel, and Angela Leis. Applying Semantic Web technologies to improve the retrieval, credibility and use of health-related web resources. *Health Informatics Journal*, 17(2):95–115, 2011. ISSN 14604582. 10.1177/1460458211405004.

Alexa T McCray, Nicholas C Ide, Russell R Loane, and Tony Tse. Strategies for supporting consumer health information seeking. In *MEDINFO 2004*, pages 1152–1156. IOS Press, 2004.

D Harrison McKnight and Charles J Kacmar. Factors and effects of information credibility. In *Proceedings of the ninth international conference on Electronic commerce*, pages 423–432, 2007.

Amy C. McPherson, Miriam L. Gofine, and Jennifer Stinson. Seeing Is Believing? A Mixed-Methods Study Exploring the Quality and Perceived Trustworthiness of Online Information About Chronic Conditions Aimed at Children and Young People. *Health Communication*, 29(5):473–482, 2014. ISSN 10410236. 10.1080/10410236.2013.768325. URL `http://dx.doi.org/10.1080/10410236.2013.768325https://doi.org/10.1080/10410236.2013.768325`.

Priyanka Meel and Dinesh Kumar Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986, 2020.

Michelle M Mello, Jeremy A Greene, and Joshua M Sharfstein. Attacks on public health officials during covid-19. *Jama*, 324(8):741–742, 2020.

Corine S. Meppelink, Hanneke Hendriks, Damian Trilling, Julia C.M. van Weert, Anqi Shao, and Eline S. Smit. Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Education and Counseling*, (xxxx), 2020a. ISSN 07383991. 10.1016/j.pec.2020.11.013. URL `https://doi.org/10.1016/j.pec.2020.11.013`.

Corine S Meppelink et al. Reliable or not? an automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Education and Counseling*, 2020b.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, 2004. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Julia Moreland, Tara L French, Grant P Cumming, et al. The prevalence of online health information seeking among patients in scotland: a cross-sectional exploratory study. *JMIR research protocols*, 4(3):e4010, 2015.

Luciana Rodrigues Alves da Mota, Carolina Cavalcanti Gonçalves Ferreira, Henrique Augusto Alves da Costa Neto, Ana Rodrigues Falbo, and Suélem de Barros Lorena. Is doctor-patient relationship influenced by health online information? *Revista da Associação Médica Brasileira*, 64:692–699, 2018.

Y. Mrabet, Mourad Sarrouti, Asma Ben Abacha, Soumya Gayen, Travis, Goodwin, Alastair R. Rae, Willie J. Rogers, and Dina Demner-Fushman. Nlm at trec 2020 health misinformation and deep learning tracks. In *TREC*, 2020.

Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on drugs. In *20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 65–74, 2014. ISBN 9781450329569. 10.1145/2623330.2623714.

Van Hoang Nguyen, Kazunari Sugiyama, Min Yen Kan, and Kishaloy Halder. Neural side effect discovery from user credibility and experience-assessed online health discussions. *Journal of Biomedical Semantics*, 11(1):1–16, 2020. ISSN 20411480. 10.1186/s13326-020-00221-1.

Aaron M Norr, Daniel W Capron, and Norman B Schmidt. Medical information seeking: impact on risk for anxiety psychopathology. *Journal of behavior therapy and experimental psychiatry*, 45(3):402–407, 2014.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 45. 2020.

Melinda Oroszlányová, Carla Teixeira Lopes, Sérgio Nunes, and Cristina Ribeiro. Predicting the quality of health web documents using their characteristics. *Online Information Review*, 42(7):1024–1047, 2018. ISSN 14684527. 10.1108/OIR-01-2017-0028.

Meeyoung Park, Hariprasad Sampathkumar, Bo Luo, and Xue Wen Chen. Content-based assessment of the credibility of online healthcare information. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, (Cdc):51–58, 2013. 10.1109/BigData.2013.6691758.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, page 12. Curran Associates Inc., Red Hook, NY, USA, 2019.

Kristen A. Peddie and Rebecca J. Kelly-Campbell. How people with hearing impairment in New Zealand use the Internet to obtain information about their hearing health. *Computers in Human Behavior*, 73:141–151, 2017. ISSN 07475632. 10.1016/j.chb.2017.03.037. URL `http://dx.doi.org/10.1016/j.chb.2017.03.037`.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Qatar, 2014. ACL.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

Marinella Petrocchi and Marco Viviani. ROMCIR 2023: Overview of the 3rd Workshop on Reducing Online Misinformation Through Credible Information Retrieval. In *Proceedings of ROMCIR 2023, European Conference on Information Retrieval*, pages 405–411. Springer, 2023.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical ner and covid-19 qa. *arXiv preprint arXiv:2004.03354*, 2020.

Sayantan Polley, Atin Janki, Marcus Thiel, Juliane Hoebel-Mueller, and Andreas Nuernberger. Exdocs: Evidence based explainable document search. In *ACM SIGIR Workshop on Causality in Search and Recommendation*, 2021.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

John Powell, Nadia Inglis, Jennifer Ronnie, and Shirley Large. The Characteristics and Motivations of Online Health Information Seekers: Cross-Sectional Survey and Qualitative Interview Study. *Journal of medical Internet research*, 13:e20, 2011. 10.2196/jmir.1600.

Ronak Pradeep, Xueguang Ma, Xinyu Zhang, H. Cui, Ruizhou Xu, Rodrigo Nogueira, Jimmy J. Lin, and D. Cheriton. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. In *TREC*, 2020.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2066–2070, 2021.

Divi Galih Prasetyo Putri, Marco Viviani, and Gabriella Pasi. A multi-task learning model for multidimensional relevance assessment. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 103–115, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85251-1.

S Pyysalo, F Ginter, H Moen, T Salakoski, and S Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44. Database Center for Life Science, Japan, 2013.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. Explaining documents' relevance to search queries. *arXiv preprint arXiv:2111.01314*, 2021.

Sivaraman Ramachandramurthy, Srinivasan Subramaniam, and Chandrasekeran Ramasamy. Distilling big data: Refining quality information in the era of yottabytes. *The Scientific World Journal*, 2015, 2015.

TS Sathyanarayana Rao and Chittaranjan Andrade. The mmr vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian journal of psychiatry*, 53(2):95, 2011.

Nisarg Raval and Manisha Verma. One word at a time: adversarial attacks on retrieval models. *arXiv preprint arXiv:2008.02197*, 2020.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China, 2019a. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019b.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Soo Young Rieh and Nicholas J Belkin. Interaction on the web: Scholars' judgment of information quality and cognitive authority. In *Proceedings of the 63rd Annual Meeting of the ASIS*, pages 25–38, 2000.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. Okapi at trec-4. *Nist Special Publication Sp*, pages 73–96, 1996.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10): 201199, 2020.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*, 2021.

David J Rothkopf. When the buzz bites back. *The Washington Post*, 11:B1–B5, 2003.

Victoria Rubin and Tatiana Lukoianova. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1):4, 2013.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Miriam Santer, Ingrid Muller, Lucy Yardley, Hana Burgess, Steven J. Ersser, Sue Lewis-Jones, and Paul Little. 'You don't know which bits to believe': Qualitative study exploring carers' experiences of seeking information on the internet about childhood eczema. *BMJ Open*, 5(4):1–5, 2015. ISSN 20446055. 10.1136/bmjopen-2014-006339.

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

Ipek Baris Schlicht, Angel Felipe Magnossão de Paula, and Paolo Rosso. Upv at trec health misinformation track 2021 ranking with sbert and quality estimators. *arXiv preprint arXiv:2112.06080*, 2021.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1245–1254, New York, NY, USA, 2011a. Association for Computing Machinery. ISBN 9781450302289. 10.1145/1978942.1979127. URL https://doi.org/10.1145/1978942.1979127.

Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1245–1254, 2011b.

Tara Kirk Sell, Divya Hosangadi, Elizabeth Smith, Marc Trotochaud, Prarthana Vasudevan, Gigi Kwik Gronvall, et al. National priorities to combat misinformation and disinformation for covid-19 and future public health threats: A call for a national strategy. baltimore, md: Johns hopkins center for health security; 2021. *URL: https://tinyurl. com/2p86c7d7 [accessed 2022-06-02]*, 2022.

Zubair Shah, Didi Surian, Amalie Dyda, Enrico Coiera, Kenneth D. Mandl, and Adam G. Dunn. Automatically appraising the credibility of vaccine-related web pages shared on social media: A twitter surveillance study. *Journal of Medical Internet Research*, 21(11):1–14, 2019. ISSN 14388871. 10.2196/14007.

Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146: 102551, 2021.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

Andrew Sih. Prey uncertainty and the balancing of antipredator and feeding needs. *The American Naturalist*, 139(5):1052–1069, 1992.

Elizabeth Sillence and Pam Briggs. Please advise: using the Internet for health and financial advice. *Computers in Human Behavior*, 23(1):727–748, 2007. ISSN 07475632. 10.1016/j.chb.2004.11.006.

Elizabeth Sillence, Pam Briggs, Peter Harris, and Lesley Fishwick. Health Websites that people can trust - the case of hypertension. *Interacting with Computers*, 19(1):32–42, 2007a. ISSN 09535438. 10.1016/j.intcom.2006.07.009.

Elizabeth Sillence, Pam Briggs, Peter Richard Harris, and Lesley Fishwick. How do patients evaluate and make use of online health information? *Social Science and Medicine*, 64 (9):1853–1862, 2007b. ISSN 02779536. 10.1016/j.socscimed.2007.01.012.

Elizabeth Sillence, Claire Hardy, Lydia C. Medeiros, and Jeffrey T. LeJeune. Examining trust factors in online food risk information: The case of unpasteurized or 'raw' milk. *Appetite*, 99:200–210, 2016. ISSN 10958304. 10.1016/j.appet.2016.01.010. URL http://dx.doi.org/10.1016/j.appet.2016.01.010.

Jaspreet Singh and Avishek Anand. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 770–773, 2019.

Karmpaul Singh and Richard J Brown. From headache to tumour: An examination of health anxiety, health-related internet use and 'query escalation'. *Journal of Health Psychology*, 21(9):2008–2020, 2016.

Mark D Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of the sixteenth ACM Conf. on Information and Knowledge Management*, pages 623–632, 2007.

Parikshit Sondhi, V. G. Vinod Vydiswaran, and Chengxiang Zhai. Reliability prediction of webpages in the medical domain. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7224 LNCS:219–231, 2012a. ISSN 03029743. 10.1007/978-3-642-28997-2_19.

Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. Reliability prediction of webpages in the medical domain. In *European conference on information retrieval*, pages 219–231. Springer, 2012b.

Shijie Song, Yuxiang Chris Zhao, Xiaokang Song, and Qinghua Zhu. The Role of Health Literacy on Credibility Judgment of Online Health Misinformation. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, pages 2019–2021, 2019. 10.1109/ICHI.2019.8904844.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management*, 58(6):102710, 2021.

Mega Subramaniam, Beth St Jean, Natalie Greene Taylor, Christie Kodama, Rebecca Follman, and Dana Casciotti. Bit by bit: using design-based research to improve the health literacy of adolescents. *JMIR research protocols*, 4(2):e62, 2015. ISSN 1929-0748. 10.2196/resprot.4058. URL http://www.ncbi.nlm.nih.gov/pubmed/26025101{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4464334.

Yalin Sun, Yan Zhang, Jacek Gwizdka, and Ciaran B Trace. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. *Journal of medical Internet research*, 21(5):e12522, 2019.

Hanna Suominen, Lorraine Goeuriot, Liadh Kelly, Laura Alonso Alemany, Elias Bassani, Nicola Brew-Sam, Viviana Cotik, Darío Filippo, Gabriela González-Sáez, Franco Luque, et al. Overview of the clef ehealth evaluation lab 2021. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 308–323. Springer, 2021.

Briony Swire-Thompson and David Lazer. Public health and online misinformation: challenges and recommendations. *Annual review of public health*, 41:433–451, 2019.

Briony Swire-Thompson, David Lazer, et al. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health*, 41(1):433–451, 2020.

Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. Where we go from here: health misinformation on social media, 2020.

Sijie Tao and Tetsuya Sakai. Realsakailab at the trec 2020 health misinformation track. In *TREC*, 2020.

Ronald W Templeton and James Franklin. Adaptive information and animal behaviour: Why motorists stop at red traffic lights. *Evolutionary Theory*, 10, 1992.

Deependra K Thapa, Denis C Visentin, Rachel Kornhaber, Sancia West, and Michelle Cleary. The influence of online health information on health decisions: A systematic review. *Patient Education and Counseling*, 104(4):770–784, 2021.

Shawn Tseng and BJ Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999.

Anthony K. H. Tung. *Rule-based Classification*, pages 2459–2462. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. 10.1007/978-0-387-39940-9_559. URL `https://doi.org/10.1007/978-0-387-39940-9_559`.

Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. Health misinformation detection in web content: A structural-, content-based, and context-aware approach based on web2vec. In *GoodIT 2021: Proceedings of the Conference on Information Technology for Social Good*, pages 19–24, 2021.

Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. An unsupervised approach to genuine health information retrieval based on scientific evidence. In *Web Information Systems Engineering – WISE 2022: 23rd International Conference, Biarritz, France, November 1–3, 2022, Proceedings*, page 119–135, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20890-4. 10.1007/978-3-031-20891-1_10. URL https://doi.org/10.1007/978-3-031-20891-1_10.

Julián Urbano, Harlley Lima, and Alan Hanjalic. Statistical significance testing in information retrieval: an empirical analysis of type i, type ii and type iii errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514, Paris, France, 2019. ACM New York, NY, USA.

Theo Van Leeuwen. What is authenticity? *Discourse studies*, 3(4):392–397, 2001.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209, 2017.

Piyush Vyas and Omar F El-Gayar. Credibility analysis of news on twitter using lstm: An exploratory study. 2020.

Charles J Walker and Bruce Blaine. The virulence of dread rumors: a field experiment. *Language & Communication*, 1991.

Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

Yunli Wang and Zhenkai Liu. Automatic detecting indicators for quality of health information on the Web. *International Journal of Medical Informatics*, 76(8):575–582, 2007. ISSN 13865056. 10.1016/j.ijmedinf.2006.04.001.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.

Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27, 2017.

Claire Wardle et al. Information disorder: The essential glossary. *Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School*, 2018.

C Nadine Wathen and Jacquelyn Burkell. 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.

Eric W Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

Elissa R Weitzman, Emily Cole, Liljana Kaci, and Kenneth D Mandl. Social but safe? quality and safety of diabetes-related online social networks. *Journal of the American Medical Informatics Association*, 18(3):292–297, 2011.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

David Wong and Man-Kuen Cheung. Online Health Information Seeking and eHealth Literacy among Patients attending a Primary Care Clinic in Hong Kong: a Cross-sectional Survey (Preprint). *Journal of Medical Internet Research*, 21, 2018. 10.2196/10831.

David Ka-Ki Wong and Man-Kuen Cheung. Online health information seeking and ehealth literacy among patients attending a primary care clinic in hong kong: a cross-sectional survey. *Journal of medical Internet research*, 21(3):e10831, 2019.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2): 80–90, 2019.

Shengli Wu. Linear combination of component results in information retrieval. *Data & Knowledge Engineering*, 71(1):114–126, 2012.

Shengli Wu, Yaxin Bi, Xiaoqin Zeng, and Lixin Han. Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Information Processing & Management*, 45(4):413–426, 2009.

Jue Xie and Frada Burstein. Using machine learning to support resource quality assessment: An adaptive attribute-based approach for health information portals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6637 LNCS:526–537, 2011. ISSN 16113349. 10.1007/978-3-642-20244-5_50.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

Qinghua Yang, Angeline Sangalang, Molly Rooney, Erin Maloney, Sherry Emery, and Joseph N Cappella. How is marijuana vaping portrayed on youtube? content, features, popularity and retransmission of vaping marijuana youtube videos. *Journal of health communication*, 23(4):360–369, 2018.

Mubashar Yaqub and Pietro Ghezzi. Adding dimensions to the analysis of the quality of health information of websites returned by google: cluster analysis identifies patterns of websites according to their classification and the type of intervention described. *Frontiers in public health*, 3:204, 2015.

Yinjiao Ye. Correlates of consumer trust in online health information: Findings from the health information national trends survey. *Journal of Health Communication*, 16(1):34–49, 2011. ISSN 10810730. 10.1080/10810730.2010.529491.

Puxuan Yu, Razieh Rahimi, and James Allan. Towards explainable search results: A listwise explanation generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 669–680, 2022.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 internet measurement conference*, pages 405–417, 2017.

Boya Zhang, Nona Naderi, Fernando Jaume-Santero, and Douglas Teodoro. Ds4dh at trec health misinformation 2021: multi-dimensional ranking models with transfer learning and rank fusion. *arXiv preprint arXiv:2202.06771*, 2022.

Shuai Zhang, Wenjing Pian, Feicheng Ma, Zhenni Ni, Yunmei Liu, et al. Characterizing the covid-19 infodemic on chinese social media: exploratory study. *JMIR public health and surveillance*, 7(2):e26090, 2021.

Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

Y Zhang, J Burkell, H Cui, and R E Mercer. An Automated Approach for Rating the Content Quality of Web Healthcare Information : A Case Study on Depression Treatment Web Pages. In *Int'l Conf. Health Informatics and Medical Systems*, pages 3–8, 2018. ISBN 1601324790.

Yan Zhang and Shijie Song. Older adults' evaluation of the credibility of online health information. *CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 358–362, 2020. 10.1145/3343413.3377997.