FIGHTING HEALTH MISINFORMATION: BUILDING AN INTERPRETABLE, CRITERIA-DRIVEN SYSTEM TO ASSIST THE PUBLIC IN ASSESSING THE QUALITY OF HEALTH NEWS

by

Xiaoyu Liu

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

December 2022

ABSTRACT

FIGHTING HEALTH MISINFORMATION: BUILDING AN INTERPRETABLE, CRITERIA-DRIVEN SYSTEM TO ASSIST THE PUBLIC IN ASSESSING THE QUALITY OF HEALTH NEWS

by

Xiaoyu Liu

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Susan McRoy

Machine learning techniques have been shown to be efficient at identifying health misinformation. However, interpreting a classification model remains challenging due to the model's intricacy. The absence of a justification for the classification result and disclosure of the model's domain knowledge may erode end-users' trust in such models. This diminished trust may also undermine the effectiveness of artificial intelligence-based initiatives to counteract health misinformation. The study objective is to address the public's need for help evaluating the quality of health news and the typical opaqueness of an AI approach.

This study employs an interpretable, criteria-based approach for automatically assessing the quality of health news on the Internet. Nine well-established criteria were chosen for building the system. To automate the evaluation of the criterion, Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest algorithms were tested. Two approaches were utilized for developing interpretable representations of the results. For the first approach, (1) word feature weights are calculated, which explains how classification models distill keywords that are relevant to the prediction; (2) then using the Local Interpretable Model Explanations (LIME) framework, keywords for visualization are selected to show how classification models

identify positive news articles; (3) and finally, the system highlights target sentences containing keywords to justify the criterion evaluation result.

For the second approach, (1) sentences that provide evidence to support the evaluation result were extracted from 100 health news articles; (2) based on these results, a typology classification model is trained at a sentence level; (3) then, the system highlights positive sentence instances for the result justification. The accuracy of both methods is measured using a small held out test set. A user study was conducted to understand how users trust in the proposed system's news evaluation result.

The performance of automatic evaluation of health news of each nine criteria ranges from highest (AUC, Precision) values of (0.89, 0.82) for Cost down to lowest values of 0.61 for AUC (Novelty) and 0.60 for Precision (Alternative). Both interpretation approaches could visually interpret the given criteria effectively. When not considering the number of sentences for visualization, the best accuracy achieved for each criterion was 100% (Cost), 66.7% (Benefit), 100% (Harm), 95% (Quality),95.45% (Mongering),90% (Conflict),65% (Alternative),68.42% (Availability) and 66.67% (Novelty). The result of the user study shows that participants have high trust in the health news quality evaluation result generated by the system. However, no statistically significant difference was observed between the study and control groups.

Results suggest one might visually interpret an automatic criterion-based health news quality evaluation successfully using either approach. This work addresses the need of interpretation in a computerized health information evaluation.

This Dissertation is Dedicated

to My Grandpa, Yongzhi Qu.

May His Gentle Soul Rest in Peace!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My advisor, Dr. Susan McRoy, is beyond deserving of my sincerest appreciation. This thesis would not have been completed without her assistance and unwavering participation at each stage of the PhD journey. My defense committee, comprised of Drs. Ajay Sethi, Amanda Simanek, and Maria Haigh, were also indispensable to the success of this endeavor. Additionally, I appreciate Dr. AkkeNeel Talsma and Dr. Teresa Johnson for the research opportunities they have provided. Working as a member of the catalyst team has been one of the most memorable experiences of my time at UWM. I appreciate you making me feel welcomed, supported, and valued. I would also like to thank Dr. Shuqian Zhang. Thank you for helping me take my first steps in the field of health informatics and for planting the desire to become a researcher in the heart of an 18-year-old girl.

I am also grateful for the support and generosity of the extraordinary people who surround me: Jan Lloren, Yi Yin, Hiba Alsghaier, Namita Singh, Qianqian Dong, Mary Ejiwale, Eman Alanazi, Vijaya Tamla Rai, Ling Tong and Jingning Ao, among others. Thank you for pursuing your own dreams while also supporting mine. You are all precious gems, illuminating my dull student life and accompanying me on this special journey.

My deepest appreciation extends to my grandparents, parents, husband, and Mrs. Beverly Morton, my American mother. You all are my primary motivation to become a better version of myself. My special thanks go to my husband, Sandeep Achuthan. Life can be very difficult at times, but I'm glad to be on this journey with you. Thank you for being there for me in times of self-doubt and grief, as well as for being my biggest cheerleader during my PhD studies in Milwaukee while I was there alone. I eagerly anticipate many more adventures with you in the future. Lastly, I want to thank me for believing in me. Congratulations, Xiaoyu Liu, you made it!

CHAPTER I: INTRODUCTION

1.1 Overview of the Problem

1.1.1 Overview of Health Misinformation on the Internet

The internet has grown in popularity as a source to learn about one's health and even investigate their health condition. It is estimated that 80% of the Internet users consult online health information before making decisions (Fox, 2011). Online media outlets such as social media feeds, forum threads, blogs, and newspapers have made information access and sharing easier. These social platforms have increased participation among health information consumers of all socioeconomic backgrounds, regardless of ethnicity, gender, or age. However, it has also accelerated the propagation of misleading information at a high speed and a wider range. With the rise of health information seeking on social media platforms, there has been an increase in concerns and health-related harmful cases with regards to misinformation. Unlike other types of misinformation, health-related misleading information, especially those that include claims of efficacy about health intervention such as medical treatments, tests, products, or procedures, can cause actual harm to real people. The general public and patients may be misled into making bad decisions that result in severe consequences regarding people's quality of life and even their risk of mortality. In the case of COVID-19, as the novel coronavirus rapidly spread throughout the world, false information about COVID-19 has developed and circulated like wildfire, leading to unprecedented levels of misinformation (Bridgman et al., 2021). Misleading and erroneous information such as conspiracy theories, poorly sourced medical advice, and information trivializing the virus has not only contributed to widespread misconceptions about the novel coronavirus but caused public panic, catastrophic consequence of public health, and even

people's distrust in public health institutions at a global level (Bridgman et al., 2021; Cui & Lee, 2020).

1.1.2 Overview of Current Countermeasures for Health Misinformation

To address this public health crisis, continuing efforts to counteract health misinformation are being carried out across a wide range of disciplines and organizations. Traditional approaches are dependent on human judgment and manual efforts, and the tasks can be separated into two categories: fact-checking and criteria-based assessments. The former approach employs experts and journalists to manually select and debunk false health claims made in online news and social media and then to promote truthful information. Criteria-based assessment, like HONcode (Team HON, n.d.), applies tailored criteria to assist patients and consumers in judging the quality of health-related online information. However, this approach also needs readers to assess health news manually by checking against the criterion one by one. Due to the human-centered nature of traditional approaches, the magnitude of misinformation limits the capacity of both to be deployed on a large scale (Botnevik et al., 2020). In more recent years, calls have been made by researchers (Dale, 2017; Hassan et al., 2015) for automating health news assessments to reduce the adverse impacts of health misinformation. As the automated system is intended to simulate manual work with machine learning techniques, datasets that are used for automatic model construction in existing studies are mostly derived from the aforementioned human-based projects. This reliance on human-based projects means that existing AI-based classification systems can be inherently divided into veracity-based works for detecting health misinformation and criteria-based approaches for assessing health information quality.

Machine learning is powerful but not a panacea. Despite the high accuracy that has been achieved by models in various fields, the fact that machine learning techniques are "Blackbox" models is often cited as a criticism of their success in the classification tasks. Without disclosing the domain knowledge (i.e., explainability or interpretability) inherent in the data, the general public's trust in and acceptance of such classification models are frequently undermined (Ayoub et al., 2021). To date, only a small body of research has incorporated explainable functionality to combat misinformation in AI-powered models (Ayoub et al., 2021; Kotonya & Toni, 2020). These previous studies on AI-based health information classification are veracity-based. Researchers have yet to construct an interpretable, criteria-driven classification system to help users evaluate the quality of health information. Also, little is known about how end users trust in such developed fake news detection models. Veracity-based fake news detectors' applicability in real life remains uncertain due to two major shortcomings. First, human-based fact-checking work involves extensive knowledge understanding, inference, and source tracking, which remains a challenge even to deep learning methods. False news content is planned to mimic the truth in order to fool readers; therefore, without cross-referencing and high-level inference, it is sometimes difficult to discern truthfulness by text analysis alone(K. Sharma et al., 2019). Second, most fake news detectors are built on linguistic cues, leading to a lack of generalizability across topics, languages, and domains (K. Sharma et al., 2019).

1.2 Purpose of The Study

To address the health misinformation problem and research gap discussed previously, this study aims to develop an AI-based interpretable, criteria-driven system that assists the public in assessing the quality of health-related news. For any given health news article that describes a health intervention, the system delivers an automatic evaluation for each of the nine well-

established criteria. In addition, the system can automatically choose and highlight pertinent phrases or words as visual signals to support the automated evaluation result.

The system provides systematic direction on how to gauge the quality of health news by using a list of well-established criteria that reminds readers of the crucial facts regarding health interventions that they must be aware of prior to making judgments. The goal is to improve the end users' critical thinking about health news through constant exposure of intervention provided by the system.

1.3 Overview of the Dataset and Criteria

The dataset used in this study was adapted from an existing resource created by HealthNewsReview.org(*HealthNewsReview.Org*, n.d.). HealthNewsReview.org is a web-based project that reviewed articles from 2005 to 2018. Their team of experts rated the claims about health care interventions to improve the quality of health care information. Their rating instrument includes ten criteria used by the Australian and Canadian Media Doctor sites, and its inter-reviewer reliability was tested using a random sample of 30 stories(Schwitzer, 2008). HealthNewsReview.org includes reviews of news stories from leading U.S (United States). media and news releases from institutes. The contents include efficacy claims about specific treatments, tests, products, or procedures. The news pieces are assessed based on a standard rating system. At least two reviewers reviewed each news story. The reviewers were selected based on their having years of experience in the health domain, spanning the fields of journalism, medicine, health services research, public health, or as a patient, and each of them signed an industry-independent disclosure agreement. For each news story or news (press) release reviewed, the criteria are scored as "Satisfactory," "Unsatisfactory," or "Not Applicable." Total scores are posted for articles with two or fewer "not applicable" ratings and are expressed as

proportions. It was acknowledged that increasing the diversity and independence of the reviewers could have reduced the potential for bias of the assessments. By the time the project ended, the website had accumulated 2616 health story reviews and 606 news release reviews. For this study, nine of the eleven criteria have been employed to build an interpretable, criteria-based system that can provide information at the sentence level. The nine criteria chosen in this study are listed below:

(1) Does the news adequately discuss the costs of the intervention?

(2) Does the news adequately quantify the benefits of the intervention?

(3) Does the news adequately explain/quantify the harms of the intervention?

(4) Does the news seem to grasp the quality of the evidence?

(5) Does the news commit disease-mongering?

(6) Does the news identify conflicts of interest?

(7) Does the news compare the new approach with existing alternatives?

(8) Does the news establish the availability of the treatment/test/product/procedure?

(9) Does the news establish the true novelty of the approach?

The criterion "Does the news release include unjustifiable, sensational language, including in the quotes of researchers?" was excluded from the study as it only applies to the news release. Another criterion "Does the story appear to rely solely or largely on a news release?", has been excluded from the current study, because of severe limitations with how this criterion was determined that make the judgements difficult to assess or reproduce. The original evaluation required the reviewers to search the web manually to find any relevant news releases and then, again manually, compare content similarities between the reviewed news and the other news sources retrieved during the search to evaluate whether one news release relies completely

or largely on another. Moreover, neither the links to those retrieved web pages or the dates they were retrieved are part of the distributed data. This is the only criterion that is not "self-contained"; the others can all be verified by having reviewers apply the same definitions to the same documents at the sentence level and verifying that there is some part of the document that supports the original evaluation result. This is the approach this study is taking.

1.4 Research Questions

To conduct the experiment in a more systematic manner, two research questions guide the experiments for the respective stage. The first stage aims to test the feasibility of automating the review of health news quality. This aim guides my first research question: *How accurately can the health news review process be modeled to predict health news quality using the data provided and annotated by HealthNewsReview.org?* This question is further broken down into the following three sub-questions:

(1) What is the best performance we can achieve for automatically reviewing health news quality per criterion?

(2) What features and algorithms contribute to predicting the review result with the highest performance per criterion?

(3) What are the criteria for which the review process can be better automated by the machine based on various evaluation metrics including AUC and Precision?

The primary aim of the second stage is to investigate how to enhance the visuality and interpretability of the expert system. This leads to the second research question: *How effectively can the automated health news review results be visualized or explained?* It consists of two sub-questions as follows:

(1) How well can interpretable A.I. techniques and typology classification approaches visualize or explain the automated health news review results compared to a simple display of black-box prediction results?

(2) Which approach demonstrates a superior performance interpreting for the same criterion?

1.5 Contribution of the Study

Notable contributions of my work are as follows:

First, this study developed two innovative methods for visualizing the interpretation of a machine learning classification system used to automate the evaluation of the quality of health news. To the best of my knowledge, the two approaches have not been presented or implemented in the interpretable intelligent machine designed to combat health misinformation.

Second, this study developed annotation schemes to extract evidence following nine criteria deemed crucial for informing internet users of medical or health news. The created datasets demonstrated their validity in developing categorization at the sentence level to support the visual interpretation of models in this work. It can also be applied to other tests and modeling tasks in the study of natural language processing, such as summarization.

Third, this study showed that supervised machine learning models trained on the datasets collected from the HealthNewsReview.org can be used to automatically evaluate health news for each criterion. Additionally, this work also shows that sentence-level text classification that are trained on supervised machine learning models are effective in interpreting document-level text classification result in the context of a criteria-based health news evaluation.

1.6 Significance of the Study

The proposed study addresses misinformation by merging principles of computer science, information science and public health to enhance the public's critical thinking about health news. Amid COVID-19, this study's motivation is consistent with ongoing anti-misinformation efforts supported by organizations that are spearheading the battle against infodemics, such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). The proposed system is designed to assist users more critically evaluate the quality of medical information found on the internet, thereby reducing the harm caused by erroneous health disinformation. Furthermore, a healthy information environment cannot be established without regulation and guidelines. Therefore, the system would serve as a tool for health-related online content authors to evaluate and standardize their article writing practices. This system will aid in the standardization of reporting health news in journalism; thus, the information will present with more rationality and less misunderstanding.

1.7 Outline

The rest of the dissertation is arranged as follows:

Chapter II summarizes and overviews fundamental theories concerning health misinformation, which helps to comprehend the panorama of health-related misinformation. This section also includes relevant initiatives that have been used in the real world or suggested in the literature on how to combat misleading health information to lessen the negative effects on public health. To provide a solid framework for chapter III, a review of numerous methodology algorithms that are pertinent to the subject topic is provided.

Chapter III comprises the specifics of study’s contributions, such as the data annotation scheme, the technical procedures applied to enable the automatic evaluation of health news for

each criterion, and the two original approaches the Typology approach and the Hybrid approach that are used to visualize the interpretation of automatic evaluation result. The two interpretation approaches form the core strengths of this work. Lastly, this chapter also presents how various techniques were employed to evaluate the performance of system components.

Chapter IV presents various results and key findings derived from the empirical experiments, including the quality of the scraped, annotated dataset for building machine learning models, the performance of models of automatic health news quality evaluation, and the performance of two visualization approaches for visualizing the interpretation of evaluation results. This chapter concludes with the results of a survey study that reveal how users' trust in an interpretable system differ from that of a non-interpretable system in the context of health news quality evaluation.

Chapter V includes the discussion of study results, which summarizes the key findings of the study, compares the proposed solution to earlier work, and acknowledges the study's limitations. The dissertation concludes with a brief summary of the study.

CHAPTER II: LITERATURE REVIEW AND RELATED WORK

2.1 Overview of Chapter II

This chapter provides a review of important theories on health misinformation, including definitions, sources, and characteristics of health misinformation, and how misinformation is perceived and disseminated by readers. Various machine learning algorithms linked with the topic are explored and reviewed in order to build the groundwork for the methodology employed in this study, which will be presented in chapter III. This chapter concludes with a review of countermeasures both implemented in the real world and proposed in the literature to combat health misinformation and reduce its negative effects on public health.

2.2 Literature Review of Fundamental Theories

2.2.1 Definitions of the Target Phenomenon

2.2.1.1 Definitions of Misinformation

Despite the extensive mention of misinformation so far, recent studies on the spread of misinformation on the internet have not made considerable use of the term "misinformation." Instead, another catchy term, "fake news," has become a popular research term after it was widely used on social media during the 2016 U.S. presidential election (Bode & Vraga, 2015). There are many definitions for "fakes news" and they have evolved rapidly. A recent definition by Gelfert (2018) noted that fake news should be "reserved for cases of deliberate presentation of (typically) false or misleading claims as news, where these are misleading by design."(Haigh & Haigh, 2020)

As more research was conducted, many researchers began to criticize the term for being woefully inappropriate and failing to address the complexity and multidimensional attributes of the phenomenon of information disorder (Wardle & Derakhshan, 2017; Habgood-Coote, 2018; Wang et al., 2019). Moreover, it has become a "politicized rhetorical device"(van der Linden, 2022). Wang et al (2019) called upon governments to stop using the term "fake news," and instead to promote the terms "misinformation" and "disinformation."

Misinformation is defined as "false, inaccurate, or misleading information that is communicated regardless of an intention to deceive"("Misinformation," 2021). The European Commission (EC) defined disinformation as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm."(McKay & Tenove, 2021) Other terms such as "propaganda," "conspiracy," "hoax," and "unreliable news," also have gained more attention and extensively

redefined and cited in an attempt to describe the nuances and subordinate relationship of various types of misinformation in more recent studies. Based on the intention to deceive and the financial or non-financial motivations of its disseminators, Verstraete et al.(2017) identified four types of fake news: satire, hoax, propaganda, and trolling (Haigh & Haigh, 2020). The research argued that both satire and hoax are purposefully false and financially motivated. However, satire differs from hoax in that hoax is designed to deceive readers. While both propaganda and trolling are intended to deceive readers, propaganda is motivated by a political purpose, whereas trolling is motivated by an attempt to gain personal humor and value.

Despite the fact that there are many terms available, no consensus among researchers has been reached on categorizing the most commonly used terms for describing this era of misinformation. Many terms are sometimes used interchangeably, "with shifting and overlapping definitions"(Persily & Tucker, 2020). Taking disinformation for example, the definition given by EU implies that disinformation is an alternative to the fake news defined by Gelfert (2018). However, this statement is not valid if we consider the definition of disinformation by Cherilyn & Julie (2018) and the four types of fake news proposed in Verstraete et al.(2017)'s study. Cherilyn & Julie (2018) defined that disinformation is "a subset of propaganda and is false information that is spread deliberately to deceive". As mentioned earlier, propaganda is one of four types of fake news that intends to deceive readers with political agendas (Verstraete et al., 2017). Given both statements, disinformation doesn't share the same meaning as fake news; rather, it inherently becomes a subtype of fake news. Additionally, Pan et al.(2021) regarded both misinformation and rumor equally as they both "focus on the unique characteristics of unverified information." Another example is the case of conflicting interpretation on the relationship between misinformation and disinformation. Pan et al. (2021) defined

disinformation as a subset of misinformation in their research. In contrast, some believe that disinformation and misinformation are two mutually exclusive categories because disinformation refers to the intentional cases while misinformation is unintentional (Fallis, 2009; Wu et al., 2019).

2.2.1.2 Definitions of Health Misinformation

As a subtopic of misinformation, the term "health misinformation" is increasingly present in our societies. Although the phrase "health misinformation" is cited extensively in the introduction chapter, definitions for such public health information disorder have not yet been agreed upon in recent studies on the dissemination of misinformation online. Existing studies handle the definition of "health misinformation" in two ways.

The first way, which is employed by the major body of existing literature, especially in earlier engineering-based studies, focuses on defining and interpreting the generic "misinformation" term. As a result, health misinformation is regarded as a type of misinformation that is themed on health topics, without probing into the nuanced differences between health misinformation and misinformation of other topics. For example, Zhao et al.(2021) adopted a definition of misinformation as "the factually incorrect information that is not backed up with evidence"(Bode & Vraga, 2015) even though the study concentrated on health-related misinformation only. A similar definition of misinformation as "communications that are not a fair representation of available evidence or communicate that evidence poorly" is given in the study by Shah et al.(2019). The study aimed to automatically appraise the credibility of web pages that are vaccine related. Some studies substitute other terms such as "fake news"(Botnevik et al., 2020), "unreliable"(Meppelink et al., 2021; Saengkhunthod et al., 2021)"propaganda"(Khanday et al., 2021), "conspiracy"(Shahsavari et al., 2020) for the term

“misinformation “with established meanings aiming to address either a specific or general health misinformation issue. Pagoto et al. (2019) defined two types of misinformation phenomena using both “health misinformation” and “health disinformation." The former is specified for describing information that is false and spread by someone who believes it to be true, and the latter refers to false information that is spread with intention. Although the authors aimed to define "health misinformation," the concept is taken from generic misinformation-related terms without placing equal focus on the context of health.

In contrast, the other way of defining health misinformation considers the "health" component in definitions. Such definitions distinguish health misinformation from misinformation regarding other topics as they embody the characteristics of health information. A researcher may consider a piece of health-related information as misleading if there is no scientific evidence to support the claim. For example, Chou et al.(2018) defined health misinformation as “a health-related claim fact that is currently false due to a lack of scientific evidence”. There is a stricter definition for health misinformation by Parfenenko et al.(2020) where medical misinformation is limited to false information that contradicts current medical knowledge. This stricter definition excludes medical information where claims lack sufficient scientific evidence.

Owing to the inherent dynamism of the social media ecosystem and wide varieties of health topics, defining “health misinformation” faces similar challenges as the general term misinformation and even becomes more elusive (Suarez-Lledo & Alvarez-Galvez, 2021). First, for narrow definitions of health misinformation, one may easily argue that news stories that contain health information do not have to be completely false to cause harm(van der Linden, 2022). Second, it is arbitrary to label information without firm evidentiary basis as

misinformation. One example is complementary and alternative medical approaches such as massage, acupuncture. They have uncomfortably coexisted with mainstream scientific medicine for decades and are widely utilized and accepted (Armstrong & Naylor, 2019; Institute of Medicine (U.S.), 2005). Lastly, what is likely true and what is likely false might change over time. This is especially the case when the professional consensus on a topic of public health is fast forming and subject to uncertainty. In the instance of COVID-19, although the media first reported that ibuprofen could exacerbate coronavirus symptoms, this report was eventually withdrawn as additional evidence emerged (van der Linden, 2022).

To minimize confusion in future narratives and to address the challenges associated with defining health misinformation, the definition of health misinformation in this study emphasizes the negative consequences of health misinformation on public health. Thus, health misinformation in this study refers to as *a broad term to refer to any untrustworthy, inaccurate, or incorrect information, which potentially misleads people to making either unnecessary or harmful health care decisions. Terms to describe various types of health-related misinformation or phenomenon such as fake news, disinformation, unreliable, incredible, or misleading information, myth, rumor, hoax, propaganda, conspiracy all fall under its umbrella.*

2.2.2 Prevalence of Health Misinformation

Despite widespread agreement among health professionals, policymakers, and the general public regarding the magnitude of health misinformation and the necessity of combating it, it is essential to identify both the most prevalent health topics and the outlets from which they are initially framed and then disseminated (Suarez-Lledo & Alvarez-Galvez, 2021).

Social media is one of the identified major online channels where health misinformation is disseminated. Waszak et al.(2018) found that 40% of the most frequently shared links

contained medical information related to related to the most common diseases and causes of death were classified as fake news. In an analysis of marijuana vaping contents shared on YouTube, the misleading information reached as high as 98.1%(Yang et al., 2018).Through the social media channel, fake news is disseminated further, faster, deeper, and more broadly, reaching more people than facts(Atehortua & Patino, 2021). Amid the coronavirus pandemic, recent evidence demonstrated that COVID-19 related false or fake news can be moved from one platform to another using shortcuts or tunnels due to the easier sharing.

Misinformation about health has been detected on different social media sites such as Twitter(Becker et al., 2016; Bonnevie et al., 2020; Hanson et al., 2013; Jamison et al., 2020), Facebook(Buchanan & Beckett, 2014; Faasse et al., 2016; Johnson et al., 2021; Seymour et al., 2015), YouTube(Abukaraky et al., 2018; Basch et al., 2017; Biggs et al., 2013; Röchert et al., 2020), Pinterest (J. Guidry et al., 2016; J. P. D. Guidry et al., 2015) and Weibo(A. Li et al., 2018; Xiao & Chen, 2020). Additionally, the spread of health-related misinformation is not confined by geography. A series of studies have reported and studied health misinformation in different geographic settings such as US(Chua & Banerjee, 2018; Radzikowski et al., 2016; Vraga & Bode, 2017), China(Chen et al., 2018; A. Li et al., 2018; Y. Li et al., 2017; Xiao & Chen, 2020),India(Leong et al., 2018) and Italy(Aquino et al., 2017; Bessi et al., 2015).

Existing studies on the spread of health misinformation on social media concern a variety of health topics. Wang et al.(2019) conducted a review study on the spread of health misinformation on social media. The study summarized two major categories communicable and non-communicable diseases with communicable disease being the largest category. Topics identified in the communicable diseases category include vaccines in general, vaccines specific to Human Papilloma Virus (HPV), Measles, Mumps, and Rubella (MMR), as well as other

infectious diseases such as Zika Virus and Middle East Respiratory Syndrome. Studies that concern chronic non-communicable diseases are identified but not limited to cancer, cardiovascular disease, psoriasis, bowl diseases. Later, a similar review study by (Yang et al., 2018) categorized existing literature into the following broad topics: vaccines, drugs, or smoking, non-communicable diseases, pandemics, eating disorders, and medical treatment, while vaccines are identified as the most examined health topic.

2.2.3 Sources of Health Misinformation

Social media receives the most critiques and attention as being the platform hosts, and accelerates the spread of health misinformation. However, those social media platforms are not always the place where health misinformation originates. Nan et.al (2021) identified five sources of health misinformation including media, industry, government and politicians, healthcare providers, and social/interpersonal groups or community networks. Following the same notion, we rearrange the sources and further categorize them into media, for-profit entities, non-profit entities &individuals.

2.2.3.1 Mass Media

Mass media is perhaps the most common source of health-related information for the public (Nan et.al.,2021). Study show that health news is ranked sixth in receiving public attention only following weather, crime, community, the environment, and politics (Kohut et al., n.d.). When it comes to health news, the media sits between public and healthcare professionals, industry, and scientists. It translates scientific knowledge about disease and new medical discoveries from scientific publications. It also transfers, disseminates health policy and health products announcements from industrial and governmental agents (D. C. Sharma et al., 2020). Therefore, mass media serves as a delivery mechanism in commutating disease prevention and

healthy lifestyles promotion to have a positively affect personal health-related changes. Such positive changes are driven by motivated people to make better-informed choices and decisions about their health, wellness, and overall quality of life (D. C. Sharma et al., 2020). However, a handful of studies expressed concerns on the quality of health news. Misinformation arises from media when reported news stories don't cover necessary information or inform readers such as potential harms of medical interventions, conflict of interest of the studies etc. In an evaluation of 500 US health news stories from mainstream news outlets in the US, including Associated Press and the three leading newsweekly magazines—TIME, Newsweek, and U.S. News & World Report, Schwitzer (2008) found that between 62%-77% of news stories failed to adequately address five essential components to inform readers including about costs, harms, benefits, the quality of evidence and the existence of alternatives. Only half of the time, the news stories disclose relevant conflicts of interest, and avoid single-source stories. In a study that targets obesity-related news stories, the study found that nearly all story journalists failed to discuss the methodology and design limitation. With the absence of necessary information, readers can be misled and ill-informed based on such news stories of low quality.

2.2.3.2 For-profit Entities

For-profit entities account for another source of health misinformation. Pharmaceutical companies exert significant influences on healthcare and treatment of diseases (Paul et al., 2010). Pharmaceutical companies have utilized the Internet more frequently as it has grown to influence patients to ask doctors to prescribe medications for them (De Freitas et al., 2013). However, they are found guilty of producing misleading, inaccurate, health information about their products (De Freitas et al., 2013). Such companies are profit-driven and have invested significantly to market and promote their products on the internet. Misinformation occurs when pharmaceutical

companies mislead users or withhold from them relevant safety and efficacy information. For example, in 2018, a pharmaceutical company named Aegerion Pharmaceuticals Inc. was sentenced to pay a $40 million penalty for illegally marketing its high cholesterol drug for conditions outside of the treatment's label ("U.S. Judge Sentences Novelion's Aegerion in Drug Marketing Case," 2018).

There are also some businesses which are not conventionally considered in the healthcare or pharmaceutical sector, but they all incur direct or indirect impacts on consumers' health. Examples are food, alcohol, tobacco, beauty and cannabis industries. The advertisements or comments related to products, or consumers' anecdotal experiences and other marketing tactics are often found attributed to diffusion of misinformation. "False advertising" activities are profit-driven and promote actions that harm health (Zgheib, 2017). Alcohol industry alone has a long history distorting scientific evidence and misleading consumers (Petticrew et al., 2020). It is reported that alcohol industry employs "dark nudges" and "sludge "strategies to change consumer behaviors to promote inappropriate alcohol consumption (Petticrew et al., 2020). Advertising non-nutritious food to children and using misleading phrases like "natural" and "light" to make health claims that are not backed by research is also spotted in global, industrialized food industry (Hindin et al., 2004). Similar advertising tactics were deployed by the tobacco industry in the past (Tan & Bigman, 2020).

2.2.3.3 Non-profit Entities & Individuals

Lastly, non-profit entities and individuals are another source of health misinformation. This group includes any institutions, organizations, or individuals that may not directly benefit from generating misleading information to promote commercial activities. However, their intention for creating, spreading health misinformation can be malicious or innocent. Entities

identified in this category are government, politicians or other public figures, healthcare providers, groups/communities, and individuals. According to a study by Brennen et al.(2020) on the types, sources, and claims of COVID-19 misinformation, top-down misinformation sourced from politicians, celebrities, and other prominent public figures made up about 20% of the studied samples but accounted for 69% of the total social media engagement. Misleading health information contained in posts and tweets published by public figures or politicians may induce dangerous and life-threatening behaviors such as the use of disinfectants, chloroquine phosphate, or the self-prescribed off-label use of medications (Atehortua & Patino, 2021). Healthcare providers including doctors, nurses, dental hygienists, medical students etc. can also be the source of health misinformation as they may also adhere to false beliefs. Few physicians have been disciplined for espousing COVID-19 claims which lack of scientific evidence. For example, in October 2021, Howard Goldman, MD of Delray Beach, Florida filed a complaint with the Florida Department of Health's Medical Quality Assurance Program about an internist, Joseph Ladapo, MD, PhD, Florida's Surgeon General and health of the Florida Department Health because the latter "spread doubt about the safety and effectiveness of COVID-19 vaccines, promoted the use of unproven and possibly dangerous medications to treat COVID-19, and questioned the value of face masks in preventing the spread of pandemic."(Rubin, 2022) Compared to other sources, health care professionals are not the primary contributors of inaccurate health information. However, they "may well be the most egregious of all because they undermine the trust at the center of the patient-physician relationship, and because they are directly responsible for people's health." said Gerald Harm, MD, president of the American Medical Association (Rubin, 2022).

Another source of misinformation come from non-governmental groups or communities, either formal or informal. Taking anti-vaccination for example, there are various known anti-vaccine organizations in the United States, such as Anti-Vaccination League of American, The Autism Community in Action, Health Freedom Idaho ("List of Anti-Vaccination Groups," 2022) and their history can be traced back to early 90s ("Anti-Vaccination Society of America," 2022). There are also seemingly less formal groups or communities that are convened on social media platforms such as Facebook groups (Kalichman et al., 2022), and medical forums(Bandari et al., 2017). Anti-vaccine websites are often found to be more effective at utilizing social interactivity than pro-vaccine websites, as they focus on social interactivities that effectively create communities of people who are affected by and are skeptical of vaccine practices(Smith & Graham, 2019). The last identified source of misinformation in this category is ordinary individuals. Posts by ordinary people seem to generate far less engagement and impacts than those of public figures and larger entities, but they are responsible for spreading most of the health misinformation on social media (Brennen et al., 2020).

2.2.4 Characteristics of Health Misinformation

Misinformation-related evidence reveal the possible characteristics a piece of health misinformation may possess, which makes it distinguishable from scientifically supported, truthful information.

The first characteristic of health misinformation is the manipulation of truth through low-quality writing styles such as oversimplification, misrepresentation, overdramatization and self-contradictory statements during the science repacking in the news reporting (Thomas et al., 2017). Such information manipulation poses high risk of misleading readers into developing wrong perceptions of health conditions. In the alcohol industry, social norming (telling

consumers that “most people” are drinking) is often included in the information and priming drinkers with various cues to drink (Petticrew et al., 2020). In addition to the content itself, the manipulation is found in the peripheral elements such as pictures, headlines, fonts to aid cognitive bias towards the health information (Petticrew et al., 2020). Research shows that COVID-19 related misinformation often aided by sensational popular media headlines and foci, fuel health-related fears and phobias (Asmundson & Taylor, 2020).

Another characteristic of health misinformation is the selective omission of critical information. In an evaluation of 500 US health news stories over 22 months, researchers found that 62%-77% of news stories failed to adequately address costs, harms, benefits and quality of evidence, and the existence of other options when covering health care products and procedures (Schwitzer, 2008). In guiding readers, the omission of potential problem definitions, explanations, evaluations, and recommendations may be just as important as their inclusion (Entman, 1993). The tendency of readers to ignore absent evidence is also identified as a key source of bias by Kahneman (2012) and such cognitive bias has been named as WYSIATI (“what you see is all there is”).

Last but not least, an unreliable source is another important characteristic of health misinformation. In an analysis of 112 million messages related to COVID-19 pandemics, the researchers from Bruno Kessler Foundation found that 40% of these messages came from unreliable resources (Lupi, n.d.). This finding is also validated in another study in which 42% of tweets about COVID-19 were found to be produced by unreliable sources (D. C. Sharma et al., 2020). The Statista Research Department identified a list of leading health misinformation websites worldwide based on the website credibility reviews and transparency criteria (Top Health Disinformation Websites 2020, n.d.). Examples include realfarmacy.com.

globalresearch.ca, collective-evolution.com, jedanews.com etc. Among all, the top-ranked health misinformation spreader Realfarmacy.com accumulated an approximate 253.6 million views between May 2019 and May 2020.

2.2.5 Sharing and Acceptance of Health Misinformation

Misleading information about health is only destructive and hazardous when it is disseminated and accepted. As of right now, numerous evidence about human cognition and behavior that are related to general misinformation have been developed across various fields, offering vital insights on identifying, deciphering, and correcting misinformation. Despite the fact that most of evidence, due to its general characteristics, is also applicable to misleading information about health and/or medical themes. Very little research specifically focuses on health misinformation. The following evidence covers various factors that contribute to people disseminating and accepting health misinformation.

2.2.5.1 Psychological Vulnerability

Similar to other types of misinformation, health misinformation can attract both malicious and normal individuals (Kahneman & Tversky, 2012; Waszak et al., 2018). Malicious spreaders create and disseminate false information on purpose, whereas normal readers may participate in the transmission of misleading content without recognizing its false and misleading nature(X. Zhou & Zafarani, 2018). In contrast to the benefit-driven motives involved in malicious users distributing misinformation, normal users may share and accept health misinformation due to psychological vulnerability. In a survey study of fake news, Zhou & Zafarani (2018) identified major theories that contribute to the psychological vulnerability and categorized them into social impact and self-impact.

The social-impact theories demonstrate how a person's perception, judgement of information may be influenced by social factors. The *bandwagon effect* (Leibenstein, 1950) suggest that people may spread false information about health care simply because others are doing so; alternatively, they may do so because they want to win others' approval and respect, as per the *normative influence theory* (Deutsch & Gerard, 1955). Social groupings have a significant impact on how people perceive themselves and adopt opinions from others. *Social identify theory* (Ashforth & Mael, 1989) demonstrates that an individual's sense of self originates from their perception of belonging to a relevant social group. The formulation of groups also contributes to cognitive distortion since members of the group are more likely to adopt popular viewpoints and this tendency is noted as "*availability cascade*"(Kuran & Sunstein, 1998) . An echo chamber can emerge in a closed discussion group when individuals exclusively share information and ideas that reflect their own; as a result, the "echo chamber effect" (Jamieson & Cappella, 2008) reinforces people's beliefs as their exposure to other viewpoints decreases. Once incorrect health standards and ideas have been established, people tend to reject new facts ("*Semmelweis reflex*") (Bálint & Bálint, 2009) or adjust their beliefs insufficiently ("*conservatism bias*") (Nickerson, 1998). For instance, if a person dislikes wearing a mask for protection, they would prefer and recall information that supports their position over evidence-based information, ultimately undermining public health guidelines and regulations (Xu et al., 2022). A similar phenomenon has been observed with the anti-vaccine movement; the decline in immunization rates in developed nations has been linked to the dissemination of misinformation about vaccines (Meppelink et al., 2019).

Additionally, Self-impact theories describe how cognitive factors derived individuals themselves affect people's sharing and acceptance behaviors. For example, *desirability bias*

shows that people prefer to read information that simply makes them feel good. Similarly, people tend to selectively read ("*selective exposure*") (Freedman & Sears, 1965) and trust ("*confirmation bias*") (Nickerson, 1998) information that confirms their preexisting beliefs and hypotheses.

2.2.5.2 Lack of Health Literacy and Critical Thinking

Today, there is a large amount of online news being published, which makes it imperative to be able to evaluate its credibility and truthfulness. The ability to do so requires a high level of information literacy and critical-thinking skills. In light of the rising tide of fake news, researchers argue that information/media literacy has become increasingly critical (Machete & Turpin, 2020). Information literacy refers to essential skillsets with which the general public are able to identify, select, understand, and use trustworthy information. In health care, such a skill set is referred to as health literacy(Machete & Turpin, 2020).

The concept of health literacy has received attention since 2000, when the WHO stated that a low level of public health literacy is a serious concern for public health (Bin Naeem & Kamel Boulos, 2021). Health literacy covers "the ability to access, comprehend, evaluate, and communicate information to promote, maintain, and improve health in a variety of settings across the life course."(Sørensen et al., 2012) It helps the general public to make use of online health resources that are beneficial to their health, while staying afloat in an era of misinformation (Abdulai et al., 2021). The notion that low health literacy is associated with the adoption of health misinformation is supported by a survey study conducted by Song et al.(2019). The study investigated the role of health literacy on credibility judgement of online health misinformation and found a significant negative coefficient of health literacy. The negative coefficient indicates that individuals with higher health literacy are relatively less likely to trust

misinformation. In response to widespread COVID-19 related misinformation, Abdulai et al. (2021) evaluated digital literacy levels among lay consumers of online COVID-19 information in Ghana, a low-income country, using a survey based on the eHealth Literacy Scale (eHEALS). The study reveals that respondents' ability to locate COVID-19 related information as well as their skills in differentiating scientific from unscientific internet-based information remain relatively low.

2.2.5.3 Health Anxiety

In the context of health misinformation perception, health anxiety can be defined as "how anxious individual is about issues directly related to the rumor"(Greenhill & Oppenheim, 2017; Pan et al., 2021a). Health anxiety occurs when perceived bodily sensations or changes are being interpreted as symptoms of being ill (Asmundson et al., 2010). Individuals with high levels of health-related anxiety have a tendency to mistake innocuous body feelings and changes as harmful; hence, they are more likely to seek information to either validate their anxiety or alleviate it (Pan et al., 2021a; Taylor, 2004). According to existing research, health anxiety increases the likelihood that individuals will accept and spread false information (Oh & Lee, 2019) and accounts for a significant portion of the variance in the misinformation acceptance (Pan et al., 2021b). Individuals with greater health concerns were more responsive to negative information and more prone to disseminating unfounded claims as a way to vent negative emotions (Oh & Lee, 2019; Pezzo & Beckstead, 2006; Xu et al., 2022).

2.3 Literature Review of Methods

2.3.1 Artificial intelligence Techniques

2.3.1.1 Overview of Artificial Intelligence

In more recent years, many attempts have been made to leverage artificial intelligence (AI) to analyze the enormous amounts of info generated daily on a scale that's impossible for humans to handle.(Marr, n.d.) AI is the simulation of human intelligence process using machines, especially computer systems.(*What Is Artificial Intelligence (AI)?*, n.d.) It is a broad field of study that encompasses several major subfields in which machine learning (ML), deep learning (DL) and natural language processing (NLP) sit. Each subfield of study is not solely independent. It may instead overlap with and inherit ideas from other related fields. For instance, ML is the study that is centered on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.(*Machine Learning: What It Is and Why It Matters | SAS*, n.d.) Among various machine-learning algorithms, DL achieves great power and flexibility compared to traditional ML models by drawing inspiration from biological neural networks to solve a wide variety of complex tasks.(Mahapatra, 2019) NLP is another essential component of artificial intelligence that studies how machines interact with human language . ML and NLP have some overlap as ML is often used to improve NLP by automating processes and delivering accurate responses. The figure below illustrates the relationships among AI, ML, DL and NLP.

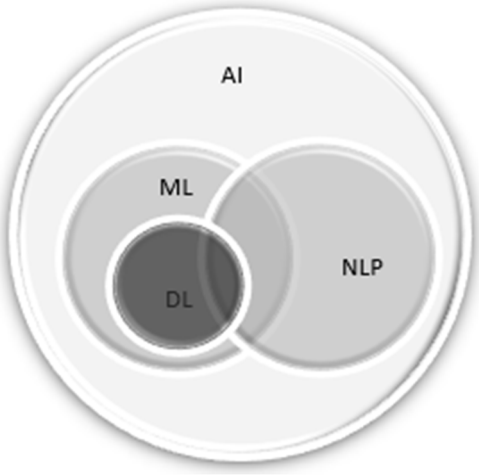

Figure 1: Relationships among AI, ML, DL and NLP

2.3.1.2 Machine Learning Algorithms

2.3.1.2.1 Random Forest

Random forest (RF) is a type of machine learning (ML) techniques that is commonly applied to classification and regression issues. It is proposed by L.Breiman in 2021 (Biau & Scornet, 2016). RF is derived from the decision tree, a tree-like representation of decisions and their potential consequences ("Decision Tree," 2022). It combines randomized decision trees and derives its prediction from the majority vote of all the trees in the forest(Biau & Scornet, 2016; Islam et al., 2019). In settings where the number of variables is substantially bigger than the number of observations, RF's approach of prediction has demonstrated good performance (Biau & Scornet, 2016). RF is versatile enough to be applied to large-scale issues and is widely employed as a "Blackbox" model in business since it generates accurate predictions across a broad range of inputs with minimal configuration(Biau & Scornet, 2016; "Random Forest," 2022). In addition, RF classifiers are appropriate for dealing with high-dimensional, noisy text data(Islam et al., 2019).

2.3.1.2.2 Logistic Regression

Logistic regression (LR) is a type of statistical model frequently used for classification and prediction analytics. LR calculates the probability that an event, such as "having disease" or "not having disease," will occur based on a given dataset of independent variables (*What Is Logistic Regression?*, n.d.). The dependent variable is bounded between 0 and 1 since the outcome is a probability. As a form of regression techniques, LR is diverse in its application to medical research and offers a potent method for analyzing the effect of a set of independent variables on a binary outcome by quantifying the contribution of each independent variable (Stoltzfus, 2011). However, compared to another regression model, liner regression, it is more difficult to understand and interpret because it uses a complex equation model for classification. In recent years, LR has also been proved robust in handling textual data in various text classification tasks (Indra et al., 2016; Pranckevičius & Marcinkevičius, 2016; Saif et al., 2018).

2.3.1.2.3 Support Vector Machine

Support vector machine (SVM) is another classical ML algorithm and can be used for both regression and classification tasks. The objective of the SVM algorithm is to locate a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifiers the data points (Suthaharan, 2016). There are many hyperplanes that can be used to divide the data points into two classes, but only the plane that creates the greatest distance between data points of both classes is deemed the optimal decision boundary (Gandhi, 2018b). SVM facilitates big data applications that operate in multiple domains. Taking its biomedical application as an example, SVM is commonly used for the automation of classifying microarray gene expression profiles (Noble, 2006). Despite its wide application, SVM is mathematically

complex and computationally expensive (Suthaharan, 2016). It is also viewed as a Blackbox model especially there is no probabilistic explanation for the classification.

2.3.1.2.4 Naïve Bayes

Naïve Bayes (NB) classifier is a probabilistic ML model that is widely used for handling classification tasks. The crux of the classifier is that it is built based on the Bayes Theorem, a mathematical formula used for calculating conditional probabilities. Conditional probability is a measure of the probability of an event occurring based on the prior knowledge of conditions (*Encyclopedia of Bioinformatics and Computational Biology*, 2018). The formula of Bayes Theorem is shown as below.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

In the formula, $P(A/B)$ represents the probability of event A occurs given that the event B has occurred. The assumption is that features are independent of each other. Hence it is called naïve (Gandhi, 2018a). In a classification task with multiple features, the Bayes theorem can be rewritten as below, where y presents the prediction target and X represents features given as $X = (x1, x2, x3, \dots, xn)$

$$P(y/X) = \frac{P(X/y)p(y)}{P(X)}$$

The probability of occurrence of event y can be calculated through further derivation and transformation of formula. Naïve Bayes classifiers are also favored in various text classifications, especially in the fields of sentiments and spam detection (Ersoy, 2021).

2.3.1.2.5 Deep Learning

DL is part of a broader family of ML family based on artificial neural networks with representation learning. As it seeks to resemble the human neural network, it is also known as a neural network. Typically, these neural networks comprise three or more layers, despite the fact that one layer can still make approximate predictions (*What Is Deep Learning?*, 2022). Additional hidden layers can aid in optimizing and refining the algorithms, allowing them to cluster data and produce extremely accurate predictions (*What Is Deep Learning?*, 2022). As the deep learning evolves, various deep learning models are proposed and advanced. For handling classification tasks, deep learning that provides discriminative function mainly include Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent neural Networks (RNN), along with their variants (Sarker, 2021). Due to their special architecture, CNN are suitable for spatial data such as images whereas RNNs are for temporal data that comes in sequences like a sequence of words (Venugopalan et al., 2015). Despite the fact that DL is known for its incredibly high accuracy, one drawback is that DL-based classifiers typically require very large training sets (P. Liu et al., 2022). Additionally, DL models often receive critiques in its inexplicability for not providing information about what made them arrive at the desired output (Horst et al., 2019).

2.3.1.3 Feature Engineering

When building ML models, feature engineering is a critical process as it has impacts on the performance of classifiers. The feature engineering process refers to the selection, manipulation and transformation of raw data into desired features, the individual measurable properties or characteristics of the target phenomenon, that can be used in supervised learning ("Feature Engineering," 2022; Zheng & Casari, 2018). It produces new features for ML through

the conversion of raw observations using statistic or machine learning approaches. The technique used for feature engineering differs based on the type of data and task the classifier handles. For text-based classification tasks, popular linguistic feature types include count vector as features; TF-IDF vectors as features; word embeddings as features; NLP-based features and topic models as features. It is also worth noting that text preprocessing is essential prior to feature engineering in a text-based classification task. Some of the textual data preprocessing steps include tokenization, lower-casing, stop words removal, stemming and lemmatization (Harshith, 2022).

2.3.1.4 Model Hyperparameter Tuning

Model hyperparameter tuning is another step that is critical to the performance of classifiers. Hyperparameters defines the model architecture. They are model arguments whose values should be determined before the learning process begins. Hyperparameter tuning refers to the process of searching for a set of optimal hyperparameters for a learning algorithm (Jordan, 2017; Liashchynskyi & Liashchynskyi, 2019). There are two popular hyperparameter tuning methods.

One method is Grid Search. Grid Search is a traditional approach to tuning hyperparameters. It simply performs a complete search over a given subset of the hyperparameters of the training algorithm, evaluating each model and selecting the architecture that produces the best result (Liashchynskyi & Liashchynskyi, 2019; Pfeiffer, 2019). Grid Search is capable of providing the best way to tune the hyperparameters as it conducts an exhaustive search. However, it has the drawback of requiring a great deal of time and space (Pfeiffer, 2019).

The other method is Random Search. Random Search provides an alternative to Grid Search's exhaustive search approach. It selects combinations of hyperparameters at random and tests them to identify the optimal hyperparameter values among the randomly selected group

(Pfeiffer, 2019). Random search is often faster because it requires fewer running iterations (Nair, 2022). It can outperform Grid Search, particularly when the number of hyperparameters for tuning is small (Liashchynskyi & Liashchynskyi, 2019). Due to the random nature of this method, its drawback lies in the high possibility of variation between runs (Pfeiffer, 2019).

2.1.3.5 Evaluation of Machine Learning Model

ML model evaluation is the process of using different evaluation metrics to understand a ML model's performance in a classification task, as well as its strengths and weaknesses (J. Zhou et al., 2021). Various evaluation metrics can be leveraged for assessing classifiers. Some of the popular ones include Accuracy, Sensitivity, Specificity, F-Measure, Precision, Recall and Area under the ROC Curve (AUC). Table 1 provides a summary of each metric's formula and explanation. The table is created based on the review study on evaluation metrics for data classification evaluations conducted by Hossin & Sulaiman (2019).

Most of these measures depend on the discrimination threshold to dichotomize the predicted output. In comparison, AUC imitates the performance of the overall classification by setting a value called the discrimination threshold to predict the probability from a binary classifier (Q. Wang & Guo, 2020). AUC is typically preferred because the AUC is insensitive to class distribution, and it doesn't depend on a cut-off value (Garrido et al., 2018). Moreover, in a study established a precise and an objective criterion for comparing evaluation measures empirically and formally, the study shows that AUC is a better measure in general and it obtained surprising and an interesting new result when it was used to evaluate machine learning models (Huang & Ling, 2005). Another study developed by Wu & Flach (2005) claimed that the performance evaluation metrics such as accuracy, recall and precision have common weakness that they are not robust to class distribution changes, specifically, in the test set when the

proportion of negative and positive instances changes, they may not perform acceptably or optimally

Table 1. Summary of ML Model Evaluation Metrics

Evaluation Metrics	Representation/Formula	Explanation
True Negative	tn	An outcome where the model correctly predicts the negative class
True Positive	tp	An outcome where the model correctly predicts the positive class
False Positive	fp	An outcome where the model incorrectly predicts the positive class
False Negative	fn	An outcome where the model incorrectly predicts the negative class
Accuracy	$\frac{tp + tn}{tp + fp + tn + fn}$	Measures the ratio of correct predictions over the total number of predictions
Sensitivity	$\frac{tp}{tp + fn}$	Measures the fraction of positive instances that are correctly predicted
Specificity	$\frac{tn}{tn + fp}$	Measures the fraction of negative instances that are correctly predicted
Precision (p)	$\frac{tp}{tp + fp}$	Measures the positive instances that are correctly predicted from the total predicted instances in a positive class
Recall (r)	$\frac{tp}{tp + tn}$	Measures the fraction of positive instances that are correctly predicted
F-Measure	$\frac{2 * p * r}{p + r}$	Represents the harmonic mean between recall and precision vales
AUC	$\frac{S_p - n_p\,(n_n\, + 1)/2}{n_p n_n}$	Reflects the overall ranking performance of a classifier

Note: S_p is the sum of all the positive instances ranked, while n_p and n_n denote the number of positive and negative instances respectively.

2.3.2 Interpretable Artificial Intelligence

2.3.2.1 Overview of Interpretable Artificial Intelligence

Machine learning has shown considerable promise for enhancing products, processes, and research. However, ML models do not provide an explanation for their predictions, which

hinders the adoption of ML. Models are interpretable when people can clearly comprehend the reasoning behind the model's predictions and decisions.

Many recent studies started to shift their focus on the interpretation of machine learning models. This area of research is often called "interpretable machine learning.", "explainable machine learning", or "explainable A.I."(Linardatos et al., 2020) Users are able to verify the model and determine if it meet the expectations. Additionally, users can discover knowledge, justify predictions, and finally improve the performance of the models (Molnar et al., 2020) by using interpretable machine learning methods. Therefore, interpretable ML improves the trust and usability of classifiers.

Over the years, researchers have developed two major model interpretability techniques, including model-agnostic and model-specific approach(Molnar, 2019). The model-agnostic approach refers to techniques that approximate any black-box machine learning model with an interpretable model to explain each prediction(Linardatos et al., 2020). Model-agnostic methods include partial dependence plot, Independent conditional expectation (ICE) plots, local interpretable model-agnostic explanations (LIME), permutation feature importance, Shapley values, etc.(Molnar, 2019) Model-specific explanation methods, on the other hand, only focus on interpreting specifically to a single model or group of models, such as tree ensembles, artificial neural networks(Molnar et al., 2020). Open-source software with implementations of various interpretable machine learning methods is also available such as SHAP (Scott Lundberg, 2018), Eli5(Mikahail Korobov & Konstantin Lopuhin, 2017), InterpretML (InterpretML Team, 2021), etc. and these tools are proved capable in handling various types of tasks including image classification, text classification. Two popular model-agnostic methodologies, LIME and Shapley, are introduced in the sections below.

2.3.2.2 LIME

LIME, proposed in 2016 by Ribeiro et al.(2016), belongs to the family of a local model-agnostic method, a type of interpretable AI method. It is used to explain individual predictions of black-box machine learning based on a surrogate model, which is trained to approximate the predictions of the underlying black-box model (Guestrin, 2016; Ribeiro et al., 2016). The intuition of LIME is based on the idea that a black box model's behavior can be learned via perturbing the input. Specifically, a modified dataset generated by LIME through permutation by removing word features, corresponding to which predictions are obtained from the black-box model. Words feature weights greater than 0 indicate that the removal of such words will affect the prediction result. For a negative case, no non-zero weight was estimated because no matter which word is removed, the predicted evaluation result remains the same. Thus, an explanation can be generated by approximating the underlying model by a more interpretable one (such as a linear model, decision tree), learned on perturbations of the original instance locally(Guestrin, 2016). Due to the local fidelity nature of LIME, LIME does not guarantee a good global approximation(Molnar, n.d.). One critique LIME often receives is that it lacks "stability"(Zafar & Khan, 2021). There are cases in which the surrogate model built by LIME can predict the instance correctly but provide wrong reasons(Guestrin, 2016).

2.3.2.3 SHAP

SHAP, short for Shapley Additive ExPlanations, is another method to explain individual predictions. It is proposed by Lundberg & Lee (2017) , providing another approach to adding model interpretability based on the game theory concept of Shapley values. The goal of SHAP is to explain the prediction of an instance by computing the contribution of each feature to the prediction (Molnar, 2022). It computes Shapley values form coalitional game theory. The feature

values of a data instance act as a player in a coalition. Shapley values guarantees a fair distribution of contribution for each of the players to the "payout", which is the prediction among the features in a ML task (Molnar, 2022; nayak, 2019). A play can be a single feature or a collection of features. Unlike LIME, SHAP is not concerned with the difference between the prediction with and without a feature. It reflects the contribution of a feature or group of features to the difference between the actual forecast and the mean prediction (nayak, 2019). SHAP gives mathematical assurances for the accuracy and consistency of model explanations, making it more theoretically stable (Dibia, 2020). However, the SHAP value calculation is highly time-consuming due to the exhaustive examination of all potential feature combinations (nayak, 2019).

2.3.2.4 Evaluation of Interpretability

There is no consensus established on the definition of interpretability of ML models, especially in the context of automatic health misinformation classification. In the study by Doshi-Velez & Kim (2017) , interpretability is defined as " the ability to explain or to present in understandable terms to a human." Interpretability is subjective, as people's perspectives of the interpretability of an ML-based system differ dependent on their personal thoughts or ideas while considering and conveying facts. Doshi-Velez & Kim (2017) proposed three main levels for the evaluation of interpretability including application level evaluation, human level evaluation and function level evaluation.

Both application-level and human-level evaluation methods require human involvement. The primary distinction is that the application level evaluation is undertaken by end users, whereas the human evaluation is conducted by laypersons, who don't have domain expertise. Function level evaluation does not require people. Instead, it uses a proxy for interpretation

quality. This approach is less costly but it is most effective when it is conducted after a human-level evaluation.

2.4 Related Work on Combating Health Misinformation

Numerous attempts have been made to combat health misinformation. Employing specialists and journalists who undertake the process of manually picking and debunking false, unfounded health claims in online news and social media is a conventional approach that has been adopted by leading organizations, including professional health organizations, governmental health departments, research/academic institutes, and third-party fact-checking organizations. Depending on how heavily human intelligence and labor are involved in anti-misinformation activities such as analyzing, evaluating, and debunking false health claims, existing work can be classified into two broad categories: human-based approaches for combating health misinformation and AI-powered tools for automatically classifying health misinformation.

2.4.1 Related Human-Based Countermeasures for Health Misinformation

2.4.1.1 Fact-checking

Fact-checking is a type of human-based approach for combating misinformation that involves working directly with the information. The purpose of fact-checking work is to give an accurate, unbiased examination of statements made in public to rectify public misconceptions and raise knowledge of key topics by experienced analysts who are thoroughly familiar with the subject background(Elhadad et al., 2020). Some fact-checking websites typically analyze claims and assign them a binary rating, e.g., true/false adopted in (Lisa Lockerd Maragakis & Gabor David Kelen, n.d.) or multicategory rating, e.g., the rating provided in (Our Fact Check Ratings,

Explained, n.d.). The context and background information provided by fact-checkers can help to clarify their assessment of a claim.

In healthcare, Quackwatch.org and Snopes.com are among the most influential fact-checking websites. Quackwatch.com is a United States-based website, self-described as a "network of people," which aims to "combat health-related frauds, myths, fads, fallacies, and misconduct" and to focus on "quackery-related information that is difficult or impossible to get elsewhere"(Kreidler, 2019). The site provides a list of archived health debunking articles and assessments organized alphabetically on topics such as Cancer Treatment Watch and Device Watch. Snopes.com is another fact-checking website created to report on urban legends. It covers a variety of topics but has entries on health care. For example, it confirmed the truthfulness of Galleri trial, a cancer-detecting blood test that was claimed to be the world's first trial by the National Health Service in England and justified its review result by offering a detailed origin description and list of sources. Similarly, health claims such as "marijuana kill cancer" were rated false, and the source of the rumor was tracked down. In the case of COVID-19, medical and health-related professional organizations, academic/research institutions, federal government such as, WHO (Mythbusters, n.d.), John Hopkins Medicine (Lisa Lockerd Maragakis & Gabor David Kelen, n.d.), the CDC (CDC, 2021) which are typically considered as the officially sanctioned sources of bona fide accurate information have also taken an active role in myth debunking.

2.4.1.2 Criteria-based Assessments

Criteria-based assessments are another group of work focused on assisting readers to assess the quality of online health information. Compared to the fact-checking website, this approach neither directly works with information nor serves to inform readers for news veracity

check. It rather aims to improve the public's critical thinking about health news or its source with well-developed instruments. The Health on the Net Foundation's HONcode (Team HON, n.d.) and the DISCERN instrument (Charnock et al., 1999)are two of the most widely cited quality evaluation tools. The HONcode is a set of eight criteria used to certify websites containing health information. The eight criteria include authoritative, complementarity, privacy, attribution, justifiability, transparency, financial disclosure, and advertising policy, with which readers can systematically analyze the reliability of health-related websites (Team HON, n.d.) . The DISCERN instrument is a sixteen-item questionnaire designed specifically for evaluating health information on treatment options, and it has been shown to have good inter-rater reliability and content validity. While some subjectivity is required for rating certain criteria, the findings show that experienced users can use the instrument, providers of health information, and patients to differentiate between high- and low-quality publications (Charnock et al., 1999). The HealthNewsReview.org website also recommends a different set of criteria. The rating instrument used includes ten criteria used by the Australian and Canadian Media Doctor sites, and its inter-reviewer reliability was tested using a random sample of 30 stories (Schwitzer, 2008) . Although it has received less attention than the other two, it was one of the few programs that provided health newsreaders with review cases to demonstrate how to apply the criteria to assess health news critically. HealthNewsReview.org monitored health news coverage by the top 50 most widely circulated newspapers in the U.S., such as ABC, CBS, and NBC. For each criterion, the story was given a rating of "satisfactory," "unsatisfactory," or "not applicable". Each piece of news was scrutinized by three reviewers, each of whom has a background in medicine, health services research, public health, or journalism (HealthNewsReview.org, n.d.-b).

2.4.1.3 Discussion of Human-based Countermeasures for Health Misinformation

Human-based approaches for combating health misinformation guarantee a high-quality review of the news or source and, consequently, certify accurate, reliable medical and health information. However, this approach is expensive in terms of the labor force, money, and time due to the massive volume of available data on social networks (Elhadad et al., 2020). As a result, organizations that rely on human fact-checkers or reviewers are often put in a difficult financial position. HealthNewsReview.Org project was shut down in 2018 due to a big funding cut(HealthNewsReview.org, n.d.-a). Snopes.com also expressed a shortage of funding as "the news and fact-checking industry is currently experiencing a downturn" (Team Snopes, n.d.). To remain efficient while the growth of online documents is accelerating, the manual review process needs to be aided, complemented, or systematically executed by automated means. Compared to the traditional fact-checking approach, artificial intelligence can automatically take a cue from articles annotated as inaccurate by domain experts. Thus, an AI-powered fake news detector takes less labor and time to separate untruthful information from legitimate information. Due to its low cost and wide availability, it has the potential to reach a huge section of a target community with minimal effort.

2.4.2 AI-based Countermeasures for Health Misinformation

2.4.2.1 Overview of AI-based Countermeasures for Health Misinformation

An increasing amount of research focuses on developing misinformation detectors to filter out unreliable and false information. The majority of techniques for accomplishing this task have been developed in the field of artificial intelligence (A.I.), primarily through the use of natural language processing (NLP) and machine learning (ML) methods. Up to this point, numerous automated systems have been proposed to weed out general misinformation. The

health topics had not been extensively incorporated in the fake news detection studies until 2019 when attention suddenly turned to detect COVID-19-related misinformation. In contrary to political and general misinformation, fact-checking or evaluating health-related information requires specific knowledge on the pertinent health topics(Kotonya & Toni, 2020). Several approaches have been proposed to detect false, misleading health news on platforms such as Twitter (Elhadad et al., 2020; Ghenai & Mejova, 2017; Khanday et al., 2021; Shah et al., 2019), websites (Dhoju et al., 2019; Meppelink et al., 2021; Saengkhunthod et al., 2021), and online forums (Kinsora et al., 2017; Parfenenko et al., 2020). Setting appropriate benchmarks for each data instance to be evaluated and annotated is unavoidable in all research when developing a detection system. Thus, based on the benchmark and objectives of the study, the retrieved research can be summarized and classified into two categories: veracity-based approaches for detecting health misinformation and criteria-based approaches for assessing health information quality.

2.4.2.2 Veracity-based System for Fake Health News Detection

Veracity-based approaches for detecting health misinformation accounts for the greater part of the literature on the automated approach on classifying misleading health news. This approach means the classifiers are trained based on datasets that are assessed and annotated by domain experts who possess the knowledge to evaluate each health-related claim or article, and the review output will conform to a certain level of facts and accuracy. Therefore, target variables usually are expressed with terms that imply the truthfulness of the information. Exemplar classification targets include but are not limited to “true/false”, “misinformative/ not misinformative”, “rumor/ not rumor”, “conspiracy/ not conspiracy”,” propaganda/ not propaganda”.

Among the retrieved empirical studies for review, a transition of health topics from miscellaneous conditions starting in 2016 to covid-19 in the most recent two years can be clearly observed. In February 2016, the World Health Organization declared the Zika outbreak a Public Health Emergency of International Concern. Since then, attention has been paid to analyze and detect health misinformation on multiple online venues. Ghenai & Mejova (2017) proposed a novel tool pipeline that combined health professionals, crowdsourcing, and machine learning to capture health-related rumors from around the world, as well as clarification campaigns from reputable health organizations. The model was built on a collection of 13 million tweets – spanning the initial reports in February 2016 and the Summer Olympics – about the Zika outbreak, as well as rumors outlined by the World Health Organization and the Snopes fact-checking website. The study discovered extremely burst behavior of rumor-related topics and demonstrated the feasibility of using automated techniques to remove rumor-bearing tweets when the questionable topic was detected. Parfenenko et al. (2020) focused on online forums instead and proposed a neural network model to classify articles about medical care into true and misinformative. During the dataset construction, the most frequently discussed health issues in the forum, such as "anemia," "diabetes," "vaccines," and "cancer," were compiled. The study compared the developed model with the state-of-the-art neural models Recursive Neural Tensor Network (RNTN) and Long short- term memory (LTSM) network with the accuracy of claim classification into true and misinformative of an 88.4% and a 90.5%, respectively which proves the effectiveness of the proposed model. The effectiveness of the proposed model was proved with an accuracy of 91%, corresponding to a 92,6% precision and 89.5% recall for misinformative instances.

In 2020 and 2021, there is a spike in the publication on Covid-19 related misinformation detection. To identify and classify four popular COVID-19 related conspiracy theories on Twitter, Gerts et al.(2021) utilized a random forest classification technique to classify conspiracy theories from a corpus of 120 million tweets. To describe COVDI-19 conspiracy beliefs over time, downstream sentiment analysis was also employed to characterize the linguistic feature. The study found that classifier performance varies with the theme of conspiracy and improves with higher specificity on topics. The four conspiracies studied had varying random forest classifier metrics (F1 scores between 0.347 and 0.857). Raju et al.(2021)introduced a new model with high classification accuracy to extract deep contextual information from online coronavirus comments based on main COVID-19 topics. A robust model for sentiment classification based on more than twenty different datasets was incorporated to detect the tweet's text, which contains misinformation. The proposed approach can create alerts regarding misinformation-containing tweets for states that need to take emergency measures to stop the virus from spreading by removing comments from their internet platforms. The study also demonstrates that using ALBERT to create sentiment ratings for COVID-19 tweets results in the final classifier's high accuracy of 0.91 in detecting falsehood. Khanday et al. (2021) used a variety of machine learning algorithms to categorize tweets as propaganda or non-propaganda. The data was obtained from Twitter and manually classified as either propaganda or not. Three textual features (TF/IDF, bag of words, and tweet length) are combined to accomplish hybrid feature engineering. The decision tree classifier outperformed all other machine learning algorithms, achieving 98.5 percent accuracy, 0.99 precision, 0.99 recall, and 0.99 F1- Score.

2.4.2.3 Criteria-based System for Health News Quality Classification

Criteria-based approaches for assessing health information quality look at misinformation based on various criteria predefined by the research. For instance, “intuitively, a news article published on the unreliable website(s) and forwarded by the unreliable user(s) is more likely to be fake news than news posted by authoritative and credible users"(X. Zhou & Zafarani, 2020). News that does not satisfy certain items in an assessment checklist for health information quality can be considered untrustworthy. In other words, this approach focuses on the characteristics of the news content, but the result does not confer the veracity of information. The target variable defined for classification in the studies typically are named "credible/not credible," "reliable/not reliable," "trustworthy/not trustworthy."

Some studies set the criteria by utilizing the sources of health information to indicate reliability. For example, (Y. Liu et al., 2019) predefined a list of reliable and unreliable websites from which health-related articles from various sources on Chinese Internet society are extracted for dataset construction. Experiments are carried out based on this dataset: machine learning classifiers using manually extracted features and text classification model FastText are tested, among which GBDT achieved the best performance with a precision of 0.8374. In a similar study carried out by Saengkhunthod et al.(2021), reliability was defined as “a reliable article must be written by experts and verifiable," based on which the team collected samples of 297 reliable and 235 unreliable articles from 7 websites in Thailand. 20 features were used in machine learning to classify the articles into reliable and not reliable class. Experimental results show that XGBoost methods were the most effective at 90.60% accuracy. Dhoju et al.(2019) defined that outlets which have been cross-checked as reliable or unreliable by credible sources. They identified 29 reliable media outlets from three sources– (1) 11 of them are certified by the

HONcode. (2) 8 from the U.S. government's health-related centers and institutions (e.g., CDC, NIH, NCBI), and (3) 10 from the most circulated broadcast mainstream media outlets (e.g., CNN, NBC). With the dataset constructed on identified sources, the team leveraged semantic patterns and built classification models to identify a health-related news article's source (reliable or unreliable) with an F-measure of 96%.

Some studies used a multi-item checklist to assess the credibility and trustworthiness of health information. Shah et al.(2019) used a 7-point checklist adapted from validated tools and guidelines to manually appraise the credibility of 474 Web pages after sampling from 143,003 unique vaccine-related Web pages shared on Twitter between January 2017 and March 2018. Using the dataset, they trained several classifiers (random forests, support vector machines, and recurrent neural networks) to predict whether the information meets each of the seven criteria. They used the follower network to estimate potential exposures relative to a credibility score defined by the 7-point checklist when estimating the credibility of all other Web pages. According to studies, the best-performing classifiers could distinguish between low, medium, and high credibility with an accuracy of 78% and labeled low-credibility Web pages with a precision of more than 96%. The seven criteria used in the study are partially adopted from DISCERN and QIMR checklists; two additional criteria specific to vaccine communities were also added. The checklist consists of (1) information presented is based on objective, scientific research; (2) adequate detail about the level of evidence offered by the research is included; (3) uncertainties and limitations in the research in focus are described; (4) the information does not exaggerate, overstate, or misrepresent available evidence; (5) provides context for the research in focus; (6) uses clear, nontechnical language that is easy to understand; and (6) is transparent about sponsorship and funding. Oroszlányová et al.(2018) the impact of web document features

on their quality, using HONcode as the ground truth, to determine whether it is possible to predict the quality of a document based on its characteristics. The study examined how the characteristics of health documents (e.g., web domain, last update, type, mention of treatment, and prevention strategies) are related to their quality. Statistical models are developed based on these characteristics to predict whether health-related web documents are of certification-level quality. To determine the significance of predictors, multivariate analysis was used. Three types of full and reduced logistic regression models were constructed and evaluated, with the most informative prediction model capable of achieving an accuracy of 89%.

2.4.2.4 Discussion of AI-based Countermeasures for Health Misinformation

Machine learning-based methods have been proved to be effective in classifying health misinformation. Compared to human-based work, a machine learning model consists of an algorithm that can learn latent patterns and relationships from data spontaneously without hard-coding fixed rules(Sarkar, 2018). However, one major challenge is interpreting how a model works when performing a classification task (Linardatos et al., 2020).

Up to date, only a small body of research has incorporated explainable functionality in their modeling among the AI-powered models for combating misinformation(Ayoub et al., 2021; Kotonya & Toni, 2020). All those studies on health information classification are veracity-based. There remains an unknown area for constructing an interpretable criteria-driven classification system to help users evaluate the quality of health information while understanding how users' trust for the model comes into play.

CHAPTER III: METHODOLOGY

3.1 Overview of Chapter III

The study consists of three components, illustrated in figure 1. In the first component, we collected reviewed health news from HealthNewsReview.org (HealthNewsReview.Org, n.d.) to build the dataset for modeling. The second component is a supervised document classification task that automates the criteria evaluation process. Each health news article will be categorized automatically at the document level using established criteria, and the output will be binary (Satisfactory/Unsatisfactory). “Satisfactory” means the entire health news meets the given criterion, and "Unsatisfactory" means the opposite.

The last component is to visualize and interpret the evaluation result provided by the health news quality evaluation system. The goal is to highlight relevant sentences that could serve as visual evidence to justify a satisfactory evaluation result. For example, for the criterion “Does the news adequately explain/quantify the harms of the intervention?”, the method should highlight sentences that describe the harms of intervention to help users quickly understand how well the criterion is being met.

Two approaches were used to achieve this goal. The first one is a hybrid approach (the Hybrid Approach). It is inspired by principles from rule-based systems, where experts specify patterns by hand. The second approach (the Typology Approach) is a supervised sentence typology classification method, where hand-labelled training data is analyzed algorithmically to build models that can detect similar patterns when applied to unseen data.

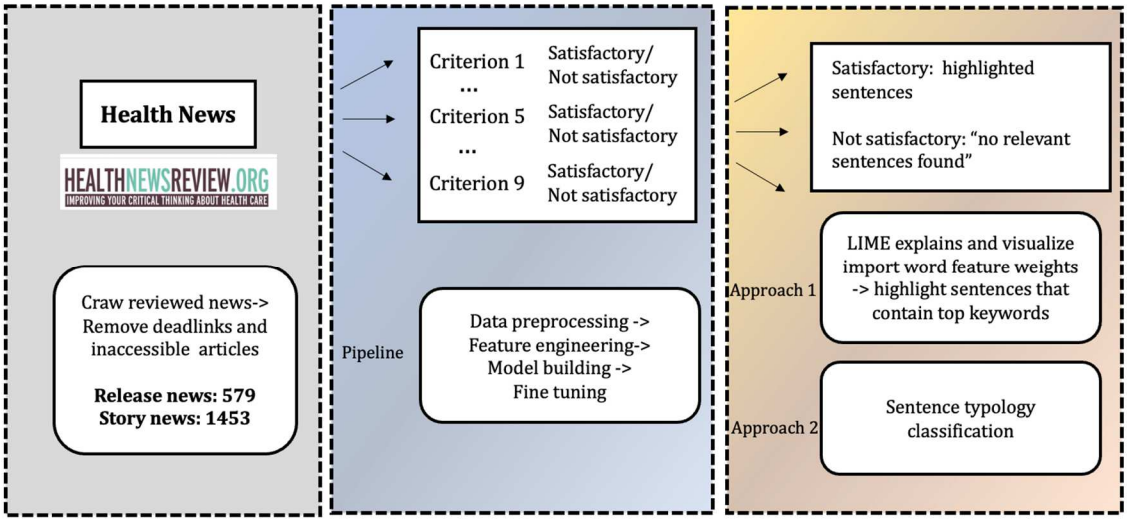

Figure 2. Overview of the study

3.2 Dataset Construction

The dataset used in the study was adapted from an existing resource created by HealthNewsReview.org(HealthNewsReview.Org, n.d.). For the study, the health story news reviews and news release reviews, as archived by HealthNewsReviews.org were crawled complying with the robots.txt. Then news contents which corresponded to the acquired reviews were also scraped to construct the initial dataset.

3.3 Automating the Criterion Evaluation: Document-level Health News Quality Classification

The obtained raw dataset needed to be cleaned before modeling. All criteria apply to both news types, so two types of news content were merged, being treated uniformly. Then health news that was scored as "not satisfactory" "not applicable" were combined and named as "unsatisfactory." The study focuses on visualizing the interpretation of a satisfactory result. No evidence is needed to justify an "unsatisfactory" or "not applicable" result because these assessments already imply a lack of evidence to support a satisfactory result. Multiple text processing techniques were applied to preprocess all news contents, including non-word

elements (numbers, assented characters, punctuation) removal, stop words removal, tokenization, stemming, and lemmatization. The textual representation was then converted into a vector space model using term frequency-inverse document frequency (TF-IDF).

Four representative algorithms, Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest algorithms (RF), were chosen for the study. The four algorithms are commonly used in health misinformation classification tasks and found effective. Over-sampling technique was adopted to address class imbalance issue. For model selection and hyperparameter tuning, Random Search was selected over Grid Search due to work efficiency considerations. Random Search was applied to determine the optimal model hyperparameters for building the classifier. For this study, the best classifier was defined as the output from the Random Search is the feature count, hyperparameter, and algorithm combination that produces the highest mean 5-cross-validated AUC score. The performance of the classifier was further evaluated through 10-repeated 10-fold-validation. Precision is also considered as another evaluation metric because a false negative (a false unsatisfactory result) is less of a concern. It is more crucial for the model to accurately identify the highest possible proportion of the positive examples that are suitable for further interpretation.

3.4 Visualizing the Interpretation of Evaluation Result

Two approaches were used to visualize the interpretation of evaluation results. The desired outcome is that all highlighted sentences are relevant to the examined criterion and provide supporting evidence to assist end-users in comprehending and validating the evaluation result. To determine what qualifies a sentence as evidence, the study strictly adhered to the criteria definitions and review guidelines provided by the HealthNewsReview.org(HealthNewsReview.org, n.d.-b) study. For example, as per the

explanation of the Harm criterion provided by HealthNewsReview.org., satisfactory health news on the Harm criterion should “include a discussion of harms and side effects, as well any measured “adverse events” in a study (Criterion #3 Does the Story Adequately Explain/Quantify the Harms of the Intervention?, n.d.) . The measured "adverse events" can be addressed by a discussion of “ both frequency of side effects and severity of side effects” and a discussion of “both major and minor side effects.”(Criterion #3 Does the Story Adequately Explain/Quantify the Harms of the Intervention?, n.d.)

3.4.1 The Hybrid Approach

The Hybrid approach is built on LIME, a popular interpretable AI technique. As mentioned in chapter ii, one critique that LIME has received is that it lacks stability in its interpretability. There are cases in which the surrogate model built by LIME can predict the instance correctly but provide wrong reasons(Guestrin, 2016). This study addresses the instability of LIME by adding manually selected keywords which reduces the risk of obtaining wrong keywords for highlighting through LIME alone. The python packages used for implementing LIME algorithms are ELI5(Lopuhin, n.d., p. 5) and LIME(Ribeiro, n.d.) API packages.

The explanation of the classification model of each criterion by the hybrid approach consists of three steps. First, a machine learning classifier classified health news into satisfactory and unsatisfactory on the chosen criterion. Then, the classification model learned the word distribution in satisfactory or unsatisfactory instances from the collection of health news document sets. LIME highlights keywords in texts that contribute to the prediction. The keywords were also ranked using a weight score indicating their contribution to the prediction. It is worth noting that the weight is a general term used to describe the weight of each feature’s

contribution to the prediction result. The actual meaning of the term and how it is dispatched to a concrete implementation based on the classifier type. For example, the word feature weights in a classification task achieved by RF algorithm signify how much discriminatory information each word contributes to the classification, but not specific to the class.

Lastly, keywords that contributed to a satisfactory prediction were combined with a list of manually picked keywords as shown in table 2. The manual selection of the keywords was based on a consensus among the annotators who had taken part in the processes of evidence extraction for the typology approach.

Table 2. Lists of manually selected keywords for the criteria

Criterion	Manually selected keywords
Cost	price, cost, charge, insurance, pay
Benefit	improve, better, effective, increase
Harm	side effect, adverse reaction, adverse event, complication, risk
Quality	journal, peer review, publication
Mongering	prevalence
Conflict	fund, sponsor, grant, spokesman, professor, director
Alternative	alternative, similar
Availability	available, market, store, sell, sale
Novelty	new, first, novel

Then, the highlighting was extended from the keyword to the sentence level to enhance the final visual representation. The sentence containing keywords with more weights was prioritized for highlighting. By default, manually selected keywords outweigh any keywords automatically picked by LIME.

3.4.2 The Typology Approach

The typology approach relies on a sentence-level text classification task. This approach is inspired by well-regarded studies in persuasive communication and rhetoric. REYNOLDS & REYNOLDS (2002) distinguished between statistical, testimonial, anecdotal, and analogical evidence; (Hoeken & Hustinx (2003) put forward four types of evidence in argumentation: individual examples, statistics, causal explanations, and expert opinions. Later studies have shown that machines can detect various typologies of evidence. For example, Fiok et al. (2021) built a classification model to automatically identify the evidence of respect in the communication through Twitter. There are two types of such sentences in the health news articles used in this study. In the case of the Harm criterion, the first type of sentences is the evidence that supports the predicted evaluation result. Sentences of this type contain a description of side effects, including the symptoms, severity, frequency of the symptoms. The other type of sentence refers to those that cannot justify why a piece of certain health news satisfies a given criterion. Thus, they are not characterized as evidence.

To implement the Typology approach, for each criterion's task, the typology approach was designed and conducted in two stages. The first stage is to build an annotated dataset of sentence evidence. Sentence evidence was extracted from health news that was evaluated to be satisfactory by HealthNewsReview.org. Three people performed the sentence extraction tasks. Training and annotation rules were completed prior to the manual annotation. The sentence extraction guideline was a full adoption of the criteria explanation by HealthNewsReivew.org (HealthNewsReview.org, n.d.-b). Two people undertook most of the extraction work. Another person worked as an independent reviewer to resolve the disagreement. When combining the extracted sentences, sentences that were picked by two extractors are characterized as evidence.

If a sentence was extracted by one extractor, but not picked by the other, an independent third person was invited to resolve the disagreement. All approved sentences were considered the positive class. To build the negative class, the same number of sentences that are irrelevant to the evaluation of the pertinent criterion were randomly selected. Inter-annotator agreement was assessed using both simple counts and the percentage of the final quantity of the evidence in total extracted items, to address the relatively small size of the sample. These measures of inter-rater agreement are also in line with expectations in studies such as (Freund & Giabbanelli, 2021). The second stage was to build a supervised machine learning classifier. The same steps were followed for automating the criterion evaluation.

For the final visual representation, the sentence classifier was applied to health news contents to identify sentence evidence. The sentence having the highest probability of being categorized as evidence by the classifier was prioritized for highlighting.

3.4.3 Evaluating and Optimizing Two Approaches

For each criterion's interpretation, two visualization approaches were evaluated to see how accurately each scheme highlighted sentences used as evidence to support the prediction result. The evaluation was conducted with 20 test cases. The selection of 20 test cases is based on the observation that the true positive health news counts in the test set (30% of the dataset) range from 20 to 70, depending on the task's criterion type. The accuracy of two highlighting schemes was measured by calculating the percentage of correctly highlighted evidence in all highlighted sentences. Three people evaluated the correctness of the highlighted sentence in accordance with each criterion's guideline. An independent reviewer was also invited to handle any disputes.

As the number of highlighted sentences may affect the highlighting accuracy and thus the final visual representation, spectrum of accuracies of both highlighting approaches were calculated when the number of highlighted sentences was increased from 1. A threshold was then selected as the lowest accuracy to determine the optimal range of sentence count for highlighting.

3.5 Evaluating the Interpretable, Criteria-driven System for Health News Quality Evaluation

For the evaluation of interpretability, a human-level evaluation in the form of a user study was conducted. The study was carried out with laypersons as the initial end-user of the proposed system is the general public. A between-subject user study was undertaken to evaluate the interpretability of the proposed system: an AI-based, Criteria-driven, Explainable system (the ACE System), using an AI-based, Criteria-driven, Unexplainable System (the ACU system) as a comparison. This study examines whether the system's interpretability facilitates users' trust in the evaluation of health news quality. The main hypothesis is that end-users would have more trust in the ACE system for evaluating health news compared to the ACU system.

Study participants were recruited using a survey distribution platform called SurveySwap (Find Survey Participants Today | SurveySwap, n.d.). During the user study, participants were provided with a link to an interactive prototype where users can check health news evaluation results as well as other functionalities without any supervision. The interactive prototype used in this user study is an early sample built to test the concept of automatic evaluation of health news quality and assess how well potential users would trust the system when provided different conditions (e.g., news of different quality levels, the system with/without explanation). It was made using an interface design tool named Figma (Figma, n.d.) where designers can design an

early model of the system or application with vector graphics, animations, and interactive widgets.

All participants were assigned to six groups randomly: three were study groups and the other three were control groups. Participants in study groups were provided with the ACE system prototype tested with health news of three quality levels (High, Medium, and Low). In this study, the high-, medium-, and low-quality news were defined as health news that meets 1-3, 4-7, and 8-10 number of criteria, respectively. Similarly, participants in control groups were provided with the ACU system prototype tested with health news of three quality levels. The only difference between ACU and ACE systems is whether there are explanations for the results. The tested news articles of three quality levels all describe different types of treatment for a disorder related to women's health (hot flashes). The respective titles for the low-, medium-, and high-quality news are "Survivors' electro-acupuncture for disrupted sleep in women with breast cancer." "You are getting cooler: Hypnosis works for hot flashes, study finds," and "Incontinence drug may cut hot flashes in breast cancer survivors." Figure 2 depicts the prototype of ACE system applied to high-quality health news. Upon the completion of prototype testing, all participants were directed to complete a survey.

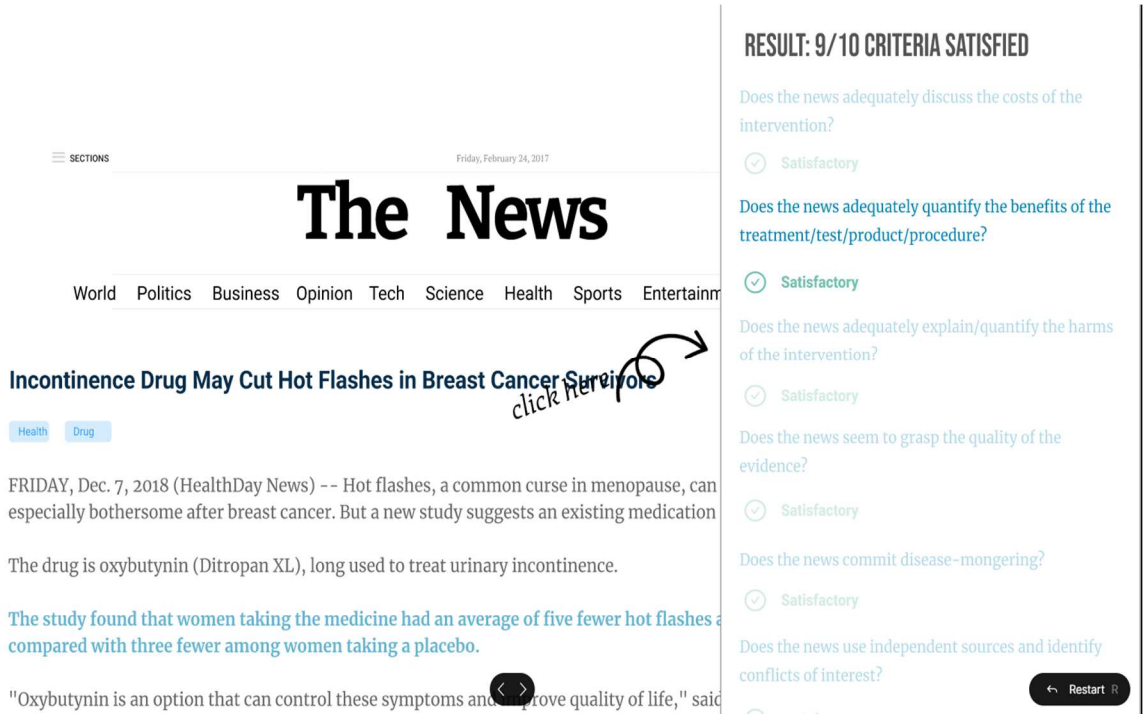

Figure 3. A snapshot of the prototype of ACE system applied to high-quality health news

There are three sections in the survey: demographic information, general trust in technology and health news sources, and users' trust in the ACE or ACU system. Questions to assess users' trust in the evaluation result were developed using a 5-point Likert scale. The survey also includes an open-ended question for participants to express their comments on other potential approaches to addressing health misinformation. Qualtrics (Qualtrics XM // The Leading Experience Management Software, n.d.), which can be integrated into SurveySwap, was used for deploying the survey. The survey questions can be found in Appendix A. The inclusion criteria for completing the survey were that participants be 18 years old and above. Reponses provided by participants who did not answer the attention question correctly were not included in the analysis.

The study obtained an Institutional Review Board (IRB) approval (IRB No.22.303) from the University of Wisconsin, Milwaukee before embarking on the user study. For the data

analysis, a chi-square test was performed to examine if there is a statistically significant relationship between the group type and any of demographic variables. This aims to identify any difference between the study group and the control group regarding the distribution of demographic data. An independent-sample T test model was then used to analyze the whether there is a statistically significant difference between the means of user trust in the system's evaluation result in two user groups. A significance level of α was set as 0.05.

CHAPTER IV: RESULT

4.1 Overview of Chapter IV

This chapter presents the primary results of the experiment. The reporting of the result consists of several components: the descriptive statistics of the dataset, the performance of the document-level classification of automatic health news quality, the performance of the interpretation schemes of two proposed approaches, the optimal sentence evidence after optimizing the interpretation approaches, user's trust in the proposed system. The chapter concludes with responses to the two groups of research questions.

4.2 Dataset Description

According to table 3, the exploratory analysis of the scraped data from HealthNewsReview.org reveals that the ratio of satisfactory to unsatisfactory instances varies by criterion. Cost is one of the 10 assessed criteria that the majority of health news articles failed to meet. 78% of the health news articles evaluated by HealthNewsReview.org failed to discuss the expenses of the intervention, insurance coverage, or even mention cost. The Benefit criterion is the second least satisfied criterion. 70% of health news articles did not appropriately quantify the intervention's benefits. Unsatisfactory news might have a lengthy description of how effective the intervention is but did support for the claims by data or some measure. Among news deemed

satisfactory, more than 70% of the health news met the Mongering criterion. For the conflict criterion, the number of satisfactory and unsatisfactory health news is comparable.

Table 3. Descriptive statistics of dataset

Statistics	News release		Story release		Total by criterion		Percentage by criterion	
Rating	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory	Satisfactory	Unsatisfactory
Cost	41	560	648	1952	689	2512	21.5245%	78.4755%
Benefit	150	451	807	1793	957	2244	29.8969%	70.1031%
Harm	121	480	854	1746	975	2226	30.4592%	69.5408%
Quality	168	433	928	1672	1096	2105	34.2393%	65.7607%
Mongering	494	107	1923	677	2417	784	75.5077%	24.4923%
Conflict	291	310	1262	1338	1553	1648	48.5161%	51.4839%
Alternative	241	360	1122	1478	1363	1838	42.5804%	57.4196%
Availability	329	272	1676	924	2005	1196	62.6367%	37.3633%
Novelty	367	234	1808	792	2175	1026	67.9475%	32.0525%

4.3 Performances of the Classification Model for Health News Quality Evaluation

After removing dead links (to inaccessible news contents), the acquired dataset yielded 1453 story news and 579 release news for training classifiers. Among all 2032 health news instances, the satisfactory/unsatisfactory instance ratio for each criterion is shown in Table 4.

Table 4. The dataset for modeling after data preprocessing

Criteria	Satisfactory	Unsatisfactory
Cost	405	1618
Benefit	634	1389
Harm	625	1398
Quality	716	1307
Mongering	1650	373
Conflict	1002	1021
Alternative	906	1117
Availability	1272	751
Novelty	1397	626

The table shows that the dataset of each criterion has an imbalanced class distribution. The conflict and the alternative datasets are relatively balanced.

Of four experimented algorithms, the base RF was found most effective in automating the assessment of 8 of the 9 criteria as shown in Appendix B. The word feature count of 2000 was found effective for classifying the Cost, Benefit, Harm, Availability and Novelty. The basic LR was found suitable in classifying satisfactory and unsatisfactory news on the Benefit criterion for producing the highest AUC (0.71) and Precision (0.70) scores. Table 5 shows the set of optimal hypermeters that Random Search picked for each criterion's classifier.

Table 5. Hyperparameters selected by Random Search for each criterion evaluation classifier

Criteria	Base Classifier	Word Feature Count	Hyperparameters
Cost	RF	2000	{'n_estimators': 1600, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 30, 'bootstrap': False}
Benefit	LR	2000	{'C': 1.021551661021462e-05, 'max_iter': 500, 'penalty': 'l2', 'solver': 'lbfgs'}
Harm	RF	2000	{'n_estimators': 1800, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': False}
Quality	RF	4000	{'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': False}
Mongering	RF	500	{'n_estimators': 600, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 60, 'bootstrap': True}
Conflict	RF	500	{'n_estimators': 1800, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 30, 'bootstrap': True}
Alternative	RF	4000	{'n_estimators': 2000, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20, 'bootstrap': True}
Availability	RF	2000	{'n_estimators': 2000, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 90, 'bootstrap': False}
Novelty	RF	2000	{'n_estimators': 800, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 100, 'bootstrap': False}

The Precision and AUC for each of the nine criteria is given in Figure 4. The measures range from highest (AUC, Precision) values of (0.89, 0.82) for Cost down to lowest values of 0.61 for AUC (Novelty) and 0.60 for Precision (Alternative).

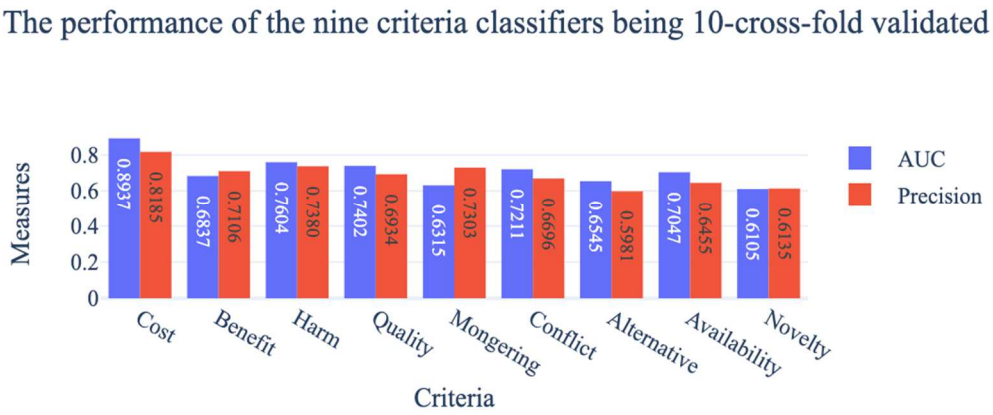

Figure 4. The performance of the nine criteria classifiers being 10-cross-fold validated with the LR algorithm for the Benefit and RF algorithm for the rest of criteria

4.4 Performances of Two Approaches for Mode Interpretation

4.4.1 The Visual Interpretation by the Hybrid Approach

LIME Text Explainer visualized how different word features that contribute to the evaluation result in each classifier. Figure 5 illustrates the top 30 bigram or unigram word features that contribute to the classification learned from the entire dataset related to a given criterion, for Cost, Harm, and Conflicts. For example, the binary word feature with the highest weight in the Harm criterion classification is "side effect". Words that directly indicate the harm of intervention, such as " risk," "concern," "bleeding," "harm," also ranked among the top features (at 3, 4, 16, & 28 respectively).

Weight	Feature
0.1070 ± 0.2493	cost
0.0289 ± 0.0956	price
0.0236 ± 0.0823	insurance
0.0194 ± 0.0686	expensive
0.0153 ± 0.0622	pay
0.0147 ± 0.0532	company
0.0132 ± 0.0409	say
0.0121 ± 0.0441	doctor
0.0112 ± 0.0367	year
0.0088 ± 0.0443	food drug
0.0085 ± 0.0361	last
0.0083 ± 0.0317	make
0.0078 ± 0.0372	drug administration
0.0074 ± 0.0246	said
0.0068 ± 0.0326	approval
0.0066 ± 0.0343	administration
0.0066 ± 0.0257	would
0.0065 ± 0.0338	last year
0.0057 ± 0.0214	much
0.0050 ± 0.0218	text
0.0049 ± 0.0249	though
0.0048 ± 0.0280	sale
0.0048 ± 0.0166	one
0.0048 ± 0.0271	mr
0.0046 ± 0.0229	approved
0.0046 ± 0.0208	several
0.0045 ± 0.0249	fda
0.0044 ± 0.0209	still
0.0044 ± 0.0178	three
0.0044 ± 0.0199	get
... 970 more ...	

Weight	Feature
0.0178 ± 0.0545	side
0.0166 ± 0.0512	side effect
0.0082 ± 0.0224	risk
0.0064 ± 0.0225	concern
0.0063 ± 0.0179	percent
0.0059 ± 0.0219	serious
0.0059 ± 0.0215	drug
0.0056 ± 0.0194	expert
0.0053 ± 0.0160	research
0.0051 ± 0.0179	effect
0.0050 ± 0.0145	may
0.0049 ± 0.0184	review
0.0049 ± 0.0163	american
0.0047 ± 0.0133	study
0.0047 ± 0.0190	safety
0.0047 ± 0.0183	bleeding
0.0045 ± 0.0139	one
0.0043 ± 0.0181	adverse
0.0043 ± 0.0202	food drug
0.0042 ± 0.0130	say
0.0042 ± 0.0161	percent patient
0.0040 ± 0.0130	year
0.0040 ± 0.0177	approved
0.0039 ± 0.0137	many
0.0039 ± 0.0163	severe
0.0039 ± 0.0116	said
0.0039 ± 0.0160	harm
0.0038 ± 0.0161	safe
0.0037 ± 0.0159	cancer institute
0.0036 ± 0.0159	national cancer
... 1970 more ...	

Weight	Feature
0.0106 ± 0.0330	university
0.0080 ± 0.0247	say
0.0073 ± 0.0244	dr
0.0060 ± 0.0237	funded
0.0054 ± 0.0214	said dr
0.0054 ± 0.0213	involved
0.0054 ± 0.0222	national
0.0053 ± 0.0177	study
0.0052 ± 0.0191	one
0.0050 ± 0.0206	professor
0.0048 ± 0.0168	research
0.0045 ± 0.0196	foundation
0.0040 ± 0.0175	many
0.0037 ± 0.0142	also
0.0036 ± 0.0142	health
0.0036 ± 0.0149	hospital
0.0036 ± 0.0152	medical
0.0035 ± 0.0161	would
0.0033 ± 0.0123	said
0.0032 ± 0.0164	grant
0.0032 ± 0.0133	may
0.0031 ± 0.0140	common
0.0031 ± 0.0133	medicine
0.0030 ± 0.0135	institute
0.0029 ± 0.0125	cause
0.0029 ± 0.0151	supported
0.0028 ± 0.0128	show
0.0028 ± 0.0123	center
0.0028 ± 0.0138	dont
0.0026 ± 0.0139	york
... 970 more ...	

Figure 5. Top 30 word features with their feature weights in Cost, Harm and Conflict classifiers that are built on RF

Similarly, words that are commonly used to describe the intervention costs and insurance coverage such as “cost”, “insurance”, “expensive” and “pay” also appear highly ranked in their contribution for the Cost criterion (at 1, 3, 4 & 5 in the ranking). For the Conflicts criterion, the words are descriptive of one’s affiliations such as “university”, “dr”, “professor” stand out (at number 1, 3, 10). The keyword “funded”, which directly discloses funding information also ranks high (at number 4).

Figure 6 shows how word features in a LR-based classification contribute to a positive or negative class for the Benefit criterion. For instance, keywords “percentage”,” patient” and “group” contribute to a positive prediction, while “Alzheimer’s”, “knees” and “memory” are found more contributive to the negative class. The word feature weights of the rest of five criteria can be found in Appendix C, D, E, F &G.

Weight[?]	Feature
+0.000	percent
+0.000	patient
+0.000	vaccine
+0.000	group
+0.000	study
+0.000	cancer
+0.000	kidney
+0.000	survival
+0.000	device
+0.000	pain
+0.000	vitamin
+0.000	stent
+0.000	test
+0.000	screening
+0.000	treatment
+0.000	melatonin
+0.000	year
+0.000	trial
+0.000	colonoscopy
+0.000	fracture
+0.000	melanoma
+0.000	radiation
… 1055 more positive …	
… 916 more negative …	
-0.000	alzheimeraos
-0.000	knee
-0.000	memory
-0.000	cognitive
-0.000	alzheimers
-0.000	mouse
-0.000	interest
-0.000	brain

Figure 6. Top 30 word features with their feature weights in the Benefit classifier that is built on LR

Figure 7 shows how LIME performs first-level visualization on a sample health news that was rated as satisfactory on the Harm criterion. The classifier predicts the sample health news with a positive result with a 65% probability. The words marked in orange were picked by LIME and explained as they have contributed to the positive classification results of the model. Certain words were also highlighted in blue despite being scarce in numbers, indicating the likelihood of an unsatisfactory prediction. Based on the prediction result, the words 'adverse,' 'reaction,' 'risk", 'adverse,' 'serious,' and 'Administration,' were ranked among the most predictive words to the satisfactory classification result. A snapshot of final visualized representation is shown in Figure 8, after highlighting sentences containing the keywords selected by LIME and human expert.

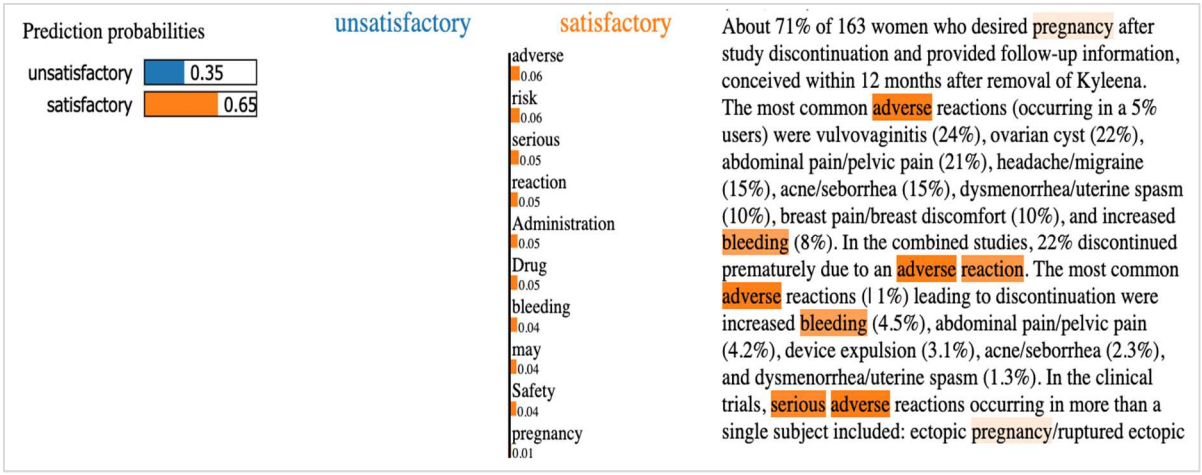

Figure 7. LIME text explainer visualizes word's contribution to a satisfactory prediction on the Harm criterion using RF Algorithm

About 71% of 163 women who desired pregnancy after study discontinuation and provided follow-up information, conceived within 12 months after removal of Kyleena.
The most common adverse reactions (occurring in a 5% users) were vulvovaginitis (24%), ovarian cyst (22%), abdominal pain/pelvic pain (21%), headache/migraine (15%), acne/seborrhea (15%), dysmenorrhea/uterine spasm (10%), breast pain/breast discomfort (10%), and increased bleeding (8%). *
In the combined studies, 22% discontinued prematurely due to an adverse reaction. *
The most common adverse reactions (> 1%) leading to discontinuation were increased bleeding (4.5%), abdominal pain/pelvic pain (4.2%), device expulsion (3.1%), acne/seborrhea (2.3%), and dysmenorrhea/uterine spasm (1.3%). *
In the clinical trials, serious adverse reactions occurring in more than a single subject included: ectopic pregnancy/ruptured ectopic pregnancy (10 subjects); pelvic inflammatory disease (6 subjects); missed abortion/incomplete spontaneous abortion/spontaneous abortion (4 subjects); ovarian cyst (3 subjects); abdominal pain (4 subjects); depression/affective disorder (4 subjects); and uterine perforation/embedded device (myometrial perforation) (3 subjects). *
Indication for Kyleena
KyleenaN (levonorgestrel-releasing intrauterine system) is a hormone-releasing IUD that prevents pregnancy for up to 5 years.

Figure 8. Example of a highlighting scheme on the Harm criterion by the Hybrid approach

4.4.2 The Visual Interpretation by the Typology Approach

The inter-annotator agreement rates, and the number of extracted sentence-level evidence for each criterion are shown in Table 6. Among all the criteria, the inter-agreement score of the Mongering is the highest with a score of 89.05%, while the Novelty is the lowest at 53.04%. The extraction for the Benefit criterion yielded 640 pieces of sentence evidence, which is the highest in number. However, the extracted evidence for the novelty evidence was scarce in number, with only 131 sentences deemed as valid evidence.

Table 6. Description of extracted evidence of each criterion

Criteria	Valid Extracted Evidence	Total Extracted Evidence	Inter-agreement Score
Cost	201	279	72.04%
Benefit	640	991	64.58%
Harm	318	440	72.27%
Quality	253	332	76.20%
Mongering	252	283	89.05%
Conflict	695	892	77.91%
Alternative	209	361	57.89%
Availability	274	475	57.68%
Novelty	131	247	53.04%

The same number of sentences as the negative class were randomly selected to build the classification datasets. Following the same approach applied to the automation of criterion evaluation, the final classification performance and hyperparameters of sentence classifiers for all criteria are detailed in Table 7.

Table 7. The performance, hyperparameter, and feature count of the evidence classifier of each criterion

Criteria	AUC	Precision	Classifier	Feature Count	Hyperparameter
Cost	0.8934	0.8261	RF	4000	{'n_estimators': 1200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 50, 'bootstrap': True}
Benefit	0.6845	0.6206	LR	1000	{'C': 0.04525351996665363, 'max_iter': 500, 'penalty': 'l2', 'solver': 'newton-cg'}
Harm	0.7242	0.6624	RF	1000	{'n_estimators': 1400, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 100, 'bootstrap': True}
Quality	0.8392	0.7737	LR	4000	{'C': 0.001477765646329587, 'max_iter': 500, 'penalty': 'l2', 'solver': 'lbfgs'}
Mongering	0.8490	0.7771	RF	4000	{'n_estimators': 1000, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 60, 'bootstrap': True}
Conflict	0.8791	0.7843	LR	1000	{'C': 0.0006442682148047918, 'penalty': 'l2', 'solver': 'lbfgs'}
Alternative	0.8305	0.7924	SVM	4000	{'kernel': 'rbf', 'gamma': 1.0, 'C': 10.0}
Availability	0.8925	0.8284	RF	4000	{'n_estimators': 400, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 60, 'bootstrap': True}
Novelty	0.8519	0.7694	RF	2000	{'n_estimators': 1000, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}

All of the sentence-level classifiers built for the typology approach could achieve an AUC score of more than 80%, indicating a good capability in identifying sentences that can serve as evidence to explain the pertinent criterion rating result. Figure 9 shows the result of applying the classifier to each sentence in the document and highlighting positive sentence instances that support a Cost criterion evaluation.

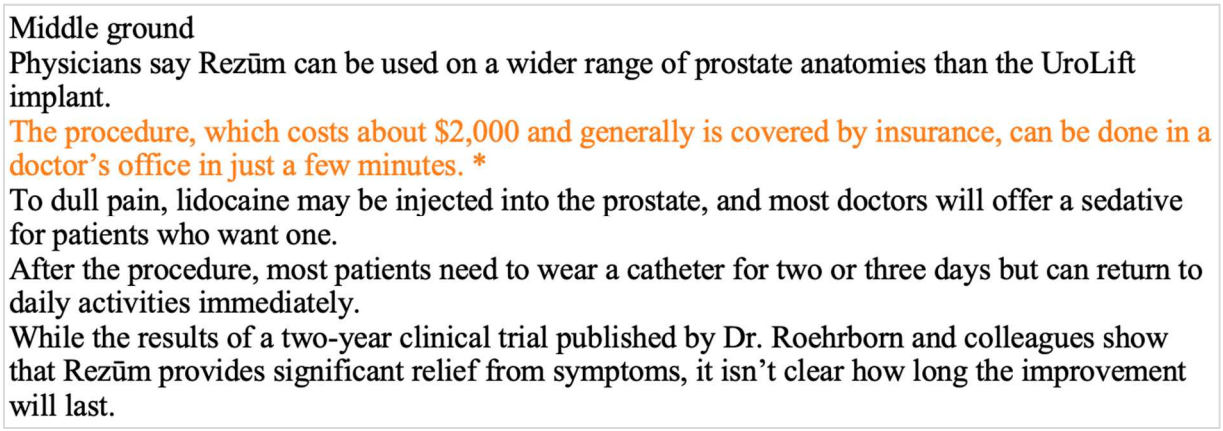

Figure 9. Example of a highlighting scheme on the Cost criterion by the Typology approach

4.4.3 The Overall Performance and Optimization of Two Approaches

To quantify the performance of two approaches, 20 health news samples that meet the given criterion but were not in the training set were randomly selected for each criterion. Each highlighted sentence was verified by at least two reviewers to determine whether it qualifies as evidence. Then, as the total number of highlighted sentences increased from 1, we calculated varying rates of accurately highlighted sentences, as shown in Table 8. The numbers with redder background colors indicate a lower highlighting accuracy produced by a given approach for a specific number of highlighting sentences.

Table 8. The accuracy of both approaches for interpreting each criterion evaluation

Criteria	Approach	Maximum Highlighting Sentence Count					
		1	2	3	4	5	6
Cost	Typology	80.00%	75.00%	60.78%	66.67%	60.00%	54.55%
	Hybrid	100.00%	92.50%	86.67%	76.25%	67.00%	59.65%
Benefit	Typology	66.67%	55.56%	53.70%	48.61%	45.56%	40.63%
	Hybrid	61.11%	50.00%	51.85%	45.83%	45.56%	43.52%
Harm	Typology	100.00%	97.50%	86.67%	76.39%	72.94%	66.67%
	Hybrid	90.00%	90.00%	90.00%	85.00%	81.00%	75.83%
Quality	Typology	95.00%	85.00%	76.67%	75.00%	69.00%	62.50%
	Hybrid	85.00%	70.00%	61.67%	55.00%	53.00%	49.17%
Mongering	Typology	95.45%	72.73%	52.38%	40.48%	33.68%	28.07%
	Hybrid	36.36%	25.00%	21.21%	18.18%	17.14%	15.00%
Conflict	Typology	75.00%	72.50%	66.67%	61.11%	55.29%	56.41%
	Hybrid	90.00%	87.50%	81.67%	68.75%	59.00%	50.93%
Alternative	Typology	65.00%	57.50%	53.33%	45.00%	40.00%	35.00%
	Hybrid	50.00%	44.44%	40.74%	37.50%	38.89%	37.96%
Availability	Typology	68.42%	50.00%	46.30%	50.00%	50.00%	56.94%
	Hybrid	47.37%	47.37%	49.12%	48.68%	46.32%	43.86%
Novelty	Typology	66.67%	58.33%	50.00%	42.65%	40.00%	36.46%
	Hybrid	55.56%	55.56%	44.44%	36.11%	33.33%	31.37%

Generally speaking, the highlighting accuracy decreases as the number of highlighted sentences increases. The performance of the interpretation approach varies by criterion. Both approaches obtained useful results (accuracy greater than 75%) in highlighting evidence to interpret Cost, Harm, Quality, and Conflict criteria evaluation. In contrast, the criteria Benefit, Alternative, Availability and Novelty are more challenging (accuracy lower than 79%) to interpret with sentences by both approaches. When comparing both interpretation approaches, the Typology approach outperformed the Hybrid for most of the criteria, especially for the Mongering criterion. The highlighting accuracy of Typology for correctly highlighting evidence

reached 95.45%, while the Hybrid only obtained an accuracy of 36.36%. The Hybrid approach was superior to the Typology approach in interpreting the Cost and Conflict criterion. When the accuracy threshold for highlighting is set to 65%, the optimal window size of sentence count for the Typology approach to get relatively better interpretation results for each criterion would be 2 (Cost), 1 (Benefit), 6 (Harm), 5 (Quality), 2 (Mongering), 3 (Conflict), 1 (Alternative), 1 (Availability), and 1 (Novelty). Similarly, the window size for the Hybrid approach would be 5 (Cost), 6(Harm), 2 (Quality), and 4 (Conflict). The Hybrid is not suitable for interpreting Benefit, Mongering, Alternative, Availability and Novelty because their highest accuracies do not exceed the threshold.

4.5 Users' Trust in the Interpretable System VS Non-Interpretable System

The first stage of the user study (Fall 2022) contained 195 responses. After removing invalid responses (incomplete surveys and surveys that failed the attention check question), 87 responses were left that could be used for further analysis. Table 9 presents the result of the chi-square analysis, in which the frequencies of major demographic variables, including age, gender, educational attainment, and major field of study, are listed and broken down by group.

Table 9. Sample demographic characteristics with Chi-Square analysis to examine the group differences

	Total N = 87		Study Group n = 48		Control Group n = 39			
	n	%	n	%	n	%	$\chi 2$	Asym.Sig.(2-sided)
Educational Attainment							4.367	0.498
Bachelor's Degree or Equivalent	37	42.5	20	54.1	17	45.9		
High School Diploma or Equivalent	9	10.3	7	77.8	2	22.2		
High School not completed or below	1	1.1	0	0.0	1	100.0		
Master's Degree or Equivalent	26	29.9	14	53.8	12	46.2		
Ph.D. Degree or Equivalent	13	14.9	6	46.2	7	53.8		
Prefer not to say	1	1.1	1	100.0	0	0.0		
Major Field of Study							7.229	0.405
Not Provided	11	12.6	8	72.7	3	27.3		
Business	13	14.9	6	46.2	7	53.8		
Health Science	2	2.3	2	100.0	0	0.0		
Humanities (Literature, History, Philosophy etc.)	5	5.7	3	60.0	2	40.0		
Law	1	1.1	0	0.0	1	100.0		
Other	5	5.7	3	60.0	2	40.0		
Social Science (Sociology, Political Science, Psychology etc.)	20	23.0	13	65.0	7	35.0		
STEM (Science, Technology, Engineering, and Math)	30	34.5	13	43.3	17	56.7		
Sex							2.931	0.402
Female	51	58.6	31	60.8	20	39.2		
Male	32	36.8	15	46.9	17	53.1		
Non-binary	3	3.4	1	33.3	2	66.7		
Prefer not to say	1	1.1	1	100.0	0	0.0		
Age							23.229	0.332
18-25	47	54.0	29	61.7	18	38.3		
26-29	16	18.4	8	50.0	8	50.0		
30-35	15	17.2	7	46.7	8	53.3		
36-39	7	8.0	4	57.1	3	42.9		
40-45	0	0.0	0	N/A	0	N/A		
46-50	1	1.1	0	0.0	1	100.0		
51-55	0	0.0	0	N/A	0	N/A		
56-60	1	1.1	0	0.0	1	100.0		
≥ 61	0	0.0	0	N/A	0	N/A		

The values of the chi-square statistic for the association between group and each of the four tested demographic variables are 4.367 (educational attainment), 7.229 (major field of study), 2.931 (sex), and 23.223 (age), respectively. The p-value of each association test is greater than the standard $\alpha = 0.05$, suggesting the none of the demographic variables is associated with the group. Statistically, there is no significant difference between the study group and the control group in terms of demographic distribution.

An independent-sample T test was conducted to analyze the survey data with a significance level of $\alpha = 0.05$. The average trust score of ACE users for the news quality result was 3.46, slightly better than 3.41, the mean trust score of ACU users. There was no statistically significant difference between users' trust in the news evaluation results of the two groups. However, when ACE users were asked to rate their level of agreement or disagreement on the statement "The highlighted sentences are helpful for me to understand the evaluation.", 65% of the participants agreed on the statements, as seen in figure 10.

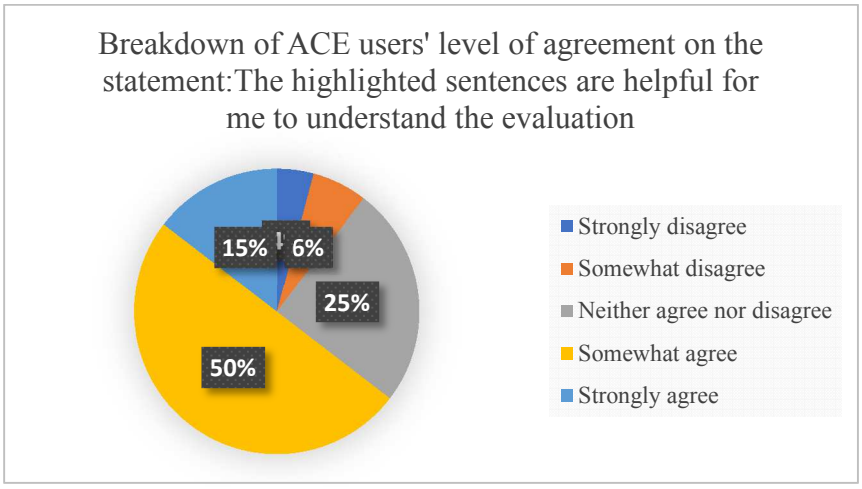

Figure 10. Breakdown of ACE users' level of agreement on the statement: The highlighted sentences are helpful for me to understand the evaluation.

In the ACU group, 73% of the participants agreed with the statement “I need more explanations to help me understand the evaluation.”. To be specific, 45% of participants chose “somewhat agree” and 28% chose “strongly agree”.

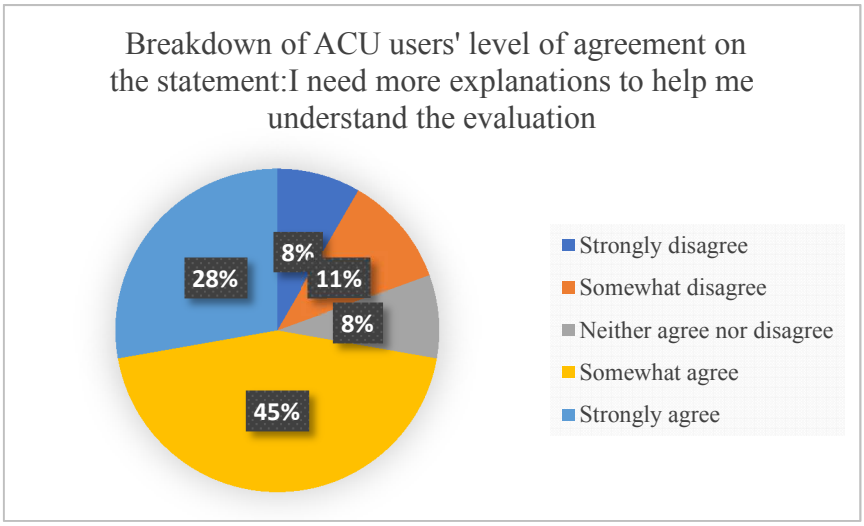

Figure 11. Breakdown of ACU users' level of agreement on the statement: I need more explanations to help me understand the evaluation

The participants of two groups were also asked to rank the importance of each criterion. As indicated in Table 10, the nine criteria were ranked in order of priority, with Quality receiving the highest average score, and Mongering receiving the lowest. All criteria were deemed important, with an average importance score of greater than 3.

Table 10. Importance of criteria

Criteria	Number	Minimum	Maximum	Mean	STD
Cost	87	1	5	3.23	1.148
Benefits	87	2	5	3.72	1.019
Harm	87	1	5	3.84	1.022
Quality	87	2	5	4.3	0.823
Mongering	87	0	5	3.02	1.067
Conflict	87	0	5	3.13	1.189
Alternative	87	1	5	3.86	0.99
Availability	87	1	5	3.97	0.958
Novelty	87	0	5	3.18	1.167

Note: 0 = Response not provided; 1 = Not important at all; 5 = Extremely important

CHAPTER V: DISCUSSION AND CONCLUSIONS

5.1 Overview of Chapter V

This chapter presents the discussion of study results, which summarizes the key findings of the study, compares the proposed solution to earlier work, and acknowledges the study's limitations. The dissertation concludes with suggestions for future research directions.

5.2 Discussion of Results

5.2.1 Principal Findings

This study experimented with two AI-based approaches to visualizing the interpretation of a criteria-based system designed to assist users in systematically evaluating the quality of health news.

The findings of the study have three folds. First, it is feasible to automate the evaluation process of health news based on the selected criteria. The classification results further validate the capability of RF, LR, SVM algorithms and TFIDF in health misinformation detection as found in past work (Hawa et al., 2021; Saengkhunthod et al., 2021; Shah et al., 2019).

Second, it is found that both Hybrid and Typology approaches could achieve the desired visualization result to justify the predicted evaluation result despite the expertise and nature of two approaches differentiate. In addition, the study determined a window size for choosing in advance the sentences to be highlighted by both ways in order to produce a better visualization for each criterion. Finally, the study determined the significance of criteria in assisting users to evaluate the quality of health news. Participants also confirmed the need for interpretation in the evaluation of the quality of health news.

5.2.2 Discussion of the Classification Model for Health News Quality Evaluation

The classifiers' performances on most of the criteria on both measures were good. The classification performance of the cost criterion, for example, achieved 89.37% and 81.85% when measured with AUC and precision, respectively. It is also the criterion that is most classifiable compared to the rest of the criteria. The classifiers of criteria including Benefit, Harm, Mongering achieved a precision of 70%, indicating the classifiers' good capability of identifying a real positive (satisfactory) case from all positive cases. On the other hand, two criteria, including the Alternative and Novelty, seem harder to classify. Both of their performances are relatively poor, with AUC and precision measures being lower than 70%.

5.2.3 Discussion of Two Interpretation Approaches

5.2.3.1 Summary of the Hybrid Approach

The hybrid approach demonstrates both good accuracy and efficiency in visualizing an automatic model's interpretation for evaluating the Cost, Harm and Conflict criteria. Compared to the typology approach, it is advantageous in terms of saving manual effort, as it does not require sentence extraction. For building the Hybrid approach, the feasibility of the rule-based strategy to enhance LIME's interpretation work is supported by our observation during evidence extraction for the typology approach that specific words or phrases such as "adverse effect," "danger," "death," and "side effect" appear repeatedly in an evaluation of the Harm criterion; keywords such as "cost"," price"," insurance" frequently appear for the Cost criterion evaluation; "spokesman" ,"funding", "sponsor" are typically used to disclose the conflicts of interests. However, the aim of the highlighting scheme in the Hybrid approach is to locate the sentences where keywords are present. Such nature leads to its two inherent drawbacks.

The first drawback of this scheme is that it sometimes fails to discern the semantic differences between sentences that share similar keywords, such as a sentence about the risk of the intervention and a sentence that describes the benefits of intervention by relieving or preventing adverse conditions. For example, in one of the test cases, the sentence "*Moreover, the study verified that long-term use of bisphosphonate drugs reduces the risk of typical osteoporosis fractures by 24 percent*." was wrongly highlighted. The sentence contains keywords, including "risk," " fractures," which are relevant to adverse symptoms. However, it introduces how bisphosphonate is expected to benefit patients by decreasing the risk of a negative outcome. Another example is to differentiate between the stock price and the intervention price. Usually, both of these lines contain a lot of terms that describe the "value" of the intervention.

The other drawback of the Hybrid approach is that it is prone to errors when interpreting a criterion evaluation result that is less keyword-reliant such as Mongering, Quality and Alternative. In contrast to the Cost, Harm, and Conflict criteria, it is difficult for LIME to extract keywords when the choice of words associated with a certain criterion are less apparent.

5.2.3.2 Summary of the Typology Approach

The typology approach, by contrast, performs better at handling expressions with more lexical variations. For example, sentences "*Last fall the Food and Drug Administration issued a "safety update" urging doctors and patients to be on the lookout for the problem*." and "*These medications are now linked to a growing number of complications, ranging in seriousness from nutrient deficiencies, joint pain and infections to bone fractures, heart attacks and dementia*." were successfully picked by the typology approach, whereas they were missed by the hybrid approach as keywords in those sentences are less commonly used to describe side effects. The typology approach distills relevant information from text documents through sentence extraction

by human experts, and the information is key to building a knowledge base for the identification of evidence. Hence, the Typology approach is generally more robust and stable than the hybrid approach when visualizing the interpretation of criteria which are less keyword reliant. When it comes to interpret the criteria such as Mongering, Alternative, Quality, the Typology approach performed noticeably better than the Hybrid technique. Typically, the supporting evidence for those criteria is subtle and highly variable in expression. No particular terms can be used to indicate the existence of the evidence.

5.2.3.3 Summary of Two Interpretation Approaches

Table 11 summarizes the advantages and disadvantages of both approaches.

Table 11. Comparison table summarizes the advantages and disadvantages of both Hybrid and Typology approaches

Approach	Advantages	Disadvantages
Hybrid	• Works better for interpreting keyword-reliant criteria such as Cost, Harm and Conflict. • Less manual efforts	• Not able to handle expression with lexical variations • Fail to discern sentences with similar lexicons with different semantics
Typology	• Works for most the criteria with better stability • Able to handle more expressions with lexical variations	• Requires more manual efforts to extract evidence

Despite the good performance of two approaches in interpreting in most of the criteria evaluation results, both approaches face two challenges in interpreting some of the criteria. The first challenge arises from the inconsistent review quality from HealthNewsReview.org. Taking the Quality criterion for example, some of health new are rated as satisfactory only because “the

study is published on a peer review journal". In other cases, reviewers from HealthNewsReview.org justified their rating by providing substantial critique to the study design. The second challenge is that some of the evaluation results' supporting evidence cannot be located in the original text; rather, they are derived from common knowledge. Interventions, such as vitamins and breastmilk, are naturally occurring. Therefore, the reviewers of HealthNewsReview.org deemed the news that discussed such interventions to be satisfactory on the Availability criterion despite the absence of an explicit description of intervention availability. There are several comparable instances regarding the evaluation of the Cost criterion. Since interventions such as meditation, yoga, and sun exposure are free, it is generally unnecessary to indicate their costs.

5.2.4 Discussion of the User Study

Through the user study, participants in the control group confirmed the need for an explanation for the news evaluation system. More than half of the participants in the study group agreed that the highlighted sentences assisted them in comprehending the criteria evaluation outcome. Nevertheless, the independent-sample T test result didn't reject the null hypothesis. The result implies that both user groups have relatively equal trust in the system's news assessment result despite the presence or absence of the interpretation. As the chi-square analysis result did not reveal any potential cofounding effects resulting from unequally distributed demographic variables between the groups, the dataset's small size and poor general population representation are likely to be the main contributors to the T test's outcome.

The recruited participants have a young profile, since 54% are between the ages of 18 and 25 and 72.4% are under the age of 30. According to studies, young people are more likely to trust technology because of positive psychological and cognitive qualities (Oksanen et al., 2020).

In contrast, older individuals have frequently demonstrated less faith in automated processes and typically have more negative views regarding AI (Hoff & Bashir, 2015; Oksanen et al., 2020). Both study and control groups exhibited a high level of trust in AI, science, and technology, with an average trust score of 4.11 (out of 5) in science and technology and an average trust score of 3.74 (out of 5) in AI. No statistically significant difference was found between groups. In a subsequent correlation analysis among the users' trust in the system evaluation result, users' trust in AI, users' trust in science and technology, the latter two variables have positive correlation with user's trust in the system. Since most of the participants are young and already have a high level of trust in AI-related technology, there remains less room for an interpretation feature to make people trust the system more.

5.2.5 Answering Research Questions

5.2.5.1 Research Question 1

The first research question is "*How accurately can the health news review process be modeled to predict health news quality using the data provided and annotated by HealthNewsReview.org?*". The answers to the first question will be arranged based on sub-questions.

(1) What is the best performance we can achieve for automatically reviewing health news quality per criterion?

The performance of the classification varies by criterion. The best classification performance, when measured with AUC score is 0.89 (Cost), 0.68 (Benefit) ,0.76 (Harm),0.74 (Quality), 0.63(Mongering),0.72(Conflict),0.65 (Alternative),0.70 (Availability) and 0.61 (Novelty).

(2) What features and algorithms contribute to predicting the review result with the highest performance per criterion?

Table 5 contains the word feature count and selected algorithm with tuned hyperparameters that contribute to the highest classification results. The exact features that contribute to the classification can be found in the word feature weights table generated by LIME for each criterion.

(3) What are the criteria for which the review process can be better automated by the machine based on various evaluation metrics including AUC and Precision?

Based on the generated results, the top criteria that can be better automated by the machine are the Cost, Harm, Quality Conflict and Availability, with a threshold AUC = 0.7. When precision = 0.7 is set as a threshold, the most classifiable criteria are Cost, Harm and Mongering.

5.2.5.2 Research Question 2

The answers to the second question, "*How effectively can automated health news review results be visualized or explained?*" are arranged as follows.

(1) How well can interpretable A.I. techniques and typology classification approaches visualize or explain the automated health news review results compared to a simple display of black-box prediction results?

Participants in the control group verified the need for interpretation to help them comprehend the black-box model news evaluation results. In addition, more than half of the participants who had access to the system with visual interpretation reported that the highlighted information aided them in assessing the outcome. Nevertheless, regardless of the existence or absence of interpretation, there is no statistically significant difference between groups in terms

of trust in the system. This may account for the tiny sample size, which is not representative of the entire population. The results will need to be revalidated once the sample size is sufficient.

(2) Which approach demonstrates a superior performance when interpreting for the same criterion?

The performances of two approaches vary by criterion and by the number of highlighting sentences. Generally speaking, the Typology approach outperformed the Hybrid for most of the criteria, especially for the Mongering criterion. The Hybrid approach was superior in efficiency and accuracy to the Typology approach in interpreting the Cost and Conflict criteria.

5.2.6 Novelty in Comparison with Prior Work

The novelty of the study arises in three aspects: generalizability, multidimensionality, and interpretability.

First, unlike the system described here, most existing veracity-based fake news detectors are built solely on manually selected linguistic cues, leading to a lack of generalizability across topics, languages, and domains (K. Sharma et al., 2019). This weakness is also apparent in the study by Gerts et al.(2021), as the team found a huge variation in the classifier performance (F1 scores between 0.347 and 0.857) on four conspiracy topics, and more narrowly defined topics could increase the performance. By comparison, our system automatically trains a criteria-based classifier that uses all words as potential features. It does not rely on a dataset with a strictly defined topic.

Secondly, our system addresses the complexity and multidimensional attributes of the phenomenon of health information disorder (Claire Wardle & Hossein Derakhshan, 2017; Joshua Habgood-Coote, 2018; Y. Wang et al., 2019). Existing studies predominately perform binary classification tasks at the document level and are primarily veracity based. This places a great

challenge upfront for identifying health misinformation, as the binary label is insufficient to represent health news's complicated evaluation process in actual practice. This is especially the case with veracity-based classification. Human-based fact-checking work involves extensive knowledge understanding, inference, and source tracking, which remains a challenge even to deep learning methods. This is because the fabricated news is intended to mirror the truth to deceive readers; as a result, without cross-referencing and high-level inference, it might be impossible to determine the authenticity of news stories by text analysis alone(K. Sharma et al., 2019). Our system addresses this challenge by evaluating health news from nine different dimensions and classifies evidence at the sentence level, which limits the amount of expertise and training required.

Finally, compared to previous interpretability work in suggested health-related misinformation detection systems, the work in adding the interpretability of a health misinformation system is innovative. To our knowledge, the current state of the art in explainable misinformation detection systems mostly looks to produce explanations for veracity predictions concerning inputs to the system. This study fills a gap in explaining a criteria-based system on health misinformation. Besides, developing the interpretable module on a criteria-based model is an advantage. The criteria-based approach inherently looks for the linguistic characteristics of health news, such as the choice of words present/absent of crucial information, whereas a veracity-based system may face a challenge to be interpreted based on the linguistic features of text alone. In addition, this study exhibits a greater level of readability of the interpretation than the interpretation work done by the existing interpretation work on health misinformation, such as the work conducted by Alharbi et al. (2021) for fake news. The interpretation level achieved in the study by Alharbi et al. (2021) remains at the word level, with

both positive and negative words highlighted and dispersed throughout the articles, whereas our study presented two approaches to achieve sentence-level visualized interpretation, which demonstrate higher levels of readability to end users.

5.2.7 Limitations and Future Work

This comes with some limitations. The first limitation is that we only considered TFIDF values of words as features for building both document level and sentence level classifiers. We acknowledge that the performance of the document-level classification model was lower compared to similar studies that adopted the same dataset from HealthNewsReview.org. The performance of our doc-level classification models for the Harm, Cost, Conflict criteria, the criteria that are most easily to be classified are 0.74, 0.82 and 0.67 when measure by Precision, and 0.76, 0.89 and 0.72 when measured by AUC. The results were calculated with 10-fold-validation. The performance is comparable to a study by Al-Jefri et al. (2020) that focuses on building health news quality classification models. The precision performance for classifying Harm, Cost and Conflict criteria is reported 74.61, 77.61 and 70.89. The study incorporated more features such as TF-IDF, Comparatives forms, NER tags, and strategically change the feature selections for different criterion classification tasks. In another study by Afsana et al.(2021) that also aimed to achieve the same research goal, the performance of their models for Harm, Cost and Conflict measure by weighted F1 was reported 0.84,0.899 and 0.835 respectively. However, their superior performance is achieved through extensive work on feature engineering with a total of 53012 features applied. Given that the key focus of this study is to build the interpretation of the system, which both prior mentioned studies lacked, the current performance of models is effective to serve the purpose of the study. In the future work, more work on the

feature engineering will be incorporated for both document-level classification, and especially the Typology approach which is embodied as a sentence-level classifier.

The second limitation is with the simple rules for the hybrid approach. The hybrid approach takes advantage of both human knowledge and an auto-generated keyword list generated by LIME. However, the existing rules provided by human experts are keyword-based and do not contain complex rules to handle various expression variants. As part of the future plan, more complex rules for the Hybrid approach will be implemented to address the weak spots of Hybrid approach to make it more able to distinguish different type of sentences when they share similar lexicons but different semantics. Additionally, the effectiveness of an ensemble approach that combines the highlighting results of the Hybrid and Typology approaches will be studied.

A further limitation of the study is the small sample size of a user study. Additionally, the major of participants recruited in the user study were young, holding higher education. In the future study, more participants will be recruited to increase the diversity of the participants. Once the sample size is sufficient, more studies will be carried out to investigate how users trust the system when the health news to be analyzed is of varying quality.

5.3 Conclusions

The study provided an interpretable, criteria-based strategy for evaluating the quality of health news. Two methods for visualizing the system's interpretation were investigated. To aid in the exploration, the experiment was developed by comparing rule-based and statistical machine learning approaches. The results suggest that either approach can successfully automate criterion-based health news quality rating with visual evidence supporting model explanation. This work has the potential to increase public trust in computer-assisted reviews of health

information. However, further investigation is needed to improve the system's performance as well as to better understand users' trust in the system.

References

Abdulai, A.-F., Tiffere, A.-H., Adam, F., & Kabanunye, M. M. (2021). COVID-19 information-related digital literacy among online health consumers in a low-income country. *International Journal of Medical Informatics*, *145*, 104322. https://doi.org/10.1016/j.ijmedinf.2020.104322

Abukaraky, A., Hamdan, A.-A., Ameera, M.-N., Nasief, M., & Hassona, Y. (2018). Quality of YouTube TM videos on dental implants. *Medicina Oral, Patologia Oral Y Cirugia Bucal*, *23*(4), e463–e468. https://doi.org/10.4317/medoral.22447

Afsana, F., Kabir, M. A., Hassan, N., & Paul, M. (2021). Automatically Assessing Quality of Online Health Articles. *IEEE Journal of Biomedical and Health Informatics*, *25*(2), 591–601. https://doi.org/10.1109/JBHI.2020.3032479

Alharbi, R., Vu, M. N., & Thai, M. T. (2021). Evaluating Fake News Detection Models from Explainable Machine Learning Perspectives. *ICC 2021 - IEEE International Conference on Communications*, 1–6. https://doi.org/10.1109/ICC42927.2021.9500467

Al-Jefri, M., Evans, R., Lee, J., & Ghezzi, P. (2020). Automatic Identification of Information Quality Metrics in Health News Stories. *Frontiers in Public Health*, *8*. https://www.frontiersin.org/articles/10.3389/fpubh.2020.515347

Anti-Vaccination Society of America. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Anti-Vaccination_Society_of_America&oldid=1067370726

Aquino, F., Donzelli, G., De Franco, E., Privitera, G., Lopalco, P. L., & Carducci, A. (2017). The web and public confidence in MMR vaccination in Italy. *Vaccine*, *35*(35, Part B), 4494–4498. https://doi.org/10.1016/j.vaccine.2017.07.029

Armstrong, P. W., & Naylor, C. D. (2019). Counteracting Health Misinformation: A Role for Medical Journals? *JAMA*, *321*(19), 1863–1864. https://doi.org/10.1001/jama.2019.5168

Ashforth, B. E., & Mael, F. (1989). Social identity theory and the organization. *The Academy of Management Review*, *14*, 20–39. https://doi.org/10.2307/258189

Asmundson, G. J. G., Abramowitz, J. S., Richter, A. A., & Whedon, M. (2010). Health anxiety: Current perspectives and future directions. *Current Psychiatry Reports*, *12*(4), 306–312. https://doi.org/10.1007/s11920-010-0123-9

Asmundson, G. J. G., & Taylor, S. (2020). Coronaphobia: Fear and the 2019-nCoV outbreak. *Journal of Anxiety Disorders*, *70*, 102196. https://doi.org/10.1016/j.janxdis.2020.102196

Atehortua, N. A., & Patino, S. (2021). COVID-19, a tale of two pandemics: Novel coronavirus and fake news messaging. *Health Promotion International*, *36*(2), 524–534. https://doi.org/10.1093/heapro/daaa140

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models. *ArXiv:2103.00747 [Cs]*. http://arxiv.org/abs/2103.00747

Bálint, P., & Bálint, G. (2009). The Semmelweis-reflex. *Orvosi Hetilap*, *150*(30), 1430–1430. https://doi.org/10.1556/oh.2009.ho2256

Bandari, R., Zhou, Z., Qian, H., Tangherlini, T. R., & Roychowdhury, V. P. (2017). A Resistant Strain: Revealing the Online Grassroots Rise of the Antivaccination Movement. *Computer*, *50*(11), 60–67. https://doi.org/10.1109/MC.2017.4041354

Basch, C. H., Zybert, P., Reeves, R., & Basch, C. E. (2017). What do popular YouTubeTM videos say about vaccines? *Child: Care, Health and Development*, *43*(4), 499–503. https://doi.org/10.1111/cch.12442

Becker, B. F. H., Larson, H. J., Bonhoeffer, J., van Mulligen, E. M., Kors, J. A., & Sturkenboom, M. C. J. M. (2016). Evaluation of a multinational, multilingual vaccine debate on Twitter. *Vaccine*, *34*(50), 6166–6171. https://doi.org/10.1016/j.vaccine.2016.11.007

Bessi, A., Zollo, F., Vicario, M. D., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Trend of Narratives in the Age of Misinformation. *PLOS ONE*, *10*(8), e0134641. https://doi.org/10.1371/journal.pone.0134641

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Biggs, T. C., Bird, J. H., Harries, P. G., & Salib, R. J. (2013). YouTube as a source of information on rhinosinusitis: The good, the bad and the ugly. *The Journal of Laryngology and Otology*, *127*(8), 749–754. https://doi.org/10.1017/S0022215113001473

Bin Naeem, S., & Kamel Boulos, M. N. (2021). COVID-19 Misinformation Online and Health Literacy: A Brief Overview. *International Journal of Environmental Research and Public Health*, *18*(15), Article 15. https://doi.org/10.3390/ijerph18158091

Bode, L., & Vraga, E. (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, *65*. https://doi.org/10.1111/jcom.12166

Bonnevie, E., Goldbarg, J., Gallegos-Jeffrey, A. K., Rosenberg, S. D., Wartella, E., & Smyser, J. (2020). Content Themes and Influential Voices Within Vaccine Opposition on Twitter, 2019. *American Journal of Public Health*, *110*(S3), S326–S330. https://doi.org/10.2105/AJPH.2020.305901

Botnevik, B., Sakariassen, E., & Setty, V. (2020). BRENDA: Browser Extension for Fake News Detection. *Proceedings of the 43rd International ACM SIGIR Conference on Research*

and Development in Information Retrieval, 2117–2120. https://doi.org/10.1145/3397271.3401396

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation* [Http://purl.org/dc/dcmitype/Text, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:178db677-fa8b-491d-beda-4bacdc9d7069

Bridgman, A., Merkley, E., Zhilin, O., Loewen, P. J., Owen, T., & Ruths, D. (2021). Infodemic Pathways: Evaluating the Role That Traditional and Social Media Play in Cross-National Information Transfer. *Frontiers in Political Science*, *3*, 20. https://doi.org/10.3389/fpos.2021.648646

Buchanan, R., & Beckett, R. D. (2014). Assessment of vaccination-related information for consumers available on Facebook. *Health Information and Libraries Journal*, *31*(3), 227–234. https://doi.org/10.1111/hir.12073

CDC. (2021, September 7). *COVID-19 Vaccine Facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/facts.html

Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology and Community Health*, *53*(2), 105–111.

Chen, B., Shao, J., Liu, K., Cai, G., Jiang, Z., Huang, Y., Gu, H., & Jiang, J. (2018). Does Eating Chicken Feet With Pickled Peppers Cause Avian Influenza? Observational Case Study on Chinese Social Media During the Avian Influenza A (H7N9) Outbreak. *JMIR Public Health and Surveillance*, *4*(1), e8198. https://doi.org/10.2196/publichealth.8198

Cherilyn, I., & Julie, P. (2018). *Journalism, fake news & disinformation: Handbook for journalism education and training*. UNESCO Publishing.

Chou, W.-Y. S., Oh, A., & Klein, W. M. P. (2018). Addressing Health-Related Misinformation on Social Media. *JAMA*, *320*(23), 2417–2418. https://doi.org/10.1001/jama.2018.16865

Chua, A. Y. K., & Banerjee, S. (2018). Intentions to trust and share online health rumors: An experiment with medical professionals. *Computers in Human Behavior*, *87*, 1–9. https://doi.org/10.1016/j.chb.2018.05.021

Claire Wardle & Hossein Derakhshan. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (p. 109). https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Criterion #3 Does the story adequately explain/quantify the harms of the intervention? (n.d.). HealthNewsReview.Org. Retrieved February 15, 2022, from https://www.healthnewsreview.org/about-us/review-criteria/criterion-3/

Cui, L., & Lee, D. (2020). CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv:2006.00885 [Cs]*. http://arxiv.org/abs/2006.00885

Dale, R. (2017). NLP in a post-truth world. *Natural Language Engineering*, *23*(2), 319–324. https://doi.org/10.1017/S1351324917000018

De Freitas, J., Falls, B. A., Haque, O. S., & Bursztajn, H. J. (2013). Vulnerabilities to misinformation in online pharmaceutical marketing. *Journal of the Royal Society of Medicine*, *106*(5), 184–189. https://doi.org/10.1177/0141076813476679

Decision tree. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Decision_tree&oldid=1106946454

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*, 629–636. https://doi.org/10.1037/h0046408

Dhoju, S., Main Uddin Rony, M., Ashad Kabir, M., & Hassan, N. (2019). Differences in Health News from Reliable and Unreliable Media. *Companion Proceedings of The 2019 World Wide Web Conference*, 981–987. https://doi.org/10.1145/3308560.3316741

Dibia, V. (2020, May 8). *ML Interpretability: LIME and SHAP in prose and code*. Cloudera Blog. https://blog.cloudera.com/ml-interpretability-lime-and-shap-in-prose-and-code/

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv*. https://arxiv.org/abs/1702.08608

Elhadad, M. K., Li, K. F., & Gebali, F. (2020). Detecting Misleading Information on COVID-19. *IEEE Access*, *8*, 165201–165215. https://doi.org/10.1109/ACCESS.2020.3022867

Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. (2018). Elsevier.

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Ersoy, P. (2021, September 21). *Naive Bayes Classifiers for Text Classification*. Medium. https://towardsdatascience.com/naive-bayes-classifiers-for-text-classification-be0d133d35ba

Faasse, K., Chatman, C. J., & Martin, L. R. (2016). A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. *Vaccine*, *34*(47), 5808–5814. https://doi.org/10.1016/j.vaccine.2016.09.029

Feature engineering. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Feature_engineering&oldid=1113283964

Figma: The collaborative interface design tool. (n.d.). Figma. Retrieved October 16, 2022, from https://www.figma.com/

Find Survey Participants Today | SurveySwap. (n.d.). Retrieved October 16, 2022, from https://surveyswap.io/

Fiok, K., Karwowski, W., Gutierrez, E., Liciaga, T., Belmonte, A., & Capobianco, R. (2021). Automated Classification of Evidence of Respect in the Communication through Twitter. *Applied Sciences*, *11*(3), Article 3. https://doi.org/10.3390/app11031294

Fox, S. (2011, February 1). Health Topics. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2011/02/01/health-topics-2/

Freedman, J. L., & Sears, D. O. (1965). Selective Exposure11The preparation of this paper was supported in part by NSF grants to the authors. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 57–97). Academic Press. https://doi.org/10.1016/S0065-2601(08)60103-3

Freund, A. J., & Giabbanelli, P. J. (2021). Are We Modeling the Evidence or Our Own Biases? A Comparison of Conceptual Models Created from Reports. *2021 Annual Modeling and Simulation Conference (ANNSIM)*, 1–12. https://doi.org/10.23919/ANNSIM52504.2021.9552054

Gandhi, R. (2018a, May 17). *Naive Bayes Classifier*. Medium. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Gandhi, R. (2018b, July 5). *Support Vector Machine—Introduction to Machine Learning Algorithms*. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Garrido, F., Verbeke, W., & Bravo, C. (2018). A Robust profit measure for binary classification model evaluation. *Expert Systems with Applications*, *92*, 154–160. https://doi.org/10.1016/j.eswa.2017.09.045

Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, *38*(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

Gerts, D., Shelley, C. D., Parikh, N., Pitts, T., Watson Ross, C., Fairchild, G., Vaquera Chavez, N. Y., & Daughton, A. R. (2021). "Thought I'd Share First" and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study. *JMIR Public Health and Surveillance*, *7*(4), e26527. https://doi.org/10.2196/26527

Ghenai, A., & Mejova, Y. (2017). Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. *ArXiv:1707.03778 [Cs]*. http://arxiv.org/abs/1707.03778

Greenhill, K. M., & Oppenheim, B. (2017). Rumor Has It: The Adoption of Unverified Information in Conflict Zones. *International Studies Quarterly*, *61*(3), 660–676. https://doi.org/10.1093/isq/sqx015

Guestrin, M. T. R., Sameer Singh, Carlos. (2016, August 12). *Local Interpretable Model-Agnostic Explanations (LIME): An Introduction*. O'Reilly Media. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Guidry, J., Jin, Y., Haddad, L., Zhang, Y., & Smith, J. (2016). How Health Risks Are Pinpointed (or Not) on Social Media: The Portrayal of Waterpipe Smoking on Pinterest. *Health Communication*, *31*(6), 659–667. https://doi.org/10.1080/10410236.2014.987468

Guidry, J. P. D., Carlyle, K., Messner, M., & Jin, Y. (2015). On pins and needles: How vaccines are portrayed on Pinterest. *Vaccine*, *33*(39), 5051–5056. https://doi.org/10.1016/j.vaccine.2015.08.064

Haigh, M., & Haigh, T. (2020). *Fighting and Framing Fake News* (pp. 303–323).

Hanson, C. L., Cannon, B., Burton, S., & Giraud-Carrier, C. (2013). An Exploration of Social Circles and Prescription Drug Abuse Through Twitter. *Journal of Medical Internet Research*, *15*(9), e2741. https://doi.org/10.2196/jmir.2741

Harper, C. A., & Baguley, T. (2019). *"You are Fake News": Ideological (A)symmetries in Perceptions of Media Legitimacy*. PsyArXiv. https://doi.org/10.31234/osf.io/ym6t5

Harshith. (2022, July 27). *Text Preprocessing in Natural Language Processing using Python*. Medium. https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8

Hassan, N., Adair, B., Hamilton, J., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015). *The Quest to Automate Fact-Checking*. https://www.semanticscholar.org/paper/The-Quest-to-Automate-Fact-Checking-Hassan-Adair/714a12ddf7d2c6ddb8a99bfb6663d2144880aa49

Hawa, S., Lobo, L., Dogra, U., & Kamble, V. (2021). Combating Misinformation Dissemination through Verification and Content Driven Recommendation. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 917–924. https://doi.org/10.1109/ICICV50876.2021.9388406

HealthNewsReview.org. (n.d.-a). *About Us*. HealthNewsReview.Org. Retrieved September 24, 2021, from https://www.healthnewsreview.org/about-us/

HealthNewsReview.org. (n.d.). HealthNewsReview.Org. Retrieved February 15, 2022, from https://www.healthnewsreview.org/

HealthNewsReview.org. (n.d.-b). *What we review and how*. HealthNewsReview.Org. Retrieved September 24, 2021, from https://www.healthnewsreview.org/about-us/how-we-rate-stories/

Hoeken, H., & Hustinx, L. (2003). The relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation*, 497–502. https://repository.ubn.ru.nl/handle/2066/82921

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Horst, F., Lapuschkin, S., Samek, W., Müller, K.-R., & Schöllhorn, W. I. (2019). Explaining the unique nature of individual gait patterns with deep learning. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-38748-8

Hossin, M., & Sulaiman, M. N. (2019). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(2), 1–11. https://doi.org/10.5281/zenodo.3557376

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Indra, S. T., Wikarsa, L., & Turang, R. (2016). Using logistic regression method to classify tweets into the selected topics. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 385–390. https://doi.org/10.1109/ICACSIS.2016.7872727

Institute of Medicine (U.S.) (Ed.). (2005). *Complementary and alternative medicine in the United States*. National Academies Press.

InterpretML Team. (2021). *InterpretML documentation*. https://interpret.ml/docs/intro.html

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019). A Semantics Aware Random Forest for Text Classification. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1061–1070. https://doi.org/10.1145/3357384.3357891

Jamieson, K. H., & Cappella, J. N. (2008). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.

Jamison, A., Broniatowski, D. A., Smith, M. C., Parikh, K. S., Malik, A., Dredze, M., & Quinn, S. C. (2020). Adapting and Extending a Typology to Identify Vaccine Misinformation on Twitter. *American Journal of Public Health*, *110*(S3), S331–S339. https://doi.org/10.2105/AJPH.2020.305940

Johnson, S. B., Parsons, M., Dorff, T., Moran, M. S., Ward, J. H., Cohen, S. A., Akerley, W., Bauman, J., Hubbard, J., Spratt, D. E., Bylund, C. L., Swire-Thompson, B., Onega, T., Scherer, L. D., Tward, J., & Fagerlin, A. (2021). Cancer Misinformation and Harmful Information on Facebook and Other Social Media: A Brief Report. *Journal of the National Cancer Institute*, djab141. https://doi.org/10.1093/jnci/djab141

Jordan, J. (2017, November 2). *Hyperparameter tuning for machine learning models.* Jeremy Jordan. https://www.jeremyjordan.me/hyperparameter-tuning/

Joshua Habgood-Coote. (2018). Stop talking about fake news!: Inquiry: Vol 62, No 9-10. *Inquiry*, *62*(9–10), 1033–1065. https://doi.org/10.1080/0020174X.2018.1508363

Kahneman, D. (2012). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (2012). Prospect Theory: An Analysis of Decision Under Risk. In *Handbook of the Fundamentals of Financial Decision Making: Vol. Volume 4* (pp. 99–127). WORLD SCIENTIFIC. https://doi.org/10.1142/9789814417358_0006

Kalichman, S. C., Eaton, L. A., Earnshaw, V. A., & Brousseau, N. (2022). Faster than warp speed: Early attention to COVD-19 by anti-vaccine groups on Facebook. *Journal of Public Health*, *44*(1), e96–e105. https://doi.org/10.1093/pubmed/fdab093

Khanday, A. M. U. D., Khan, Q. R., & Rabani, S. T. (2021). Identifying propaganda from online social networks during COVID-19 using machine learning techniques. *International Journal of Information Technology*, *13*(1), 115–122. https://doi.org/10.1007/s41870-020-00550-5

Kinsora, A., Barron, K., Mei, Q., & Vydiswaran, V. G. V. (2017). Creating a Labeled Dataset for Medical Misinformation in Health Forums. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 456–461. https://doi.org/10.1109/ICHI.2017.93

Kohut, A., Doherty, C., Dimock, M., & Keeter, S. (n.d.). *KEY NEWS AUDIENCES NOW BLEND ONLINE AND TRADITIONAL SOURCES*. 129.

Kotonya, N., & Toni, F. (2020). Explainable Automated Fact-Checking for Public Health Claims. *ArXiv:2010.09926 [Cs]*. http://arxiv.org/abs/2010.09926

Kreidler, M. (2019, October 10). *Home Page | Quackwatch*. https://quackwatch.org/

Kuran, T., & Sunstein, C. R. (1998). *Availability Cascades and Risk Regulation* (SSRN Scholarly Paper No. 138144). https://papers.ssrn.com/abstract=138144

Leibenstein, H. (1950). Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. *The Quarterly Journal of Economics*, *64*(2), 183–207. https://doi.org/10.2307/1882692

Leong, A. Y., Sanghera, R., Jhajj, J., Desai, N., Jammu, B. S., & Makowsky, M. J. (2018). Is YouTube Useful as a Source of Health Information for Adults With Type 2 Diabetes? A South Asian Perspective. *Canadian Journal of Diabetes*, *42*(4), 395-403.e4. https://doi.org/10.1016/j.jcjd.2017.10.056

Li, A., Huang, X., Jiao, D., O'Dea, B., Zhu, T., & Christensen, H. (2018). An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. *Asia-Pacific Psychiatry*, *10*(1), e12314. https://doi.org/10.1111/appy.12314

Li, Y., Zhang, X., & Wang, S. (2017). Fake vs. Real health information in social media in China. *Proceedings of the Association for Information Science and Technology*, *54*(1), 742–743. https://doi.org/10.1002/pra2.2017.14505401139

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS* (arXiv:1912.06059). arXiv. https://doi.org/10.48550/arXiv.1912.06059

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021, Vol. 23, Page 18*, *23*(1), 18. https://doi.org/10.3390/E23010018

Lisa Lockerd Maragakis & Gabor David Kelen. (n.d.). *Covid-19—Myth Versus Fact*. Johns Hopkins Medicine. Retrieved September 22, 2021, from

https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/2019-novel-coronavirus-myth-versus-fact

List of anti-vaccination groups. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=List_of_anti-vaccination_groups&oldid=1105904574

Liu, P., Wang, L., Ranjan, R., He, G., & Zhao, L. (2022). A Survey on Active Deep Learning: From Model Driven to Data Driven. *ACM Computing Surveys*, *54*(10s), 221:1-221:34. https://doi.org/10.1145/3510414

Liu, Y., Yu, K., Wu, X., Qing, L., & Peng, Y. (2019). Analysis and Detection of Health-Related Misinformation on Chinese Social Media. *IEEE Access*, *7*, 154480–154489. https://doi.org/10.1109/ACCESS.2019.2946624

Lopuhin, M. K., Konstantin. (n.d.). *eli5: Debug machine learning classifiers and explain their predictions* (0.11.0) [Python; OS Independent]. Retrieved February 15, 2022, from https://github.com/eli5-org/eli5

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Lupi, V. (n.d.). *COVID-19 and fake news in the social media*. FBK. Retrieved October 4, 2022, from https://www.fbk.eu/en/press-releases/covid-19-and-fake-news-in-the-social-media/

Machete, P., & Turpin, M. (2020). The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible Design, Implementation and Use of*

Information and Communication Technology (pp. 235–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_20

Machine Learning: What it is and why it matters | SAS. (n.d.). Retrieved October 26, 2021, from https://www.sas.com/en_us/insights/analytics/machine-learning.html

Mahapatra, S. (2019, January 22). *Why Deep Learning over Traditional Machine Learning?* Medium. https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

Marr, B. (n.d.). *Fake News Is Rampant, Here Is How Artificial Intelligence Can Help*. Forbes. Retrieved December 15, 2021, from https://www.forbes.com/sites/bernardmarr/2021/01/25/fake-news-is-rampant-here-is-how-artificial-intelligence-can-help/

McKay, S., & Tenove, C. (2021). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, *74*(3), 703–717. https://doi.org/10.1177/1065912920938143

Meppelink, C. S., Hendriks, H., Trilling, D., van Weert, J. C. M., Shao, A., & Smit, E. S. (2021). Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Education and Counseling*, *104*(6), 1460–1466. https://doi.org/10.1016/j.pec.2020.11.013

Meppelink, C. S., Smit, E. G., Fransen, M. L., & Diviani, N. (2019). "I was Right about Vaccination": Confirmation Bias and Health Literacy in Online Health Information Seeking. *Journal of Health Communication*, *24*(2), 129–140. https://doi.org/10.1080/10810730.2019.1583701

Mikahail Korobov, & Konstantin Lopuhin. (2017). *ELI5 documentation*. https://eli5.readthedocs.io/en/latest/index.html

Misinformation. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Misinformation&oldid=1049600253

Molnar, C. (n.d.). *Interpretable Machine Learning*. Retrieved February 15, 2022, from https://christophm.github.io/interpretable-ml-book/

Molnar, C. (2019). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Leanupb.

Molnar, C. (2022). *9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/shap.html

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Communications in Computer and Information Science*, *1323*, 417–431. https://doi.org/10.1007/978-3-030-65965-3_28

Mythbusters. (n.d.). Retrieved September 28, 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

Nair, A. (2022, May 2). *Grid Search VS Random Search VS Bayesian Optimization*. Medium. https://towardsdatascience.com/grid-search-vs-random-search-vs-bayesian-optimization-2e68f57c3c46

nayak, ashutosh. (2019, December 22). *Idea Behind LIME and SHAP*. Medium. https://towardsdatascience.com/idea-behind-lime-and-shap-b603d35d34eb

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), Article 12. https://doi.org/10.1038/nbt1206-1565

Oh, H. J., & Lee, H. (2019). When Do People Verify and Share Health Rumors on Social Media? The Effects of Message Importance, Health Anxiety, and Health Literacy. *Journal of Health Communication*, *24*(11), 837–847. https://doi.org/10.1080/10810730.2019.1677824

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.568256

Oroszlányová, M., Teixeira Lopes, C., Nunes, S., & Ribeiro, C. (2018). Predicting the quality of health web documents using their characteristics. *Online Information Review*, *42*(7), 1024–1047. https://doi.org/10.1108/OIR-01-2017-0028

Our Fact Check Ratings, Explained. (n.d.). Snopes.Com. Retrieved September 22, 2021, from https://www.snopes.com/fact-check-ratings/

Pagoto, S., Waring, M. E., & Xu, R. (2019). A Call for a Public Health Agenda for Social Media Research. *Journal of Medical Internet Research*, *21*(12), e16661. https://doi.org/10.2196/16661

Pan, W., Liu, D., & Fang, J. (2021a). An Examination of Factors Contributing to the Acceptance of Online Health Misinformation. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.630268

Pan, W., Liu, D., & Fang, J. (2021b). An Examination of Factors Contributing to the Acceptance of Online Health Misinformation. *Frontiers in Psychology*, *12*, 630268. https://doi.org/10.3389/fpsyg.2021.630268

Parfenenko, Y., Verbytska, A., Bychko, D., & Shendryk, V. (2020). Application for Medical Misinformation Detection in Online Forums. *2020 International Conference on E-Health and Bioengineering (EHB)*, 1–4. https://doi.org/10.1109/EHB50910.2020.9280120

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, *9*(3), Article 3. https://doi.org/10.1038/nrd3078

Persily, N., & Tucker, J. A. (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.

Petticrew, M., Maani, N., Pettigrew, L., Rutter, H., & Van Schalkwyk, M. C. (2020). Dark Nudges and Sludge in Big Alcohol: Behavioral Economics, Cognitive Biases, and Alcohol Industry Corporate Social Responsibility. *The Milbank Quarterly*, *98*(4), 1290–1328. https://doi.org/10.1111/1468-0009.12475

Pezzo, M. V., & Beckstead, J. W. (2006). A Multilevel Analysis of Rumor Transmission: Effects of Anxiety and Belief in Two Field Experiments. *Basic and Applied Social Psychology*, *28*(1), 91–100. https://doi.org/10.1207/s15324834basp2801_8

Pfeiffer, A. (2019, June 11). *Comparing Grid and Randomized Search Methods in Python*. Medium. https://betterprogramming.pub/comparing-grid-and-randomized-search-methods-in-python-cd9fe9c3572d

Pranckevičius, T., & Marcinkevičius, V. (2016). Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification. *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 1–5. https://doi.org/10.1109/AIEEE.2016.7821805

Qualtrics XM // The Leading Experience Management Software. (n.d.). Retrieved October 16, 2022, from https://www.qualtrics.com/

Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L. (2016). The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance*, *2*(1), e5059. https://doi.org/10.2196/publichealth.5059

Raju, R., Bhandari, S., Mohamud, S. A., & Ceesay, E. N. (2021). Transfer Learning Model for Disrupting Misinformation During a COVID-19 Pandemic. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 0245–0250. https://doi.org/10.1109/CCWC51732.2021.9376066

Random forest. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1109842872

REYNOLDS, R. A., & REYNOLDS, J. L. (2002). Evidence. In *The Persuasion Handbook: Developments in Theory and Practice* (pp. 427–445). SAGE Publications, Inc. https://doi.org/10.4135/9781412976046

Ribeiro, M. T. (n.d.). *lime: Local Interpretable Model-Agnostic Explanations for machine learning classifiers* (0.2.0.1). Retrieved February 15, 2022, from http://github.com/marcotcr/lime

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. http://arxiv.org/abs/1602.04938

Röchert, D., Neubaum, G., & Stieglitz, S. (2020). Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models. In *... on Disinformation in Open Online ...*. Springer. https://link.springer.com/chapter/10.1007/978-3-030-61841-4_8

Rubin, R. (2022). When Physicians Spread Unscientific Information About COVID-19. *JAMA*, *327*(10), 904–906. https://doi.org/10.1001/jama.2022.1083

Saengkhunthod, C., Kerdnoonwong, P., & Atchariyachanvanich, K. (2021). Detection of Unreliable Medical Articles on Thai Websites. *2021 13th International Conference on Knowledge and Smart Technology (KST)*, 102–107. https://doi.org/10.1109/KST51265.2021.9415756

Saif, M. A., Medvedev, A. N., Medvedev, M. A., & Atanasova, T. (2018). Classification of online toxic comments using the logistic regression and neural networks models. *AIP Conference Proceedings*, *2048*(1), 060011. https://doi.org/10.1063/1.5082126

Sarkar, D. (DJ). (2018, December 13). *The Importance of Human Interpretable Machine Learning*. Medium. https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Schwitzer, G. (2008). How Do US Journalists Cover Treatments, Tests, Products, and Procedures? An Evaluation of 500 Stories. *PLOS Medicine*, *5*(5), e95. https://doi.org/10.1371/journal.pmed.0050095

Scott Lundberg. (2018). *SHAP documentation*. https://shap.readthedocs.io/en/latest/index.html

Seymour, B., Getman, R., Saraf, A., Zhang, L. H., & Kalenderian, E. (2015). When advocacy obscures accuracy online: Digital pandemics of public health misinformation through an antifluoride case study. *American Journal of Public Health*, *105*(3), 517–523. https://doi.org/10.2105/AJPH.2014.302437

Shah, Z., Surian, D., Dyda, A., Coiera, E., Mandl, K. D., & Dunn, A. G. (2019). Automatically Appraising the Credibility of Vaccine-Related Web Pages Shared on Social Media: A Twitter Surveillance Study. *Journal of Medical Internet Research*, *21*(11), e14007. https://doi.org/10.2196/14007

Shahsavari, S., Holur, P., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the Time of Corona: Automatic detection of Covid-19 Conspiracy Theories in Social Media and the News. *ArXiv:2004.13783 [Cs]*. http://arxiv.org/abs/2004.13783

Sharma, D. C., Pathak, A., Chaurasia, R. N., Joshi, D., Singh, R. K., & Mishra, V. N. (2020). Fighting infodemic: Need for robust health journalism in India. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1445–1447. https://doi.org/10.1016/j.dsx.2020.07.039

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques. *ArXiv:1901.06437 [Cs, Stat]*. http://arxiv.org/abs/1901.06437

Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327. https://doi.org/10.1080/1369118X.2017.1418406

Song, S., Zhao, Y., Song, X., & Zhu, Q. (2019). The Role of Health Literacy on Credibility Judgment of Online Health Misinformation. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–3. https://doi.org/10.1109/ICHI.2019.8904844

Sørensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., & Brand, H. (2012). Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health*, *12*, 80. https://doi.org/10.1186/1471-2458-12-80

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of Health Misinformation on Social Media: Systematic Review. *Journal of Medical Internet Research*, *23*(1), e17187. https://doi.org/10.2196/17187

Suthaharan, S. (2016). Support Vector Machine. In S. Suthaharan (Ed.), *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning* (pp. 207–235). Springer US. https://doi.org/10.1007/978-1-4899-7641-3_9

Tan, A. S. L., & Bigman, C. A. (2020). Misinformation About Commercial Tobacco Products on Social Media—Implications and Research Opportunities for Reducing Tobacco-Related Health Disparities. *American Journal of Public Health*, *110*(S3), S281–S283. https://doi.org/10.2105/AJPH.2020.305910

Taylor, S. (2004). Understanding and treating health anxiety: A cognitive-behavioral approach. *Cognitive and Behavioral Practice*, *11*(1), 112–123. https://doi.org/10.1016/S1077-7229(04)80015-4

Team HON. (n.d.). *Health On the Net, promotes transparent and reliable health information online through HONcode certification.* Retrieved September 24, 2021, from https://www.hon.ch/en/

Team Snopes. (n.d.). *Save Our Snopes! Contribute Now*. Snopes.Com. Retrieved September 24, 2021, from https://www.snopes.com/sos/

Thomas, R. J., Tandoc, E. C., & Hinnant, A. (2017). False Balance in Public Health Reporting? Michele Bachmann, the HPV Vaccine, and "Mental Retardation." *Health Communication*, *32*(2), 152–160. https://doi.org/10.1080/10410236.2015.1110006

Top health disinformation websites 2020. (n.d.). Statista. Retrieved October 4, 2022, from https://www.statista.com/statistics/266916/top-health-disinformation-websites-visits/

Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *WIREs Climate Change*, *11*(5), e665. https://doi.org/10.1002/wcc.665

U.S. judge sentences Novelion's Aegerion in drug marketing case. (2018, January 30). *Reuters*. https://www.reuters.com/article/us-novelion-therape-settlement-idUSKBN1FJ2ZG

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), Article 3. https://doi.org/10.1038/s41591-022-01713-6

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). *Sequence to Sequence—Video to Text*. 4534–4542.

https://openaccess.thecvf.com/content_iccv_2015/html/Venugopalan_Sequence_to_Sequence_ICCV_2015_paper.html

Verstraete, M., Bambauer, D. E., & Bambauer, J. (2017). *Identifying and Countering Fake News*. https://doi.org/10.2139/SSRN.3007971

Vraga, E. K., & Bode, L. (2017). Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication*, *39*(5), 621–645. https://doi.org/10.1177/1075547017731776

Wang, Q., & Guo, A. (2020). An efficient variance estimator of AUC and its applications to binary classification. *Statistics in Medicine*, *39*(28), 4281–4300. https://doi.org/10.1002/sim.8725

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine*, *240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media – The pilot quantitative study. *Health Policy and Technology*, *7*(2), 115–118. https://doi.org/10.1016/j.hlpt.2018.03.002

What is Artificial Intelligence (AI)? - AI Definition and How it Works. (n.d.). SearchEnterpriseAI. Retrieved October 26, 2021, from https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence

What is Deep Learning? (2022, March 30). https://www.ibm.com/cloud/learn/deep-learning

What is Logistic regression? | IBM. (n.d.). Retrieved October 12, 2022, from https://www.ibm.com/topics/logistic-regression

What is Natural Language Processing? | IBM. (n.d.). Retrieved October 26, 2021, from https://www.ibm.com/cloud/learn/natural-language-processing

Wu, S., & Flach, P. (2005). *A scored AUC Metric for Classifier Evaluation and Selection*.

Xiao, L., & Chen, S. (2020). Misinformation in the Chinese Weibo. *International Conference on Human-Computer …*. https://link.springer.com/chapter/10.1007/978-3-030-49570-1_28

Xu, S., Coman, I. A., Yamamoto, M., & Najera, C. J. (2022). Exposure Effects or Confirmation Bias? Examining Reciprocal Dynamics of Misinformation, Misperceptions, and Attitudes Toward COVID-19 Vaccines. *Health Communication*, *0*(0), 1–11. https://doi.org/10.1080/10410236.2022.2059802

Yang, Q., Sangalang, A., Rooney, M., Maloney, E., Emery, S., & Cappella, J. N. (2018). 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. https://doi.org/10.1080/10810730.2018.1448488

Zafar, M. R., & Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525–541. https://doi.org/10.3390/make3030027

Zgheib, P. W. (2017). *Advertising Deceit: Manipulation of Information, False Advertising, and Promotion* [Chapter]. Advertising and Branding: Concepts, Methodologies, Tools, and Applications; IGI Global. https://doi.org/10.4018/978-1-5225-1793-1.ch068

Zhao, Y., Da, J., & Yan, J. (2021). Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, *58*(1), 102390. https://doi.org/10.1016/j.ipm.2020.102390

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), Article 5. https://doi.org/10.3390/electronics10050593

Zhou, X., & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. In *ArXiv preprint arXiv:1812.00315*. academia.edu. https://www.academia.edu/download/61063969/reza_zafarani20191029-68261-1clpnpz.pdf

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, *53*(5), 1–40. https://doi.org/10.1145/3395046

Appendix A. Survey

Information Consent

I am a Ph.D. Candidate at the University of Wisconsin Milwaukee. I am conducting a user study on a system, which is designed to assist people in evaluating the quality of health news systematically using a variety of criteria. I would like to hear your input on it.

This survey contains credits to get free survey responses at SurveySwap.io. However, no credit would be awarded if a participant 1. declines the consent 2. is younger than 18 3. fails the attention check question

University of Wisconsin-Milwaukee
Informed Consent to Participate in Research

Study title: "Do you trust me?" Understanding users' trust for evaluating explainability in a health news quality evaluation system

Researcher[s]: Xiaoyu Liu, Ph.D. student in Biomedical and Health Informatics, Department of Computer Science

I am inviting you to take a survey for research. This survey is completely voluntary. There are no negative consequences if you don't want to take it. If you start the survey, you can always change your mind and stop at any time.

What is the purpose of this study?
I am conducting a user study on an AI-based application that is designed to assist people in evaluating the quality of health news systematically using a variety of criteria. I would like to know your feedback on the system.

What will I do?
I'll begin by inviting you to freely explore and test all of the prototype system's functions. Following that, you will be asked to complete a survey about your experience with the prototype. The survey is divided into three sections: demographic information, questions regarding your experience and trust with the evaluation result provided by the system, and open-ended questions to let you share your experiences and thoughts about potential approaches to counteract health misinformation. The entire process will take approximately 5-7 minutes.

Risks

- The study risk is minimal. However, risks associated with the study may include tiredness, and stress while reading the survey. To mitigate this risk, you can skip any questions or quit the survey at any time.
- Some questions may be personal or upsetting. You can skip them or quit the survey at any time.
- Online data being hacked or intercepted: Anytime you share information online there are risks. We're using a secure system to collect this data, but we can't completely eliminate this risk.
- Breach of confidentiality: There is a chance your data could be seen by someone who shouldn't have access to it. We're minimizing this risk in the following ways:
  - Data is anonymous.
  - We'll store all electronic data on a password-protected, encrypted computer.

Possible benefits: The study will help me figure out if the system's explanations can increase end users' trust in and acceptance of the system. The findings of the study may also aid in the design, and development of AI-based systems aimed at counteracting health misinformation in a more trustworthy and effective manner. As study participants, you will learn how to think critically about health news and get familiar with a set of criteria that will assist you in analyzing health news systematically.

Estimated number of participants: 600

How long will it take? 5-7 mins

Costs: None

Compensation: None

Future research: Your data won't be used or shared for any future research studies.

Confidentiality and Data Security
No identifying information will be collected from this study.

Where will data be stored? On the researchers' computers and On the servers for the online survey software (Qualtrics).

How long will it be kept? Until Jun 09 2025.

Who can see my data?

- The researcher will have access to your responses to all questions. This is so we can analyze the data and conduct the study.
- Agencies that enforce legal and ethical guidelines, such as
  - The Institutional Review Board (IRB) at UWM
- We may share our findings in publications or presentations. If we do, the results will be presented using grouped data, with no individual results. If we quote you, we'll use pseudonyms (fake names).

Questions about the research, complaints, or problems: Contact Xiaoyu Liu, 6154243800, liu267@uwm.edu

Questions about your rights as a research participant, complaints, or problems: Contact the UWM IRB (Institutional Review Board) at 414-662-3544 / irbinfo@uwm.edu.

Please print or save this screen if you want to be able to access the information later.
IRB #:2.303
IRB Approval Date: Jun 10 2022

Agreement to Participate
Your participation is completely voluntary, and you can withdraw at any time.
To take this survey, you must be:

- **At least 18 years old**
- **In the US**

If you meet these criteria and would like to take the survey, click the button below to start.

- ◯ I consent, begin the study
- ◯ I do not consent, I wish to quit the study

Demographic Information

Age

Sex

- ◯ Male
- ◯ Female
- ◯ Non-binary
- ◯ Prefer not to say

Compensation: None

Future research: Your data won't be used or shared for any future research studies.

Confidentiality and Data Security
No identifying information will be collected from this study.

Where will data be stored? On the researchers' computers and On the servers for the online survey software (Qualtrics).

How long will it be kept? Until Jun 09 2025.

Who can see my data?
- The researcher will have access to your responses to all questions. This is so we can analyze the data and conduct the study.
- Agencies that enforce legal and ethical guidelines, such as
  - The Institutional Review Board (IRB) at UWM
- We may share our findings in publications or presentations. If we do, the results will be presented using grouped data, with no individual results. If we quote you, we'll use pseudonyms (fake names).

Questions about the research, complaints, or problems: Contact Xiaoyu Liu, 6154243800, liu267@uwm.edu

Questions about your rights as a research participant, complaints, or problems: Contact the UWM IRB (Institutional Review Board) at 414-662-3544 / irbinfo@uwm.edu.

Please print or save this screen if you want to be able to access the information later.
IRB #:2.303
IRB Approval Date: Jun 10 2022

Agreement to Participate
Your participation is completely voluntary, and you can withdraw at any time.
To take this survey, you must be:
- **At least 18 years old**
- **In the US**

If you meet these criteria and would like to take the survey, click the button below to start.

◯ I consent, begin the study

◯ I do not consent, I wish to quit the study

Demographic Information

Age

Sex

◯ Male

◯ Female

◯ Non-binary

◯ Prefer not to say

	Never	Rarely	Sometimes	Frequently	Always
	1	2	3	4	5
Mobile phone					
Email					
Internet					
Social media (Facebook, Twitter, etc.)					
Talk with family, friends or colleagues					

Please tell us how much you trust the health information from each source.

	Not at all	Not very much	Somewhat	Quite a lot	Completely
	1	2	3	4	5
Daily newspaper					
TV news					
Radio news					
Mobile phone					
Email					
Internet					
Social media (FacebooK, Twitter etc.)					
Talk with family, friends or colleagues					

Please read statements below and indicate how much you agree or disagree with each of these statements

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Science and technology are making our lives healthier, easier, and more comfortable.	○	○	○	○	○
Artificial intelligence (AI) is making our lives healthier, easier, and more comfortable.	○	○	○	○	○

Trust in ACE

Suppose there is a lady suffering from hot flashes - a frequent side effect of menopause or breast cancer. To help her, you looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation: https://www.figma.com/proto/NkHTmcDtRUo03JoR82Qz2S/ACE-L?node-id=0%3A1
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the **bottom right of this page** to continue the survey.

Suppose there is a lady suffering from hot flashes - a frequent side effect of menopause or breast cancer. To help her, you looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation: https://www.figma.com/proto/l0oUZG90l4FfVfvzDWayOI/ACE-M?node-id=0%3A1.
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the **bottom right of this page** to continue the survey.

Suppose there is a lady suffering from hot flashes, a frequent side effect of menopause or breast cancer. To help her, you looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation: https://www.figma.com/proto/9DiXBxl9ZFlE3M5DrXAlaw/ACE-H?node-id=0%3A1
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the **bottom right of this page** to continue the survey.

How many criteria does the system prototype employ to help users evaluate health news quality?

- ○ 5
- ○ 10

Please read statements below and indicate how much you agree or disagree with each of these statements

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I trust the system's evaluation for this article.	○	○	○	○	○
The highlighted sentences are helpful for me to understand the evaluation.	○	○	○	○	○

Please rate the importance of each criterion in terms how well it assists you in gauging the quality of health news.

	Not important at all 1	2	3	4	Extremely important 5
1. The costs of the intervention					
2. The benefits of the intervention					
3. The harms of the intervention					
4. The quality of the evidence					
5. Commit disease-mongering					
6. Conflict of interest					
7. Comparing Existing alternatives					
8.Availability of the treatment					

	Not important at all 1	2	3	4	Extremely important 5
9. True novelty of the intervention					
10. Unjustifiable, Sensational languages					

What else do you think can help reduce health misinformation?

Trust in ACU

Suppose there is a lady suffering from hot flashes - a frequent side effect of menopause or breast cancer. To help her, you looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation:
https://www.figma.com/proto/jMv3t6KKHED46oCOVzhuE7/ACU-L?node-id=0%3A1
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the bottom right of this page to continue the survey.

Suppose there is a lady suffering from hot flashes, a frequent side effect of menopause or breast cancer. To help her, you looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation:
https://www.figma.com/proto/vYcVrRRQpVrWxLljh7bf1T/ACU-M?node-id=0%3A1
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the bottom right of this page to continue the survey.

Suppose there is a lady suffering from hot flashes, a frequent side effect of menopause or breast cancer. To help her, You looked up treatment options on the internet and found an article about a drug that may cut hot flashes. Please click the link to read this article and check out the news quality evaluation:
https://www.figma.com/proto/4SnW69kZWATTailKY3JRci/ACU-H?node-id=0%3A1
Note: No need to read the entire news. For the best experience, please open the link on a desktop.

After your exploration, you will be asked how you think about the evaluation result.

Please click → at the bottom right of this page to continue the survey.

How many criteria does the system prototype employ to help users evaluate health news quality?

- ○ 5
- ○ 10

Please read statements below and indicate how much you agree or disagree with each of these statements

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I trust the system's evaluation for this article.	○	○	○	○	○
I need more explanations to help me understand the evaluation	○	○	○	○	○

Please rate the importance of each criterion in terms how well it assists you in gauging the quality of health news.

	Not important at all 1	2	3	4	Extremely important 5
1. The costs of the intervention					
2. The benefits of the intervention					
3. The harms of the intervention					
4. The quality of the evidence					
5. Commit disease-mongering					
6. Conflict of interest					
7. Comparing Existing alternatives					

	Not important at all 1	2	3	4	Extremely important 5
8.Availability of the treatment					
9. True novelty of the intervention					
10. Unjustifiable, Sensational languages					

What else do you think can help reduce health misinformation?

Appendix B. Performance of different base classifiers for automating nine criteria evaluation

Word Feature Count	Base Classifier	Measure	Cost	Benefit	Harm	Quality	Mongering	Conflict	Alternative	Availability	Novelty
500	RF	AUC	0.8570	0.6757	0.7078	0.7034	**0.6082**	**0.7016**	0.6095	0.6789	0.6026
		Precision	0.8188	0.6438	0.7114	0.6751	0.7430	0.6514	0.5850	0.6682	0.6210
	SVM	AUC	0.8224	0.6899	0.6881	0.6451	0.5267	0.6371	0.5901	0.6364	0.5767
		Precision	0.8338	0.6927	0.6844	0.6398	0.7179	0.6028	0.5726	0.6258	0.6152
	NB	AUC	0.7743	0.6639	0.6528	0.6069	0.5680	0.5850	0.5937	0.6726	0.5925
		Precision	0.8055	0.6704	0.6709	0.6109	0.7350	0.5680	0.5742	0.6578	0.6290
	LR	AUC	0.8306	0.6973	0.6976	0.6587	0.5511	0.6418	0.5930	0.6565	0.5911
		Precision	0.8345	0.6939	0.6851	0.6467	0.7255	0.5990	0.5675	0.6411	0.6221
1000	RF	AUC	0.8748	0.7019	0.7247	0.7013	0.5890	0.6925	0.6078	0.6867	0.5905
		Precision	0.8193	0.6914	0.7166	0.6850	0.7446	0.6463	0.5936	0.6577	0.6278
	SVM	AUC	0.8229	0.6972	0.6942	0.6528	0.5561	0.6135	0.5910	0.6329	0.5825
		Precision	0.8296	0.6988	0.6890	0.6394	0.7148	0.5809	0.5628	0.6204	0.6198
	NB	AUC	0.7730	0.6666	0.6588	0.6050	0.5826	0.5679	0.6016	0.6662	0.5936
		Precision	0.8039	0.6598	0.6753	0.6122	0.7393	0.5430	0.5807	0.6411	0.6254
	LR	AUC	0.8255	0.7074	0.6982	0.6669	0.5702	0.6271	0.6008	0.6590	0.5944
		Precision	0.8219	0.6929	0.6951	0.6420	0.7257	0.5820	0.5652	0.6322	0.6336
2000	RF	AUC	**0.8750**	0.6969	**0.7313**	0.7037	0.5909	0.6980	0.6007	**0.6925**	**0.6098**
		Precision	0.8224	0.6770	0.7215	0.6838	0.7347	0.6547	0.5793	0.6557	0.6270
	SVM	AUC	0.8246	0.6937	0.6795	0.6574	0.5713	0.6233	0.6093	0.6251	0.5912
		Precision	0.8287	0.6855	0.6856	0.6430	0.7219	0.5820	0.5791	0.6143	0.6198
	NB	AUC	0.7847	0.6762	0.6572	0.6106	0.5880	0.5808	0.6093	0.6629	0.5970
		Precision	0.7997	0.6753	0.6688	0.6037	0.7387	0.5674	0.5791	0.6320	0.6270
	LR	AUC	0.8277	**0.7032**	0.6937	0.6685	0.5793	0.6368	0.6115	0.6588	0.6036
		Precision	0.8170	0.7043	0.6914	0.6481	0.7333	0.6037	0.5720	0.6326	0.6380
4000	RF	AUC	0.8715	0.6996	0.7213	**0.7038**	0.5798	0.6924	**0.6146**	0.6912	0.5970
		Precision	0.8173	0.6961	0.7315	0.6573	0.7536	0.6642	0.5678	0.6553	0.6343
	SVM	AUC	0.8165	0.6920	0.6793	0.6602	0.5917	0.6307	0.5984	0.6256	0.5806
		Precision	0.8132	0.6864	0.6868	0.6538	0.7383	0.5979	0.5653	0.6026	0.6289
	NB	AUC	0.7855	0.6809	0.6532	0.6159	0.6062	0.5881	0.6060	0.6645	0.5915
		Precision	0.8012	0.6660	0.6672	0.6147	0.7439	0.5635	0.5804	0.6347	0.6259
	LR	AUC	0.8246	0.7031	0.6931	0.6665	0.6050	0.6425	0.6131	0.6603	0.5926
		Precision	0.8227	0.7006	0.6960	0.6503	0.7425	0.6044	0.5802	0.6330	0.6355

Appendix C. Top 30 word features with their feature weights in the Quality classifier that is built on RF

Weight	Feature
0.0127 ± 0.0474	study
0.0111 ± 0.0484	group
0.0090 ± 0.0453	medicine
0.0088 ± 0.0403	evidence
0.0080 ± 0.0332	one
0.0070 ± 0.0358	placebo
0.0069 ± 0.0299	researcher
0.0069 ± 0.0328	trial
0.0066 ± 0.0323	data
0.0065 ± 0.0361	england
0.0064 ± 0.0369	new england
0.0061 ± 0.0362	journal medicine
0.0053 ± 0.0261	result
0.0053 ± 0.0317	england journal
0.0051 ± 0.0288	editorial
0.0050 ± 0.0306	randomly
0.0046 ± 0.0258	known
0.0045 ± 0.0230	medical
0.0044 ± 0.0283	assigned
0.0042 ± 0.0227	get
0.0041 ± 0.0201	may
0.0036 ± 0.0218	yet
0.0036 ± 0.0253	randomly assigned
0.0035 ± 0.0190	published
0.0035 ± 0.0200	still
0.0034 ± 0.0218	randomized
0.0034 ± 0.0198	whether
0.0034 ± 0.0202	early
0.0032 ± 0.0205	would
0.0028 ± 0.0191	typically
… 3970 more …	

Appendix D. Top 30 word features with their feature weights in the Mongering classifier that is built on RF

Weight	Feature
0.0158 ± 0.0384	benefit
0.0097 ± 0.0245	health
0.0095 ± 0.0244	study
0.0093 ± 0.0269	case
0.0082 ± 0.0268	survival
0.0079 ± 0.0247	group
0.0079 ± 0.0240	level
0.0073 ± 0.0217	medicine
0.0073 ± 0.0206	researcher
0.0068 ± 0.0219	brain
0.0060 ± 0.0205	approach
0.0056 ± 0.0200	month
0.0055 ± 0.0200	public
0.0054 ± 0.0151	may
0.0054 ± 0.0161	research
0.0052 ± 0.0135	new
0.0052 ± 0.0148	year
0.0051 ± 0.0160	university
0.0051 ± 0.0147	one
0.0049 ± 0.0150	said
0.0048 ± 0.0161	medical
0.0047 ± 0.0150	risk
0.0047 ± 0.0170	important
0.0046 ± 0.0133	also
0.0045 ± 0.0152	american
0.0045 ± 0.0176	death
0.0043 ± 0.0129	patient
0.0043 ± 0.0156	author
0.0042 ± 0.0147	result
0.0042 ± 0.0140	published
… 470 more …	

Appendix E. Top 30 word features with their feature weights in the Alternative classifier that is built on RF

Weight	Feature
0.0050 ± 0.0211	effective
0.0033 ± 0.0151	year
0.0032 ± 0.0148	one
0.0029 ± 0.0133	treatment
0.0029 ± 0.0143	better
0.0029 ± 0.0165	yet
0.0028 ± 0.0118	study
0.0023 ± 0.0118	need
0.0022 ± 0.0112	university
0.0022 ± 0.0122	medication
0.0021 ± 0.0134	two year
0.0021 ± 0.0104	result
0.0020 ± 0.0113	case
0.0020 ± 0.0119	standard
0.0020 ± 0.0106	compared
0.0020 ± 0.0107	say
0.0020 ± 0.0099	used
0.0019 ± 0.0114	include
0.0019 ± 0.0098	patient
0.0019 ± 0.0105	two
0.0019 ± 0.0123	effectiveness
0.0018 ± 0.0105	way
0.0018 ± 0.0091	may
0.0018 ± 0.0104	early
0.0018 ± 0.0100	available
0.0016 ± 0.0103	development
0.0016 ± 0.0082	found
0.0016 ± 0.0086	finding
0.0016 ± 0.0090	benefit
0.0016 ± 0.0104	option
… 3970 more …	

Appendix F. Top 30 word features with their feature weights in the Availability classifier that is built on RF

Weight	Feature
0.0095 ± 0.0299	drug
0.0092 ± 0.0319	company
0.0071 ± 0.0211	time
0.0071 ± 0.0206	year
0.0058 ± 0.0282	food drug
0.0054 ± 0.0262	drug administration
0.0050 ± 0.0168	called
0.0046 ± 0.0232	administration
0.0046 ± 0.0142	test
0.0044 ± 0.0188	approved
0.0042 ± 0.0201	fda
0.0041 ± 0.0145	still
0.0039 ± 0.0159	taken
0.0038 ± 0.0125	professor
0.0037 ± 0.0136	would
0.0036 ± 0.0149	doctor
0.0035 ± 0.0102	new
0.0035 ± 0.0128	without
0.0035 ± 0.0121	say
0.0034 ± 0.0102	people
0.0034 ± 0.0093	may
0.0034 ± 0.0136	tested
0.0034 ± 0.0188	approval
0.0033 ± 0.0087	study
0.0033 ± 0.0154	diet
0.0031 ± 0.0091	said
0.0031 ± 0.0118	make
0.0031 ± 0.0090	research
0.0031 ± 0.0139	made
0.0030 ± 0.0083	also
… 1970 more …	

Appendix G. Top 30 word features with their feature weights in the Novelty classifier that is built on RF

Weight	Feature
0.0049 ± 0.0215	new england
0.0046 ± 0.0162	side
0.0046 ± 0.0201	journal medicine
0.0045 ± 0.0162	healthy
0.0044 ± 0.0148	trial
0.0044 ± 0.0188	england journal
0.0041 ± 0.0142	drug
0.0041 ± 0.0123	say
0.0040 ± 0.0146	another
0.0039 ± 0.0116	patient
0.0038 ± 0.0113	help
0.0038 ± 0.0135	genetic
0.0037 ± 0.0134	last
0.0037 ± 0.0148	follow
0.0036 ± 0.0145	either
0.0035 ± 0.0157	england
0.0035 ± 0.0112	effect
0.0034 ± 0.0121	symptom
0.0033 ± 0.0132	side effect
0.0032 ± 0.0093	study
0.0030 ± 0.0099	researcher
0.0030 ± 0.0116	information
0.0030 ± 0.0120	expert
0.0030 ± 0.0106	evidence
0.0029 ± 0.0088	also
0.0029 ± 0.0091	new
0.0029 ± 0.0092	year
0.0028 ± 0.0086	one
0.0028 ± 0.0087	said
0.0028 ± 0.0111	fda
… 1970 more …	