European Journal of Computer Science and Information Technology,13(25),63-85,2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Leveraging AI/NLP to Combat Health Misinformation and Promote Trust in Science

Sree Priyanka Uppu
University of Southern California, Los Angeles, USA

doi: https://doi.org/10.37745/ejcsit.2013/vol13n256385 Published May 21, 2025

Citation: Uppu, SP (2025) Leveraging AI/NLP to Combat Health Misinformation and Promote Trust in Science, *European Journal of Computer Science and Information Technology*,13(25),63-85

Abstract: *The proliferation of health misinformation online poses a significant threat to public well-being and erodes trust in scientific consensus. Artificial Intelligence and Natural Language Processing offer powerful tools for identifying and countering such misinformation across digital platforms. By examining techniques like concept clustering and bot detection as applied to e-cigarette discussions on social media, this paper illuminates how these technologies can detect problematic content and proactively promote accurate scientific information. The analysis reveals patterns in how misinformation spreads through automated accounts, emotional triggers, and network effects. Beyond detection capabilities, AI can generate accessible scientific content, tailor communication to address public concerns, and personalize health messaging for diverse audiences. Despite promising applications, implementation faces challenges including distinguishing nuance from falsehood, addressing algorithmic bias, balancing free expression with harm prevention, ensuring system transparency, adapting to evolving tactics, and integrating human oversight effectively. Developing ethical AI solutions for health communication requires balancing technological capabilities with human expertise while safeguarding fundamental rights.*

Keywords: artificial intelligence, bot detection, health misinformation, information ecosystems, sentiment analysis

INTRODUCTION

The digital transformation of the 21st century has fundamentally altered how health information is created, disseminated, and consumed. While democratizing access to health knowledge, digital platforms have simultaneously become vectors for an unprecedented proliferation of misinformation. The World Health Organization (WHO) has acknowledged that alongside the COVID-19 pandemic, societies worldwide have been fighting an "infodemic"—an overabundance of information, some accurate and some not, that spreads alongside the disease [1]. During the early phase of the COVID-19 pandemic, approximately 800 people died, and around 5,800 were hospitalized globally because of misinformation related to unfounded

treatments, demonstrating the severe consequences of this phenomenon [1]. Effective responses to health misinformation require precision in defining what constitutes misinformation, as Vraga and Bode argue that misinformation is bounded by both expertise and evidence, providing a framework that helps distinguish between deliberate falsehoods, outdated information, and matters of scientific uncertainty [11]. This "infodemic" has manifested in various forms beyond COVID-19, from vaccine hesitancy campaigns to the promotion of unverified treatments that have been linked to adverse health outcomes in vulnerable populations.

The stakes of this information crisis extend beyond individual health decisions. During the COVID-19 pandemic, misinformation flooded various platforms, with recent studies showing that the potential reach of COVID-19 misinformation on Twitter amounted to over 100 million users [2]. Health misinformation can spread with remarkable speed on social media platforms, with engagement rates for health-related false content sometimes exceeding those of accurate information. Approximately 65% of social media users who engage with health content report encountering misleading health information within their networks, according to findings reported in the literature [2]. Furthermore, the erosion of trust in scientific institutions threatens to undermine collective public health responses to emerging challenges, as evidenced during recent global health emergencies.

The scale of this problem has overwhelmed traditional fact-checking mechanisms. Manual content moderation cannot keep pace with the volume of potential health misinformation spreading across digital platforms. This challenge is compounded by the "information cascade" phenomenon, whereby misinformation that reaches a critical mass of shares can appear credible simply due to its popularity, regardless of its factual accuracy [1]. The exponential growth of user-generated content across multiple platforms, languages, and formats has created an environment where manual intervention alone is increasingly ineffective.

Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies offer promising solutions to this complex challenge. These computational approaches can operate at scale, analyzing millions of posts, comments, and articles in real-time to identify patterns indicative of misinformation. Digital tools that leverage machine learning algorithms can be employed to enhance monitoring and analysis of online discussions, particularly in identifying emerging health topics and potential misinformation trends [2]. By utilizing these advanced technologies, systems can be trained to recognize not only known falsehoods but also to detect emerging narratives and novel misinformation tactics as they evolve.

This paper examines the potential of AI/NLP methodologies in addressing health misinformation, with particular attention to their application in social media environments where such content often originates and spreads most rapidly. Drawing on methodologies employed in the analysis of e-cigarette discussions on social platforms, we will explore how techniques such as concept clustering, sentiment analysis, and automated bot detection can be harnessed to identify, track, and counter misleading health information. Analysis of platforms like Twitter has already shown promise in tracking public discourse around health

topics, with researchers successfully using natural language processing to categorize tweets and identify content patterns related to various health conditions [2].

Beyond detection capabilities, we will investigate how these technologies can be deployed to actively promote science literacy and evidence-based health communication. This includes the development of AI-powered tools for generating accessible scientific content, real-time fact-checking systems integrated into digital platforms, and personalized educational interventions designed to build critical evaluation skills among users. The WHO has recognized the importance of such technological approaches, noting that they can help health authorities disseminate reliable information more efficiently while countering the spread of harmful content [1].

However, the implementation of these technologies raises important ethical considerations that must be carefully addressed. Questions regarding algorithmic bias, the boundaries between misinformation and legitimate scientific debate, potential limitations on free expression, and the transparency of automated moderation systems all require thoughtful examination. Studies have shown that balanced technological approaches must be combined with community engagement and education to effectively combat health misinformation [2]. This paper will outline a framework for the responsible development and deployment of AI/NLP tools in this sensitive domain, emphasizing the importance of human oversight, cross-disciplinary collaboration, and ongoing evaluation of both efficacy and ethical implications.
As health misinformation continues to evolve in sophistication and reach, the integration of AI/NLP approaches represents not merely a technological solution but a necessary evolution in our collective response to this pressing public health challenge. By combining computational methods with human expertise and ethical guidelines, we can work toward creating more resilient information ecosystems that prioritize evidence-based health communication.

The Digital Infodemic and E-Cigarette Misinformation

Characterizing the Health Infodemic Phenomenon

The term "infodemic" has gained prominence in recent years to describe the rapid and widespread dissemination of information—both accurate and inaccurate—that occurs during health crises or around emerging health concerns. This phenomenon creates an environment where distinguishing reliable health information from misinformation becomes increasingly challenging for the average consumer. Electronic cigarettes (e-cigarettes) represent a particularly illustrative case study of this information crisis, as they exist at the intersection of emerging technology, consumer health products, and evolving scientific understanding.The COVID-19 pandemic has served as a case study in how health misinformation spreads online, with analysis of Twitter conversations revealing distinct patterns of misinformation narratives involving treatments, transmission, and severity claims that evolved in real-time alongside the disease itself [12]. Social media platforms have become central battlegrounds in the information war surrounding e-cigarettes. A computational analysis of Twitter conversations about e-cigarettes identified 1,669,123 tweets

generated between 2014 and 2016, demonstrating the substantial volume of online discourse surrounding these products [3]. Within this vast corpus of content, researchers discovered that the Twitter discourse around e-cigarettes is heavily influenced by commercial interests, with advertising and promotional content constituting a significant portion of the overall conversation. This digital ecosystem has become a complex mixture of evidence-based information, commercial messaging, personal anecdotes, and potentially misleading claims.

The spectrum of e-cigarette information includes significant discussion about health aspects and cessation. Research indicates that approximately 37% of Twitter users who post about e-cigarettes discuss these products in relation to smoking cessation, suggesting that perceptions of e-cigarettes as quitting aids are widespread despite mixed evidence regarding their effectiveness for this purpose [3]. The potential for misinformation is particularly high in this domain, as users may share personal success stories that do not necessarily reflect the scientific consensus or broader population-level outcomes.

Sources and Propagation Mechanisms

Understanding the origins and propagation mechanisms of e-cigarette information is essential for developing effective monitoring and intervention strategies. Research has identified several key actors contributing to the digital discourse surrounding these products. A significant finding is the presence of social bots actively participating in e-cigarette discussions. In one comprehensive analysis, researchers determined that 29.35% of Twitter users discussing e-cigarettes were likely bots rather than human users, based on established detection algorithms [3]. This high proportion of automated accounts suggests orchestrated efforts to shape public perception of e-cigarettes.

The role of these automated accounts extends beyond mere participation in the conversation. Analysis reveals that bot accounts were more likely to post content about e-cigarettes that was commercial or promotional in nature. Specifically, these automated accounts were responsible for 32.01% of tweets promoting e-cigarettes directly or indirectly, compared to non-bot accounts that generally posted more diverse content, including personal experiences and opinions [3]. The disproportionate focus of bot accounts on promotional content raises concerns about potential commercial manipulation of online discourse.

Beyond automated accounts, commercial entities generate significant volumes of content promoting e-cigarettes, sometimes using sophisticated marketing techniques. A related analysis of e-cigarette marketing on Instagram found that e-cigarette companies primarily use the platform to promote their brands through high-quality lifestyle posts and images, with companies posting an average of 5.8 times per week [4]. This frequent posting schedule helps maintain visibility and shapes perceptions of e-cigarettes as lifestyle products rather than health-related devices.

Content Analysis and Concept Clustering

The application of concept clustering and other computational text analysis techniques has proven invaluable for understanding the landscape of e-cigarette discussions online. These methods allow researchers to process vast quantities of social media content and identify prominent themes, narratives, and information clusters without requiring predetermined categories.

When applied to Twitter discourse surrounding e-cigarettes, natural language processing and topic modeling revealed several distinct narrative threads. Analysis identified five major topics dominating the conversation: (1) personal experiences and opinions (20.93% of tweets); (2) advertisements and promotions (20.29%); (3) policy and government regulations (14.76%); (4) cessation and health (17.64%); and (5) flavors and specific components (14.45%) [3]. This distribution highlights the multifaceted nature of e-cigarette discussions, with personal and commercial content outweighing policy and health considerations. Content analysis also revealed geographic patterns in e-cigarette discourse. Geotagged tweets about e-cigarettes were identified from all 50 U.S. states, with varying levels of activity. States with the highest volume of e-cigarette-related tweets included California (17.91%), New York (12.12%), Texas (6.33%), and Florida (5.59%), suggesting potential regional differences in interest or marketing activity [3]. Such geographic variations may reflect differences in regulatory environments, market penetration, or cultural attitudes toward these products.

Table 1. Major Topics in E-Cigarette Twitter Discourse [3]

Topic Category	Percentage of Tweets
Personal experiences/opinions	20.93%
Advertisements/promotions	20.29%
Cessation and health	17.64%
Policy and government regulations	14.76%
Flavors and specific components	14.45%

The Role of Automated Accounts and Marketing Strategies

Bot detection represents another crucial application of AI in understanding the e-cigarette information ecosystem. By analyzing behavioral patterns, posting frequencies, and content similarities, machine learning algorithms can distinguish between human users and automated accounts with increasing accuracy. This distinction is critical because bot networks can artificially amplify certain narratives, creating the illusion of widespread belief or consensus where none may exist. Analysis of temporal patterns in bot activity related to e-cigarettes revealed interesting trends. Researchers found that automated accounts were more active during weekdays than weekends, with peak activity occurring on Wednesdays, when they accounted for 35.71% of all e-cigarette-related tweets [3]. This pattern differs from human posting behavior, which tends to be more evenly distributed throughout the week, and suggests coordinated, possibly commercial, messaging strategies rather than organic conversation.

On Instagram, a platform increasingly important for e-cigarette marketing, content analysis revealed specific strategies employed to promote these products. Research showed that 41.8% of promotional posts featured vapor clouds, while 27.0% incorporated price promotions or discounts [4]. The visual nature of Instagram allows marketers to showcase product aesthetics, with 57.3% of posts featuring devices with unique or visually appealing designs. Additionally, 34.4% of posts contained smoking cessation or health claims, suggesting a deliberate attempt to position e-cigarettes as health-enhancing products [4].

The marketing strategies extend beyond visual content to community building. Analysis found that 83.6% of Instagram posts included brand-specific hashtags to create recognizable communities, while 35.2% used Instagram-specific strategies like contests to drive engagement [4]. These tactics help create brand loyalty and social identity around e-cigarette products, potentially influencing perceptions and usage patterns among followers.

Through the combined application of concept clustering, sentiment analysis, and bot detection, researchers have begun to develop a more nuanced understanding of how e-cigarette information and misinformation spreads online. The identification of automated accounts as major contributors to the discourse, along with the documentation of specific marketing strategies employed on visual platforms like Instagram, provides crucial context for understanding how public perceptions of these products are shaped. These insights are essential for designing targeted interventions that can effectively monitor potentially misleading narratives while promoting evidence-based information about these controversial products.

European Journal of Computer Science and Information Technology,13(25),63-85,2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

AI/NLP Techniques for Combating Health Misinformation

The rapid evolution of artificial intelligence (AI) and natural language processing (NLP) has created unprecedented opportunities for identifying, analyzing, and countering health misinformation at scale. These technologies provide sophisticated mechanisms to process vast quantities of textual data, extract meaningful patterns, and implement targeted interventions. While technological solutions are essential for addressing the scale of health misinformation, Swire-Thompson and Lazer emphasize that effective interventions must be designed with an understanding of the psychological factors that make people vulnerable to misinformation, including cognitive biases, health literacy gaps, and information processing tendencies [13]. This section examines the specific AI/NLP techniques that show particular promise in addressing the health misinformation challenge.

Bot Detection: Identifying Automated Amplifiers

Automated accounts represent a significant force in the propagation of health misinformation across digital platforms. The detection and classification of social bots has become an essential component in understanding how health information and misinformation spread online. A systematic approach to bot detection involves analyzing account behavior, content patterns, and network characteristics to differentiate between human and automated users. Research indicates that bots can constitute between 9% and 15% of active accounts during health-related discussions, making their identification crucial for understanding information dissemination patterns [5]. The identification and characterization of social bots represents a critical challenge in misinformation research, as Ferrara et al. demonstrate that sophisticated bots can mimic human behavior with increasing precision, employing temporal patterns, content diversity, and network structures that evade simple detection methods [16].

Modern bot detection systems employ sophisticated machine learning algorithms that evaluate multiple dimensions of account behavior simultaneously. Standard approaches include the Botometer (formerly BotOrNot) system, which analyzes over 1,000 features related to user profile, friend networks, temporal activity patterns, language, and sentiment. In applications to health-related content, such systems have achieved detection accuracy rates between 85% and 95%, though performance varies based on the sophistication of the bots being analyzed [5].

Table 2. Automated Account Influence in Online Health Discourse [3, 5]

Metric	Value
Bot accounts in e-cigarette discussions	29.35%
Bot share of promotional e-cigarette tweets	32.01%
Peak bot activity (Wednesdays)	35.71%
Bot detection accuracy for health content	85-95%
Reduction in extreme content after bot removal	11-13%

The practical impact of bot detection in health contexts is substantial. Studies examining vaccination discussions on social media have found that after removing suspected bot accounts from analysis, the proportion of content exhibiting extreme anti-vaccination sentiments decreased by approximately 11-13%, suggesting that automated accounts may disproportionately promote contentious or polarizing health messages [5]. This finding highlights the importance of distinguishing between organic human discourse and artificially amplified messaging when analyzing health communication patterns.

Natural Language Understanding: Claim Extraction and Verification

Beyond identifying the sources of misinformation, NLP techniques enable the automated extraction and assessment of specific health claims from unstructured text. Natural Language Understanding (NLU) systems can parse social media posts, news articles, and forum discussions to identify assertions that can be evaluated against scientific evidence.

Claim detection represents a foundational capability for addressing health misinformation. Recent advances in this domain include the development of deep learning models that can identify both explicit statements and implicit claims within health-related text. These systems typically process textual input through multiple analytical layers, including syntactic parsing, named entity recognition, and semantic role labeling, to extract structured representations of health assertions. In evaluations against expert-annotated datasets, modern claim extraction systems have demonstrated precision rates of 79-87% and recall rates of 72-81% for identifying health-related claims in social media content [6].

The verification of extracted claims requires comparison against reliable knowledge sources. Automated approaches to this challenge include knowledge graph-based reasoning, where claims are matched against structured databases of scientific findings, and evidence retrieval systems that search medical literature for relevant information. A comprehensive analysis of automated fact-checking approaches for health content found that hybrid systems combining knowledge graphs with natural language inference models achieved accuracy rates of 76% for ruling on the veracity of health claims, though performance dropped to 61-67% for claims requiring specialized domain knowledge [6].

Sentiment Analysis: Understanding Emotional Drivers

Misinformation often spreads through appeals to emotion rather than reason, making sentiment analysis a valuable tool for understanding and countering such content. Advanced sentiment analysis goes beyond simple positive/negative classification to identify specific emotional states and rhetorical strategies that may signal problematic content.

Research examining the emotional characteristics of health information has revealed distinct patterns associated with misinformation. A systematic analysis of vaccine-related content found that posts containing misinformation exhibited significantly higher emotional intensity scores (mean 0.57 on a 0-1 scale) compared to factual information (mean 0.41), with particularly elevated levels of fear and anger

expressions [6]. This emotional loading may contribute to the "virality" of misinformation, as content triggering strong emotional responses tends to receive greater engagement and sharing.

The application of fine-grained emotion detection models to health content has proven particularly valuable for identifying potential misinformation. These systems typically employ lexicon-based approaches combined with machine learning classifiers to categorize content according to specific emotional categories. When applied to a corpus of COVID-19 discussions, such analysis identified that content later verified as misinformation displayed distinct emotional signatures, with 27% higher fear expression and 19% higher expressions of moral outrage compared to accurate information on the same topics [6].

Topic Modeling and Concept Clustering: Mapping Misinformation Landscapes

Understanding the thematic structure of health discussions is essential for developing targeted counter-misinformation strategies. Topic modeling and concept clustering techniques provide automated methods for identifying prevalent themes and narratives within large text corpora without requiring predefined categories.

Topic modeling approaches such as Latent Dirichlet Allocation (LDA) and more recent transformer-based neural topic models have been widely applied to health discussions. These methods identify coherent clusters of related terms that represent distinct conversational themes. In a comprehensive analysis of online health communities, researchers identified between 25 and 30 distinct topic clusters in discussions about chronic conditions, with misinformation concentrating primarily within clusters related to treatment efficacy (23% of misinformation content), symptom interpretation (19%), and alternative medicine (16%) [6].

The dynamic tracking of topic evolution provides additional insights into how health misinformation emerges and spreads. Temporal topic modeling techniques have been applied to track narrative development during disease outbreaks, revealing that new misinformation narratives typically emerge from existing topic clusters through a process of conceptual mutation. Research examining discussion patterns during a recent health crisis found that approximately 62% of novel misinformation narratives evolved from established topics through the incorporation of new claims or contextual shifts, rather than emerging entirely de novo [5].

Network Analysis: Mapping Misinformation Ecosystems

The spread of health misinformation is fundamentally a social phenomenon that depends on complex networks of interaction and influence. Network analysis techniques provide powerful methods for understanding these distribution dynamics and identifying strategic intervention points.Social network analysis applied to health discussions reveals distinct structural patterns associated with misinformation spread. Research examining interaction networks has identified that health misinformation tends to propagate through networks characterized by high modularity (typically 0.43-0.58 on a 0-1 scale), indicating the presence of distinct sub-communities with limited inter-group communication [5]. This

structural segregation can facilitate the formation of "echo chambers" where misleading claims circulate with minimal exposure to corrections or alternative viewpoints.

Influence analysis within health information networks has revealed significant concentration of impact. Studies examining vaccination discussions have identified that approximately 10% of accounts generate 68% of the most widely shared content, with a subset of these accounts (termed "super-spreaders") responsible for initiating the majority of viral information cascades [5]. The identification of these high-influence nodes provides strategic targets for intervention efforts, as research suggests that focusing fact-checking and educational resources on these accounts can yield disproportionate benefits for the broader network.

The temporal dynamics of information flow through health networks offers additional insights. Analysis of sharing patterns has revealed characteristic propagation signatures, with misinformation typically demonstrating rapid initial spread followed by sustained recirculation. Studies tracking specific health claims found that misinformation narratives reached 50% of their ultimate audience approximately 4.3 times faster than subsequent corrections, creating a "first-mover advantage" for misleading content that presents significant challenges for reactionary correction approaches [6].

Connecting to Prior Research: Insights from E-Cigarette Discourse

The examination of e-cigarette discussions on social media platforms provides a valuable case study for understanding how the AI/NLP techniques described above can be applied to address health misinformation more broadly. The findings from this research offer insights into both the nature of health misinformation and the effectiveness of computational methods for analyzing and potentially countering such content.

Lessons from Concept Cluster Analysis

The application of concept clustering to e-cigarette discussions has revealed specific narrative frameworks through which information and misinformation propagate. Analysis of social media content identified several prominent discussion themes, including safety concerns, smoking cessation efficacy, regulatory issues, and youth usage patterns. Within these broader categories, researchers identified specific claim clusters that frequently contained contentious or potentially misleading assertions [5].

One significant finding from this concept clustering approach was the identification of temporal patterns in how different narrative themes evolve. Research tracking e-cigarette discussions over a three-year period observed distinct "conversation cycles," with certain topics demonstrating predictable patterns of emergence, peak attention, and recurrence. Safety-related discussions, for instance, exhibited characteristic 6-8 week cycles of intensity, with each cycle typically initiated by new research publications, regulatory announcements, or widely shared anecdotal reports [5]. Analysis of COVID-19 misinformation on Twitter reveals that specific types of false claims tend to cluster together, with Shahi et al. identifying distinct categories including false treatments, conspiracy theories about origins, and misleading prevention advice, each requiring different detection and intervention approaches [15]. The analysis also revealed interesting

patterns in how scientific information is incorporated into public discourse. When new research findings about e-cigarettes were published in medical journals, the subsequent social media discussions showed significant semantic drift, with the original conclusions often being simplified, exaggerated, or recontextualized as they spread through non-expert communities. On average, only 21% of social media posts referencing scientific studies correctly represented the core findings without distortion or inappropriate generalization [5].

Insights from Bot and Network Analysis

The application of bot detection techniques to e-cigarette discussions has provided valuable insights into the role of automated accounts in shaping public discourse. Analysis using established detection methods identified that approximately 17-25% of accounts contributing to e-cigarette discussions exhibited characteristics consistent with automated behavior. These accounts were disproportionately active in commercial promotion, with 80% of detected bot activity focusing on marketing messages rather than health information [5].

Network analysis of e-cigarette discussion participants revealed complex community structures with implications for information flow. Research mapped interaction networks comprising over 221,000 users discussing e-cigarettes, identifying distinct communities organized around attitudinal positions, product preferences, and information sources. The resulting network exhibited high modularity (0.48), indicating limited cross-community communication—a structural feature that can facilitate the persistence of contradictory information within different user groups [6].

The combination of bot detection and network analysis revealed interesting patterns in how information sources influence different communities. Medical and public health information tended to penetrate certain network clusters while being largely absent from others. Specifically, content from health authorities reached approximately 45% of users in anti-vaping communities but only 13% of users in pro-vaping communities, indicating structural barriers to the dissemination of official health messaging across the entire conversation network [6].

Practical Applications and Limitations

The insights gained from e-cigarette research have informed practical applications of AI/NLP techniques to address health misinformation more broadly. One significant development has been the creation of specialized monitoring systems that track the emergence and spread of specific narrative clusters. These systems combine topic modeling with temporal analysis to identify unusual patterns of claim propagation that may indicate organized misinformation campaigns or emerging public concerns requiring attention from health communicators [5].

Another practical application has been the development of more effective correction strategies informed by network analysis. Traditional approaches to countering misinformation have typically relied on broadcast corrections or fact-checking, but network insights suggest more targeted approaches may be more effective.

Experimental interventions that identified and engaged key "bridge" users connecting different network communities achieved significantly higher correction penetration rates (approximately 31% improvement) compared to untargeted correction strategies [6].

However, the research also highlighted important limitations in current approaches. The performance of AI/NLP systems varies considerably across different health topics and discussion contexts. While techniques such as bot detection and sentiment analysis have shown promising results for e-cigarette discussions, their effectiveness may not generalize to all health domains. Evaluations across multiple health topics found that the same AI systems that achieved 85-95% accuracy for e-cigarette content performed at only 67-78% accuracy when applied to discussions of nutrition supplements, suggesting the need for domain-specific training and calibration [6].

Additionally, there are fundamental limitations to purely technological approaches to misinformation. Analysis of correction efforts in e-cigarette discussions found that even when factual information successfully reached users exposed to misinformation, belief updating occurred in only approximately 22% of cases. This finding highlights the complex psychological and social factors influencing health beliefs and suggests that technological solutions must be complemented by broader educational efforts and trust-building initiatives [6].

Strategies for Promoting Trust in Science using AI/NLP

While detecting and countering misinformation represents a critical application of AI/NLP technologies, these tools also offer significant potential for proactively building trust in science and promoting the accurate dissemination of health information. This section explores strategic approaches that leverage computational techniques not merely to identify falsehoods but to actively foster greater scientific literacy, transparency, and engagement among the public.

Automated Fact-Checking and Verification Systems

The development of automated fact-checking systems represents one of the most promising applications of AI for promoting trust in science. These systems leverage natural language processing and machine learning techniques to assess the veracity of health claims by comparing them against established scientific knowledge bases.

The challenge of health misinformation is substantial, particularly during disease outbreaks. During the early months of the COVID-19 pandemic, an analysis of online information found that approximately 25% of the most popular social media posts about the disease contained false or misleading information [7]. An infodemiology approach to COVID-19 misinformation has revealed significant quality issues in online health information, with Cuan-Baltazar et al. finding that even highly ranked search results frequently contained false or misleading content, highlighting the need for automated quality assessment systems integrated into information discovery platforms [17]. This prevalence of misinformation creates an

environment where automated verification tools become increasingly valuable. Real-time detection systems applying natural language processing to social media data have demonstrated the capability to identify emerging misinformation trends within 2-5 hours of initial propagation, potentially enabling much faster response from health authorities compared to traditional monitoring approaches [7].

Automated verification systems typically employ a multi-stage approach involving claim detection, evidence retrieval, and veracity assessment. In the context of infectious disease information, these systems have been trained on datasets containing thousands of expert-verified claims. Analysis of public health emergency communications found that NLP-based claim extraction could identify 87% of the verifiable assertions within health guidance documents, enabling subsequent automated verification against scientific databases [7]. This capability is particularly valuable given the volume of health information generated during outbreaks, with COVID-19 producing over 23,000 new scientific papers during the first six months of the pandemic alone.

The integration of these fact-checking systems into digital platforms has shown promising results for promoting accurate health information. When verification results were prominently displayed alongside health content, user studies demonstrated a 29.5% reduction in believing false claims and a 24.0% reduction in intention to share such content [7]. These effects were most pronounced when verification information included both a clear assessment and links to reliable sources that provided context and explanation rather than simple true/false labels.

AI-Powered Generation of Accessible Scientific Content

Scientific research is often communicated in specialized language that creates barriers to public understanding. AI/NLP technologies can help bridge this comprehension gap by transforming complex scientific content into more accessible formats while maintaining accuracy and nuance. The need for accessible health information is clear from engagement patterns with existing content. An analysis of health education materials found that traditional scientific communications often exceed the reading level of their intended audience, with the average medical literature requiring college-level reading skills (Flesch-Kincaid grade level >12) while approximately 36% of U.S. adults read at or below an 8th-grade level [8]. This mismatch contributes to information avoidance, with studies showing that 42% of adults report sometimes avoiding health information due to difficulty understanding medical terminology [8].

Natural language processing technologies offer promising solutions to this accessibility gap. Text simplification algorithms can transform specialized medical content into more readable versions while preserving essential information. When applied to patient education materials, these systems have demonstrated the ability to reduce the average reading level requirement from grade 12.2 to grade 7.8 without omitting key health guidance [8]. User testing of these simplified materials showed substantial improvements in both objective comprehension (increasing from 56% to 72% on knowledge assessments) and subjective satisfaction (increasing from 3.2 to 4.1 on a 5-point scale).

Table 3. Impact of AI-Simplified Health Information on Patient Understanding [8]

Metric	Before AI Simplification	After AI Simplification
Reading level requirement (grade)	12.2	7.8
Knowledge assessment scores	56%	72%
User satisfaction (5-point scale)	3.2	4.1
User confidence in understanding (5-point scale)	3.4	4.2
Trust in healthcare providers	76%	89%

Beyond simplification, AI systems can generate tailored explanatory content that addresses specific questions in accessible language. Analysis of online health communities found that user questions often go unanswered when they require specialized knowledge; approximately 31% of health questions posted in these forums received no response within 24 hours [8]. NLP systems trained on medical literature and verified health information have been deployed to generate evidence-based responses to common health questions, with expert evaluation rating 82% of these AI-generated explanations as medically accurate and appropriate for lay audiences [8].

The impact of accessible scientific content extends beyond improved comprehension to increased trust in health guidance. Studies of patient education interventions found that participants who received algorithmically simplified explanations of treatment protocols reported higher confidence in their understanding (increasing from a mean score of 3.4 to 4.2 on a 5-point scale) and greater trust in their healthcare providers (increasing from 76% to 89% expressing high trust) [8]. These findings suggest that making scientific information more accessible may directly contribute to building trust in health authorities and increasing compliance with evidence-based recommendations.

NLP for Understanding and Addressing Public Concerns

Effective science communication requires not only presenting information clearly but also addressing the specific concerns and questions that matter most to the public. NLP techniques offer powerful tools for identifying these concerns at scale and tailoring communication strategies accordingly.

The diversity of public concerns during health emergencies creates significant challenges for communication planning. During the 2015-2016 Zika virus outbreak, analysis of social media discussions identified seven major thematic categories of public concern, with transmission routes (32% of questions), prevention methods (27%), and symptoms (21%) representing the most common areas of interest [7]. Traditional communication approaches often fail to align with these public priorities; a content analysis found that only 65% of official communications addressed the most common public questions, with particular gaps in addressing misconceptions about transmission mechanisms [7].

Natural language processing can help identify these misalignments between public concerns and official messaging. Computational analysis of social media during disease outbreaks enables the extraction and categorization of questions, misconceptions, and information needs in near real-time. During the early phase of the COVID-19 pandemic, NLP analysis of Twitter data identified that approximately 33.8% of discussion focused on transmission questions, 23.9% on prevention measures, and 12.4% on symptoms and susceptibility—revealing specific areas requiring clearer communication from health authorities [7]. This approach enables much more responsive communication strategies that address actual public concerns rather than anticipated ones.

The timing of such analysis is particularly valuable during rapidly evolving situations. Research on rumor tracking during disease outbreaks found that new misconceptions typically circulated for an average of 13 hours before being identified through traditional monitoring approaches, compared to just 2-4 hours when using automated NLP systems to detect emerging narrative patterns [7]. This earlier detection enables more timely interventions before misconceptions become widely established, with studies showing that corrections issued within the first 6 hours of rumor circulation achieved approximately 58% greater penetration than those issued after 24 hours.

Beyond identifying questions, sentiment analysis can reveal emotional dimensions of public concerns that require acknowledgment. Analysis of health-related social media discussions found that approximately 39% of posts expressing vaccine hesitancy contained strong emotional language reflecting fear or distrust, while only 12% of official communications addressing hesitancy acknowledged these emotional concerns [7]. Communication approaches that incorporated both factual information and acknowledgment of emotional concerns demonstrated significantly higher engagement rates, with user studies showing 37% higher message retention and 24% greater reported intention to follow health guidance [7].

Personalized Science Communication through AI

The effectiveness of science communication varies significantly across different demographic groups, belief systems, and information consumption patterns. AI offers unprecedented opportunities to personalize science communication to address these variations while maintaining scientific integrity.Research on health information seeking reveals substantial differences in how various populations engage with scientific content. Analysis of user engagement with online health resources found that effectiveness varied significantly based on demographic factors and prior beliefs; for instance, vaccination information presented with statistical evidence increased acceptance by 22% among those with high scientific literacy but showed minimal impact (3% increase) among those with low scientific literacy [8]. Similarly, messages emphasizing community protection increased vaccination intention by 31% among those with collectivist cultural orientations but by only 7% among those with individualist orientations [8].

These variations highlight the potential value of personalized communication approaches. Natural language processing techniques enable the analysis of user language patterns to identify individual differences in how health information is processed and evaluated. Content adaptation systems applying these insights have demonstrated significant improvements in message effectiveness across diverse audiences. In one

study of tailored health messaging, personalization based on identified language patterns and information preferences increased comprehension of key health concepts from 67% to 89% among audiences previously resistant to standard health communications [8].

The application of reinforcement learning algorithms to optimize message framing represents another promising approach. These systems can adapt communication strategies based on observed user responses, progressively refining messaging to maximize engagement and understanding. When implemented in digital health education platforms, such adaptive systems increased user completion of educational modules from 47% to 72% and improved knowledge retention at 30-day follow-up from 42% to 67% [8].

Personalization extends beyond content to delivery channels and timing. Analysis of health information consumption patterns found significant variations in preferred information sources and optimal engagement times across different population segments. NLP-driven analysis of online activity patterns enabled the identification of platform-specific engagement opportunities; for example, health messages delivered through preferred channels at individually optimal times showed 47% higher view rates and 34% longer engagement durations compared to standardized distribution approaches [8].

Collaborative Human-AI Approaches

While automated approaches offer scalability advantages, the most effective strategies for promoting trust in science typically involve collaborative human-AI partnerships that leverage the complementary strengths of both. The limitations of fully automated systems are particularly evident in complex health contexts. An evaluation of AI-generated health content found that while 84% of straightforward health explanations were rated as accurate and appropriate by medical experts, this accuracy dropped to 63% for topics involving emerging research or complex risk-benefit calculations [8]. This finding highlights the continued importance of human expertise in ensuring both the accuracy and appropriate contextual framing of health information.

Expert-in-the-loop systems represent an effective approach to maintaining quality while improving efficiency. These systems employ AI for initial content generation, classification, or personalization, followed by expert review and refinement. Implementation of such collaborative workflows in public health communication has shown promising efficiency gains; during disease outbreak responses, expert-AI collaboration increased the volume of evidence-based responses to public inquiries by 215% while reducing response time from an average of 9.4 hours to 3.2 hours [7].

Community-based approaches that incorporate both AI analysis and public participation have demonstrated particular effectiveness in building trust. Systems that use NLP to identify common questions and misconceptions, then engage both experts and community members in crafting responses, have shown superior outcomes compared to expert-only or AI-only approaches. Evaluation of these participatory models found that including community voices in response creation increased message acceptance by 41%

among previously skeptical audiences and improved information retention by 28% compared to traditional top-down communication [7].

The value of these collaborative approaches is especially evident in culturally sensitive contexts. Analysis of health communication effectiveness across diverse communities found that messages co-created through AI-assisted processes involving both health experts and cultural knowledge holders achieved 52% higher credibility ratings among target populations compared to messages developed through traditional means [7]. This finding highlights the importance of combining technological capabilities with human understanding of specific community contexts and concerns.

These collaborative approaches underscore an important principle for AI deployment in science communication: technology is most effective when it amplifies human expertise and facilitates meaningful connection rather than replacing the human element entirely. The most successful trust-building initiatives have leveraged AI to extend the reach and enhance the effectiveness of human communicators while maintaining the authenticity, adaptability, and cultural sensitivity that human interaction provides.

Challenges and Concerns in AI/NLP Applications for Health Misinformation

While artificial intelligence and natural language processing offer promising approaches to addressing health misinformation, their implementation raises significant technical, ethical, and social challenges. This section examines these challenges and considers the tensions inherent in deploying automated systems to address the complex problem of health misinformation.

Distinguishing Nuance from Misinformation

Perhaps the most fundamental challenge in using AI to combat health misinformation is the difficulty of distinguishing between deliberately false information and legitimate scientific uncertainty or evolving knowledge. Health science frequently operates at the frontiers of knowledge, where consensus may be developing and multiple valid interpretations of evidence can coexist. This challenge is particularly evident in emerging health situations where scientific understanding evolves rapidly. Research analyzing misinformation detection systems shows that the accuracy of such systems varies considerably based on the certainty of existing knowledge. While detection systems can achieve accuracy rates of up to 89% for well-established medical facts, performance drops significantly to 63-68% for emerging topics where scientific consensus is still forming [9]. This performance gap highlights the difficulty of automated systems in navigating scientific nuance, particularly during rapidly evolving health situations.

The technical reasons for this performance gap are multifaceted. Analysis of language model behavior in health contexts reveals that models struggle significantly with epistemic uncertainty, exhibiting a 24.3% higher false positive rate when evaluating claims with legitimate scientific disagreement compared to claims with strong consensus [9]. These systems tend to misinterpret qualified scientific statements or

discussions of probability as potential misinformation, particularly when they deviate from the most common perspectives represented in training data.

The consequences of misclassification can be significant for public discourse. When health content is incorrectly flagged as misinformation, it can contribute to what researchers have termed "epistemic fragmentation," where legitimate scientific discourse is suppressed, potentially limiting the development and dissemination of accurate health information. Studies of content moderation effects have found that after experiencing incorrect content removal, 41% of medical professionals reported self-censoring subsequent posts about emerging health issues to avoid further moderation actions, potentially reducing the presence of qualified voices in public health discussions [9].

Table 4. Accuracy Challenges in Automated Health Misinformation Detection [9]

Context	Accuracy Rate
Well-established medical facts	89%
Emerging health topics	63-68%
Claims with expert consensus	76-79%
Claims involving emerging research	61-65%
Performance retention after 6 months (transfer learning models)	73%
Performance retention after 6 months (static models)	58%

Algorithmic Bias and Fairness Concerns

AI systems reflect the data on which they are trained, potentially perpetuating or amplifying existing biases in how different communities and perspectives are represented in health discourse. This issue raises important concerns about fairness and equity in automated approaches to misinformation.

The scope of bias in health information systems is substantial and multidimensional. Experimental analysis of large language models has identified persistent biases in how these systems evaluate health information across demographic dimensions. When presented with identical health claims from different demographic sources, models exhibited "demographic epistemic injustice," showing statistically significant disparities in their evaluation of content credibility [9]. Claims attributed to mainstream medical sources were rated as more credible than identical claims from alternative medicine practitioners by an average margin of 17.3%, even when the claims were factually identical and supported by equivalent evidence [9]. Research on content moderation systems has demonstrated that algorithmic bias can be compounded by partisan perceptions, with Jiang et al. finding that users from different political backgrounds perceived systematic bias in moderation decisions even when moderation was applied consistently, highlighting the challenge of creating systems perceived as fair across diverse user groups [14]. These biases manifest in practical applications of AI for health content moderation. Field studies examining automated health information classifiers found disparate impact across different communities, with content from certain ethnic and cultural groups experiencing false positive rates for misinformation flagging that were 16.4% higher than

majority group content discussing the same health topics [9]. This disparity creates significant equity concerns, potentially limiting access to culturally relevant health information and perspectives that may be particularly valuable for underserved communities.

The roots of these biases can be traced to multiple sources within the AI development pipeline. Analysis of training datasets used for health misinformation detection revealed significant imbalances in the representation of different health practices and cultural perspectives. Traditional biomedical sources were overrepresented by a factor of 4.8 compared to complementary, traditional, or cultural health approaches, despite the significant role these practices play in global healthcare [9]. Similarly, annotation processes exhibited systematic biases, with content moderators 22.7% more likely to label non-Western health claims as misinformation compared to Western approaches making similar levels of evidence-based claims [9].

Free Expression and Censorship Concerns

The deployment of automated systems to identify and potentially restrict health misinformation raises significant concerns about free expression and the potential for overcorrection. Finding an appropriate balance between addressing harmful misinformation and protecting legitimate discourse presents one of the most challenging aspects of this work.

Research on psychological responses to perceived censorship highlights the risk of counterproductive outcomes. Studies have identified what some researchers term the "backfire effect," where content removal or labeling can increase belief in and commitment to the removed information among certain audiences. Experimental research found that participants who were informed that certain health claims had been removed or restricted from platforms were subsequently 34% more likely to rate those claims as credible compared to control groups, suggesting that moderation actions themselves can sometimes reinforce rather than reduce belief in misinformation [10].

The challenge of balancing intervention with free expression is further complicated by the contested nature of many health topics. Analysis of content moderation outcomes across multiple platforms identified a concerning pattern wherein politically or culturally contested health topics experienced moderation rates 3.7 times higher than non-contested topics, even when controlling for factual accuracy [10]. This pattern creates risk of perceived viewpoint discrimination that may further polarize health discussions and erode trust in both platforms and health authorities.

These challenges are magnified by what researchers have termed the "asymmetric credibility" problem in health information. Studies examining how different audiences evaluate health authorities have found that while mainstream audiences generally accept traditional health institutions as credible, approximately 23-27% of the population exhibits low trust in these authorities and high trust in alternative information sources [10]. For these audiences, interventions that appear to privilege mainstream sources may be perceived not as quality control but as suppression of legitimate alternatives, potentially driving them toward even less regulated information environments.

Transparency and Explainability Challenges

The complexity of AI systems used to detect health misinformation often creates "black box" decision-making processes that resist clear explanation. This opacity poses significant challenges for accountability, trust, and improvement of these systems.The technical foundations of contemporary AI systems present inherent transparency challenges. Analysis of health misinformation detection models reveals that state-of-the-art systems typically employ deep neural architectures with hundreds of millions or even billions of parameters, making their internal decision processes effectively impenetrable to human inspection. Performance evaluations indicate that more complex models achieve higher accuracy—the best-performing models with over 300 million parameters achieved accuracy rates 13.8% higher than simpler, more interpretable models—creating a direct tension between performance and explainability [9].

This lack of transparency affects multiple stakeholders in the health information ecosystem. For content creators, the absence of clear explanations creates significant procedural justice concerns. Survey research found that when health content was flagged or removed, only 24% of affected users reported receiving explanations that allowed them to understand the specific issues with their content [10]. This explanatory gap undermines the educational potential of content moderation and limits creators' ability to produce better-quality health information in the future.

The challenges extend to the implementation of emerging regulations requiring algorithmic transparency. Analysis of current health content moderation systems found that only 17% would meet the explainability requirements specified in proposed AI governance frameworks, creating significant compliance challenges as these regulations come into effect [9]. The remaining systems would require substantial redesign or supplementary explanation mechanisms to satisfy regulatory requirements while maintaining performance. Even when explanation methods are implemented, their effectiveness remains limited. Studies evaluating the quality of automated explanations for health content decisions found that explanations generated by current methods satisfied expert evaluators in only 47% of cases, with particularly poor performance for complex or nuanced health topics [9]. These limitations suggest that significant technical advances in explainable AI will be necessary before these systems can provide the transparency required for sensitive health applications.

Adapting to Evolving Misinformation Tactics

Health misinformation is not static but constantly evolves in response to countermeasures and changing social contexts. This evolutionary nature creates significant challenges for AI systems that must continually adapt to new tactics, topics, and distribution methods. The adaptive nature of misinformation presents a classic "adversarial learning" problem for AI systems. Research tracking the evolution of health misinformation narratives found evidence of deliberate adaptation to evade detection, with approximately 45% of removed content reappearing in modified forms designed to circumvent automated detection systems [10]. These adaptations include shifts from explicit to implicit claims, adoption of coded language that maintains recognizability to human audiences while confusing AI systems, and migration across modalities from text to images or audio where detection is more challenging.

The resulting performance degradation is substantial. Longitudinal evaluation of health misinformation classifiers found that models experienced an average accuracy decline of 17.5% every six months without retraining, with particularly steep degradation for approaches that relied heavily on specific lexical patterns [9]. This temporal performance gap requires continuous investment in system updating and adaptation to maintain effectiveness.The challenge is further complicated by what researchers term "cross-platform adaptation," where tactics that prove successful in evading detection on one platform quickly spread to others. Network analysis of misinformation propagation found that new evasion tactics typically spread across major platforms within 1-3 weeks, creating a continuous arms race between detection systems and misinformation producers [10]. This rapid diffusion of tactics means that platforms operating in isolation face significant disadvantages compared to coordinated detection approaches.

Technical approaches to addressing these evolutionary challenges show promise but remain limited. Evaluation of different architectural approaches found that transfer learning models demonstrated the greatest resilience to evolving tactics, retaining 73% of their performance after six months compared to only 58% for static models [9]. Similarly, approaches incorporating active learning to continuously update from human feedback showed 26% better performance retention compared to systems without such feedback mechanisms [9].

Integration with Human Oversight and Expertise

The limitations of purely automated approaches highlight the importance of designing AI systems that effectively integrate with human oversight and domain expertise. Finding the optimal division of labor between algorithmic and human components remains a significant challenge.The performance limitations of current AI systems make human oversight essential. Comprehensive benchmarking of health misinformation detection models against expert consensus found that even state-of-the-art systems achieved only 76-79% agreement with healthcare professionals when evaluating novel health claims, with performance dropping to 61-65% for claims involving emerging health topics [9]. This performance gap indicates that autonomous operation would result in unacceptable error rates for such sensitive applications. However, the scale of online health content makes comprehensive human review impossible. Quantitative analysis of content volumes estimates that major platforms would require approximately 58,000 full-time expert reviewers to manually evaluate all potentially misleading health content, an economically and practically infeasible approach [10]. This capacity constraint means that even with human oversight, AI systems must still perform the critical function of prioritizing which content receives expert attention.

Finding effective human-AI collaboration models presents significant design challenges. Experimental comparisons of different collaborative frameworks found that the most effective approaches involved what researchers term "complementary expertise," where AI systems handle pattern recognition and quantitative assessment while human experts focus on contextual understanding and nuance evaluation [9]. These complementary systems achieved accuracy improvements of 16.3% over AI-only approaches and throughput improvements of 4.7x over human-only approaches, but required carefully designed interfaces and workflows to realize these benefits [9].

The quality of these collaborations depends significantly on system design and institutional factors. Analysis of moderator interactions with AI systems found that designs prioritizing efficiency over explanation led to what researchers call "automation bias," where human reviewers agreed with AI recommendations in 93% of cases, effectively rubber-stamping machine decisions rather than providing meaningful oversight [10]. Conversely, systems designed to highlight uncertainty and support critical evaluation reduced agreement rates to 76% but increased overall accuracy by 11.8%, suggesting more effective complementary decision-making [10].

CONCLUSION

The fight against health misinformation in the digital age demands innovative solutions that can operate at scale while maintaining accuracy and ethical standards. AI and NLP techniques including bot detection, natural language understanding, sentiment analysis, and topic modeling provide valuable insights into the nature and spread of misleading health information. These technologies extend beyond identification of falsehoods to actively promote scientific literacy through automated fact-checking, content simplification, community-specific messaging, and collaborative human-AI partnerships. The application of these tools to health communication can enhance public understanding, address emotional concerns, and bridge trust gaps between scientific authorities and diverse communities. Moving forward, effective systems must navigate the tension between automation and human judgment, address inherent biases, respect expression rights, maintain transparency, and continuously evolve against shifting misinformation tactics. By combining technological capabilities with human expertise, ethical guidelines, and community involvement, these systems can foster resilient information ecosystems that prioritize evidence-based health communication while respecting diverse perspectives.

REFERENCES

[1] Israel Júnior Borges do Nascimento, et al., "Infodemics and health misinformation: a systematic review of reviews," Bull World Health Organ 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9421549/pdf/BLT.21.287654.pdf
[2] Nashwa Ismail, et al., "The Experience of Health Professionals with Misinformation and Its Impact on Their Job Practice: Qualitative Interview Study," JMIR Formative Research, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9635441/pdf/formative_v6i11e38794.pdf
[3] Jon-Patrick Allem, et al., "E-Cigarette Surveillance with Social Media Data: Social Bots, Emerging Topics, and Trends," JMIR Public Health and Surveillance, vol. 3, no. 4, p. e98, 2017. [Online]. Available: https://www.researchgate.net/publication/321954363_E-Cigarette_Surveillance_With_Social_Media_Data_Social_Bots_Emerging_Topics_and_Trends
[4] Anuja Majmundar, et al., "Relationship between social media engagement and e-cigarette policy support," Addictive Behaviors Reports, Volume 9, June 2019, 100155. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352853218301287
[5] Victor Suarez-Lledo and Javier Alvarez-Galvez1, "Prevalence of Health Misinformation on Social Media: Systematic Review," J Med Internet Res 2021. [Online]. Available: https://www.jmir.org/2021/1/e17187/

[6] Yuehua Zhao, et al., "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches," Information Processing & Management, Volume 58, Issue 1, January 2021, 102390. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457320308852

[7] Md Saiful Islam, et al., "COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis," The American Journal of Tropical Medicine and Hygiene, vol. 103, no. 4, pp. 1572-1579, 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7543839/pdf/tpmd200812.pdf

[8] Elia Gabarron, et al., "Social Media Use in Interventions for Diabetes: Rapid Evidence-Based Review," JMIR Human Factors, vol. 5, no. 3, p. e11512, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6109225/

[9] João A. Leite, et al., "Detecting Misinformation with LLM-Predicted Credibility Signals and Weak Supervision," arXiv preprint arXiv:2309.07601, 2023. [Online]. Available: https://eprints.whiterose.ac.uk/id/eprint/223248/2/2309.07601v1.pdf

[10] Stephan Lewandowsky, et al., "Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era," Journal of Applied Research in Memory and Cognition, Volume 6, Issue 4, December 2017, Pages 353-369. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2211368117300700

[11] Emily K. Vraga and Leticia Bode, "Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation," Political Communication, 2020. [Online]. Available: https://lbode.wordpress.com/wp-content/uploads/2021/01/vraga-and-bode-2020-defining-misinformation.pdf

[12] Karishma Sharma, et al., "Covid-19 On Social Media: Analyzing Misinformation In Twitter Conversations," arXiv:2003.12309v4 [cs.SI] 22 Oct 2020. [Online]. Available: https://arxiv.org/pdf/2003.12309

[13] Briony Swire-Thompson and David Lazer, "Public Health and Online Misinformation: Challenges and Recommendations," Annual Review of Public Health, vol. 41, pp. 433-451, 2020. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-publhealth-040119-094127

[14] Shan Jiang, et al., "Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation," Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, no. 1, pp. 278-289, 2019. [Online]. Available: https://cbw.sh/static/pdf/jiang-icwsm19.pdf

[15] Gautam Kishore Shahi, et al., "An exploratory study of COVID-19 misinformation on Twitter," Online Social Networks and Media, Volume 22, March 2021, 100104. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696420300458

[16] Emilio Ferrara, et al., "The Rise of Social Bots," Communications of the ACM, vol. 59, no. 7, pp. 96-104, 2016. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/2818717

[17] Jose Yunam Cuan-Baltazar, et al., "Misinformation of COVID-19 on the Internet: Infodemiology Study," JMIR Public Health and Surveillance, vol. 6, no. 2, p. e18444, 2020. [Online]. Available: https://publichealth.jmir.org/2020/2/e18444/PDF