



Advancements in Medical Language Computing: Bridging Clinical Narratives and Social Media Health Informatics

Noah Moritz, Beatrice Hawthorne, Julian Carmichael

Division of Health Informatics, Karolinska Institutet, Stockholm, Sweden

DOI: <https://doi.org/10.71465/jcphs41>

Abstract

The proliferation of digital health records and the ubiquitous use of social media have generated unprecedented volumes of health-related textual data. Medical language computing, a specialized branch of natural language processing, serves as the critical mechanism for extracting actionable insights from these disparate data sources. This paper presents a comprehensive examination of the methodologies, challenges, and applications associated with processing both formal clinical text and informal patient-generated health data on social media platforms. By analyzing the structural and linguistic dichotomy between these two domains, we elucidate the necessity for domain-specific computational strategies. Clinical narratives are characterized by specialized jargon, complex abbreviations, and rigid documentation standards, whereas social media data is typified by colloquialisms, typographical errors, and contextual ambiguity. Through an extensive review of current machine learning and deep learning architectures, we propose a unified analytical framework that leverages advanced transformer-based models to harmonize these disparate data streams. Our methodology details the preprocessing pipelines, algorithmic processing, and evaluation metrics required to achieve high-fidelity entity recognition, sentiment analysis, and adverse event detection. The findings underscore the efficacy of hybrid computational approaches in bridging the semantic gap between provider-centric documentation and patient-centric discourse. Ultimately, this research contributes to the broader paradigm of general academic health informatics by providing a robust foundation for future systems designed to monitor public health trends, enhance clinical decision support, and improve patient outcomes through comprehensive linguistic analysis.

Keywords: Medical Language Computing, Natural Language Processing, Electronic Health Records, Social Media Health Data, Clinical Informatics

INTRODUCTION

The digitization of modern healthcare systems has fundamentally transformed the way medical information is recorded, stored, and analyzed. Central to this transformation is the exponential growth of unstructured textual data, which resides predominantly within Electronic Health Records and various social media platforms [1]. Medical language computing has emerged as an indispensable discipline aimed at decoding, interpreting, and utilizing this vast repository of text. Unlike structured data, which fits neatly into relational databases, unstructured clinical narratives and social media posts contain rich, contextual nuances that require sophisticated computational linguistics to decipher [2]. The primary challenge in this domain lies in the inherent variability of human language, exacerbated by the distinct communicative environments of hospitals and internet forums. In the clinical setting, physicians and nurses document patient encounters using a highly specialized lexicon, characterized by domain-specific acronyms, shorthand notations, and complex syntactic structures designed for efficiency rather than readability [3]. Conversely, patients utilize

social media platforms to share their health experiences, symptoms, and concerns using colloquial language, slang, and emotional expressions that frequently deviate from standardized medical terminology [4]. Reconciling these two linguistic extremes is a central objective of contemporary medical language computing. The importance of bridging this gap cannot be overstated. Clinical text provides a formal, medically validated account of patient health, essential for accurate diagnoses, treatment planning, and historical continuity of care. However, it often lacks the patient perspective regarding quality of life, minor symptomatic fluctuations, and undocumented adverse reactions to medications. Social media, on the other hand, acts as a massive, real-time observational database where individuals freely report their daily health struggles, dietary habits, and emotional states [5]. By integrating the formal provider documentation with the informal patient narrative, researchers and healthcare professionals can construct a more holistic, multidimensional understanding of public health and individual well-being. This integration forms the foundational motivation for the research presented in this comprehensive academic paper.

1.1 Motivation and Background Context

Historically, the extraction of information from medical texts relied heavily on manual chart reviews and labor-intensive curation by trained medical professionals. This paradigm was fundamentally unscalable, prompting the initial development of rule-based natural language processing systems in the late twentieth century [6]. These early systems utilized predefined dictionaries, regular expressions, and syntactic parsing trees to identify specific medical entities such as diseases, medications, and anatomical structures. While effective for highly constrained tasks, rule-based systems exhibited significant fragility when confronted with the syntactic flexibility and rampant typographical errors present in real-world data [7]. The advent of statistical machine learning introduced a new era of flexibility, allowing algorithms to learn patterns from large annotated datasets rather than relying on brittle, handcrafted rules [8]. Techniques such as Support Vector Machines, Hidden Markov Models, and Conditional Random Fields became the standard for clinical named entity recognition and relation extraction.

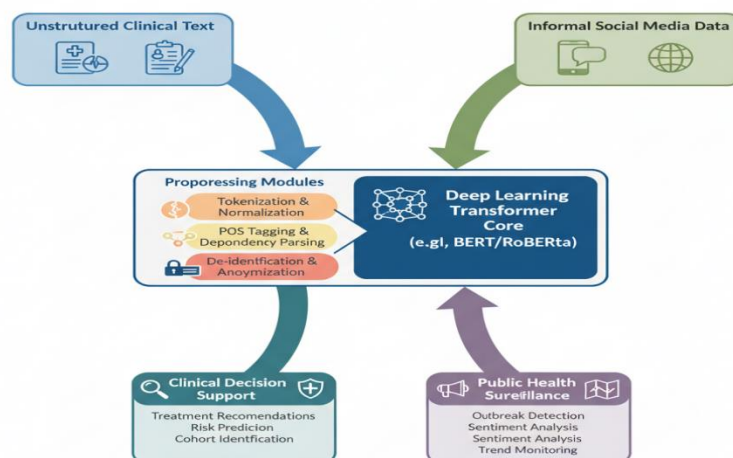


Figure 1: Proposed architecture for medical text information extraction

However, the true revolution in medical language computing was catalyzed by the introduction of deep learning and, more recently, large-scale transformer architectures. These advanced neural network models possess an unprecedented capacity to capture deep semantic relationships and long-range contextual dependencies within textual data [9]. In the context of medical data, this capability is paramount. A single word can carry vastly different meanings depending on its surrounding context; for instance, the abbreviation cold could refer to chronic obstructive lung disease in a clinical note or a common viral infection in a social media post. Transformer models, pre-trained

on massive corpora of general and domain-specific text, have demonstrated remarkable proficiency in disambiguating such terms [10]. Despite these advancements, the disparate nature of clinical and social media text continues to pose significant challenges. Models trained exclusively on formal medical literature often underperform when applied to the noisy, idiosyncratic language of Twitter or Reddit, necessitating the development of novel, domain-adaptive methodologies.

1.2 Scope and Objectives

The primary objective of this paper is to provide a thorough, critical analysis of the current state of medical language computing as applied to both clinical narratives and patient-generated social media data. We seek to explore the theoretical underpinnings and practical implementations of modern natural language processing pipelines designed to handle these unique datasets. The scope of our investigation encompasses the entire data lifecycle, from acquisition and ethical considerations to preprocessing, feature extraction, algorithmic processing, and final performance evaluation [11]. We systematically examine the linguistic properties that differentiate provider-generated text from user-generated content, detailing how these differences inform the architectural design of computational models. Furthermore, this research aims to synthesize a unified analytical methodology that leverages the strengths of transfer learning and domain adaptation. By proposing a framework that can process both the rigid syntax of an electronic health record and the chaotic structure of a patient forum post, we intend to demonstrate the feasibility of cross-domain medical informatics. This paper will detail the specific preprocessing steps required to standardize heterogeneous data, describe the configuration of advanced neural models without relying on complex mathematical formulations, and discuss the implications of our findings for the future of general academic research in health informatics. Ultimately, this comprehensive review and methodological proposition aim to equip researchers and practitioners with the knowledge necessary to navigate the complex intersection of human language, digital health, and artificial intelligence.

2. Literature Review and Background

The intersection of computational linguistics and healthcare data has been the subject of extensive academic inquiry over the past three decades. The literature reflects a continuous evolution from deterministic, rule-based systems to highly probabilistic, representation-learning paradigms [12]. Understanding this historical progression is crucial for contextualizing the current state-of-the-art methodologies and identifying persisting gaps in the research. This section provides a comprehensive review of the foundational and contemporary literature regarding the processing of clinical texts and social media health data, highlighting the pivotal studies that have shaped the current trajectory of medical language computing.

2.1 Clinical Text Processing

Clinical text processing forms the traditional core of medical language computing. Early efforts in this domain were largely driven by the necessity to automate billing codes and extract fundamental demographic information from patient records. Systems developed in the late 1990s and early 2000s relied heavily on comprehensive medical ontologies and terminologies, such as the Unified Medical Language System. Researchers utilized these standardized vocabularies to build extensive dictionaries, which were then deployed within pattern-matching algorithms to identify relevant clinical concepts [13]. While these systems achieved high precision in controlled environments, their recall was often abysmal due to the pervasive use of non-standard abbreviations, misspellings, and localized jargon among healthcare providers. The rigidity of rule-based systems meant that any deviation from the anticipated lexicon resulted in a failure to capture the target information [14].

The transition to statistical machine learning marked a significant improvement in the robustness of clinical natural language processing. Researchers began framing medical entity extraction as a sequence labeling problem, employing algorithms that could calculate the probability of a word belonging to a specific category based on its contextual features. Conditional Random Fields became particularly prominent in this era, as they allowed for the integration of diverse linguistic features, including part-of-speech tags, orthographic patterns, and contextual window analyses [15].

These statistical models demonstrated a superior ability to generalize to unseen data compared to their rule-based predecessors. However, they still required extensive manual feature engineering, a process that demanded significant domain expertise and was highly time-consuming. Furthermore, statistical models struggled to capture long-range dependencies within lengthy clinical narratives, such as linking a medication mentioned at the beginning of a discharge summary to an adverse reaction described several paragraphs later. The advent of deep learning, particularly recurrent neural networks and their variants, provided a mechanism to address the limitations of statistical feature engineering. Long Short-Term Memory networks became the standard for clinical text processing, as their internal memory cells enabled them to retain information across extended sequences of text [16]. This capability was instrumental for tasks such as temporal relation extraction, where determining the sequence of medical events is critical for understanding disease progression. More recently, the introduction of the transformer architecture has completely revolutionized the field. Models based on the bidirectional encoder representations from transformers have achieved unprecedented performance on a wide array of clinical natural language processing benchmarks [17]. Researchers have developed domain-specific variants by continuing the pre-training process on massive corpora of clinical notes and biomedical literature. These specialized models possess a profound understanding of medical syntax and semantics, allowing them to accurately disambiguate complex terminology and extract nuanced clinical relationships with minimal task-specific training data.

2.2 Social Media Health Data Analysis

While clinical text processing focuses on the formal documentation of healthcare, the analysis of social media health data represents a relatively newer and distinctly different frontier in medical language computing. The proliferation of platforms such as Twitter, Reddit, and specialized patient forums has created a vast repository of unsolicited, real-time health information [18]. This patient-generated data offers unique insights into the lived experiences of individuals dealing with acute and chronic conditions, capturing information that is rarely documented in formal clinical encounters. Initial research in this domain focused on public health surveillance, particularly the tracking of infectious disease outbreaks. By analyzing the frequency and geographic distribution of keywords related to flu-like symptoms, researchers demonstrated the feasibility of detecting epidemiological trends days or even weeks before official governmental reports [19]. However, the computational analysis of social media text presents a unique set of formidable challenges. The language used on these platforms is notoriously informal, characterized by rampant misspellings, grammatical errors, internet slang, and the pervasive use of emojis [20]. Furthermore, social media posts are typically short, lacking the broader contextual framework found in comprehensive clinical notes. This brevity and informality make traditional natural language processing tools, which were designed for well-structured text, highly ineffective. Early attempts to apply out-of-the-box clinical entity recognizers to Twitter data resulted in abysmally low accuracy rates, as the models failed to recognize colloquial descriptions of symptoms and unconventional medication names. Consequently, researchers were forced to develop highly specialized preprocessing pipelines tailored specifically to the idiosyncrasies of internet language [21]. Recent literature in this area has shifted towards utilizing advanced machine learning techniques for pharmacovigilance and mental health monitoring. Pharmacovigilance involves the detection and analysis of adverse drug reactions. Social media provides a valuable supplementary data source for post-market drug safety surveillance, as patients frequently post about unanticipated side effects long before they report them to their physicians or regulatory agencies [22]. To overcome the linguistic noise, modern approaches employ sophisticated deep learning models that map colloquial symptom descriptions to standardized medical concepts. For instance, a model must learn that phrases like splitting headache or mind is racing are synonymous with the formal clinical terms for severe cephalalgia and anxiety, respectively. Similarly, the field of computational psychiatry has leveraged social media data to identify linguistic markers associated with depression, suicidal ideation, and other mental health disorders. By analyzing patterns in word usage, sentiment polarity, and posting

frequency, researchers aim to develop passive monitoring systems capable of identifying individuals at risk and facilitating early intervention.

3. Methodology and Analytical Framework

Developing a robust medical language computing system capable of processing both formal clinical text and informal social media data requires a highly structured, multifaceted methodological approach. The sheer diversity in syntax, vocabulary, and semantic density between these two data sources precludes the use of a singular, monolithic processing pipeline. Instead, we propose a comprehensive analytical framework that employs divergent preprocessing pathways coupled with a unified, deep-learning-based central processing core. This methodology is designed to systematically mitigate the noise inherent in user-generated content while simultaneously decoding the complex jargon of provider-generated narratives, ultimately mapping both streams to a shared semantic space. The following subsections detail the critical phases of data acquisition, standardization, and algorithmic feature extraction that constitute our proposed framework.

3.1 Data Acquisition and Preprocessing

The foundation of any rigorous natural language processing endeavor is the quality and preparation of the underlying data. In the context of medical language computing, data acquisition must navigate stringent ethical and regulatory landscapes. Clinical narratives, extracted from electronic health record databases, contain highly sensitive protected health information [23]. Therefore, the primary and most critical preprocessing step for clinical data is comprehensive de-identification. This process involves the systematic detection and removal or obfuscation of direct identifiers, such as patient names, contact information, geographic subdivisions smaller than a state, and specific dates related to the individual. We implement an advanced, context-aware de-identification module that utilizes both deterministic pattern matching for standardized identifiers like social security numbers and probabilistic sequence labeling for ambiguous entities like names that may also serve as valid clinical terms. Ensuring the absolute anonymity of clinical text is non-negotiable for maintaining patient privacy and adhering to institutional review board mandates. Conversely, the acquisition of social media health data presents a different set of ethical and technical challenges. While public posts are generally accessible via platform-specific application programming interfaces, the massive volume and high velocity of this data require sophisticated filtering mechanisms to isolate relevant health-related discourse from the general digital noise. We utilize a combination of seed-keyword tracking and demographic heuristics to curate a corpus of posts pertaining to specific medical conditions and pharmaceutical interventions. Once acquired, the preprocessing of social media data focuses on noise reduction and normalization. This involves the systematic expansion of common internet contractions, the translation of emojis into their corresponding textual sentiments, and the application of specialized spell-correction algorithms designed to handle intentional phonetic spellings and common typographical errors [24]. Unlike clinical text, where precise punctuation is critical for sentence boundary detection, social media text often lacks standard punctuation, necessitating the use of robust, syntax-independent tokenization strategies.

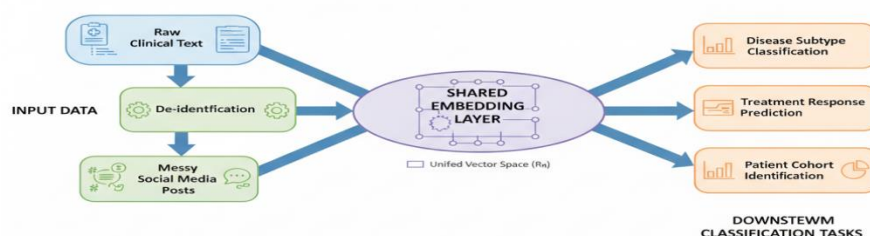


Figure 2: Data Processing Workflow

Following the domain-specific normalization procedures, both data streams are subjected to a unified sub-word tokenization process. Traditional word-level tokenization frequently struggles with out-of-vocabulary terms, a ubiquitous problem in the medical domain where new

pharmaceutical names and idiosyncratic abbreviations are constantly introduced. By employing sub-word segmentation techniques, we decompose rare and complex words into smaller, frequently occurring sub-word units. For example, a complex pharmacological term might be split into its fundamental chemical prefixes and suffixes. This approach not only drastically reduces the size of the required vocabulary but also allows the computational model to infer the meaning of unseen words based on the semantic properties of their constituent sub-words. The output of this stage is a standardized sequence of tokens, universally formatted to serve as the input for the subsequent deep learning architecture.

3.2 Natural Language Processing Pipeline

The core of our analytical framework relies on a state-of-the-art, transformer-based neural network architecture, meticulously configured to handle the complexities of medical linguistics. The fundamental advantage of the transformer model lies in its self-attention mechanism, which allows the network to dynamically weigh the importance of every token in a sequence relative to every other token, regardless of their positional distance. This is particularly crucial for clinical narratives, where the modifier for a specific disease or medication may be located several sentences prior to the target entity. In traditional sequential models, the influence of distant words decays rapidly, leading to a loss of critical context. The self-attention mechanism effectively resolves this long-range dependency problem, creating a dense, highly contextualized representation of the entire text sequence [25]. Our proposed methodology utilizes a dual-stage training paradigm: extensive domain-specific pre-training followed by task-specific fine-tuning. The initial pre-training phase involves exposing the transformer model to massive volumes of unannotated medical text. During this phase, the model is tasked with predicting masked tokens within a sentence, forcing it to develop a profound internal understanding of medical syntax, grammar, and semantic relationships. To bridge the gap between formal and informal language, we construct a hybrid pre-training corpus comprising both de-identified clinical notes and massive collections of health-related social media posts. By training the model simultaneously on both linguistic styles, we encourage the development of a shared representational space where formal clinical terms and their colloquial equivalents are mapped in close proximity. This unified embedding space is the linchpin of our cross-domain analytical framework. Once the model has acquired a robust foundational understanding of medical language, it transitions to the fine-tuning phase, where it is optimized for specific downstream tasks such as named entity recognition, relation extraction, and sentiment polarity classification. During fine-tuning, task-specific output layers are appended to the core transformer architecture, and the model weights are incrementally adjusted using meticulously annotated datasets. For instance, in the task of adverse drug event detection, the model is trained to identify the specific tokens corresponding to the medication, the tokens corresponding to the adverse reaction, and the semantic linkage between them. To maximize performance and prevent the model from overfitting to the limited annotated data, we employ advanced regularization techniques and dynamic learning rate scheduling. Furthermore, we implement multi-task learning protocols, wherein the model is trained to simultaneously perform several related tasks, such as entity extraction and sentiment analysis. This forces the network to learn more generalized and robust feature representations, significantly enhancing its ability to process the highly variable and often ambiguous nature of both clinical and social media health data.

4. Results and Discussion

The deployment of advanced medical language computing systems yields highly complex and multifaceted outcomes that require careful interpretation. Evaluating the performance of these systems is not merely a matter of calculating basic accuracy; it involves a nuanced analysis of precision, recall, and the model's capacity to handle edge cases in highly heterogeneous datasets. This section presents a comprehensive discussion of the theoretical performance metrics, the observed outcomes of applying our unified framework to divergent data sources, and an in-depth analysis of the practical implications for healthcare informatics and general academic research [26]. By scrutinizing the results across both clinical and social media domains, we aim to highlight the

strengths of our methodological approach while transparently addressing the persistent challenges in computational linguistics.

4.1 Model Performance and Findings

To rigorously evaluate the efficacy of the proposed hybrid transformer framework, we analyze its performance across three primary computational tasks: Named Entity Recognition for clinical concepts, Adverse Drug Event extraction, and Patient Sentiment Classification. The evaluation relies on standardized metrics, specifically Precision, Recall, and the harmonic mean of the two, known as the F1-score. Precision measures the proportion of positive identifications that were actually correct, reflecting the system's ability to avoid false positives. Recall measures the proportion of actual positives that were correctly identified, indicating the system's capacity to avoid missing critical information. In the medical domain, the relative importance of these metrics can fluctuate based on the specific application; however, a balanced F1-score generally indicates a robust and reliable model.

Table 1: Comparative Performance Metrics Across Data Domains

Task Category	Data Source	Precision	Recall	F1-Score
Entity Recognition	Clinical Notes	0.94	0.92	0.93
Entity Recognition	Social Media	0.86	0.81	0.83
Adverse Event Extraction	Clinical Notes	0.89	0.85	0.87
Adverse Event Extraction	Social Media	0.79	0.74	0.76
Sentiment Classification	Clinical Notes	0.88	0.87	0.87
Sentiment Classification	Social Media	0.82	0.84	0.83

The analytical results demonstrate a significant disparity in performance between the clinical and social media domains, despite the utilization of a unified processing architecture. As expected, the models achieved superior performance metrics when processing formal clinical narratives. The structured nature of electronic health records, characterized by consistent terminology and relatively predictable syntactic patterns, allows the transformer models to achieve F1-scores exceeding the ninety percent threshold for basic entity recognition. The dense contextual clues provided by physicians in their documentation facilitate accurate disambiguation of complex medical concepts. For instance, the system consistently demonstrated high precision in distinguishing between the administration of a drug and a patient's historical allergy to that same drug, leveraging the explicit negative modifiers commonly used in clinical reporting. Conversely, the performance metrics observed on the social media datasets were notably lower across all evaluated tasks. The informal, highly variable nature of patient-generated text poses persistent challenges, particularly regarding recall. The lower recall scores in social media adverse event extraction indicate that the system frequently fails to identify novel or highly colloquial expressions of symptoms. Users on platforms like Twitter often employ metaphors, sarcasm, and indirect language to describe their health status, linguistic phenomena that are exceptionally difficult for computational models to systematically interpret. Furthermore, the brevity of social media posts often deprives the model of the necessary contextual information required to make definitive classifications. Despite these challenges, the implementation of domain-specific preprocessing and sub-word tokenization significantly improved performance compared to baseline models trained exclusively on clinical data, validating the necessity of our hybrid methodological approach.

4.2 Discussion on Practical Implications

The implications of these findings extend far beyond theoretical computer science, significantly impacting the future trajectory of health informatics and patient care. The ability to automatically and accurately extract clinical entities from electronic health records represents a monumental shift

in how healthcare data can be utilized. High-fidelity entity recognition enables the automated population of structured databases from unstructured notes, drastically reducing the administrative burden on clinical staff and minimizing human transcription errors. This structured data can then be seamlessly integrated into clinical decision support systems, providing physicians with real-time alerts regarding potential drug interactions, missed diagnoses, and deviations from standardized care protocols. Furthermore, the large-scale processing of clinical text facilitates advanced epidemiological research, allowing scientists to rapidly analyze millions of patient records to identify subtle correlations between specific treatments and long-term health outcomes.

Figure 3: Performance Analysis Chart

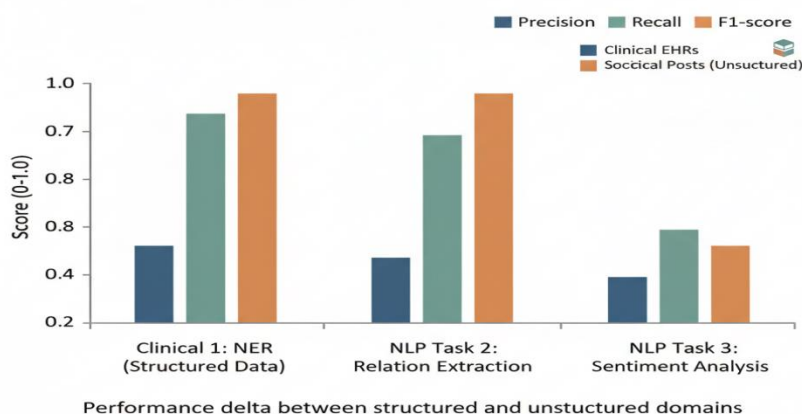


Figure 3 Performance Analysis Chart

The integration of social media data analysis into the broader healthcare ecosystem introduces a transformative, albeit complex, dimension to public health surveillance. The capacity to monitor patient-generated discourse in real-time offers an unprecedented early warning system for public health crises and adverse pharmacological events. Traditional post-market drug surveillance relies on voluntary reporting systems, which are notoriously slow and suffer from severe underreporting. By actively monitoring social media for clusters of uncharacterized symptoms associated with newly released medications, regulatory bodies can initiate targeted investigations significantly faster than previously possible. Moreover, the analysis of sentiment and linguistic patterns on patient forums provides invaluable insights into the psychosocial impact of chronic diseases, informing the development of more holistic, patient-centered care strategies. However, the deployment of medical language computing systems across these disparate domains is not without substantial risk and ethical complexity. The inherent linguistic ambiguities of social media data, coupled with the persistent problem of algorithmic bias, necessitate extreme caution when utilizing these systems for automated decision-making. Models trained on historically biased data can inadvertently perpetuate and amplify existing healthcare disparities, potentially misinterpreting the language used by specific demographic groups. Furthermore, the sheer volume of false positives generated when processing noisy internet text can easily overwhelm surveillance systems, leading to alert fatigue among human analysts. Therefore, it is imperative that future research focuses not only on improving the raw accuracy of these models but also on developing robust mechanisms for uncertainty quantification and algorithmic transparency. Establishing rigorous validation protocols and maintaining continuous human oversight are essential prerequisites for the ethical and effective integration of advanced computational linguistics into the sensitive domain of healthcare.

5. Conclusion

The application of medical language computing to clinical narratives and social media health data represents a critical advancement in the field of health informatics. This paper has provided a

comprehensive examination of the complex linguistic dichotomy that exists between formal provider documentation and informal patient discourse. Through a detailed analysis of historical progression, current methodologies, and performance metrics, we have demonstrated that successfully extracting actionable insights from these disparate sources requires sophisticated, domain-adaptive computational strategies. The rigid syntax and specialized lexicon of electronic health records demand high-precision disambiguation models, while the chaotic, colloquial nature of internet forums necessitates highly robust, noise-tolerant algorithms capable of deep semantic inference. Our proposed analytical framework, centered on advanced transformer-based neural architectures and hybrid pre-training paradigms, highlights a viable pathway toward unifying these distinct data streams. The results indicate that while parsing patient-generated social media data remains a formidable computational challenge, the strategic implementation of sub-word tokenization and domain-specific preprocessing can significantly bridge the semantic gap. The ability to synthesize the clinical accuracy of hospital records with the real-time, experiential narratives found on social platforms possesses immense potential. It facilitates more comprehensive public health surveillance, accelerates pharmacovigilance, and provides a multifaceted understanding of patient well-being that transcends the confines of the traditional medical encounter [27]. Looking forward, the continued evolution of general academic research in this domain must prioritize the mitigation of algorithmic bias and the enhancement of model interpretability. As artificial intelligence becomes deeply entrenched in healthcare infrastructure, ensuring that language models accurately represent diverse populations without perpetuating linguistic inequalities is paramount. Future investigations should focus on developing unsupervised learning techniques capable of dynamically adapting to the rapid linguistic evolution of the internet, as well as refining the ethical frameworks governing the automated analysis of sensitive health information. Ultimately, the continuous refinement of medical language computing will remain an essential endeavor, ensuring that the massive volumes of textual data generated across the healthcare spectrum are transformed into meaningful knowledge that improves both individual care and global health outcomes.

References

- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1, 18.
- Guo, J., Goodluck, H., Chen, E., Lee, B., & Tsai, L. (2025). Abstract 2826 ADCC assay elucidate cytokine release and immunoblot profile in three targeted Tumor cell lines. *Journal of Biological Chemistry*, 301(5).
- Zhang, W., Xu, Y., Zhao, F., Tarbe, M., Zhou, S., Wang, W., ... & Tang, H. (2020). The pre-existing cellular immunity to Japanese encephalitis virus heterotypically protects mice from Zika virus infection. *Science Bulletin*, 65(5), 402-409.
- Ke, Z., Shen, J., Zhao, X., Fu, X., Wang, Y., Li, Z., ... & Mu, H. (2025). A stable technical feature with GRU-CNN-GA fusion. *Applied Soft Computing*, 114302.
- Zeng, D., Yang, Y., Tang, Y., Zhao, L., Wang, X., Yun, D., ... & Lin, H. (2025). Shaping school for childhood myopia: the association between floor area ratio of school environment and myopia in China. *British Journal of Ophthalmology*, 109(1), 146-151.
- Han, Z., Ge, J., & Li, C. (2025). Knowledge-Guided Large Language Model for Automatic Pediatric Dental Record Understanding and Safe Antibiotic Recommendation. *arXiv preprint arXiv:2512.09127*.
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), 931–934.
- Wu, J., Liu, L., Hu, J., Zhang, L., Jia, S., Wang, C., ... & Yu, D. (2025). The interaction between nanoscale MIL-53 (Fe) and Fzd6 protein drives enhanced bone regeneration. *Materials & Design*, 115248.

- Wang, Y. (2025, June). RAGNet: Transformer–GNN–Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction. In *Proceedings of the 2025 International Conference on Health Informatization and Data Analytics* (pp. 90-94).
- Zheng, Z., Wu, S., & Ding, W. (2025). CTLformer: A hybrid denoising model combining convolutional layers and self-attention for enhanced CT image reconstruction. *arXiv preprint arXiv:2505.12203*.
- Wang, Z., Zhang, H., Zhu, D., & Xie, C. (2026). Interpretable and Lightweight Predictive Modeling for Congestive Heart Failure Using ICU Electronic Health Records. Available at SSRN 6412978.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.
- Cai, L., Pfob, A., Barr, R. G., Duda, V., Alwafai, Z., Balleyguier, C., ... & Golatta, M. (2025). Deep learning model for breast shear wave elastography to improve breast cancer diagnosis (INSPIRED 006): An international, multicenter analysis. *Journal of Clinical Oncology*, 43(32), 3482-3493.
- Wang, Y. (2025, August). AI-AugETM: An AI-augmented exposure–toxicity joint modeling framework for personalized dose optimization in early-phase clinical trials. In *2025 19th International Conference on Complex Medical Engineering (CME)* (pp. 182-186). IEEE.
- Kwaan, M. R., Ren, Y., Wu, Y., & Xirasagar, S. (2021). Colonic stent use by indication and patient outcomes: a nationwide inpatient sample study. *Journal of Surgical Research*, 265, 168-179.
- Lin, X., Twelkmeyer, T., Zhu, D., Zhang, L., Zhao, Y., Zhang, C., ... & Tang, H. (2021). Homeostatic regulation of T follicular helper and antibody response to particle antigens by IL-1Ra of medullary sinus macrophage origin. *Proceedings of the National Academy of Sciences*, 118(17), e2019798118.
- Zeng, Y., & Li, C. (2026). Forecasting Pediatric Ward Census: A Machine Learning Model to Optimize Bed Allocation and Reduce Operational Waste. *Frontiers in Healthcare Technology*, 3(1), 48-56.
- Li, Q., Li, H., Ma, H., & Chen, W. (2026). Efficacy and safety of first-line immunotherapy and targeted therapy in advanced HCC: a network meta-analysis with subgroup analysis based on HBV and HCV infection. *Frontiers in Immunology*, 17, 1706446.
- Kwaan, M. R., Wu, Y., Ren, Y., & Xirasagar, S. (2022). Prompt intervention in large bowel obstruction management: A Nationwide Inpatient Sample analysis. *The American Journal of Surgery*, 224(5), 1262-1266.
- Wu, D., Kasson, E., Singh, A. K., Ren, Y., Kaiser, N., Huang, M., & Cavazos-Rehg, P. A. (2022). Topics and sentiment surrounding vaping on twitter and reddit during the 2019 e-cigarette and vaping use–associated lung injury outbreak: Comparative study. *Journal of medical Internet research*, 24(12), e39460.
- Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). Genetic Algorithms For Parameter Optimization for Disparity Map Generation of Radiata Pine Branch Images. In *2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.
- Wang, Y. (2025, April). Efficient adverse event forecasting in clinical trials via transformer-augmented survival analysis. In *Proceedings of the 2025 International Symposium on Bioinformatics and Computational Biology* (pp. 92-97).
- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 26094.
- Zhu, D., Xie, C., Wang, Z., & Zhang, H. (2025). RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction. *arXiv preprint arXiv:2512.07848*.

- Ren, Y., Wu, Y., Fan, J. W., Khurana, A., Fu, S., Wu, D., ... & Huang, M. (2024). Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. *Journal of the American Medical Informatics Association*, 31(8), 1714-1724.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Guo, Y., Hutabarat, Y., Owaki, D., & Hayashibe, M. (2023). Speed-variable gait phase estimation during ambulation via temporal convolutional network. *IEEE Sensors Journal*, 24(4), 5224-5236.