

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: October 05, 2024 at 2:30 PM

Speaker Information

Bushra Hossain (Std No. 0422054008) is a full time Ph.D. student in the department of CSE, BUET. She completed her B.Sc. in CSE from Military Institute of Science and Technology (MIST) in 2010 and M.Sc. in CSE from BUET in 2018. Her research interest covers Natural Language Processing, Human Computer Interaction and Health Informatics. She is currently doing her Ph.D. under the supervision of Dr. Mohammed Eunus Ali. She will discuss her ongoing research in this talk.



Subreddit to Symptomatology: A Lexicon-based Approach to Extract and Analyze Disease Symptoms from Social Media

Complex medical conditions such as Polycystic Ovary Syndrome (PCOS), autoimmune disorders, and longCOVID affect millions globally and often go undiagnosed due to their enigmatic nature. Consequently, individuals frequently turn to social media platforms, such as Reddit, to share their experiences and seek support for managing their conditions. While some studies have explored NLP methods and medical information extraction tools, these typically focus on generic symptoms in clinical notes and struggle to identify disease-specific, subtle symptoms from the informal language used on social media. In this paper, we propose a lexicon-based symptom extraction (LSE) method to identify a comprehensive list of disease symptoms, including subtle symptom mentions from online health discourse. In a real-world PCOS subreddit use case, we find that LSE significantly outperforms state-of-the-art baselines, achieving at least 41% and 16% higher F1 scores than automatic medical extraction tools and large language models, respectively. Notably, LSE ensures broad coverage of symptoms reported in major health guidelines and uncovers interesting insights, such as social determinants of the disease and patterns in symptom occurrence. Additionally, we perform a thematic analysis of peer interactions to identify self-management strategies and reveal knowledge gaps, including rumors and misinformation. As a by-product of our study, we provide a labeled dataset of PCOS Reddit interactions and a PCOS symptom lexicon to support future health informatics research.