Understanding Pre-eclampsia on Reddit

Supervisors: VijayaChitra Modhukur (Medical School, University of Tartu, Estonia), Rajesh Sharma (CSAI, Plaksha University).

Objective: Examine Reddit conversations about pre-eclampsia to pinpoint main topics (such as symptoms and treatments) and sentiment patterns (positive, negative, neutral) through the use of natural language processing (NLP). This initiative modifies techniques from a study on endometriosis discussions on social media (Chiu et al., 2022) to investigate patient experiences with pre-eclampsia.

Significance: Pre-eclampsia is a significant pregnancy complication impacting 2-8% of pregnancies, prediminatly due ti high blood pressure. It can present serious risks to both mother and baby. Reddit provides a wealth of patient-generated content that can reveal emotional and informational patterns. No previous research has utilized sentiment analysis and topic modeling for pre-eclampsia discussions on Reddit, making this a unique student project with the **potential for publication.**

Methodology

1. **Data Collection**: Use Reddit API Wrapper to scrape posts/comments from subreddits like, r/preeclampsia. Use keywords: "pre-eclampsia," "preeclampsia."
2. **Preprocessing**: Clean text (remove URLs, emojis), tokenize, lemmatize using NLTK/spaCy.
3. **Topic Modeling**: Apply Latent Dirichlet Allocation (LDA) with Gensim to identify 5-10 topics (e.g., "symptoms: swelling, headache").
4. **Sentiment Analysis**: Use VADER for sentiment scoring (positive, negative, neutral).
5. **Visualization**: Create word clouds, sentiment charts, and topic visualizations using Matplotlib/Seaborn/pyLDAvis.

Timeline (8 Weeks)

- **Week 1**: Set up project, review literature, configure Reddit API.
- **Week 2**: Collect and preprocess data.
- **Weeks 3-4**: Clean data, run initial LDA and VADER models.
- **Weeks 5-6**: Refine models, generate visualizations.
- **Weeks 7-8**: Analyze results, write report, prepare presentation.

Deliverables

- Python scripts for data collection, analysis, and visualization.
- Anonymized dataset of Reddit posts/comments.
- Visualizations (word clouds, sentiment charts, topic maps).
- 8-page report (in IEEE format) summarizing findings and implications.

References

1. Goel, R., Modhukur, V., Täär, K., Salumets, A., Sharma, R., & Peters, M. (2023). Users' concerns about endometriosis on social media: Sentiment analysis and topic modeling study. *Journal of Medical Internet Research, 25*, e45381. https://doi.org/10.2196/45381
2. Emanuel, R. H. K., Docherty, P. D., Lunt, H., Campbell, R. E., & Moeller, K. (2024). Extracting features and sentiment from text posts and comments relating to polycystic ovary syndrome. *IFAC-PapersOnLine, 58*(24), 19–24. https://doi.org/10.1016/j.ifacol.2024.11.005
3. Dhankar, A., & Katz, A. (2023). Tracking pregnant women’s mental health through social media: An analysis of Reddit posts. *JAMIA Open, 6*(4), ooad094. https://doi.org/10.1093/jamiaopen/ooad094

This project offers an opportunity to identify insights to understand patient perspectives on pre-eclampsia.