**Topic No. (2)**

* **Social Media Analytics for Healthcare Surveillance using Text Mining**

**Motivation**

* Social media produce massive amounts of data on healthcare-related topics, including personal health issues, symptoms, and side effects [[1](#_ENREF_1)]. Twitter data has been found to be useful for public health applications [[2](#_ENREF_2)], including: (1) monitoring diseases, (2) public reaction, (3) outbreaks or emergencies, (4) prediction, (5) lifestyle, and (6) geolocation of disease surveillance [[3](#_ENREF_3)]. Public health surveillance is a systematic collection, analysis, and monitoring of health-related data for public goods, mainly for the purpose of disease prevention and control. For example, syndrome monitoring can be used to track and early detect infectious diseases to indicate potential outbreaks, support disease modeling, or detect cases of biological terrorism. An ontology-based text mining system can also be used to identify and track the distribution of infectious disease outbreaks from linguistic signals on Twitter [[4](#_ENREF_4)]. Twenty-five percent of patients with chronic condition post their experience on social media, and enormous data is generated continuously at an unprecedented scale. However, social media data is known to be notoriously. As a result, a novel and practical framework combining Natural Language Processing (NLP) and Machine Learning (ML) is necessary for extracting signals from the user-generated healthcare content on social media. Furthermore, a predictive model with such extracted signals should be developed as a useful tool for public health surveillance.

**Problem Description**

1. Healthcare-related textual information should be extracted and modeled for the purpose of healthcare surveillance.
2. Robust predictive models are required for accurate forecasting of disease outbreaks and hospital emergency visits based on ML techniques.

**Methodology**

1. A student uses social media APIs (e.g., Twitter APIs) to scrape a large volume of healthcare-related data.
2. A student is required to carry out text preprocessing and feature reduction for effectively reducing lexical nose and maintaining a reasonable feature size.
3. A student performs a feature extraction to transform textual data into numerical vectors for classification using ML techniques. Topic distribution from Latent Dirichlet Allocation (LDA) [[5](#_ENREF_5)] is applied as document level representation.
4. Unified Medical Language System (UMLS) is utilized to process information about the disease and medical terminologies [[6](#_ENREF_6)].
5. A student develops feature augmentation based domain adaption algorithms for efficient signal extractions.
6. A student is also recommended to develop a hospital emergency visit prediction model based on the collected Twitter data.

**Resources (Implementation Framework and Platform)**

* Machine Learning for Language Toolkit, MALLET
* Unified Medical Language System (UMLS)

|  |  |
| --- | --- |
| **Research Design and Methods** | |
| **Items** | **Useful Resources** |
| **Overview**  To take full advantage of Big Data, NLP, ML, there is an urgent need to enrich our understanding of those techniques and use them to provide insights for health providers and patients. | https://bit.ly/32jrztg |
| **Research Question(s)**  Two intriguing questions about healthcare social media and disease management/prevention must be addressed:   1. How can we use NLP and ML for a particular disease surveillance (for instance, asthma or sepsis) to enable health providers to respond promptly? 2. How can we apply NLP and ML to enhance accurate forecasting of disease outbreaks and hospital emergency visits? | \*Important remark:  Based on students’ interest, one disease should be identified for the research topic.  -top diseases (acute, chronic, communicable and non-communicable) |
| **Sources of Data (Healthcare related Twitter tweets)**  **(Option\_1)**   * <http://healthcare-twitter-analysis.com/> * <https://snap.stanford.edu/data/twitter7.html> * <https://lionbridge.ai/datasets/12-best-social-media-datasets/>   **(Option\_2)**   * Perform web scraping to collect Social Q&A sites * For instance, (patientslikeme.com) * <https://lionbridge.ai/datasets/12-best-social-media-datasets/> * (Yahoo! Answer) |  |
| **Collection of Data**  Web scraping should be applied to collect social media datasets or data preprocessing should be carried out to store those public available dataset (JSON file) into database system mySQL.  . | - Python  - R |
|  | |
| **Timeframes (need to be discussed)**   1. Literature review 2. Web Scraping for social media data or data preprocessing 3. Analytics for textual healthcare dataset 4. Prototype development of data visualization 5. Time-series Analytics for textual healthcare dataset 6. Building predictive models for forecasting hospital visits 7. Hypothesis testing with STATA |  |

**References:**

1. Lee, K., A. Agrawal, and A. Choudhary. *Real-time disease surveillance using twitter data: demonstration on flu and cancer*. in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013. ACM.

2. Denecke, K. and W. Nejdl, *How valuable is medical social media data? Content analysis of the medical web.* Information Sciences, 2009. **179**(12): p. 1870-1880.

3. Andreu-Perez, J., et al., *Big data for health.* IEEE journal of biomedical and health informatics, 2015. **19**(4): p. 1193-1208.

4. Collier, N., et al., *BioCaster: detecting public health rumors with a Web-based text mining system.* Bioinformatics, 2008. **24**(24): p. 2940-2941.

5. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003. **3**(Jan): p. 993-1022.

6. Bodenreider, O., *The unified medical language system (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(suppl\_1): p. D267-D270.