Morgan Whitman
University of Illinois Urbana-Champaign
CS410 - Text Information Systems
Technology Review

Text Analysis in Healthcare and Biomedicine

"Our health care system has failed when a doctor fails to treat an illness that is treatable." This quote from Kevin Alan Lee urges for more efficient modes of healthcare but can be expanded to finding efficiencies utilizing data science and applying this within biomedicine. Text analysis helps solve the unique challenges presented by biomedical data and provide interesting applications to treat disease and improve the quality of people's lives.

As the amount of biomedical and healthcare data continues to grow, developing models, relationships, and insights has become more challenging. Biomedical data is largely unstructured and disorganized data. In addition to this, the data is high dimensional and difficult to handle. This information is held within scientific papers, electronic health records, submitted health claims, diagnostic imaging, case summaries, patient notes, and scientific research papers. Extracting information from this text is difficult, but utilizing text analysis techniques helps model, discover and extract relationships, patterns, and trends from this raw data.

A unique feature of biomedical text analysis is the ontological and terminological resources in biomedicine. These resources were developed to standardize the diverse scientific language for research and communication (Flórez-Arango, and Castrillón-Betancur, 2015[1]). This structure supplements the development in utilizing natural language processing to create algorithms to effectively parse, normalize, classify, and "understand" the data. Biomedical

[1] Flórez-Arango,, J. and Castrillón-Betancur, J., 2021. *Terminologies and classification systems in biomedical sciences*. [ebook] Available at: <http://scielo.sld.cu/pdf/rcim/v7n1/rcim09115.pdf>

natural language processing that structures the originally unstructured biomedical data is improved upon with the standardized scientific vocabularies.

One successful technique of text analysis takes unstructured data and outputs precise topic labels using a hybrid inverse document frequency and machine learning fuzzy k-means clustering algorithm (J. Rashid et al, 2019[2]). This proposed algorithm uses the bag-of-words model to generate word frequencies, local through the term frequency and global using a hybrid inverse documents frequency with principal component analysis. Thereafter, k-means clustering is utilized with a discriminant analysis classifier. This technique proves high classification and clustering performance in addition to a stable execution time over a range of topics.

Interesting applications of text analysis is Data integration. Data integration utilizes data sets with supervision to discover unique patterns between experiments. In these cases, narrowed datasets and noise reduction techniques supplemented successful correlation to identify biological signals of disease.

Overall, text analysis takes the challenges presented by the vast amounts of unstructured, complicated biomedical data and presents great opportunity to utilize text analysis within the biomedical sciences. This opportunity is ripe with productive and altruistic results to be a part of improving lives and treating “treatable” diseases.

[2] "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," in *IEEE Access*, vol. 7, pp. 146070-146080