IDENTIFYING PATTERNS IN DRUG EFFICACY BY ANALYZING DRUG REVIEWS THROUGH A CLUSTERING APPROACH

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the face of growing review platforms and discussion forums, consumers are increasingly relying on the tools to when coming up with the purchasing decisions (Dinh, Chakraborty, & McGaugh, 2020). Review sites allow current consumers to make comments on products and services that they have utilized, thus helping the potential buyers in making smart purchasing decisions (Dinh et al., 2020). According to Qiu and Zhang (2024), 95% of shoppers read online reviews before making a purchase. Additionally, a study by Biswas, Sengupta, and Ganguly (2022) found that products that have at least five reviews have a 270% more chance of being purchased compared to products that do not have reviews. Therefore, the studies demonstrated that customer review was essential in purchasing a product, particularly in determining the effectiveness of drugs. This is because reviews provide useful information about a drug's effectiveness, side effects, and overall patient experience, allowing healthcare providers and potential users to make informed decisions. Generally, online review about drugs consists of two major parts which are ratings and text reviews (Dinh et al., 2020). While ratings provide a numerical assessment of a customer's overall experience, text reviews offer more detailed insights into a drug's effectiveness and side effects that ratings unable to provide with (Dinh et al., 2020). By referring to Sridharan and Sivaramakrishnan (2024), medical review is critical to optimize medication therapy and minimize medication errors. Besides that, drug reviews provide valuable insights to the pharmacology by optimizing drugs and improving results of medical cases (J. Liu, Zhou, Jiang, & Zhang, 2020). Additionally, understanding patients' medical conditions through drug reviews can help patients to choose a better medicine when medical advice is limited (Zeroual, Harrou, Dairi, & Sun, 2020). Thus, sentiment analysis of drug reviews offers valuable insights into drug effects and benefits that may not be fully addressed in clinical trials. Therefore, this thesis tends to apply Large Language Models and clustering techniques to analyse drug reviews and identify patterns in drug efficacy. By extracting meaningful insights from patient feedback, the project aims to enhance the understanding about the performance of drugs across various conditions.

1.2 Problem Background

Randomized controlled trials (RCTs) are usually considered as the gold standard in evaluating drug efficacy due to the monitored procedures that aim to eliminate bias (Hariton & Locascio, 2018). The randomization process in RCTs ensures that distribution of age, gender and health status are evenly distributed across different group of treatments. This strengthened the ability of experiment to determine that observed effects were due to the drug itself as the bias had been minimize. Hence, the controlled structured make RCTs reliable for the treatment outcomes. However, despite their quality, RCTs fail to provide a complete overview of a drug's effectiveness in the real world because limitations exist in its generalizability and applicability to the larger patient population. RCTs frequently include strict eligibility requirements that exclude individuals with various health issues. As a result, the outcomes cannot fully generalize in a more diverse population (Kostis & Dobrzynski, 2020). RCTs results not accurately reflect real world situations of drug efficacy regarding long-term side effects and the improvement of symptoms across different patient groups. Hence, although RCTs offer valuable information about drug effectiveness in optimum conditions, it’s still have drawbacks in evaluating its efficacy in uncontrolled situations (Kaul, Bose, Kumar, Ilahi, & Garg, 2021). In this context, patient reviews are critical in bridging the gap between RCT findings and real-world scenarios.

Several studies have been conducted to address this gap in understanding. Shahid, Singh, Gupta, and Sharma (2022) proposed a deep learning-based medical recommendation system with N-Gram and patient review data. They apply sentiment analysis to recommend treatments based on patient feedback (Shahid et al., 2022). Furthermore, Rathod, Patel, Goswami, Degadwala, and Vyas (2023) studied the application of machine learning methods in sentiment analysis of drug reviews to extract insightful information from unstructured data collected on the Internet. Their study examined the performance of different machine learning algorithms in sentiment analysis and feature engineering techniques, and can accurately capture the sentiment of drug reviews, achieving high accuracy and F1 scores (Kostis & Dobrzynski, 2020; Rathod et al., 2023). While traditional clinical trials give valuable information on drug performance in controlled conditions, they fail to concern on the wide range of patient experiences and long-term effects that may occur with actual use in real world. Clustering drug reviews able to discover underlying patterns of drug efficacy by analysing the personal experiences of consumers. Hence, by detecting positive or negative patterns of reviews will help in clinical decision-making.

1.3 Problem Statement

As RCTs frequently conducted with controlled conditions and select the individuals with specific disease, hence limiting its generalizability for a more diverse patient population. Medical professionals lack the comprehensive knowledge about drug performance in various scenarios. However, patient drug reviews with the real-world experiences provide unreported side effects and varying efficacy results. Thus, the problem statement of this study is the lack of comprehensive information from RCTs limit the understanding of drug efficacy in diverse patients' populations.

1.4 Research Questions

This thesis aims to visualize the patterns in drug efficacy by utilizing Large Language Models (LLMs) in extracting the relevant keywords and clustering patient reviews based on the keywords.

The research questions are:

(a) How do LLMs extract meaningful information from drug reviews?

(b) What clustering techniques perform well in categorizing the extracted keywords from drug reviews?

(c) What insights can be derived from clustering drug reviews?

1.5 Research Goal

The aim of the project is to identify patterns in drug efficacy to enhance the understanding of drug performance across diverse patient populations by utilizing LLMs and clustering techniques in patient drug reviews.

1.5.1 Research Objectives

The objectives of the research are:

(a) To conduct a preprocessing of the drug reviews datasets for drug efficacy analysis

(b) To extract relevant keywords from the pre-processed dataset by using Large Language Models

(c) To implement clustering techniques to categorize the extracted keywords and visualize the findings

1.6 Scope

The scopes of the research are:

(a) The data will be collected from UCI Irvine Machine Learning Repository

(b) The programming languages used is Python

(c) Concentrate on the sentiment analysis of the patient drug review, aiming to extract insights related to drug efficacy, side effects and overall patient satisfaction

1.7 Significance of Research

The significance of this research is the potential to bridge the gap between RCTs and the real-world scenarios. The research will contribute to a better understanding of drug performance across diverse patient populations by utilizing LLMs and clustering techniques to extract and categorize the meaningful information from patient feedback. Additionally, evaluating the drug efficacy across various patient populations will enhance the clinical decision-making process and provide valuable insights for medical professionals in improving patient care and treatment outcomes.

REFERENCES

Al-Hadhrami, S., Vinko, T., Al-Hadhrami, T., Saeed, F., & Qasem, S. N. (2024). Deep learning-based method for sentiment analysis for patients' drug reviews. *PeerJ Computer Science, 10*, e1976.

Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2024). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics, 20*(1/2), 142-161.

Ampel, B., Yang, C.-H., Hu, J., & Chen, H. (2024). Large language models for conducting advanced text Analytics Information Systems Research. *ACM Transactions on Management Information Systems*.

Anderson, P., Higgins, V., Courcy, J. d., Doslikova, K., Davis, V. A., Karavali, M., & Piercy, J. (2023). Real-world evidence generation from patients, their caregivers and physicians supporting clinical, regulatory and guideline decisions: an update on Disease Specific Programmes. *Current Medical Research and Opinion, 39*(12), 1707-1715.

Belal, M., She, J., & Wong, S. (2023). Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Benatti, A., & Costa, L. d. F. (2024). Agglomerative clustering in uniform and proportional feature spaces. *arXiv preprint arXiv:2407.08604*.

Bhardwaj, A., Pandey, A., & Dahiya, S. (2022). *Review based on Variations of DBSCAN algorithms.* Paper presented at the 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS).

Biswas, B., Sengupta, P., & Ganguly, B. (2022). Your reviews or mine? Exploring the determinants of "perceived helpfulness" of online reviews: a cross-cultural study. *Electronic Markets, 32*(3), 1083-1102.

Bu, K., Liu, Y., & Ju, X. (2024). Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems, 283*, 111148.

Bushra, A. A., & Yi, G. (2021). Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. *IEEE Access, 9*, 87918-87935.

Chong, B. (2021). K-means clustering algorithm: a brief review. *vol, 4*, 37-40.

Cimino, A., Culbertson, C., Watkins, E., Li, J., & Wangeshi, S. (2024). RWD119 A Methodological Approach Using Sentiment Analysis of Online Medical Platforms As a Real-World Data Source of Patient Experiences. *Value in Health, 27*(6), S381.

Dinh, T., Chakraborty, G., & McGaugh, M. (2020). *Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis and Data Mining Models.* Paper presented at the SAS 2020 Global Forum.

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). *Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning.* Paper presented at the Proceedings of the 2018 international conference on digital health.

Gui, C., Han, D., Gao, L., Zhao, Y., Wang, L., Xu, X., & Xu, Y. (2024). Application of Enhanced K-Means and Cloud Model for Structural Health Monitoring on Double-Layer Truss Arch Bridges. *Infrastructures, 9*(9), 161.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software, 91*, 1-30.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology, 125*(13), 1716.

Hu, L., Jiang, M., Dong, J., Liu, X., & He, Z. (2024). Interpretable Clustering: A Survey. *arXiv preprint arXiv:2409.00743*.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622*, 178-210.

Ji, Z., & Wang, C.-L. (2021). *Accelerating DBSCAN algorithm with AI chips for large datasets.* Paper presented at the Proceedings of the 50th International Conference on Parallel Processing.

Jiang, K., Lai, X.-x., Yang, S., Gao, Y., & Zhou, X.-H. (2024). A Practical Analysis Procedure on Generalizing Comparative Effectiveness in the Randomized

Clinical Trial to the Real-world Trialeligible Population. *arXiv preprint arXiv:2406.04107*.

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. (2024). Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 100059.

Kaul, P., Bose, B., Kumar, R., Ilahi, I., & Garg, P. K. (2021). The strength of a randomized controlled trial lies in its design—randomization. *Supportive Care in Cancer*, 1-3.

Kim, K. S., Chan, A.-W., Belley-Côté, E. P., & Drucker, A. M. (2022). Noninferiority Randomized Controlled Trials. *Journal of Investigative Dermatology, 142*(7), 1773-1777.

Kostis, J. B., & Dobrzynski, J. M. (2020). Limitations of randomized clinical trials. *The American journal of cardiology, 129*, 109-115.

Kubota, Y., & Narukawa, M. (2023). Randomized controlled trial data for successful new drug application for rare diseases in the United States. *Orphanet Journal of Rare Diseases*, *18*(1), 89.

Lee, D.-g., Kim, M., & Shin, H. (2022). *Drug Repositioning with Disease-Drug Clusters from Word Representations*. Paper presented at the 2022 IEEE International Conference on Big Data and Smart Computing (BigComp).

Liakos, A., Pagkalidou, E., Karagiannis, T., Malandris, K., Avgerinos, I., Gigi, E., . . . Tsapas, A. (2024). A Simple Guide to Randomized Controlled Trials. *The International Journal of Lower Extremity Wounds*, 15347346241236385.

Liu, J., Zhou, Y., Jiang, X., & Zhang, W. (2020). Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews. *BMC medical informatics and decision making, 20*, 1-13.

Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience, 2022*(1), 3762431.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(6), e1219.

Newell, D. J. (2020). Randomised controlled trials in health care research. In *Researching Health Care* (pp. 47-61): Routledge.

Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

Oti, E. U., & Olusola, M. O. (2024). OVERVIEW OF AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS. *Technology, 7*(2), 14-23.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review, 56*(7), 6439-6475.

Pratama, M. A. Y., Hidayah, A. R., & Avini, T. (2023). Clustering K-Means untuk Analisis Pola Persebaran Bencana Alam di Indonesia. *Jurnal Informatika Dan Tekonologi Komputer (JITEK), 3*(2), 108-114.

Qiu, K., & Zhang, L. (2024). How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors. *Data and Information Management, 8*(2), 100058.

Rangapur, A., & Rangapur, A. (2024). The Battle of LLMs: A Comparative Study in Conversational QA Tasks. *arXiv preprint arXiv:2405.18344*.

Rathod, D., Patel, K., Goswami, A. J., Degadwala, S., & Vyas, D. (2023). *Exploring drug sentiment analysis with machine learning techniques*. Paper presented at the 2023 International Conference on Inventive Computation Technologies (ICICT).

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Řezanková, H. (2018). *Different approaches to the silhouette coefficient calculation in cluster evaluation*. Paper presented at the 21st international scientific conference AMSE applications of mathematics and statistics in economics.

Shahapure, K. R., & Nicholas, C. (2020). *Cluster quality analysis using silhouette score*. Paper presented at the 2020 IEEE 7th international conference on data science and advanced analytics (DSAA).

Shahid, T., Singh, S., Gupta, S., & Sharma, S. (2022). *Analyzing Patient Reviews for Recommending Treatment Using NLP and Deep Learning-Based Approaches*. Paper presented at the International Conference on Advancements in Interdisciplinary Research.

Sridharan, K., & Sivaramakrishnan, G. (2024). Unlocking the potential of advanced large language models in medication review and reconciliation: a proof-of-concept investigation. *Exploratory Research in Clinical and Social Pharmacy, 15*, 100492.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods, 23*, 16094069241231168.

Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Xu, Q., Gu, H., & Ji, S. (2024). Text clustering based on pre-trained models and autoencoders. *Frontiers in Computational Neuroscience, 17*, 1334436.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, solitons & fractals, 140*, 110121.