Social Media as a Sensor: Analyzing Twitter Data for Breast Cancer Medication Effects Using Natural Language Processing

Seibi Kobara[1], Masoud Nateghi[1], and Alireza Rafiei[1]

[1]Department of Computer Science and Informatics, Emory University, Atlanta, GA, USA

Breast cancer remains a significant cause of cancer-related deaths across the globe [1]. Despite extensive research efforts leading to the development of various interventions tailored to different breast cancer types and endotypes, the prevalence of this disease in the United States continues to rise [2]. The use of natural language processing (NLP) has been explored for detecting breast cancer recurrence and cataloging patient-reported symptoms linked to cancer treatments [3]. However, the practicality of leveraging Twitter data to discern medication usage and its consequent side effects among breast cancer patients has not been extensively studied. Uncovering the potential commonality of drug-related side effects through analysis of social media could broaden our understanding of breast cancer's impact, offering insights that may surpass those obtained from traditional cohort studies.

In this study, our first objective is to develop a classifier capable of differentiating between Twitter posts related to breast cancer and unrelated content. The second goal is to create a comprehensive lexicon of breast cancer-related medications and side effects, and subsequently design an NLP framework that can identify language pertaining to medications and side effects. Lastly, we aim to conduct a descriptive analysis to examine the prevalence of medication-induced side effects using statistical testing. For the first objective, we will use the existing annotation file of breast cancer posts to develop different machine learning and deep learning-based classifiers. The developed models will be evaluated using the corresponding test dataset. For the second objective, we will use the collection of approved breast cancer medication in the National Cancer Institute [4]. We expect that posts on Twitter may use abbreviated drug names or colloquially unique names; thus, we will create a manual, standard document to define the potential words for each drug. After curating the manual annotation, we will design a rule-based NLP model to detect the use of breast cancer medications in the posts. Additionally, we aim to investigate the side effects associated with breast cancer drugs. We will first collect reported side effects in previous epidemiologic research on breast cancer. With these symptoms, we will create a manual annotation set, which we will use to discover posts with side effects. For this objective, we will use a partial annotation dataset to test the model performance. Finally, we will perform a descriptive analysis of the side effects associated with breast cancer medications. To test the difference in the distribution of the side effects, we will compare the proportion of different side effects for a particular drug. To validate this finding, we also plan to use a manually curated collection of data to analyze the differences in the proportion of side effects for a certain drug.

References

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68, 394–424 (2018).
2. Ellington, T. D. et al. Trends in breast cancer incidence, by race, ethnicity, and age among women aged≥ 20 years—United States, 1999–2018. Morbidity and Mortality Weekly Report 71, 43 (2022).
3. Al-Garadi, M.A., et al. Automatic Breast Cancer Cohort Detection from Social Media for Studying Factors Affecting Patient-Centered Outcomes. In Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18 (pp. 100-110). Springer International Publishing.
4. National Cancer Institute. Drugs Approved for Breast Cancer 2023. https : / / www . cancer . gov/about-cancer/treatment/drugs/breast.