Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Social Computing (Eiji Aramaki (Professor))		
Student ID	2111442	Submission date	2023 / 7 / 24
Name	ZHANG ZHOUQING		
Thesis title	Poster–patients relationships show differing content and language on Chinese Weibo: Text classification, sentiment analysis, and topic modeling of posts on breast cancer		

Abstract

Background:
Breast cancer affects the lives of not only those diagnosed, but also the people around them. Many are open to sharing about their experiences as narratives on social media. However, these narratives may differ according to who the poster is, and their relationship with the patient is. In China, Weibo is one of the most popular social media platforms, and breast cancer–related posts can be frequently found here. We thus use this as a resource to examine how posts differ according to the different poster–patient relationships.

Objective:
With the goal of understanding the different experiences of those affected by breast cancer in China, we aim to explore how content and language used in relevant posts differ according to who the poster is, and their relationship with the patient. Accordingly, we examine the emotional content of these posts, and the common topics they mention. Our goal is to study if there are differences in emotional expression and topic content if the patient is the poster themselves, or a friend, family, relative or acquaintance.

Method:
We scraped a total of N=10322 relevant Weibo posts. Using a 2–step analysis method, we first fine–tune two Chinese RoBERTa models, on dataset annotated with poster–patient relationships. These models were lined in sequence, first a binary classifier (no patient or patient), and multiclass classifier (post user, Family members, Friends relatives, Acquaintances, Heard relation) to classify patient relationships. Next, we used the LIWC lexicon to conduct sentiment analysis on 5 emotion categories (positive and negative emotions, anger, sadness, and anxiety), followed by topic modeling (BERTopic).

Results:
Our binary model (F1 = 0.93) and multiclass model (F1 = 0.83) were largely able to classify patient relationships accurately. Subsequent sentiment analyses showed significant differences in emotion categories across all patient relationships. Notably, negative emotions and anger were higher for the 'no patient' class, but sadness and anxiety were higher for the 'family member' class. Focusing on the top 30 topics, we also noted that topics about fears and rage towards cancer were higher in the 'no patient' class, but topics about cancer treatment were higher in the 'family member' class.