Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing

Swati Rajwal[1]*, Avinash Kumar Pandey[2]*, Zhishuo Han[2], Abeed Sarker[1]
[1]Dept. of Biomedical Informatics, Emory University, [2]Goizueta Business School, Emory University
{swati.rajwal, avinash.kumar.pandey, zhishuo.han, abeed.sarker}@emory.edu
* Shared First Author.

Abstract

We leveraged a dataset of ∼1.5 million Twitter (now X) posts to develop a framework for analyzing breast cancer (BC) patients' concerns and possible reasons for treatment discontinuation. Our primary objectives were threefold: (1) to curate and collect data from a BC cohort; (2) to identify topics related to uncertainty/concerns in BC-related posts; and (3) to conduct a sentiment intensity analysis of posts to identify and analyze negatively polarized posts. RoBERTa outperformed other models with a micro-averaged F_1 score of 0.894 and a macro-averaged F_1 score of 0.853 for (1). For (2), we used GPT-4 and BERTopic, and qualitatively analyzed posts under relevant topics. For (3), sentiment intensity analysis of posts followed by qualitative analyses shed light on potential reasons behind treatment discontinuation. Our work demonstrates the utility of social media mining to discover BC patient concerns. Information derived from the cohort data may help design strategies in the future for increasing treatment compliance.

Keywords: Natural Language Processing, Breast Cancer, Social Media, Concerns, Topic Modeling

1. Introduction

1.1. Background

In 2020, there were 1,603,844 new cases of breast cancer (BC) and 602,347 died of BC in the United States as per the Centers for Disease Control and Prevention (2021). While advances in treatment have improved survival rates, a critical challenge persists in the continuity of long-term therapies which are crucial for reducing the risk of cancer recurrence. Unfortunately, many patients discontinue their treatments prematurely which is often linked to adverse patient-centered outcomes (PCOs). PCOs encompass a range of patient-specific experiences and even extend to broader socio-economic concerns related to treatments that are inherently difficult to measure and are underrepresented in electronic health records (EHRs). Motivated by this, our objective is to verify if social media BC cohort contains information about patient concerns, sentiments, and potential reasons for treatment non-compliance or discontinuation.

1.2. Related Work

Social media data has long been used for a range of tasks such as sentiment and opinion mining (Pak and Paroubek, 2010; Ananth et al., 2017), medication-/drug-related information analysis (Klein et al., 2024; Nikfarjam et al., 2019; Weissenbacher et al., 2021), mental health-related research (Amir et al., 2019; Le Glaz et al., 2021), substance use and recovery (Kepner et al., 2022; Balsamo et al., 2023; Yang et al., 2023), public health (Antonius and Rich, 2013) and many others. Additionally, researchers are actively using social media platforms to discover targeted cohorts (Krauss et al., 2015; Sarker et al., 2017; Al-Garadi et al., 2020). Studies have shown that people often use social media to share their health-related experiences, including for BC (Attai et al., 2015; Nzali et al., 2017; Sarker, 2017). Thus, social media is a promising resource for capturing patient experiences and sentiments, or PCOs, provided the target cohort is accurately identified. In this work, we aim to utilize a dataset collected in prior work by Al-Garadi et al. (2020). The dataset contains N = 1,454,638 tweets from 583,962 unique users. We developed a BC self-report identification system utilizing supervised machine learning models and RoBERTa (Liu et al., 2019). We used data from the cohort to identify BC-related uncertainty or concern topics. We performed a sentiment intensity analysis of patients who voiced dissatisfaction and identification of treatment discontinuation in the self-reported posts category.

2. Methodology

2.1. Self-Report Classifiers

We used manually-annotated data (Train=3513, Dev=302, Test=1204) to compare decision tree, logistic regression, random forest, naïve Bayes, and RoBERTa (Liu et al., 2019) for the task of BC self-report classification.

2.2. GPT-4 & Topic Modelling Framework

We used GPT-4 to generate an initial list of 20 seed words (A.1) related to uncertainty/fear. GPT-4's suggestions were often formal and not popu-

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 264–270
20 May, 2024.

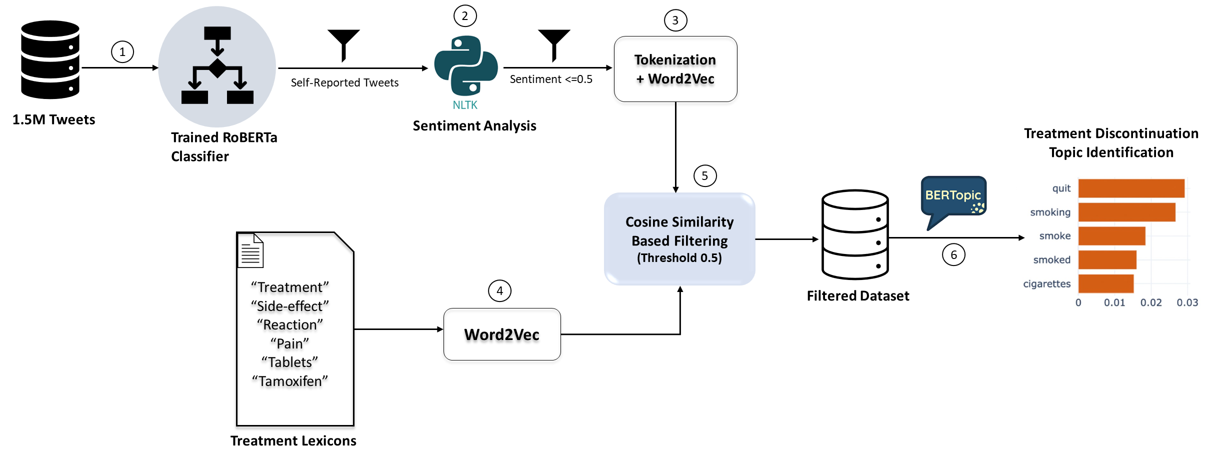

Figure 1: NLP pipeline for treatment discontinuation topic identification in a breast cancer Twitter cohort.

larly used by people to express their fear. Therefore, we utilized the Word2Vec word embedding model (Mikolov et al., 2013), particularly leveraging the GoogleNews-vectors-negative300 pretrained model[1] to identify similar keywords. The Word2Vec model's output, after applying cosine similarity, helped us in curating our final list of keywords (A.2). Finally, we applied BERT topic modeling (Grootendorst, 2022) to all posts containing these terms in order to identify the key issues that are of concern to BC patients.

2.3. Sentiment Analysis Framework

We calculated the sentiment score for each Tweet using NLTK's Sentiment Analyzer.[2] Then we applied BERTopic to a corpus of posts that had negative sentiment (score<=0.5) and contained specific terms indicative of treatments, side effects, and symptoms associated with BC (A.3). The lexicon was derived from a PubMed paper on BC symptoms (Koo et al., 2017).

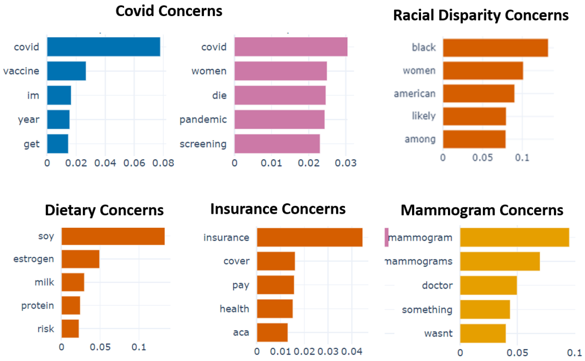

Figure 2: Key concerns among breast cancer patients discovered via topic modeling.

3. Results and Discussion

3.1. Classifier Evaluation & Results

RoBERTa significantly outperformed all models with a micro-averaged F_1 score of 0.894, macro-averaged F_1 score of 0.853 and, lowest log loss of 0.332 (Table 1, Appx. A.4). In the ~1.5 million dataset, the best-performing (RoBERTa) classifier identified 154,571 posts as self-reported BC. Such posts were further used to identify topics for treatment discontinuation as shown in Figure 1.

3.2. Breast Cancer Concerns

BERTopic revealed insightful topics associated with the public's concerns. In Fig. 2, the Y-axis shows topics (cluster of terms) that have been identified in the text. X-axis represents the scores/weights of each term within a specific topic. Each bar's length indicates the strength of association between the term and its respective topic. The different colors for each topic are for visual clarity only. Below are the key issues associated with each topic:

- **Impact of COVID on Screenings**: A major topic cluster was around the impact of COVID-19 on BC patients. Concerns were raised about seeking medical help during the pandemic which could lead to delayed diagnoses and an increase in mortality rates. These findings are in line with a report by the CDC (2023).
- **Mammogram Anxiety**: Anxiety associated with mammogram screening was another notable topic. Discussions revealed anxiety towards mammogram screenings, exacerbated by the pandemic and fear of discovering BC.
- **Insurance Issues**: Posts (Table A.1) reflected frustration over the lack of coverage and the financial burden placed on patients, highlighting the systemic barriers to accessing care.

[1] huggingface.co/fse/word2vec-google-news-300
[2] nltk.org/howto/sentiment.html

Model	Hyperparameter	F_1 micro	F_1 macro	F_2 micro	F_2 macro	Log loss
Decision Tree	criterion='gini', max_depth=10	0.778	0.608	0.778	0.596	0.734
Logistic Reg.	C=10, penalty='l2'	0.772	0.576	0.772	0.570	0.464
Naïve Bayes	alpha=0.1	0.745	0.427	0.745	0.468	0.568
Random forest	n_estimators=50	0.752	0.476	0.752	0.498	0.652
RoBERTa	epochs=20, batch_size=16	**0.894**	**0.853**	**0.894**	**0.841**	**0.332**

Table 1: Self-reported breast cancer tweet classification results across multiple evaluation metrics.

- **Soy Consumption**: Dietary habits and its link to BC is discussed. The chatter pointed to confusion regarding the consumption of soy and its alleged estrogenic effects.
- **Racial Disparity in Mortality**: The conversations highlighted concerns around the disproportionate impact of BC on African-American women including higher mortality rates.

3.3. Sentiment Analysis and Treatment Discontinuation

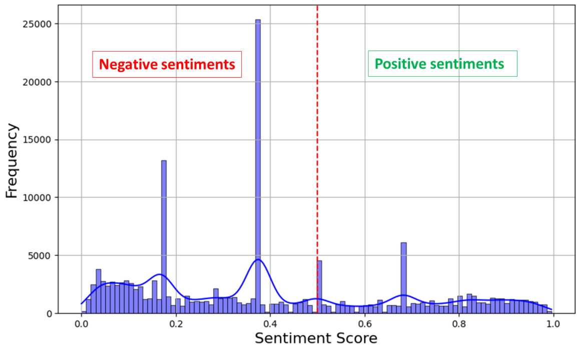

Figure 3: Sentiment Analysis of Breast Cancer Tweets via RoBERTa reveals more negative posts.

Sentiment analysis of self-reported BC posts revealed a pronounced skew towards negative sentiments, with the majority of sentiment scores falling at or below the 0.5 threshold (Figure 3). This trend suggests that the conversations are predominantly centered around the challenges faced by individuals, potentially reflecting the adverse effects of BC treatments. The BERTopic analysis of negative sentiment posts using keywords related to the side effects and treatment experiences of BC patients highlighted following main themes potentially linked to treatment discontinuation (Figure 4):

- **Hair/Baldness**: Concerns over hair loss reveal the psychological impact and social implications of chemotherapy-induced alopecia.
- **Smoking Restrictions**: Discussion around "quit," "smoking," suggests that patients might face difficulty adhering to treatment protocols that necessitate smoking cessation.
- **Insurance Issues**: High frequency of words like "insurance," "aca,", "cost," and "bill" indicate that financial burdens and insurance coverage limitations are significant barriers.
- **Bowel/Colon Issues**: References to terms such as "colon," "colonoscopy," "bowel," "colorectal," & "polyps" suggest gastrointestinal side effects or procedures related to treatment may be intolerable for some patients.
- **Sexual Abuse Concerns**: Presence of terms like "sexual," "abuse," "rape," "assault," and "abuser" may reflect traumatic personal histories that intersect with treatment experiences.
- **Mental/Emotional Health**: A significant number of mentions of "cry," "crying," "emotional," "emotions," and "cried" highlight the emotional and psychological toll of BC and treatment.

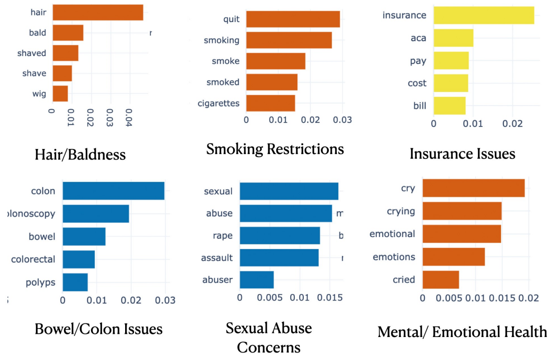

Figure 4: Treatment discontinuation topics from negative sentiment self-reported posts.

3.4. Post Classification Analyses

3.4.1. Error Analysis

Lack of context, ambiguous references, and the use of informal language were the primary reasons for classification errors. For instance, Tweet 1 in A.6 expresses fear and mentions a family history of BC without directly stating a diagnosis. This poses challenges for the model to accurately classify such posts as self-reports. Moreover, the context in which BC is discussed varies widely. For

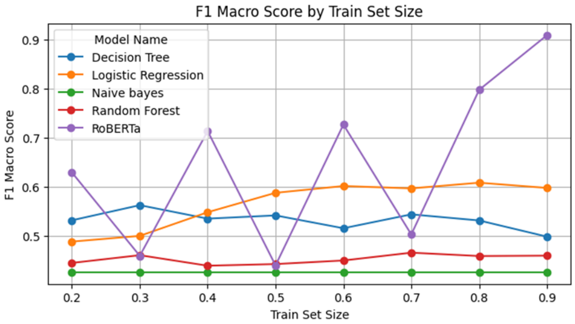

Figure 5: F_1-score at different training set sizes for different classifiers.

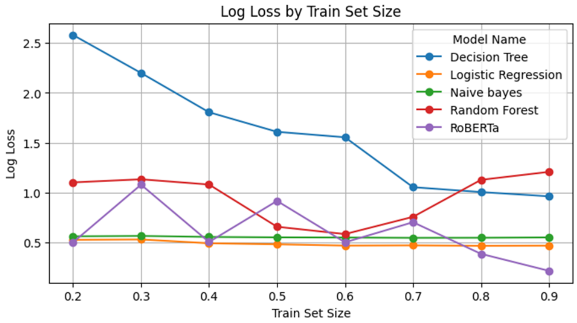

Figure 6: Log loss at different training set sizes for different classifiers.

example, Tweet 2 (A.6) combines personal success with a history of overcoming cancer, and Tweet 3 (A.6) talks about a writing project, neither of which may align with the classifier's training on more direct reports of BC. In case of false positives, we found that the model may be overly sensitive to certain keywords typically associated with personal experiences of BC, such as "battle with BC," "treatment," "therapy," and mentions of specific treatments like "Trastuzumab" or "radio therapy." The model also often misinterprets the sharing of news articles (Tweets in A.5) as personal reports of BC. Lastly, the model sometimes misclassifies supportive messages and discussions around BC as self reports. These findings show areas of possible improvement for the cohort creation step.

3.4.2. Classification Performance at Different Training Data Sizes

To assess if the cohort creation process (i.e., self-report detection for BC) can be improved in the future by increasing the training set size, we conducted training set size *vs.* performance experiments. As depicted in Fig. 5 and 6, RoBERTa shows a continuously increasing trend in F_1 score and decreasing log loss as training size increases. This suggests that further improvement in the cohort creation process is possible by annotating more data, which can enable us to more accurately detect the BC cohort in the future. Though small, the logistic regression model also shows gradual improvement with an increase in training size.

3.5. Limitations and Future Directions

Ideally, for comprehensive longitudinal sentiment analysis, a sufficient number of posts for each unique user ID is required. However, due to the scarcity of posts per user in our dataset, we performed a consolidated sentiment analysis. Additionally, in topic modeling, we identified discussions of other health-related experiences (such as COVID-19) with BC scenarios. This overlap affected classifier's precision in isolating BC-specific conversations from other health concerns. These issues could be potentially mitigated by conducting experiments on a larger dataset containing posts from multiple social media sources rather than one. In the future, it will be interesting to study how current framework performs similar tasks while using data from various other social media platforms such as Reddit and Facebook, and from other time periods. It is also important to collect long-term data from this cohort to potentially discover temporal trends.

4. Conclusion

We investigated the potential use of Twitter (X) dataset for unveiling concerns among a cohort of BC patients. Specifically, we focused on (1) utilizing ML and transformer-based models for training an automatic classifier, (2) identifying BC-related concerns, and (3) performing sentiment analysis on self-reported posts and identifying potential reasons for treatment discontinuation of patients. Our experimental results highlight RoBERTa as the best-performing model for cohort identification. Our topic modeling framework of BC patient discussions reveals that concerns extend beyond traditional PCOs based solely on treatment side effects. Patients express a range of issues, including anxiety related to diagnostic mammogram procedures, barriers to screening due to COVID-19 pandemic, challenges with insurance coverage, and significant socio-emotional distress, such as racial discrimination and concerns about sexual abuse. Our findings reveal broader systemic, medical, and social challenges that need to be addressed for targeted public health messaging and inclusive community support aimed at alleviating fears and ensuring equitable access to healthcare resources.

5. Data and Code Availability

Code is openly available here: github.com/swati-rajwal/BreastCancer_tweets_project.

6. Bibliographical References

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, Jie Lin, Sabrina Li, Angel Xie, Whitney Hogg-Bremer, Mylin Torres, Imon Banerjee, and Abeed Sarker. 2020. Automatic breast cancer cohort detection from social media for studying factors affecting patient-centered outcomes. In *Artificial Intelligence in Medicine*, pages 100–110, Cham. Springer International Publishing.

Silvio Amir, Mark Dredze, and John W. Ayers. 2019. Mental health surveillance over social media with digital cohorts.

SVSolai Ananth, Chandu Pmss, PG Scholar, and Assistant Professor. 2017. Live twitter knowledge as a corpus for sentiment analysis and opinion mining. *Global Journal of Pure and Applied Mathematics*, 13.

Nicky Antonius and L. Rich. 2013. Discovering collection and analysis techniques for social media to improve public safety. *The International Technology Management Review*, 3.

Deanna J. Attai, Michael S. Cowher, Mohammed Al-Hamadani, Jody M. Schoger, Alicia C. Staley, and Jeffrey Landercasper. 2015. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey.

Duilio Balsamo, Paolo Bajardi, Gianmarco De Francisci Morales, Corrado Monti, and Rossano Schifanella. 2023. The pursuit of peer support for opioid use recovery on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 17.

CDC. Cancer and COVID-19 | CDC — cdc.gov. `https://www.cdc.gov/cancer/dcpc/about/covid-19.htm`. [Accessed 03-11-2024].

Centers for Disease Control and Prevention. 2021. Uscs data visualizations.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Wayne Kepner, Meredith C. Meacham, and Alicia L. Nobles. 2022. Types and sources of stigma on opioid use treatment and recovery communities on reddit. *Substance Use and Misuse*, 57.

Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024. Overview of the 8th Social Media Mining for Health Applications (SMM4H) shared tasks at the AMIA 2023 Annual Symposium. *Journal of the American Medical Informatics Association*, page ocae010.

Minjoung Monica Koo, Christian von Wagner, Gary A. Abel, Sean McPhail, Greg P. Rubin, and Georgios Lyratzopoulos. 2017. Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. *Cancer Epidemiology*, 48:140–146.

Melissa J. Krauss, Shaina J. Sowles, Megan Moreno, Kidist Zewdie, Richard A. Grucza, Laura J. Bierut, and Patricia A. Cavazos-Rehg. 2015. Hookah-related twitter chatter: A content analysis. *Preventing Chronic Disease*, 12.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine learning and natural language processing in mental health: Systematic review. *J Med Internet Res*, 23(5):e15708.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Azadeh Nikfarjam, Julia D Ransohoff, Alison Callahan, Erik Jones, Brian Loew, Bernice Y Kwong, Kavita Y Sarin, and Nigam H Shah. 2019. Early detection of adverse drug reactions in social health networks: A natural language processing pipeline for signal detection. *JMIR Public Health Surveill*, 5(2):e11264.

Mike Donald Tapi Nzali, Sandra Bringay, Christian Lavergne, Caroline Mollevi, and Thomas Opitz. 2017. What patients can tell us: topic analysis for social media on breast cancer. *JMIR medical informatics*, 5(3):e7779.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.

Abeed Sarker, Pramod Chandrashekar, Arjun Magge, Haitao Cai, Ari Klein, and Graciela Gonzalez. 2017. Discovering cohorts of pregnant women from social media for safety surveillance and analysis.

Graciela Gonzalez-Hernandez; Karen O'Connor; Guergana Savova; Abeed Sarker. 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of Medical Informatics*, 26(01):214–227. Publisher: Georg Thieme Verlag KG.

Davy Weissenbacher, Suyu Ge, Ari Klein, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. 2021. Active neural networks to detect mentions of changes to medication treatment in social media. *Journal of the American Medical Informatics Association*, 28(12):2551–2561.

Yuan-Chi Yang, Mohammed Ali Al-Garadi, Jennifer S. Love, Hannah L. F. Cooper, Jeanmarie Perrone, and Abeed Sarker. 2023. Can accurate demographic information about people who use prescription medications nonmedically be derived from twitter? *Proceedings of the National Academy of Sciences*, 120(8):e2207391120.

A. Appendices

A.1. GPT-4 Generated Seed Words

"uncertain", "doubtful", "ambiguous", "hesitant", "perplexed", "confused", "questionable", "unsure", "indecisive", "bewildered", "vague", "puzzled","skeptical", "inconclusive", "wavering", "distrustful", "baffled", "ambivalent", "hazy", "tentative"

A.2. Final List of Fear-related Keywords

"scared", "nervous", "worried", "suspicion", "uncertain", "reluctant", "confused", "doubtful", "unsure", "skeptical", "awkward", "insurance", "soy", "covid", "black", "bra", "concerned"

A.3. List of Treatment Keywords

"treatment", "medication", "medicine", "med", "tablets", "side effect", "reaction", "drug", "tamoxifen", "chemo", "mental", "emotion", "lump", "pain", "sleep", "docetaxel", "oncologist", "doc", "stop"

A.4. Hyperparameter search space

For the *Naïve Bayes classifier*, 'alpha' range = [0.01, 0.1, 1, 10, 100]. For *Decision Tree model*, we tested the 'max_depth' = [None, 3, 5, 10, 15, 20]. For the *Random Forest model*, 'n_estimators' =[10, 50, 100] and 'max_depth'= [None, 5, 10] were tested. For the *Logistic Regression model*, the 'C' (inverse of regularization strength) range = [0.01, 0.1, 1, 10], and the penalties = ['l1', 'l2'].

A.5. False positive Examples

Tweet 1: "Cam's Corner her Battle with Stage 3 Breast Cancer"
Tweet 2: "Daily Mirror - Breast Cancer: Difficult Treatment Decisions.
Tweet 3: "Effect at One Year of Adjuvant Trastuzumab for HER2+ Breast Cancer Combined with Radiation or an Anthracycline on Left Ventricular Ejection Fraction"
Tweet 4: "@user No. Breast cancer reconstruction."
Tweet 5: "@user Breast cancer: hope the best of the best, and hope you can feel better"
Tweet 6: "<number> and held years of breast cancer for my mother in france . several surgeries, chemo and radio therapy. <repeat> cost her..."
Tweet 7: "they pulled out of utah now <user> is back. tonight on <user> at <number>– why a breast cancer survivor feels let down and why the charity says they never left. pic.twitter.com / <number>lkl<number>gj"

A.6. False Negatives Examples

Tweet 1: "I just found a rather large, hurtful lump in my breast. As a woman, I dont care if its nothing at all, the thought still terrifies me. Breast cancer is prominent in my family. Iam only 24 yo"
Tweet 2: "I'd be honored to tell you amazing women more about my story of being the first black woman to own a tequila brand and overcoming pancreatic and breast cancer."
Tweet 3: "If you're wondering what I've been doing writing-wise lately: I've been creating a breast cancer journal."
Tweet 4: "I should NOT have to tell people, wear a mask because I have metastatic breast cancer. Give a damn about people like me who have immune issues. #resiliencechat"
Tweet 5: "@user I have medical tattoos, 2 tiny blues when i had radiation for breast cancer."
Tweet 6:"b<user> selfish stockpiling - i order shopping every week. order weds for delivery fri or sat. yesterday no slots available. we are pensioners, i am just finishing breast cancer treatment so not easy to get out for "big" shop..."

Topic	**Tweet Examples**
Covid-19 affected Breast Cancer Patients	“my **fear of being locked up again** is one of two contributors to my refusal to seek medical help for possible **breast cancer**.”
	“**pandemic leads women to delay mammograms**; experts fear future rise in breast cancer deaths <url> via <user>
	“<user> always had a normal **mammogram**. this year she was **scared to go in due to covid**. she pushed it back a few months. when doctors found stage <number> breast cancer.”
Anxiety associated with Mammogram Screening	“thousands of women under <age> with breast cancer showing up on their **mammograms**. i think another **huge issue is lumps being fobbed off as hormonal changes** - natasha was only referred because she put her foot down”
	“i had this conversation with my mom in the summer when she was **scared to get her mammogram**. she went and **she did have breast cancer** that fortunately was caught early.”
	“same with **mammograms**. I am **so scared to get them** and i have <allcaps> to because my **family has a history of breast cancer**”
	“also, this isn’t doctors, it’s the insurance, but i’m pissed that **i can’t get a mammogram** when my granny literally has breast cancer. they’ll do an ultrasound **but not a mammogram**. the f**k is that?”
Concerns regarding Insurance Policies	“the **last thing a breastcancer patient needs** to worry about is **if their insurance will cover treatment**. unfortunately, this is the reality for most facing this disease. help us eliminate barriers to care”
	“I am <number> just found a lump in my breast, **got a referral for a mammogram** from my pcp, got a **letter from insurance** promptly <allcaps> **telling me ,not all of this will be covered** by us just fyi, now i have to worry… & maybe have to pay out of pocket to find out”
	“so i,am mad because **my insurance doesn’t cover** a well women, **No exam nor a mammogram** till i,am <number> i mean **breast cancer can affect women of all ages** you don’t think late teens deserve to check on their health.”
	“im **supposed to have mammograms on the reg** bc of family history but **insurance barely or just doesnt pay for them**. im lucky to have insurance at all but I m terrified my broke ass is gonna get breast cancer; not be able to do shit about it”
Confusions regarding dietary habits	“<user> <user> i’d like to switch to **soy milk**. i’m worried that it’s been **linked to breast cancer** due to the **presence of estrogen**”
	“proponents say that **soy can help prevent** heart disease, **breast cancer**, and more. detractors worry that **soy might interfere with thyroid function** and block nutrient absorption. who’s right?”
Racial Disparity Concerns	“i visited the cdc website to learn about cancer disparities between black and white american women, and i was extremely shocked to learn about the development of **breast cancer** so present in **young black women** when it compares to **young white women**.”
	“**black women** have a <number>% **higher death rate** from **breast cancer**. we shouldn’t be surprised that communities with unequal health outcomes have unequal trust in vaccines.”
	“black women, specifically **young black women**, are more **susceptible to breast cancer**. Don’t be afraid to self-exam often!”

Table A.1: Table showing sample tweets with different topics of concerns/uncertainty amongst BC patients