International Journal of Community Medicine and Public Health
Sinharoy S. Int J Community Med Public Health. 2024 May;11(5):1783-1790
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040
DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20241164

Original Research Article

Predicting financial toxicity in cancer patients using machine learning: a Twitter or X-based approach

Sanaya Sinharoy[1,2]*

[1]Central Bucks HS-East, Doylestown, Pennsylvania, United State of America
[2]Pafin Health LLC, Doylestown, PA18902, USA

Received: 29 February 2024
Revised: 05 April 2024
Accepted: 06 April 2024

***Correspondence:**
Sanaya Sinharoy,
E-mail: sanayasm97@gmail.com

ABSTRACT

Background: Financial strain resulting from cancer treatment correlates with reduced quality of life, treatment nonadherence, bankruptcy, and maladaptive behaviours. This study aims to explore the potential of a supervised machine learning algorithm in predicting financial toxicity in cancer patients based on their Tweets.
Methods: A dataset of Tweets related to cancer and financial toxicity was constructed using Twitter's API. The dataset was curated, and synthetic Tweets were generated to augment the final dataset. A supervised machine learning algorithm, specifically Multinomial Naïve Bayes, was trained and tested to predict financial toxicity in cancer patients.
Results: The model demonstrated high accuracy (0.97), precision (0.95), recall (0.99), specificity (0.96), F-1 score (0.97) and area-under-the-receiver-operating-characteristics (0.98) in predicting financial toxicity from Tweets. Wordcloud visualizations illustrated distinct linguistic patterns between Tweets related to financial toxicity and those unrelated to financial toxicity. The study also outlined potential proactive strategies for leveraging social media platforms like Twitter to identify and support cancer patients experiencing financial toxicity.
Conclusions: This study marks the first attempt to construct a dataset of Tweets related to financial toxicity in cancer patients and to evaluate a predictive model trained on this dataset. The findings highlight the predictive capabilities of the model and its potential utility in guiding health systems and cancer center financial navigators to alleviate economic burdens associated with cancer treatment.

Keywords: Cancer, Financial toxicity, Machine learning, Naïve bayes, Synthetic data, Twitter

INTRODUCTION

Cancer treatment stands out as one of the most expensive healthcare burdens in India, with individuals facing a six-fold higher risk of impoverishment from the substantial out-of-pocket costs compared to those associated with infectious diseases.[1] In the United States, average cost of medical care and drugs surpass $42,000 in the year following a cancer diagnosis.[2] To compound the financial strain, more than 80% of cancer patients exit the workforce during their initial treatment, leading to significant economic challenges.[3] Consequently, more than 40% of patients deplete their entire life savings within the first two years of treatment.[3] Additionally, approximately 30% of Americans with a history of cancer report encountering difficulties paying their medical bills, resorting to borrowing money, or even filing for bankruptcy protection due to their cancer-related expenses.[4] The term financial toxicity refers to the combination of high costs of cancer treatment and detrimental consequences such as diminished health-related quality of life, treatment nonadherence, employment disruption, bankruptcy, premature mortality,

and maladaptive behaviours.[5-9] This burden is pervasive, with 48%-73% of cancer survivors reporting experiencing some deleterious effects of financial toxicity.[10]

Considering the range of emotions endured by cancer patients such as fear, anxiety, anger, depression, despair, and helplessness, it is understandable that they utilize the connectivity of social media platforms for health-related discussions and to express personal sentiments.[11] Cancer patients are notably active on Twitter and their tweets predominantly focus on emotional exchanges, including greetings, treatment discussions, and expressions of emotions.[12-14] Unlike other social media platforms such as Facebook, Instagram, or Snapchat, Twitter exhibits a more relaxed attitude towards the sharing of negative emotions, possibly due to the platform's anonymity in interactions.[15] This characteristic may be particularly beneficial for cancer patients, providing them with a comfortable outlet to express negative thoughts and feelings.[16]

While the public disclosure of official medical records is often met with reluctance by most individuals, many patients display a readiness to share health-related content on social media platforms.[17] Additionally, these patients exhibit a willingness to authorize the integration of their social interactions with personal electronic medical record (EMR) data.[18] This growing acceptance of social media as a platform for health-related networking reflects patients' inclination towards engaging through this medium.[19] Earlier investigations have established the foundation for analysing sentiment in social media within the realm of healthcare research.[20-23] Consequently, this study investigates a novel approach for detecting financial strain in cancer patients by examining social media content, such as Tweets. This method holds promise for providing customized financial assistance to vulnerable population cohorts who express their financial challenges related to cancer treatment through Tweets, yet may be reluctant to communicate these issues with their healthcare providers. However, not unlike other emotional expressions on social media, indication of financial toxicity in Tweets may be subtle, making them not readily apparent to human readers. These subtle cues may be reflected in the nuances of the Twitter user's language and tone, which can be discerned by machine learning (ML) algorithms. ML, a branch of artificial intelligence, utilizes automated mathematical model creation to iteratively learn from input data, enabling the identification and prediction of future states.[24-27] By comparing Tweets of cancer patients with and without financial toxicity during the training process, ML algorithms can effectively identify financial toxicity.

The primary objective of this study was to investigate the potential of a supervised ML algorithm (Multinomial Naïve Bayes) to predict financial toxicity in cancer patients based on their Tweets. These predictions can guide financial navigators in health systems and cancer centres to mitigate financial toxicity during cancer treatment or facilitate tailored financial assistance for affected patients.

METHODS

This study was conducted between November 2023 and February 2024. A visual representation of the methodology employed in this study is depicted (Figure 1).

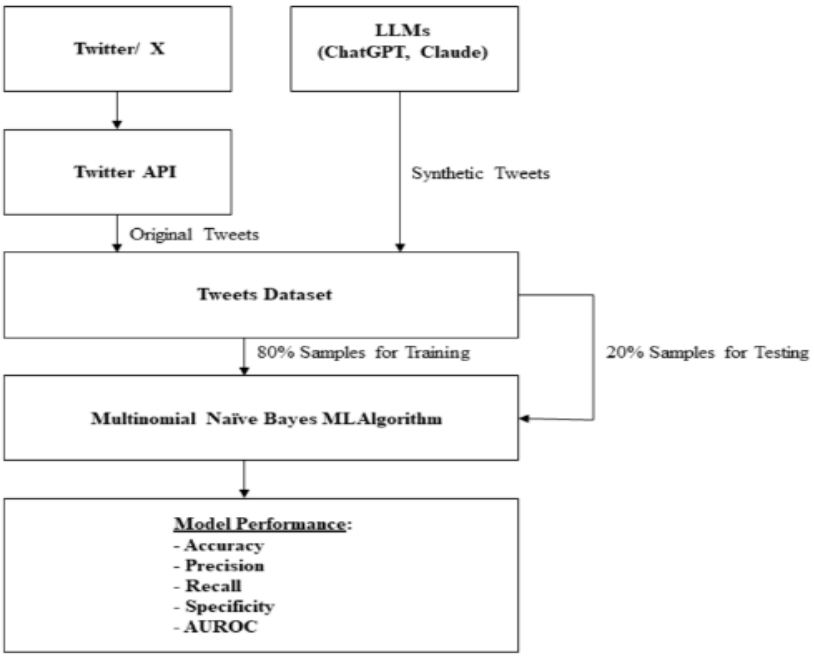

Figure 1: Methodology employed in this study.

Tools

Python libraries including pandas, scikit-learn, statsmodel, matplotlib, wordcloud, and seaborn were utilized to build the ML classification model. ChatGPT 3.5, OpenAI's large language model (LLM) and Claude, Anthropic's LLM were used in the generation of synthetic Tweets.

Creation of the Tweets dataset

The initial data collection process involved utilizing the Twitter application programming interface (API), necessitating the establishment of a Twitter developer subscription account to extract tweets. Access token, access token secret, API key, and API secret key were generated through the developer account to facilitate the authorization process for collecting Tweets. A combination of keywords namely cancer, debt, and help were employed to retrieve Tweets concerning both cancer and financial toxicity (henceforth referred as "financial toxic" Tweets). Conversely, a combination of keywords namely cancer, and treatment were utilized to extract Tweets related to cancer but not associated with financial toxicity (henceforth referred as "non-financial toxic" Tweets).

The author meticulously assessed each Tweet and categorized them as either exhibiting financial toxicity or not. Tweets that did not fall into either category, along with repeat Tweets were excluded. Each Tweet was sanitized to eliminate hyperlinks, retweets, emojis, and hashtags, ensuring that Tweets were appropriate inputs for the ML classification model. Following the sanitization process, 20 financial toxic Tweets and 57 non-financial toxic Tweets related to cancer were selected (henceforth, collectively referred to as "Original Tweets").

The Tweet collection process proved to be slow due to the request rate limits imposed by Twitter. Throughout the collection process, the occurrence of Tweets indicative of financial toxicity was relatively low compared to the abundance of those not related to financial toxicity. Access to historical Tweets was constrained by the limitations of the Twitter developer subscription utilized in this study, which restricted sampling to Tweets published within the previous 7 days. Although the data collection process could have been prolonged for several additional months, a prolongation was deemed suboptimal for this limited budget study due to its potential economic, resource and time implications. Hence, it was decided to augment and balance the Original Tweets data using synthetic Tweets. The background and methodology for generating synthetic data is elaborated below.

From the process of data annotation, to dataset generation, synthetic data offers unprecedented flexibility in the models trained. LLMs have been employed in previous research to annotate data directly in zero-shot scenarios.[28-30] In scenarios with limited resources, low volume and imbalanced datasets, supplementing datasets with synthetic data has demonstrated the potential to enhance model performance across various natural language processing (NLP) tasks, including sarcasm detection, translation and sentiment analysis; for a comprehensive overview, refer to Feng et al.[31-34] Recent research has expanded beyond data augmentation to the creation of entirely synthetic datasets.[35]

In this study, synthetic Tweets representing both financial toxic and non-financial toxic categories to augment the dataset were created by providing specific prompts to ChatGPT 3.5 and Claude. The prompts utilized for generating synthetic Tweets are outlined (Table 1). The prompt used to create synthetic financial-toxic Tweets involved adapting the elements of the cancer-specific COmprehensive Score for financial Toxicity (COST) questionnaire, but modifying the elements to mimic a cancer patient experiencing financial hardship.[36,37] Furthermore, this prompt leveraged the prevalent financial difficulty themes faced by individuals dealing with cancer.[38]

After generation of synthetic tweets, a simple program was employed to examine the text of all synthetic Tweets and filter out any repetitive ones. The author meticulously reviewed each Tweet to confirm that it exhibited financial toxicity or not. Furthermore, each synthetic Tweet underwent a sanitization process to remove index numbers, and hashtags if present, ensuring suitability for input into the ML classification model.

A total of 929 Tweets comprising of synthetic Tweets and Original Tweets were aggregated into a single .csv file. The final Tweets dataset comprised 403 financial toxic Tweets, designated with a polarity of 1, and 526 non-financial toxic Tweets, assigned a polarity of 0 are shown (Figure 2). A selection of illustrative financial toxic and non-financial toxic Tweets is presented (Table 2).

Table 1: Prompts used to generate synthetic Tweets.

Prompt for financial toxic Tweets	Prompt for non-financial toxic Tweets
Generate [specify number] different tweets as if written by cancer patients. You can base each of the tweets by selecting randomly from one of the below: **1. I know that I do not have enough money in savings, retirement, or assets to cover the costs of my treatment** **2. My out-of-pocket medical expenses are more than I thought they would be**	Generate [specify number] different tweets related to cancer and as if written by cancer patients. Include a good mix of tweets related to

Continued.

Prompt for financial toxic Tweets	Prompt for non-financial toxic Tweets
3. I worry about the financial problems I will have in the future because of my illness or treatment **4. I feel I have no choice about the amount of money I spend on care** **5. I am frustrated that I cannot work or contribute as much as I usually do** **6. I am not satisfied with my current financial situation** **7. I am unable to meet my monthly expenses** **8. I feel financially stressed** **9. am concerned about keeping my job and income, including paid work at home** **10. My cancer or treatment has reduced my satisfaction with my present financial situation** **11. I do not feel in control of my financial situation** **12. My illness has been a financial hardship to my family and me** **In addition, you can leverage the seven themes below to help generate the tweets: (1) the burden of travel, (2) a willingness to pursue treatment despite financial risk, (3) fear of destitution, (4) financial toxicity equaling physical toxicity, (5) changes in food spending, (6) reluctance to confide in the study investigator or financial navigator about financial toxicity, and (7) difficulty navigating financial aid**	diagnosis, treatment, recovery, strides being made in cancer research and other cancer associated topics. Tweets should not be related to financial burden of cancer. Do not repeat tweets that have been previously generated or repeat any tweets within this cluster. Do not number the tweets as 1, 2, 3, etc. Do not include hashtags. Be creative.

Machine learning classification

Term frequency-inverse document frequency technique

This study leverages the Term Frequency-Inverse Document Frequency (TF-IDF) technique in NLP. TF-IDF operates by evaluating the importance of a term within a document ("document" corresponds to an individual Tweet) relative to a larger corpus of documents ("corpus" corresponds to the collection of all the Tweets). It accomplishes this by computing two key components: term frequency (TF), which measures the frequency of a term within a document, and inverse document frequency (IDF), which assesses the rarity of a term across the entire corpus. By combining these metrics, TF-IDF assigns higher weights to terms that are prevalent within a document and infrequent across the corpus, effectively highlighting terms that are both relevant and distinctive. In the realm of sentiment analysis, TF-IDF proves particularly beneficial for binary classification tasks. In this study, TF-IDF is utilized to convert textual Tweet data into numerical feature vectors, that are subsequently used by the Multinomial Naïve Bayes classification algorithm as discussed below in Section 2.3.2

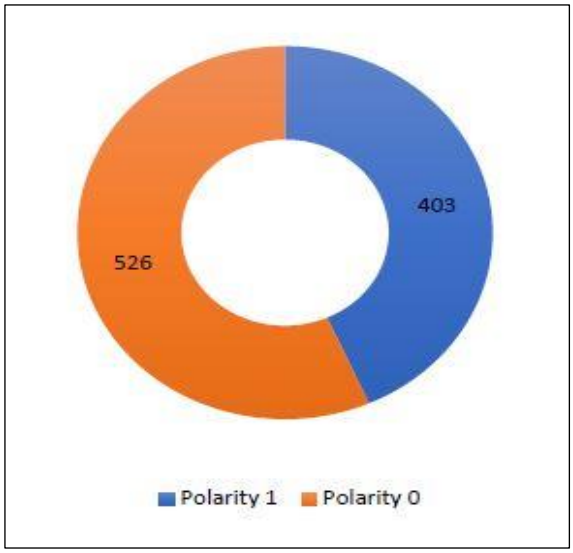

Figure 2: Distribution of (a) financial toxic (polarity 1) Tweets, and (b) non-financial toxic (polarity 0) Tweets in final Tweets dataset.

Table 2: An illustrative selection of financial toxic Tweets and non-financial toxic Tweets.

Financial toxic Tweets	Non-financial toxic Tweets
Cancer has not only taken a toll on my health but also on my finances. It's a double hardship for me and my family.	Cancer may be a part of your journey, but it does not define your destination. Keep moving forward with courage and hope.
Changes in food spending have become necessary adjustments during my treatment. It's disheartening to see how illness affects every aspect of life, even the grocery budget.	Behind every statistic is a story—a life impacted by cancer. That's why we must continue to invest in research and support for all those affected.
The amount of money I spend on care feels out of my control. I'll do whatever it takes to beat cancer, even if it means going into debt.	No texting and driving cancer cells! New med blocks signals to keep tumors from distracting multi-task division!

Continued.

Financial toxic Tweets	Non-financial toxic Tweets
The burden of travel for treatment is overwhelming. Not only do I worry about medical expenses, but also the costs of transportation and accommodation.	Stop copying me, cancer! New drug blocks genetic Xerox machine to keep tumors from spreading their mutations!
Despite the financial strain, I'll continue to pursue treatment. My health is worth any cost, even if it means sacrificing financially.	To all the caregivers supporting loved ones through cancer: your unwavering love and support make all the difference.
I used my entire 401k to pay for cancer treatment last year. Now I have nothing saved for retirement and I'm only 50.	We've got cancer on the run! Researchers sprint ahead developing cutting-edge immunotherapies.

Multinomial Naïve Bayes classification

Multinomial Naïve Bayes is a probabilistic classification algorithm that is widely used in text classification and sentiment analysis. It is based on Bayes' theorem, with a "naive" assumption that features are independent of each other, given the class label. Despite its simplicity and the strong independence assumptions, Naïve Bayes demonstrates computational efficiency and practical effectiveness, particularly with short texts like Tweets.[39-42] Consequently, Multinomial Naïve Bayes was selected for this analysis to predict financial toxicity in Tweets.

Data were randomly divided into training and test samples using an 80:20 ratio. The assessment of model performance employed various metrics, including accuracy, precision, recall, specificity, and F-1 score as outlined (Table 3). In addition, the area-under-the-receiver-operating-characteristics-curve (AUROC) was also determined.

Table 3: Definition of model performance metrics.

Performance Metric	Definition
Precision	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$
Recall (sensitivity)	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}}$
Accuracy	$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Negative} + \text{True Negative} + \text{False Positives}}$
F-1 Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

RESULTS

Wordcloud visualization

Wordcloud, a data visualization tool that represents word frequency through graphical depiction, was employed to visualize the frequent terms in financial toxic tweets and non-financial toxicity tweets (Figure 3 and Figure 4). The Wordcloud associated with financial toxicity displays terms including financial, cost, debt, saving (Figure 3). Conversely, the Wordcloud that illustrates a high frequency of terms typically unrelated to financial toxicity, such as strength, fight, love, cancer-free, resilience is shown (Figure 4).

Model performance

Accuracy, precision, recall, specificity, F-1 score, and AUROC were determined to be 0.97 (95% CI: 0.94-0.99), 0.95 (95% CI: 0.91-0.97), 0.99 (95% CI: 0.96-1.00), 0.96 (95% CI: 0.93-0.98), 0.97 (95% CI: 0.93-0.99) and 0.98 respectively.

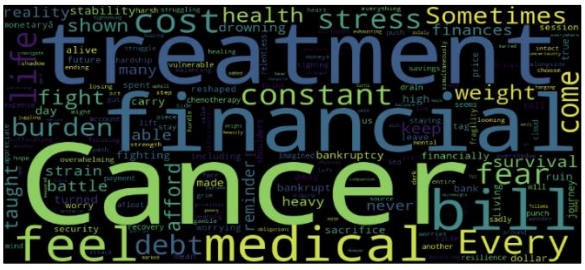

Figure 3: Wordcloud visualization of financial toxic Tweets.

Figure 4: Wordcloud visualization of non-financial toxic Tweets.

DISCUSSION

The findings of this study underscore the ability to accurately predict financial toxicity from Tweets using a Multinomial Naïve Bayes machine learning model. This supervised ML methodology demonstrated discriminatory abilities across various performance metrics. These results suggest potential avenues for leveraging such predictions to support financial navigators within healthcare systems and cancer centres. This utilization could help to alleviate the economic burdens associated with cancer treatment or facilitate tailored financial assistance for affected patients, particularly for those who may express their financial challenges through social media but hesitate to discuss them with healthcare providers. Moreover, the impact of this study could be amplified by patients' willingness to authorize the integration of their social interactions with personal EMR data.

With ASCO emphasizing the inclusion of financial factors in cancer treatment planning[43], health systems and cancer centres can adopt proactive social platform strategies, leveraging Twitter, such as: (i) Engaging in discussions on financial toxicity via hospital/cancer centre Twitter accounts and utilizing algorithms like the one employed in this study to identify potential indications of financial toxicity among cancer patients, or their family members. Financial navigators could then offer guidance on available resources or encourage direct communication for copay assistance and charity programs, (ii) Leveraging algorithms, as used in this study, to predict financial toxicity from patients' Tweets for those who expressly consent to integrate their social interactions with personal EMR data.

Beyond healthcare institutions, the prediction of financial toxicity could be integrated into the Twitter channels of pharmaceutical manufacturers that post help-seeking direct-to-consumer-advertising (DTCA). Predictive algorithms like those employed in this study could flag patients experiencing financial strain, enabling patient assistance program navigators at the manufacturers to provide education on eligibility and availability of support programs while adhering to privacy, FDA, and other governmental regulations. Charitable programs and patient advocacy groups could similarly utilize such algorithms to identify vulnerable patients, and to extend guidance and assistance to such patients. Additionally, large Twitter-based communities such as Breast Cancer Social Media (#BCSM) could discreetly identify and support vulnerable patients by employing similar predictive algorithms.

This study presents several limitations that should be acknowledged. Firstly, due to financial and time constraints, the data collection process via Twitter was limited to a short duration. Future studies with greater resources could address this limitation by extending data collection over a longer period, possibly spanning several months, to substantially increase the volume of Twitter data collected. Additionally, synthetic data could be utilized to balance the dataset as needed. This approach would help mitigate the reliance on a small dataset, which can influence model performance. As larger datasets become available, alternative machine learning classification models and potentially ensemble techniques will be utilized in future efforts. Secondly, the generation of synthetic Tweets was facilitated by ChatGPT 3.5 and the subscription-free version of Claude. To improve the authenticity of synthetic data, future efforts will incorporate advanced language model versions such as ChatGPT-4 and Claude pro. These models can possibly better capture the complexity and variability observed in real-world Tweets, thereby mitigating potential overfitting challenges associated with synthetic data. Thirdly, the prompt settings used for the language models were limited, and only the first output for synthetic Tweets was considered. Exploring alternative prompts may yield improved synthetic Tweets in future research endeavours.

CONCLUSION

In conclusion, this study demonstrates the potential of machine learning algorithms, specifically the Multinomial Naïve Bayes model, to accurately predict financial toxicity in cancer patients based on their Tweets. To the author’s knowledge, this represents the first endeavor to construct a dataset of Tweets related to financial toxicity in cancer patients and subsequently utilize it to effectively evaluate a model trained on this dataset. The findings underscore the discriminatory capabilities of the model, suggesting its utility in supporting financial navigators within healthcare systems and cancer centers to alleviate the economic burdens associated with cancer treatment and facilitate tailored financial assistance for vulnerable patients. The integration of social media data into healthcare decision-making processes presents opportunities for enhancing patient support and engagement, ultimately improving patient outcomes and the quality of cancer care.

ACKNOWLEDGEMENTS

Author would like to thank Judith Elinow of Central Bucks HS-East for her encouragement.

Funding: No funding sources
Conflict of interest: Author is Founder of Pafin Health LLC
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Yadav J, Menon GR, John D. Disease-specific out-of-pocket payments, catastrophic health expenditure and impoverishment effects in India: an analysis of

national health survey data. Appl Heal Econom Health Policy. 2021;19:769-82.
2. Mariotto AB, Enewold L, Zhao J, Zeruto CA, Yabroff KR. Medical care costs associated with cancer survivorship in the United States. Cancer Epidemiol Biomar Prevent. 2020;29(7):1304-12.
3. Gilligan AM, Alberts DS, Roe DJ, Skrepnek GH. Death or debt? National estimates of financial toxicity in persons with newly-diagnosed cancer. Ame J Med. 2018;131(10):1187-99.
4. Banegas MP, Guy GP, de Moor JS, Ekwueme DU, Virgo KS, Kent EE, et al. For working-age cancer survivors, medical debt and bankruptcy create financial hardships. Health Affairs. 2016;35(1):54-61.
5. Zafar SY, Peppercorn JM, Schrag D, Taylor DH, Goetzinger AM, Zhong X, et al. The financial toxicity of cancer treatment: a pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. Oncolog. 2013;18(4):381-90.
6. Bestvina CM, Zullig LL, Rushing C, Chino F, Samsa GP, Altomare I, et al. Patient-oncologist cost communication, financial distress, and medication adherence. J Oncol Pract. 2014;10(3):162-7.
7. Dowling EC, Chawla N, Forsythe LP, de Moor J, McNeel T, Rozjabek HM, et al. Lost productivity and burden of illness in cancer survivors with and without other chronic conditions. Cancer. 2013;119(18):3393-401.
8. Ramsey S, Blough D, Kirchhoff A, Kreizenbeck K, Fedorenko C, Snell K, et al. Washington State cancer patients found to be at greater risk for bankruptcy than people without a cancer diagnosis. Health affairs. 2013;32(6):1143-52.
9. Ramsey SD, Bansal A, Fedorenko CR, Blough DK, Overstreet KA, Shankaran V, et al. Financial insolvency as a risk factor for early mortality among patients with cancer. J Clin Oncol. 2016;34(9):980.
10. Gordon LG, Merollini KM, Lowe A, Chan RJ. A systematic review of financial toxicity among cancer survivors: we can't pay the co-pay. Patient-Patient-Cent Outc Res. 2017;10:295-309.
11. Slevin ML, Nichols SE, Downer SM, Wilson P, Lister TA, Arnott S, et al. Emotional support for cancer patients: what do patients really want?. Brit J Can. 1996;74(8):1275-9.
12. Murthy D, Gross A, Longwell S. Twitter and e-health: A case study of visualizing cancer networks on Twitter. In: Proceedings of Information Society (i-Society). International Conference. London, UK; 2011.
13. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. BMC Res Not. 2012;5:1-9.
14. Myrick JG, Holton AE, Himelboim I, Love B. # Stupidcancer: exploring a typology of social support and the role of emotional expression in a social media community. Health Communi. 2016;31(5):596-605.
15. Vermeulen A, Vandebosch H, Heirman W. # Smiling,# venting, or both? Adolescents' social sharing of emotions on social media. Comput Hum Behav. 2018;84:211-9.
16. Wang J, Wei L. Fear and hope, bitter and sweet: Emotion sharing of cancer community on twitter. Soci Med Soci. 2020;6(1):2056305119897319.
17. Paul M, Dredze M. You are what you tweet: Analyzing twitter for public health. InProceedings of the International AAAI Conference on Web and Social Media 2011;5(1):265-72.
18. Padrez KA, Ungar L, Schwartz HA, Smith RJ, Hill S, Antanavicius T, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. BMJ Qual Safe. 2016;25(6):414-23.
19. Fisher J, Clayton M. Who gives a tweet: assessing patients' interest in the use of social media for health care. World Evidence-Based Nurs. 2012;9(2):100-8.
20. Huesch MD, Currid-Halkett E, Doctor JN. Public hospital quality report awareness: evidence from National and Californian Internet searches and social media mentions, 2012. BMJ open. 2014;4(3):e004417.
21. Wong CA, Sap M, Schwartz A, Town R, Baker T, Ungar L, Merchant RM. Twitter sentiment predicts Affordable Care Act marketplace enrollment. J Med Inter Res. 2015;17(2):e51.
22. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Ame Medi Informat Associat. 2014;21(6):1098-103.
23. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, et al. Tweets about hospital quality: a mixed methods study. BMJ Qual Safe. 2014;23(10):838-46.
24. Sidey-Gibbons C, Pfob A, Asaad M, Boukovalas S, Lin YL, Selber JC, et al. Development of machine learning algorithms for the prediction of financial toxicity in localized breast cancer following surgical treatment. JCO. 2021;5(5):338-47.
25. Liu Y, Chen PH, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. Jama. 2019;322(18):1806-16.
26. Alpaydin E. Introduction to Machine Learning. 4th ed. Cambridge, MA, The MIT Press;2020.
27. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19:1-8.
28. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proceed Nat Acad Sci. 2023;120(30):e2305016120.
29. Josifoski M, Sakota M, Peyrard M, West R. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. arXiv preprint arXiv:2303.04132; 2023.

30. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science?. Computat Linguist. 2024:1-55.
31. Abaskohi A, Rasouli A, Zeraati T, Bahrak B. Utnlp at semeval-2022 task 6: A comparative analysis of sarcasm detection using generative-based and mutation-based data augmentation. arXiv preprint arXiv:2204.08198; 2022.
32. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709; 2015.
33. Maqsud U. Synthetic text generation for sentiment analysis. InProceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 2015; pp. 156-161.
34. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A survey of data augmentation approaches for NLP. arXiv preprint arXiv:2105.03075. 2021 May 7.
35. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic Data Vault. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada. 2016:399-410.
36. De Souza JA, Yap BJ, Wroblewski K, Blinder V, Araújo FS, Hlubocky FJ, et al. Measuring financial toxicity as a clinically relevant patient-reported outcome: the validation of the COmprehensive Score for financial Toxicity (COST). Cancer. 2017;123(3):476-84.
37. De Souza JA, Yap BJ, Hlubocky FJ, Wroblewski K, Ratain MJ, Cella D, et al. The development of a financial toxicity patient-reported outcome in cancer: the COST measure. Cancer. 2014;120(20):3245-53.
38. Williams C, Meyer L, Kawam O, Leventakos K, DeMartino ES. The faces of financial toxicity: a qualitative interview study of financial toxicity in advanced cancer patients in phase i oncology trials. Mayo Cli Proceed Innovat Qual Out. 2023;7(6):524-33.
39. Chakrabarti S, Roy S, Soundalgekar MV. Fast and accurate text classification via multiple linear discriminant projections. VLDB J. 2003;12:170-85.
40. Ting SL, Ip WH, Tsang AH. Is Naive Bayes a good classifier for document classification. Int J Softw Engin Applicat. 2011;5(3):37-46.
41. Gamallo P, Garcia M. Citius: A naive-bayes strategy for sentiment analysis on english tweets. InProceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014). 2014; 171-175.
42. Wang SI, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. InProceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012; 90-94.
43. Gilligan T, Coyle N, Frankel RM, Berry DL, Bohlke K, Epstein RM, et al. Patient-clinician communication: American Society of Clinical Oncology consensus guideline. Obstet Gynecol Surv. 2018;73(2):96-7.

Cite this article as: Sinharoy S Predicting financial toxicity in cancer patients using machine learning: a Twitter/X-based approach. Int J Community Med Public Health 2024;11:1783-90.