RESEARCH

# Cancer Health Disparities drivers with BERTopic modelling and PyCaret Evaluation

Mary Adewunmi[1,2]; Saksham Kumar Sharma[3]; Nistha Sharma[4]; N Sudha Sharma[3]; Bayangmbe Mounmo[5]

[1]University of UTAS, Australia, [2]NACETEM, Nigeria, [3]Maharaja Surajmal Institute of Technology, India, [4]Pune Institute of Computer Technology, India, [5]Kanda Weather Group, USA.

*Corresponding author: Nistha Sharma email: sharmanistha2000@gmail.com

## ABSTRACT

The complex interplay of social, behavioral, lifestyle, environmental, health system, and natural health variables contribute to disparities in cancer treatment across racial and ethnic groups. Consequently, it is necessary to identify the variables contributing to cancer health inequalities and develop strategies to achieve health equality. PubMed abstract on Cancer health disparities was scraped with a bio.Entrez python package. Preprocessed data with regex and Natural tool kit (NLTK), topic modeling with BERTopic embeddings, and c-TF-IDF to construct dense clusters and analyze top topics linked with Cancer health disparities. Model evaluation with PyCaret coherence score and web app deployment with Streamlit. The results showed that Topic 32, with the terms obese, female, male, school, survey, student, poet, and discrepancy, had the best coherence score of 0.3687. In contrast, topic 8, with terms prevalence, adult, income, high, usage, diabetes, education, elderly, change, and low received the lowest coherence score of 0.3255. The model classifies each Subject Word score based on the scores, the granular topic concerns, and trends related to cancer health disparities, investigates the connection between drivers of cancer health disparities, and evaluates the model with their coherence score values.

KEYWORDS: Cancer health disparity, BERTopic, c-TF-IDF, PyCaret.

Citation: Adewunm M et al (2022) Cancer Health Disparities drivers with BERTopic modelling and PyCaret Evaluation. Cancer Health Disparities 6: e1-e12. doi:10.9777/chd.2022.1005

## INTRODUCTION

Cancer disparities occur across geographic regions, socioeconomic classes, and racial and ethnic groupings. For example, rural areas had higher lung, cervical, and colorectal cancer death rates than urban ones, owing to poverty, risky behavior, and lower vaccination and screening rates [1]. This is consistent with the growing divide in life expectancy between rural and urban areas (Torre et al., 2015). Low educational attainment is a marker of socioeconomic deprivation and correlates with increased all-cause mortality in the general population. Fifty per cent (50%) of all premature deaths may be avoided if all sectors of the U.S. population experienced college graduates' death rates. Socioeconomic status is also a significant predictor of cancer death. Around a quarter of cancer deaths may be avoided if all Americans obtain a college education (Withrow et al., 2021) [2]. Furthermore, cancer survival improves with increasing socioeconomic position across all racial and ethnic groups in the United States [3]. Nonetheless, socioeconomic disparities in cancer mortality have shifted dramatically over time [4]. Until the 1980s, the socioeconomic position was positively connected with cancer mortality rates in the United States, indicating that the wealthy face a greater risk of cancer. However, this link has shifted in the opposite direction, with affluent Americans now having a lower risk of dying from cancer, owing to advancements in disease prevention, early cancer detection, and cancer therapy that benefit people with private health insurance. Socioeconomic disparities are the primary cause of excess mortality from lung, colorectal, cervical, stomach, and liver cancers among Americans living in deprived areas [3]. While prostate cancer mortality did not differ significantly by socioeconomic class in the past, an inverse socioeconomic gradient presently exists [5] [6]. Additionally, neighborhood socioeconomic hardship is associated with shorter telomere length, a marker of premature ageing, and deadly malignancy [7].

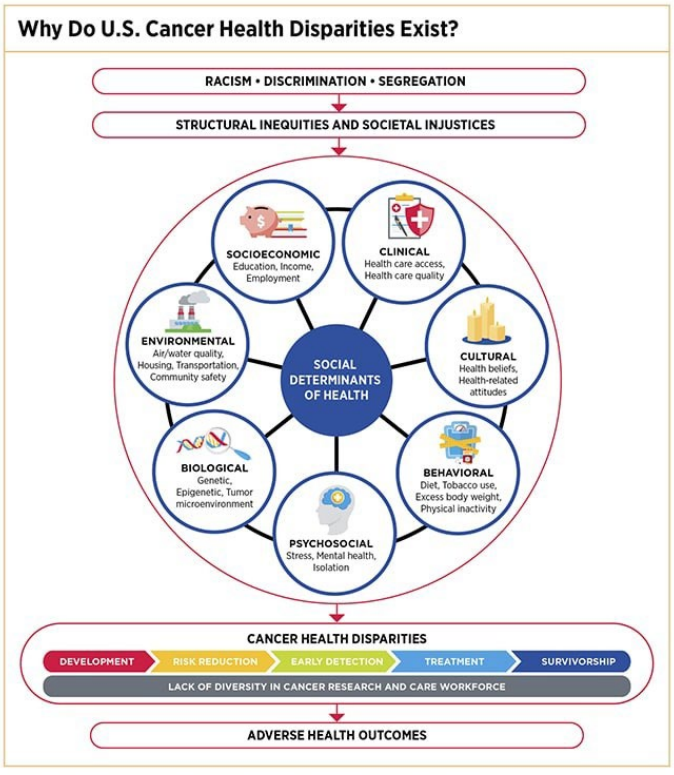
Global disparities in cancer incidence and mortality rates are observed across the board for most cancer sites, indicating socioeconomic disparities and considerable differences in risk factor exposure [8]. Breast, colorectal, and prostate cancer rates differ significantly between high- and low-income countries, geographic regions, and race/ethnic groupings. As migration studies for breast and other cancers have demonstrated, differences in health care and modifiable risk factor exposure are significant drivers of these global disparities [9] [10]. Lung cancer is the most important cause of cancer death globally but is significantly underrepresented in Sub-Saharan Africa due to low smoking rates. Prostate cancer is the most frequent cancer in men worldwide, but its prevalence varies considerably by geography, with low rates in East Asia and high speeds in Western countries. The incidence disparity has lessened as East Asia's habits have become more westernized [11]. Notably, prostate cancer is the leading cause of cancer death among men in Sub-Saharan Africa and the Caribbean [12], leading to the theory that males of Sub-Saharan African ancestry. This may pre- dispose to prostate cancer and a more aggressive illness due to ancestral genetic characteristics. Cervical cancer is the leading cause of cancer death in women in Sub-Saharan Africa and Southeast Asia, owing to human papillomavirus infections and late disease identification [13]. Other cancers with a high incidence and fatality rate in Eastern Asia include stomach and oesophagal cancer (Lin et al., 2021). Helicobacter pylori infection and a diet high in salt are significant risk factors for stomach cancer [14]. This cancer is more prevalent on the Korean peninsula due to regional dietary risk factors and chronic Helicobacter pylori infections [15].

In contrast, Malawi in Eastern Africa is mainly affected by oesophagal cancers [16], with the highest global disease rates due to unknown risk factors. Finally, liver cancer is most prevalent in Northern and Western Africa and Southeast Asia [17]. For example, it is the leading cause of cancer mortality in Mongolia [18]. Chronic hepatitis B and C virus infections and aflatoxin exposure are significant causes of disease in these regions. In contrast, heavy alcohol use and non-alcoholic fatty liver disease are tremendous contributors to the rising incidence of liver cancer in several high-income countries [19]. The remaining sections summarise prior research on topic modelling for language, project workflow, modelling experiment, results, evaluation, discussion and conclusion.

The objectives of this study are:

- Find out granular topics related to cancer health disparities with topic modelling.
- Narrow down the knowledge structure of the high topics word scores and the associated trends.
- Validate the discovered trends.

Figure 1. Why do U.S. Cancer health disparities exist (AACR Cancer Disparities Progress Report 2022)




Adapted from *American Association for Cancer Research® (AACR) Cancer Disparities Progress Report 2022*

## Reviewed Methods

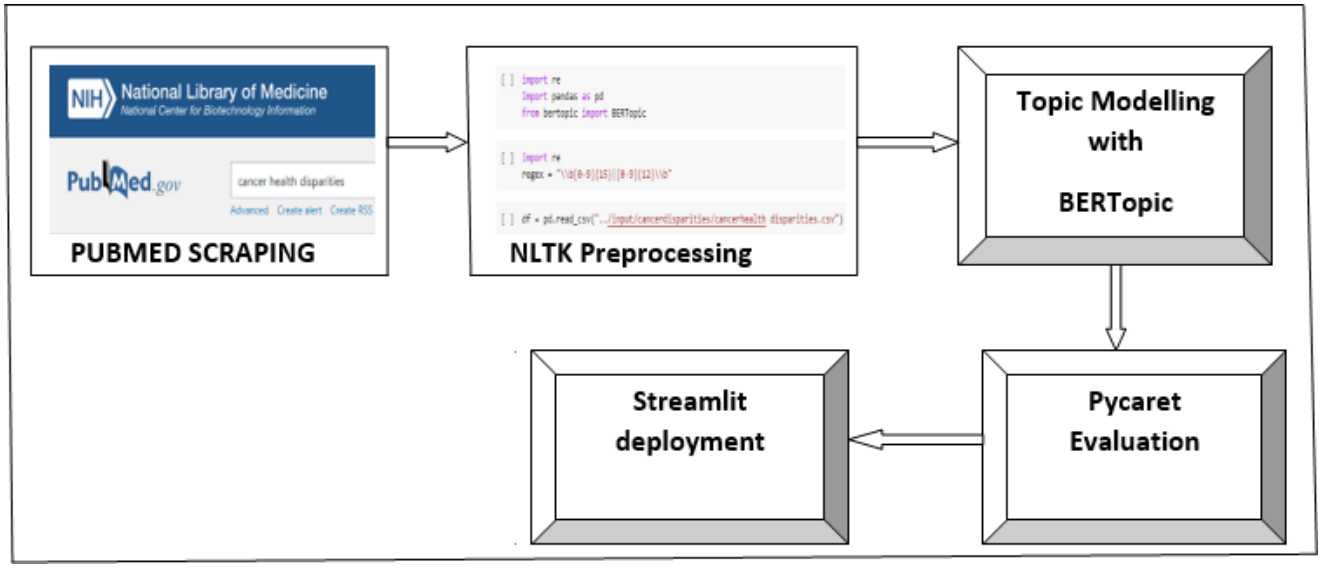
### Topic Modelling with BERTopic and PyCaret

This is an unsupervised machine learning method for discovering abstract subjects in substantial text collections. It aids in organizing, comprehending, and summarising vast quantities of textual material and locating hidden issues that differ across documents within a particular corpus. The objective of topic modelling is to group documents and words with similar meanings. It has several critical applications, including Natural language processing (NLP) and information retrieval (I.R.). It uses unsupervised machine learning algorithms to identify themes inside document sets. The

Probabilistic Latent Semantic Analysis (PLSA) was initially suggested in 1999 [20], followed by the Latent Dirichlet Allocation (LDA) in 2003 [21], which has since become one of the most used topic modelling methodologies. In addition, the development of Pre-trained Language Models (PLMs) has contributed to the subject modelling problem. For instance, BERTopic [22] is a topic modelling technique that employs BERT embeddings and a class-based TF-IDF to create dense clusters.

Additionally, it uses the Uniform Manifold Approximation and Projection (UMAP) technique to reduce the dimensionality of the embeddings before clustering the documents [23]. Initial studies with the BERTopic approach yielded promising results; consequently, this work aims to conduct experiments with the BERTopic technique utilizing various PLMs and compare their results to well-established methods such as LDA. PyCaret [24] is an open-source Python toolkit for low-code machine learning that streamlines ma- chine learning operations. It is an end-to-end machine learning and model management application that exponentially accelerates the trial cycle and increases your productivity. Compared to other open-source machine learning libraries, PyCaret is an alternative low-code library that can replace hundreds of lines of code with only a few lines. This makes experiments significantly quicker and more productive. PyCaret is a wrapper for several machine learning libraries and frameworks, including sci-kit-learn, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray, and a few more.

Figure 2. Project Workflow




## Methods

This article analyzes PubMed abstracts using embedding clustering-based models with LDA [21]. History's typical conventional topic model produces subjects using document-topic and topic-word distributions. In this study, we examine the topic of modelling text corpora and other discrete data sets by extracting the causes of health disparities in cancer care.

### BERTopic Modelling

BERTopic model exploits BERT+UMAP+HDBSCAN [22], a clustering-based method that uses HDBSCAN (McInnes and Healy, 2017) to cluster Sentences BERT embeddings. Uniform Manifold Approximation

Projection (UMAP) [25] and a class- based Term Frequency Inverse Document Frequency (c-TF-IDF) to reduce embedding dimensions. This method identifies themes on cancer health inequalities and narrows down the knowledge structure of high-scoring themes and related trends. The objective is to discover concise descriptions of the cancer health disparity col- lection that facilitate the efficient processing of large groups while preserving the essential statistical relationships beneficial for fundamental tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

# Experiments

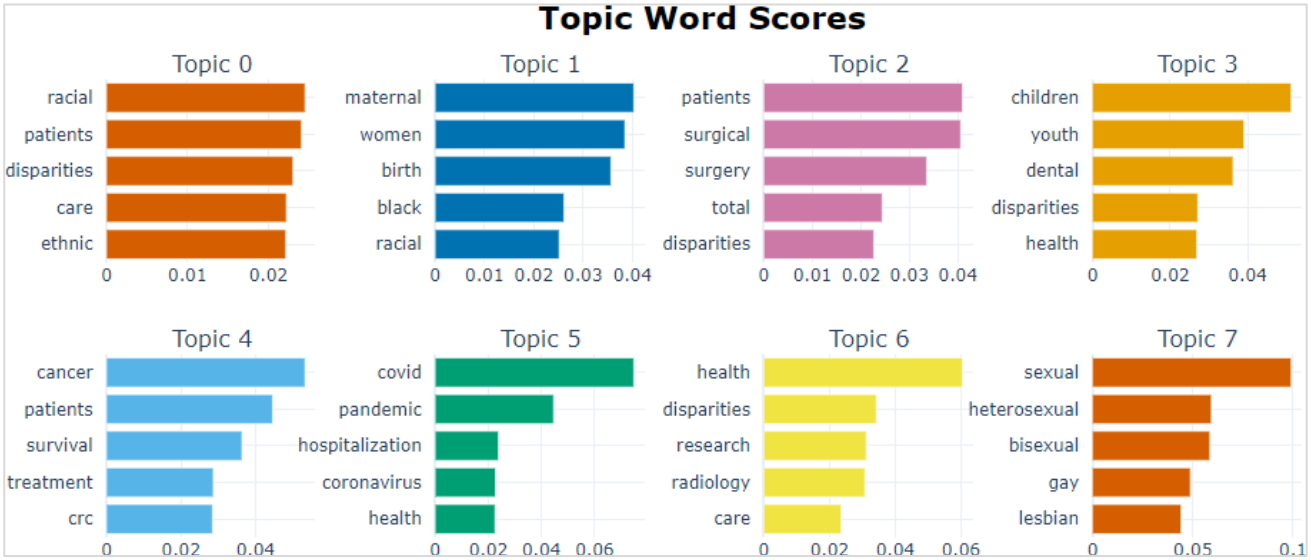
## Datasets

We scraped abstracts on Colon cancer disparities from the PubMed database with Entrez Global Query Cross-Database Search System, using the keyword 'Cancer health disparity' and preprocessed the data with Natural Language Tool Kit (NLTK) us-ing the regex method.

## Evaluation Metrics

We assess the subject's quality in terms of topic diversity and topic coherence: Topic Diversity (T.U.)(Nan et al., 2019) quantifies the originality of words across all subjects. Normalized Point- wise Mutual Information (NPMI) (Newman et al., 2010) measures topic coherence internally using a sliding window to count word co-occurrence patterns. Topic Coherence (CV) (Roder et al., 2015) is a variant of NPMI that uses one-set segmentation to calculate word co-occurrences and cosine similarity as the similarity measure.

Figure 3. Topic Word Scores




Visualizing the chosen keywords for a few themes in Figure 2. The relative c-TF-IDF scores across and within articles provides insight [26]. Additionally, it is simple to compare subject representations to one another. We can view the top words for each topic and the topic word scores. So, in Topic 0, the top term racial care and ethnic. Related topics, for other words, can be analyzed. From the topic similarity scores, the top 10 words in the document are patients, surgical, surgery, pain, disparities, total, white, racial, care, and black. Patients have a word score of 0.049; surgical has a word score of 0.039; surgery is 0.032; disparities are 0.025, the total has a word score of 0.025, white has a word score of 0.023, racial has a word score of 0.021, care has a

word score of 0.021, and black has a word score of 0.0. These words have a similarity score of 0.48.

## Frequent topics

In Figure 4, Topic -1, disparities_health_care_ patients, are the commonly ignored outliers, but this topic seems relevant to the researched theme.

Figure 4. Frequent Topics generated

```
[('maternal', 0.040108673190743266),
 ('women', 0.03795595766510634),
 ('birth', 0.03547814056356476),
 ('black', 0.02562747044322597),
 ('racial', 0.024061832063230797),
 ('pregnancy', 0.02383824507330423),
 ('hispanic', 0.020972620404514464),
 ('prenatal', 0.02096463661662166),
 ('white', 0.02065252835924124),
 ('risk', 0.020214363457865897)]
```

Figure 5. Topics and the probability

| | Topic | Count | Name |
|---|---|---|---|
| 0 | -1 | 245 | -1_disparities_health_care_patients |
| 1 | 0 | 63 | 0_maternal_women_birth_black |
| 2 | 1 | 54 | 1_cancer_patients_survival_treatment |
| 3 | 2 | 53 | 2_patients_surgical_surgery_total |
| 4 | 3 | 50 | 3_children_youth_dental_disparities |

As observed from the hierarchical clustering of the data, the closely clustered racial health disparity is linked to the youth's mental health. Therefore, the impact of racial health disparities on the mental health of individuals is considerable. Furthermore, the healthcare sector and cancer disparity are clustered, suggesting the need for more thorough research in this field. Therefore, increasing awareness and research on cancer disparity will substantially affect the healthcare sector.

Figure 6. Hierarchical Clusters of Topics Most Similar Topics

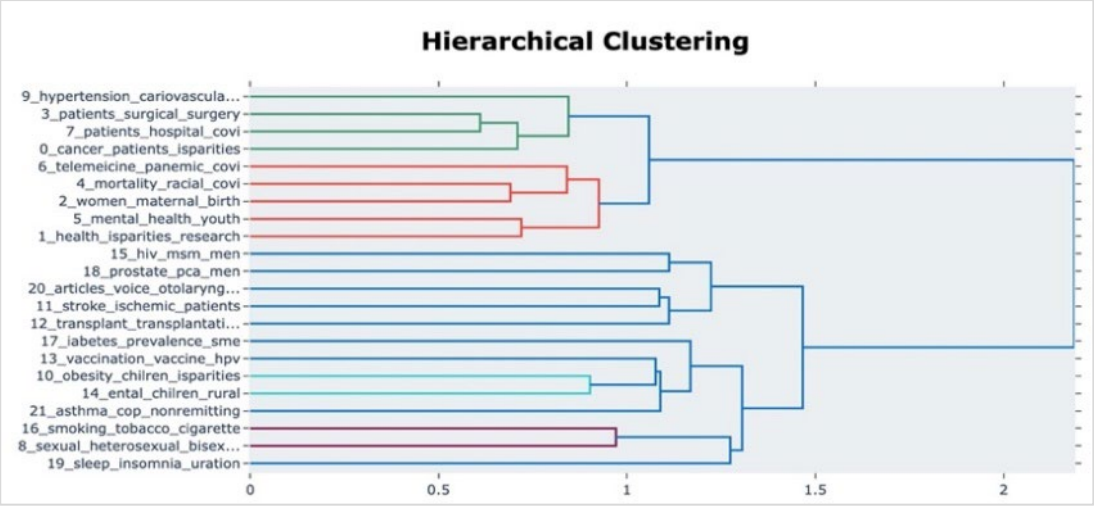



Figure 7. Most Similar topics

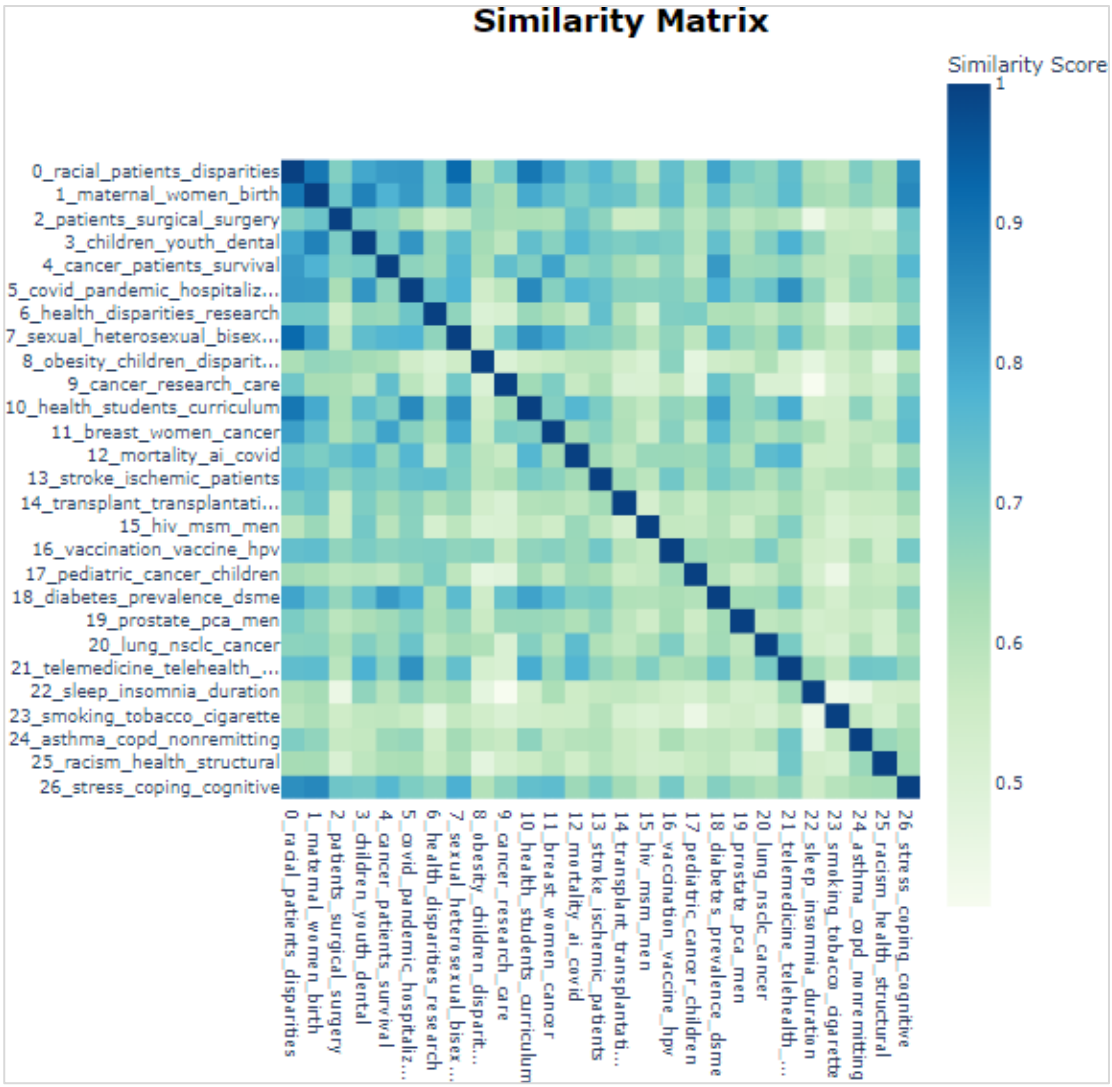
```
Most Similar Topic Info:
[('patients', 0.046773193659976404), ('surgical', 0.044184916820197116), ('surgery', 0.036751940158855544), ('total', 0.02
7718933687618064), ('isparities', 0.02517922291925961), ('care', 0.023341206009093597), ('white', 0.02213933525079471),
('racial', 0.02196061166758708), ('hospital', 0.021762003032270456), ('black', 0.020503228628651327)]
Similarity Score: 0.48779489027342576
```

Fig 7. shows the related topics to cancer health disparity and its similarity score. The highest is cancer with 0.0508, and the least is the incidence, with a similarity score of 0.5603.

## Heatmap Matrix

Build a heatmap of the similarity matrix for the subject. A heatmap depicting the similarity of topics generated based on the cosine similarity matrix between topic embeddings. For example, the similarity score among racial-patients-disparities, maternal-women-birth, patients-surgical-surgery, children-youth- dental, cancer-patients-survival, covid-pandemic-survival, health- disparities-research and sexual-heterosexual-bisexual are very high amongst others as displayed in fig.8.

Figure 8. Similarity Matrix heatmap




## Model Evaluation with PyCaret coherence

Probabilistic topic models like LDA are popular text analysis techniques because they provide both a corpus's predictive and latent topic representation. However, because of the unsupervised training process, there is a longtime expectation that the latent space identified by these models is typically relevant and valuable and that testing such assumptions is difficult. Furthermore, there is no universally accepted list of themes against which all com- pared corpora. However, it is equally vital to determine whether a trained model is excellent terrible and compare different models/methods. Various methods have been employed in many

practical applications to determine if "the right thing" has been learned about the word corpus. The Coherence Score is the evaluation method used in the proposed research to evaluate Topic Models. Topic coherence assesses the semantic similarity between high- scoring terms in a topic's score. These metrics aid in the distinction between semantically interpretable issues and statistical inference artefacts. According to Fig. 9, topic 32 words have the highest coherence score, i.e., 0.3687, and topic 8 words have the lowest, i.e., 0.3255. From Fig. 10 obesity, female, male, school, survey, student, post, disparity, gender, and increase in topic 32.

Figure 9. Topic with the highest coherence value.

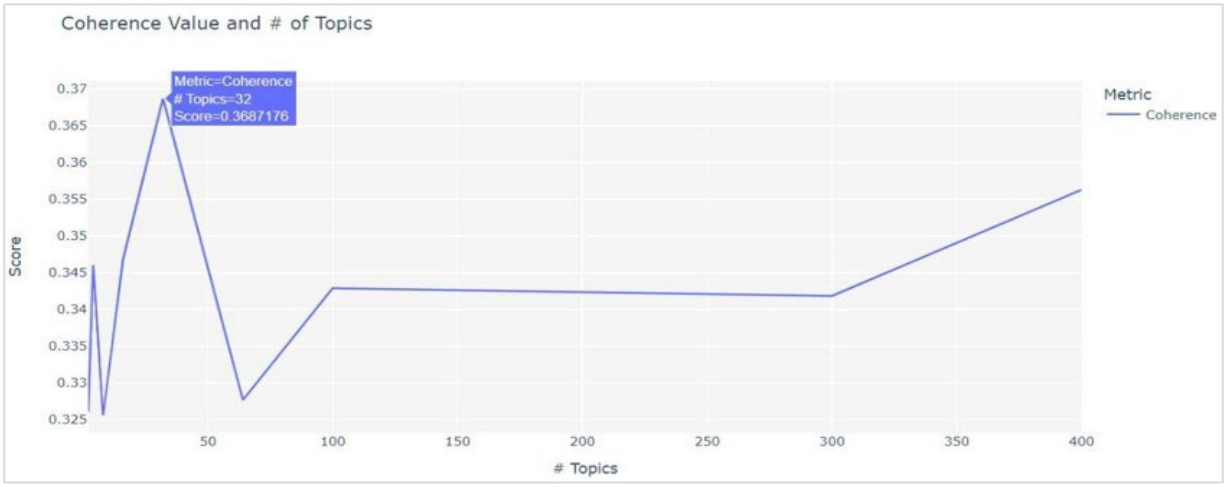
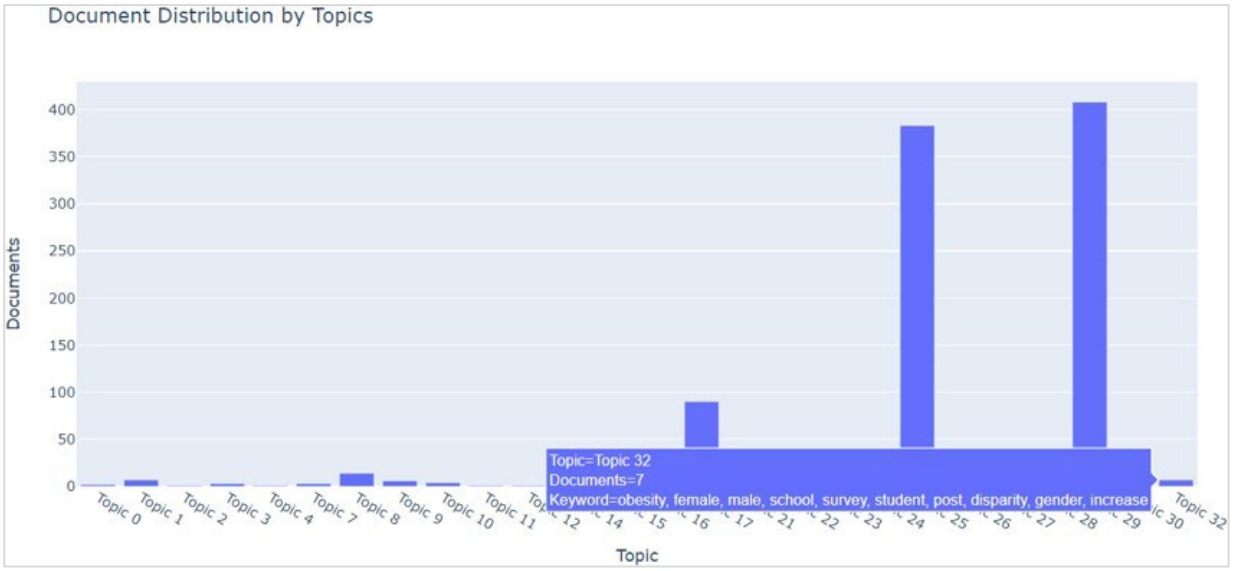



Figure 10. Words in Topic 32




In contrast, the coherence score drops sharply until topic 64, then rises until topic 100. Then, from 100 to 300, the coherence score is nearly steady, with just a minor decline. After topic 300, the coherence score begins to rise again as displayed in fig 11 and 12.

Figure 11. Topics with the lowest coherence value

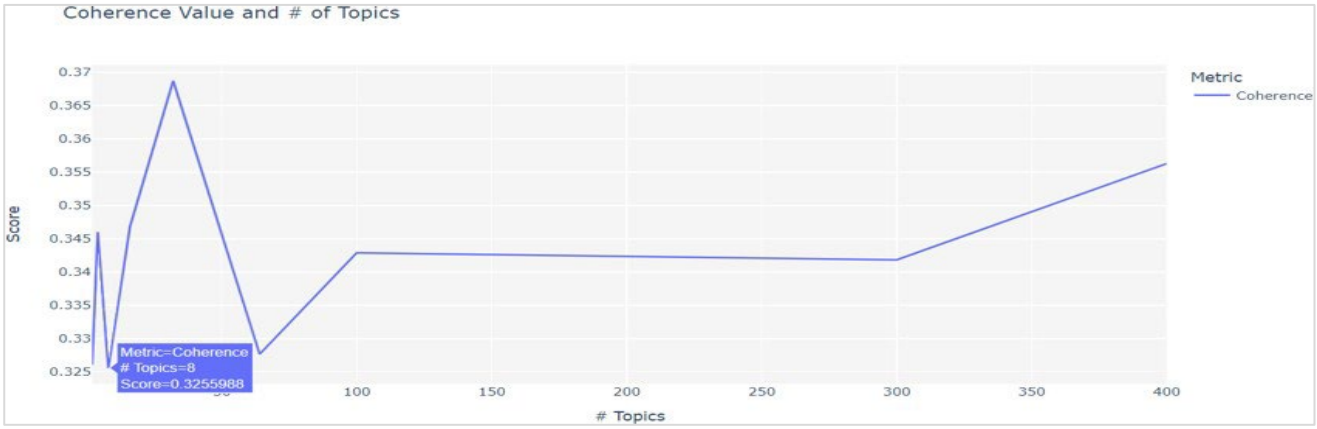
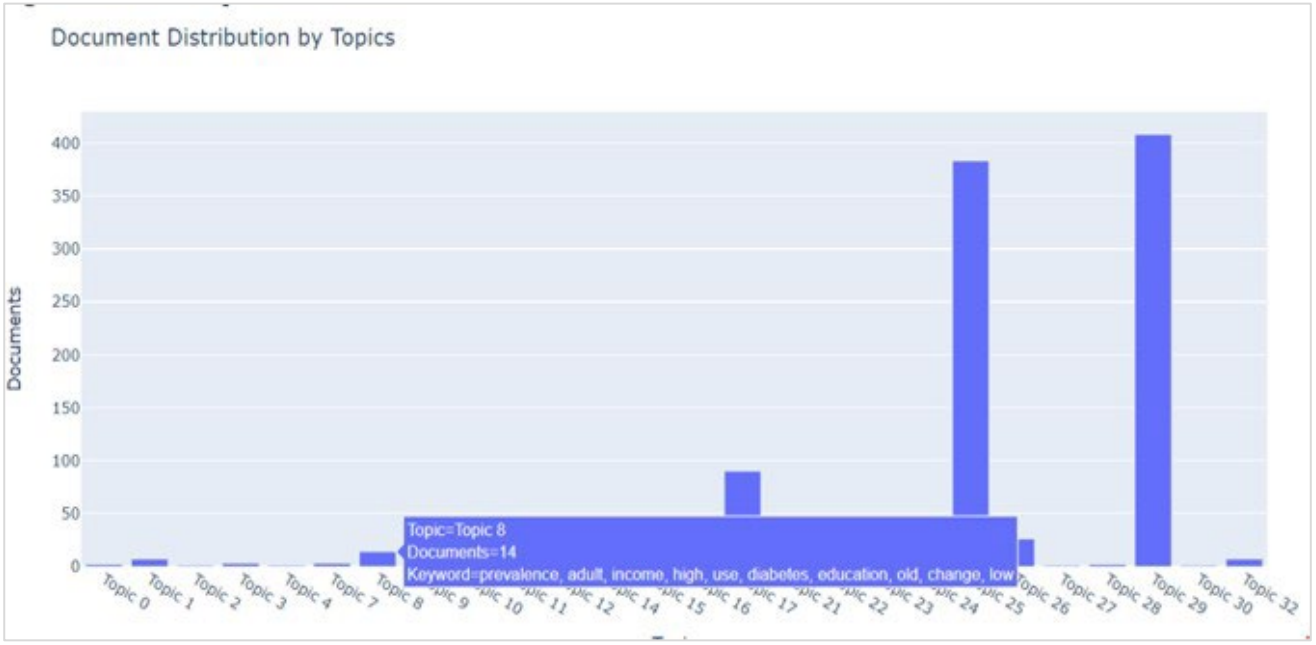


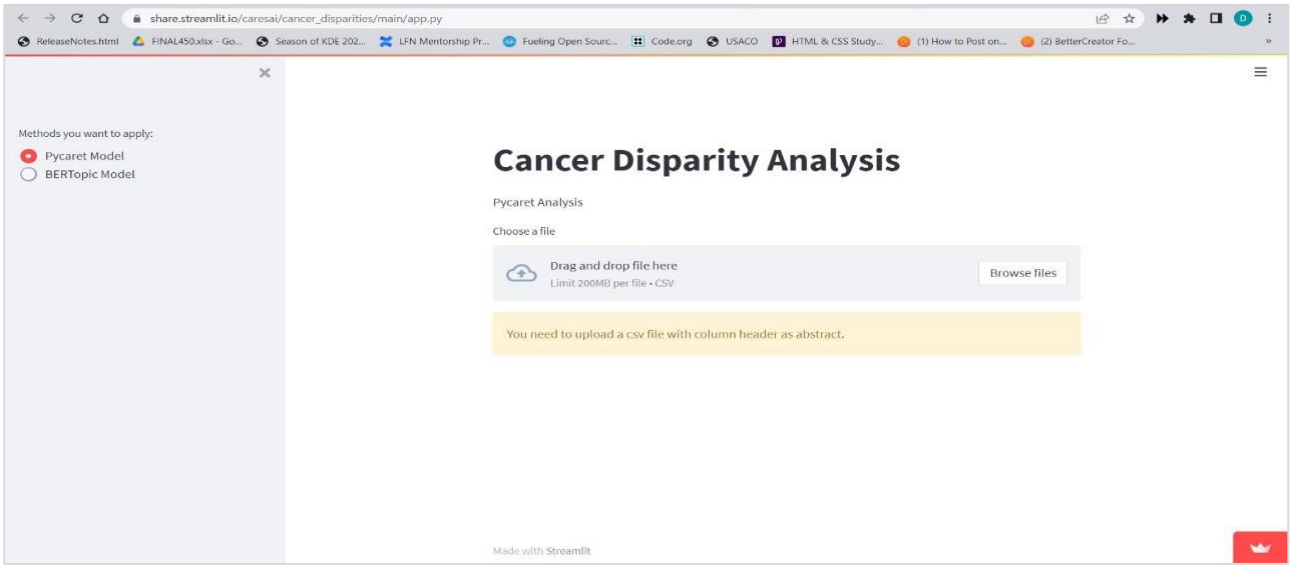

Figure 12. Words in Topic 8




## Streamlit Deployment

Streamlit is used to deploy data visualisations generated by topic modelling using PyCaret and BERTopic. In addition, it is an open- source Python framework for building Machine Learning and Data Science web applications. Streamlit enables the application development by simplifying the interactive coding cycle and displays results in a web application. Figure 10 depicts the web interface of the Streamlit web.

Figure 13. Streamlit display of Cancer disparity web app




## Conclusion

Identifying granular topics contributing to cancer health disparities is necessary and developing measures for achieving health equity. First, textual data was extracted from PubMed abstracts, then preprocessed the data with Natural Language ToolKit (NLTK) libraries, topic modelling using the PyCaret and BERTopic libraries and evaluation with PyCaret. Both models produced outcomes which have been discussed in the previous sections. Following a thorough examination, phrases such as racial, health, care, black, ethnic, white, population, socioeconomic status, sexual and others appeared in most papers on cancer disparities. Interestingly, rare topics like obesity, dental, children, school and discrepancy emerge from the models. As a result, it is safe to assume that these racial and ethnic minority groups, countries' economically disadvantaged classes, and, in general, those with less representation or resources become the targets of cancer disparities. The paper also includes figures from PyCaret and BERTopic, such as hierarchical clustering, similarity matrix, and Topic word scores. These analyses can be valuable in developing additional ways for more standardized cancer treatment, reducing existing disparities among minorities.

## Future Recommendation

To eliminate cancer disparities, government, private, non-profit institutions and individuals need to be actively involved in policies guiding cancer research, prevention, and treatment. Also, other rare terms discussed in the conclusion need further research, even though it is a challenging goal. Furthermore, a thorough examination of the various individuals affected by cancer and how they differ in their treatment compared to the more privileged and minority classes need to be carried out. Most importantly, Artificial Intelligence can be used to detect these subjects and act as a catalyst in many cancer treatment processes starting from a cancer diagnosis.

## Funding

The paper was not funded.

## Declaration Of Interest

All authors of this manuscript are part-time members of a newly founded cancer research AI

group(caresAI). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

## Acknowledgement

Special thanks to OMDENA for creating a platform for us to connect and share similar interest on cancer research using AI pathways.

## Authors' contributions

MA – conceived the idea of using BERTopic for extracting topics on cancer health disparity as a use case, designed the workflow,scraped the abstract data form PubMed, part of implementation phase, manuscript writeups and overall editing.

SKS – Evaluated the model using Pycaret coherence score, deployed it on streamlit webapp, results interpretation on the manuscript.

NS – Part of implementation phase using BERTopic for topic extraction of the use case and result interpretation on the manuscript

NSS - Part of implementation phase using BERTopic for topic extraction of the use case and result interpretation on the manuscript

BM – Part of implementation phase using BERTopic for topic extraction of the use case.

.

## REFERENCES

1. Zahnd WE, James AS, Jenkins WD, Izadi SR, Fogleman AJ, Steward DE, et al. Rural-urban differences in cancer incidence and trends in the United States. Cancer Epidemiology and Prevention Biomarkers. 2018;27:1265–1274.
2. Leive AA, Ruhm CJ. Education gradients in mortality trends by gender and race. Journal of Human Capital. 2022;16(1):47– 72.
3. Berkman AM, Andersen CR, Puthenpura V, Livingston JA, Ahmed S, Cuglievan B, et al. Impact of Race, Ethnicity, and Socioeconomic Status over Time on the Long-term Survival of Adolescent and Young Adult Hodgkin Lymphoma Survivors. Cancer Epidemiology and Prevention Biomarkers. 2021;30(9):1717–1725.
4. Cooper RM, Chung J, Hogan T, Haque R. Patterns of overall mortality by race/ethnicity and socioeconomic status in insured cancer patients in Southern California. Cancer Causes & Control. 2021;32(6):609–616.
5. Singh GK, Jemal A. Socioeconomic inequalities in cancer incidence and mortality. The American Cancer Society's Principles of Oncology: Prevention to Survivorship. Hoboken: Wiley; 2018.
6. Tweed EJ, Allardice GM, Mcloone P, Morrison DS. Socioeconomic inequalities in the incidence of four common cancers: a population-based registry study. Public Health. 2018;154:1– 10.
7. Castro S, Sosa E, Lozano V, Akhtar A, Love K, Duffels J, et al. The impact of income and education on lung cancer screening utilization, eligibility, and outcomes: A narrative review of socioeconomic disparities in lung cancer screening. Journal of Thoracic Disease. 2021;13(6):3745–3745.
8. Nejatinamini S, Godley J, Minaker LM, Sajobi TT, Mccormack GR, Cooke MJ, et al. Quantifying the contribution of modifiable risk factors to socioeconomic inequities in cancer morbidity and mortality: a nationally representative population-based cohort study. International Journal of Epi- demiology. 2021;50(5):1498–1511.
9. Minas TZ, Kiely M, Ajao A, Ambs S. An overview of cancer health disparities: new approaches and insights and why they matter. Carcinogenesis. 2021;42(1):2–13.
10. Gillmann C, Pajor G, Ramadori P, Albers P, Mons U, Steindorf K, et al. Solving problems is smart, preventing them is wise: Lessons learned from the 2nd International DKFZ Con- ference on Cancer Prevention. International Journal of Cancer. 2021;148(12):3086–3096.
11. Hamdi Y, Abdeljaoued-Tej I, Zatchi AA, Abdelhak S, Boubaker S, Brown JS, et al. Cancer in Africa: the untold story. Frontiers in Oncology. 2021; p. 11–11.
12. Seraphin TP, Joko-Fru WY, Kamaté B, Chokunonga E, Wabinga H, Somdyala NIM, et al. Rising prostate cancer incidence in sub-Saharan Africa: a trend analysis of data from the African Cancer Registry Network. Cancer Epidemiology and Prevention Biomarkers. 2021;30(1):158–165.

13. Bassette, Emma, et al. "Perceptions of Cervical Cancer Screening in Rural Guatemala." Hispanic Health International 20.2(2022): 156-163.
14. Collatuzzo G, Pelucchi C, Negri E, López-Carrillo L, Tsugane S, Hidaka A, et al. Exploring the interactions between Helicobacter pylori (Hp) infection and other risk factors of gastric cancer: A pooled analysis in the Stomach cancer Pooling (StoP) Project. International Journal of Cancer. 2021;149(6):1228–1238.
15. Lee, Dong-Hae, Jong-Hun Ha, Jeong-Ih Shin,Kyu-Min Kim, Jeong-gyu Choi, Seorin Park, Jin-Sik Park et al."Increased Risk of Severe Gastric Symptoms by Virulence Factors vacAslc, alpA, babA2, and hopZ in helicobacter pylori Infection."(2021):368-379.
16. Huang J, Koulaouzidis A, Marlicz W, Lok V, Chu C, Ngai CH, et al. Global burden, risk factors, and trends of esophageal cancer: an analysis of cancer registries from 48 countries. Cancers. 2021;13(1):141–141.
17. Rumgay H, Ferlay J, Martel CD, Georges D, Ibrahim AS, Zheng R, et al. Global, regional and national burden of primary liver cancer by subtype. European Journal of Cancer. 2022; 161:108–118.
18. He WQ, Gao X, Gao L, Ma Y, Sun D, Sun J. Contrasting Trends of Primary Liver Cancer Mortality in Chinese Mongol and Non-Mongol. Asian Pacific Journal of Cancer Prevention. 2021;22(9):2757–2763.
19. Spearman CW, Desalegn H, Ocama P, Awuku YA, Ojo O, Elsahhar M, et al. The sub-Saharan Africa position statement on the redefinition of fatty liver disease: from NAFLD to MAFLD. Journal of Hepatology. 2021;74(5):1256–1258.
20. Oneata D. Probabilistic latent semantic analysis. Proceedings of the Fifteenth Conference on Uncertainty. 1999; p. 1–7.
21. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research. 2003;3:993–1022.
22. Grootendorst M; 2020. Available from: Https://Doi.Org/10.5281/Zenodo.
23. Meehan C, Meehan S, Moore W; 2020. Available from: Https://Www.Mathworks.Com/Matlabcentral/Fileexchange/71902.
24. Ali, Moez."PyCaret: An open source, low-code machine learning library in Python."PyCaret version 2(2020)
25. McInnes, Leland, John Healy, and James Melville."Umap:Uniform manifod approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426(2018).
26. Zeng Z, Hua B. Uncovering Topics of Public Cultural Activities: Evidence from China. Data Intelligence. 2017; p. 1–19.