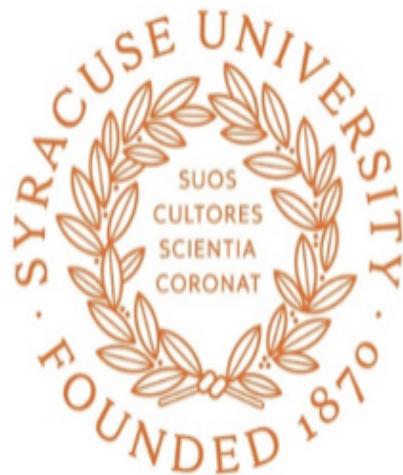


**IST 736 Text Mining Final Project**  
**Breast Cancer Posts and Comments Analysis.**



**Syracuse University, Syracuse,  
New York – 13210**

**Submitted by**

**Abirami Rajalingam**

**Sravan Kumar Mangalagiri**

**Venkata Sai Mani Lakshmi Kavya Darsi**

# Breast Cancer Posts and Comments Analysis.

## Abstract:

In the era of digital communication, social media platforms like Reddit have become pivotal in shaping public discourse on health issues, including breast cancer. This study aims to systematically analyze the discussions, questions, and concerns about breast cancer on such platforms to assist people in understanding and addressing public needs more effectively. By examining the types of questions asked and the responses provided within these online communities, we seek to identify key themes and areas of public interest or misunderstanding regarding breast cancer and its treatment. This analysis not only highlights the most common inquiries and misconceptions but also provides insights into the emotional and practical challenges faced by individuals affected by breast cancer. The findings of this study are intended to aid healthcare professionals in tailoring their communication strategies, educational materials, and patient engagement approaches. Furthermore, this research contributes to the broader goal of enhancing shared decision-making processes between health professionals and people by illuminating the information needs and concerns of the latter. Ultimately, this study underscores the importance of leveraging social media data to bridge gaps in understanding between medical professionals and the public, thereby fostering more effective health communication and improved patient care in the context of breast cancer.

## Data Description:

The process of data consolidation in this study was meticulously executed by integrating two distinct datasets. The first dataset consisted of Reddit questions, while the second comprised the associated comments. To achieve this integration, the unique identifiers (IDs) associated with each question were utilized in conjunction with the corresponding parent IDs from the comment's dataset. This methodological approach was crucial in merging both datasets into a singular, cohesive file. It ensured that the data representation was both structured and comprehensive, allowing for an effective analysis of the discussions and concerns surrounding breast cancer on social media platforms.

After Merging the dataset, we have 65553 rows in data where Title, Question, Comments, Score and Controversiality as the main columns.

Title	Question	Comments
8544	8544	65553

**Title:** It depicts the Main theme of a Post (Question)

**Ex:** How did your hair grow back after chemotherapy?

**Question:** It defines what the question is about in a detailed manner.

**Ex:** I am just curious.... what happened to your hair after chemotherapy? Mine, before, was stick straight. It had darkened over the years, and was a very dirty blonde, with a little grey, but I covered that up with L'Oreal. When it grew back (very slowly, I might add), it was CURLY! Tiny pin-curls, all over my head. Ringlets in the back. And, it was the exact same color I had when I was a kid. That lasted for about 8 or 9 months, and the new growth is stick straight, dirty blonde, with much more grey. :(

**Comments:** The list of people who commented on the post and what they have commented.

**Ex:** My mom always had coarse, thick, wavy hair. After her chemo treatments, her hair grew back with magnificent curls! Really soft and shiny.

**Score:** The difference of number of Upvotes and the number of downvotes.

**Controversiality:** Based on the Upvotes and Downvotes, the comment is towards the controversiality or not.

## Data Modelling:

### Data Profiling:

The dataset, comprising titles, selftext (questions), and bodies (comments), required preprocessing to enhance its quality for analysis. This preprocessing involved cleansing the text of hyperlinks, converting all characters to lowercase, and excising punctuation. Subsequently, tokenization and lemmatization procedures were applied to distill the text to its lexical roots, followed by the removal of stopwords to eliminate common but non-informative words. Additionally, all conventional salutations were expunged from the text to maintain a focus on the substantive content. These steps collectively refined the dataset, rendering it more suitable for the subsequent stages of our analytical process.

### Word Vectorizing:

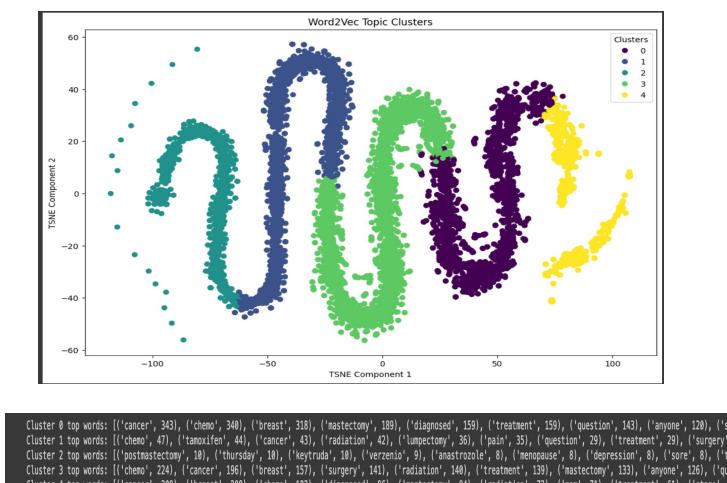
Vectorization was executed utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) method with the configuration parameters set to a maximum of 500 features and English stop words excluded to refine the feature set.

Concurrently, a Word2Vec model was developed, parameterized with a vector size of 100, a context window of 5, a minimum token count threshold of 1, and a skip-gram to 0. Subsequent to vectorization, clustering algorithms were applied to the derived vector spaces from both TF-IDF and Word2Vec models. The efficacy of the clustering solutions was quantitatively assessed using the silhouette score metric.

The performance analysis revealed that the TF-IDF vectorization achieved a superior silhouette score of **0.0033**, indicating more cohesive and separated clusters compared to the Word2Vec model, which yielded a silhouette score of **0.09327**. These scores reflect the relative adequacy of each vectorization technique in capturing the inherent structure of the dataset for clustering purposes.

## Word2Vec:

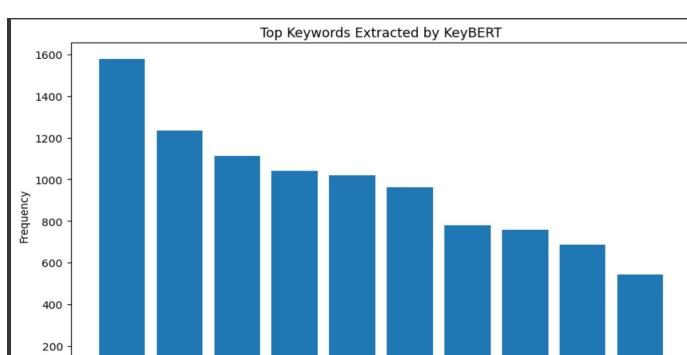
We have performed an elbow method to find the optimal k value and performed TSNE to change the dimensionality reduction to 2D. We performed KMeans using Word2Vec and the results showed that the clusters are near to each other. Some of the topics were repeated in each cluster which we cannot define what a particular cluster means and it is predicted according to the topic. In our text clustering study, we determined the optimal number of clusters using the elbow method and applied Word2Vec for text vectorization followed by K-Means clustering. Dimensionality reduction was performed using t-SNE to visualize the data in two dimensions. The visualization revealed closely positioned clusters, indicating similar topics within the clusters and challenging the distinct definition of each topic. So, for this reason we are exploring TF-IDF even though it has low silhouette Score. The output below has the k-Means Cluster on using Word2Vec in which we can see Cancer in every cluster.



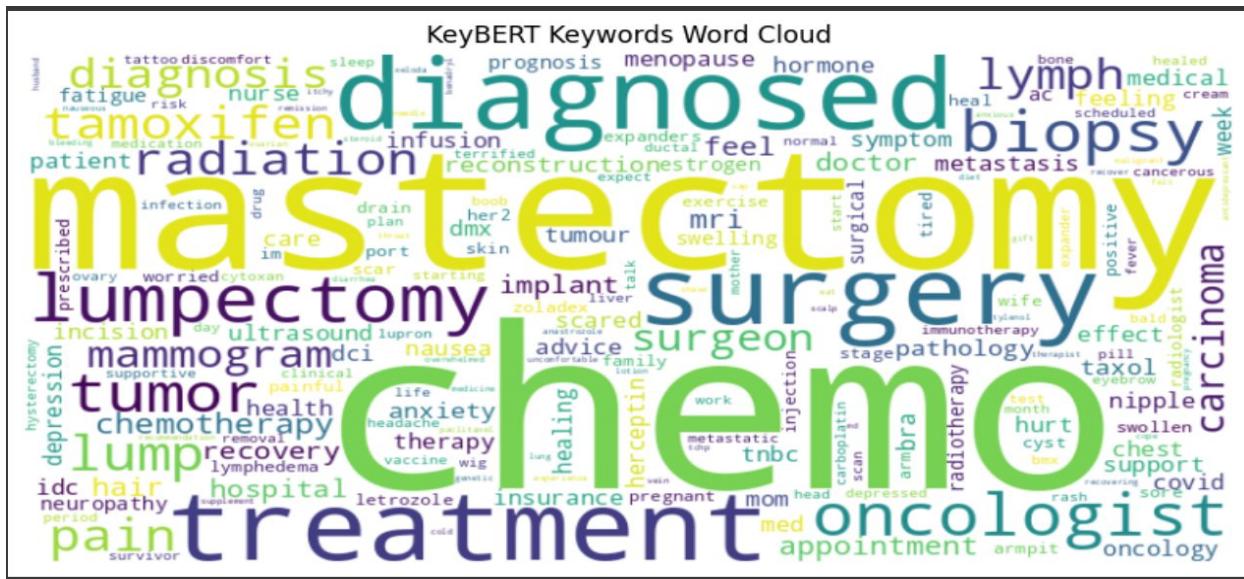
## KeyWord Extraction Using KeyBERT:

KeyBERT is utilized for keyword extraction to distill texts to their most relevant components, reducing dimensionality and computational load for models, and enhancing the interpretability and accuracy of the model's insights by focusing on contextually significant words by using yake and rake packages.

```
0     [donate, cancer, donates, chemotherapy, fund]
1         [hair, blonde, grey, ringlet, color]
2     [lump, mammogram, nipple, swollen, cancer]
3     [mammogram, radiologist, breast, chest, lymph]
4     [lump, biopsy, mammogram, breast, ultrasound]
5     [mammogram, lump, cancer, breast, malignant]
6         [statistic, cancer, phd, breast, rate]
7     [family, cancer, mom, sister, lumpectomy]
8     [cancer, romantic, relationship, love, breast]
9     [breastcancer, cancer, breast, treatment, diag...]
10    [cancer, chemo, survivor, breast, uplifting]
11    [lymph, breast, cancer, advice, support]
12    [wig, cancer, scarf, breast, hair]
13    [cancer, writing, expense, breast, experience]
14    [increaseddecrease, carcinoma, neopugen, herce...]
15    [lump, mammogram, breast, cyst, chest]
16    [ovary, ovarian, removing, removed, remove]
17    [cancer, lump, tumor, wife, family]
18    [cancer, pathology, mastectomy, mutation, lymph]
19    [tumor, cancer, malignant, diagnosed, chemo]
20    [cancer, mastectomy, mammogram, lymphoma, breast]
21    [cancer, men, breast, diagnosed, risk]
22    [rubber, sure, insert, purchase, shouldnt]
```



## Word Cloud:



## Latent Dirichlet Allocation(lda):

We are using LDA for our dataset to analyze the co-occurrence of words within these text features, thereby facilitating a deeper understanding of the prevalent themes and subjects across the collection of texts. The fitted LDA model was then used to transform the TF-IDF matrix into topic distributions for each document. Evaluation metrics such as log likelihood and perplexity were calculated to gauge the model's explanatory power on the observed data. A further comparative analysis was conducted by iterating the LDA model across a range of one to four topics, documenting the log likelihood and perplexity to assess and compare the performance of the models at varying levels of topic granularity.

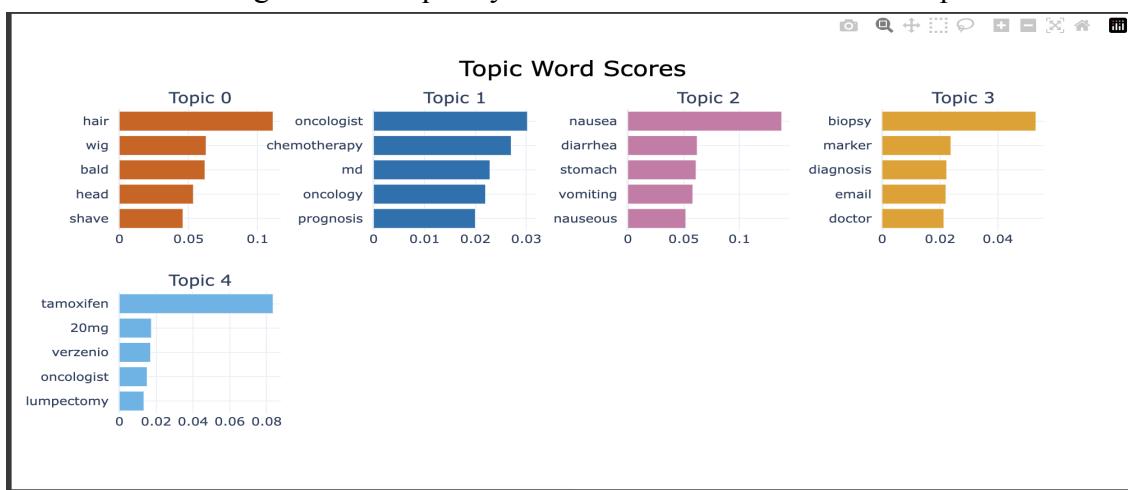
Log Likelihood values become less negative, which could indicate improving model fit. However, the Perplexity also increases, which would generally suggest a decrease in the model's predictive power.

```
Topic 0:  
breast mastectomy surgery nipple bra dci scared radiation pain cancer  
Topic 1:  
chemo treatment cancer infusion taxol therapy anxiety pain nausea herceptin  
Topic 2:  
insurance hurt dmx chemo tamoxifen treatment prescribed letrozole exercise metastatic  
Topic 3:  
doctor hair chemo feel feeling effect treatment tired cancer im  
Topic 4:  
cancer diagnosed lumpectomy biopsy mastectomy tumor oncologist surgery breast chemo  
Log Likelihood: -126503.38461513004  
Perplexity: 1430.9049924556778  
1 Topics - Log Likelihood: -114383.495691571  
1 Topics - Perplexity: 713.3155887012034  
2 Topics - Log Likelihood: -117538.59621291686  
2 Topics - Perplexity: 855.037761898644  
3 Topics - Log Likelihood: -121140.49633236378  
3 Topics - Perplexity: 1051.5604280966713  
4 Topics - Log Likelihood: -124209.72635613257  
4 Topics - Perplexity: 1254.2837221985833
```

## BERT:

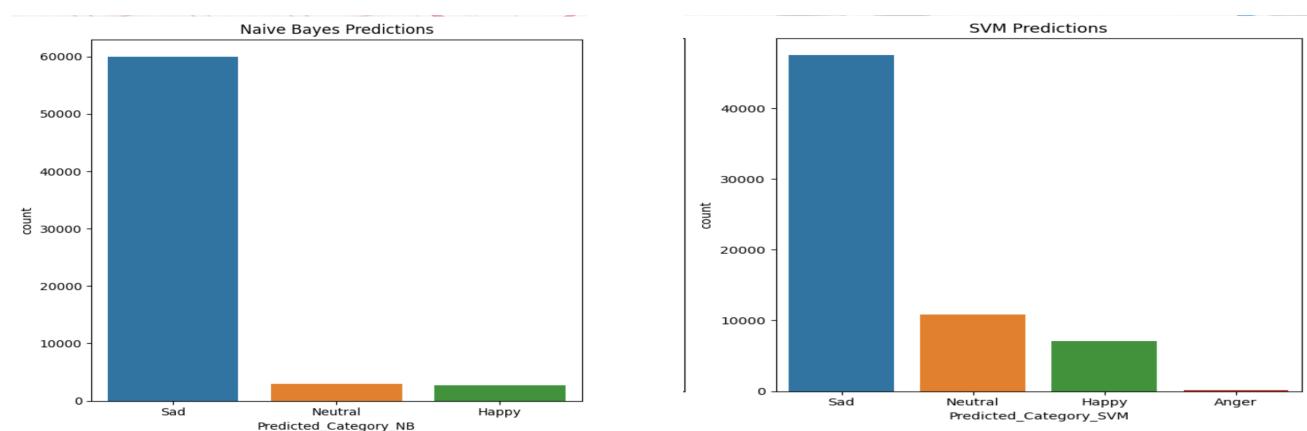
Incorporating BERT in this process leverages its advanced capability to understand context and semantics, enhancing the precision of topic detection and keyword relevance. This visual depiction allows for an immediate grasp of the data's thematic composition, streamlining strategic decision-making and insights derivation.

The bar chart illustrates the findings from a BERTopic analysis, where each bar signifies the weighted significance of keywords within identified topics (Topic 0-4), as determined by the model. The choice of topics is tailored to the dataset's intricacy and the desired detail level. The bars' lengths reflect the c-TF-IDF scores of keywords, showcasing their pertinence to the topics, with scores factoring in term frequency and distinctiveness across the corpus.



## Emotion Analysis:

We have unlabeled data so far, for the Emotion Analysis we used Twitter data which is labeled with emotions and using it with the Reddit data. The sentence embeddings using a pre-trained DistilBERT model from the Sentence Transformers library, enhancing the capability to capture emotional subtleties within the comments, followed by another round of K-Means clustering to classify the data into nuanced emotional categories. By using this we get mixed emotions which leads to using SVM and NaiveBayes for Emotion Analysis.



## **Q & A Chat Bot:**

### **Introduction:**

Our aims are to successfully developed a Q & A chatbot designed to assist doctors, patients, and the public in comprehending complex medical terminology related to breast cancer. Utilizing data sourced from Reddit—a platform where individuals openly share their experiences, concerns, and insights on breast cancer—our chatbot aims to enhance understanding and raise awareness about this critical health issue. Given the sensitive nature of breast cancer and the numerous inquiries surrounding it, our chatbot serves as an accessible resource, drawing from the wealth of communal knowledge on Reddit to provide support and information to those seeking it.

### **Preprocessing:**

We used (cosine\_similarity, TfidfVectorizer), for chatbot. And developed an intelligent chatbot by first extracting and consolidating URLs from Reddit discussions on breast cancer into a single string for analysis. Our preprocessing involved converting text to lowercase, removing URLs, punctuation, and stop words, then performing tokenization and lemmatization to streamline the text for our AI models. After preprocessing the questions, we chose Tfidf vectorization—using max\_df of 0.95 and min\_df of 2—for its superior performance over word2vec. We then applied BERT Topic modeling to discern key topics and employed K-means clustering to organize comments by these topics, resulting in a dataset primed for the chatbot to deliver topic-specific assistance.

### **Streamlined Workflow of a breast cancer information Chatbot:**

The Q & A chatbot processes and responds to inquiries about breast cancer. It normalizes and simplifies user input through text preprocessing, identifies greetings and farewells to engage in natural dialogue, and searches for related topics and reference links within a dataset. Using a scoring filter to improve chatbot accuracy by removing irrelevant responses.

Below are Chatbot Snapshots: Demonstrating Utility and Effectiveness in Information Delivery:

```
Hello! How can I help you with?
You: tamoxifen
Bot: Top results for the topic 'tamoxifen':
is there an alternative to tamoxifen? it doesn't mix with a medicine i am already on. and i'm supposed to go on it for five years. i asked m
i don't think so. if you're premenopausal.
i started taking tamoxifen yesterday and, to be honest, am a bit afraid because of all these horror stories i've heard. i need a few stories
i took tamoxifen for 7 years without significant side effects. it was much less troublesome than the ai med i'm on now.
hi all. first post in here. i'm 10-months-post-lumpectomy with second surgery due to complications, but got clear margins. i've been on tam
i'm perimenopausal, but was still regular and even on the tamoxifen. however, my periods became even heavier and the cramping was agonizing.
so i'm trying lupron as the last ditch before a total hysterectomy (yeah, the cramps are that bad). two shots in, most recently 6 days ago.
is this possibly a new side effect of the lupron? i'm still on tamoxifen (and doubling the hormone blockers plus still menstruating is a gai
Related URLs:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3230769/
https://lifemath.net
https://www.onclive.com/view/breast-cancer-guidelines-carve-out-category-for-low-er-expression)
https://breast.predict.nhs.uk/tool(https://breast.predict.nhs.uk/tool)
https://breast.predict.nhs.uk/tool(https://breast.predict.nhs.uk/tool)
https://breast.predict.nhs.uk/tool(https://breast.predict.nhs.uk/tool)
https://breast.predict.nhs.uk/tool(https://breast.predict.nhs.uk/tool)
```

```
You: node
Bot: Top results for the topic 'node':
i had stage 1 and not in lymph nodes. i decided to get a bi-lateral mastectomy just so i wouldn't need to worry about it! i was out of work fo
i am a single mom with young children at the time.
it was tough the first two weeks. after that it got better with time.
hope everything goes well!
i have one chemo left and then i find out my results prior to surgery. i am also tn, stage 2b, grade 3 idc. my tumor was 4.3 cm when diagnos
i am (was) stage 3 tnbc and inflammatory breast cancer (double whammy). i'm just on the tail end of my treatment now and doing well! i did !
it'll be a rough few months, but there have been advances in research for tnbc and there are aggressive treatment plans built to combat the
yes, focus on the wonderful sleep and drugs! my surgery is tomorrow too - lumpectomy and lymph nodes. we got this!!!!
i had a lumpectomy in august (so two incisions: one at 12:00 and one under my arm for removal/inspection of two lymph nodes). i had minor so
Related URLs:
https://imgur.com/a/V6PtyOf
https://www.areyoudense.org/l\(https://www.areyoudense.org/\)\)
https://www.areyoudense.org/l\(https://www.areyoudense.org/\)\)
https://imgur.com/a/JL1gpFU
https://en.wikipedia.org/wiki/Lymphedema
https://i.imgur.com/IIF5vd0.jpeg
https://www.annualsofoncology.org/article/S0923-7534\(19\)35464-X/fulltext
You: good bye
Bot: Goodbye! Have a great day.
```

## Emotional Analysis on chatbot :

In our project, we've incorporated emotion analysis to gauge the sentiment of chatbot responses. After preprocessing similar to earlier steps, we trained both Naive Bayes and SVM models to accurately identify and understand the underlying emotions in the chatbot's replies. This allows for a more nuanced and empathetic user interaction.

Emotional Analysis Snapshots: Showcasing the Chatbot's Sentiment Detection Capabilities for Improved Communication:

Response: Bot: Top results for the topic 'node':  
i had stage 1 and not in lymph nodes. i decided to get a bi-lateral mastectomy just so i wouldn't need to worry about it! i was out of work for  
i am a single mom with young children at the time.  
it was tough the first two weeks. after that it got better with time.  
hope everything goes well!

i have one chemo left and then i find out my results prior to surgery. i am also tn, stage 2b, grade 3 idc. my tumor was 4.3 cm when diagnosed:  
i am (was) stage 3 tnbc and inflammatory breast cancer (double whammy). i'm just on the tail end of my treatment now and doing well! i did :

it'll be a rough few months, but there have been advances in research for tnbc and there are aggressive treatment plans built to combat the  
yes, focus on the wonderful sleep and drugs! my surgery is tomorrow too - lumpectomy and lymph nodes. we got this!!!!  
i had a lumpectomy in august (so two incisions: one at 12:00 and one under my arm for removal/inspection of two lymph nodes). i had minor sc

Related URLs:

<https://imgur.com/a/V6Pty0f>  
<https://www.areyoudense.org/l/>(<https://www.areyoudense.org/>)  
<https://www.areyoudense.org/l/>(<https://www.areyoudense.org/>)  
<https://imgur.com/a/JL1gpFU>  
<https://en.wikipedia.org/wiki/Lymphedema>  
<https://i.imgur.com/IIF5vd0.jpeg>  
[https://www.annalsofoncology.org/article/S0923-7534\(19\)35464-X/fulltext](https://www.annalsofoncology.org/article/S0923-7534(19)35464-X/fulltext)

Predicted Emotion: Sad

Response: Bot: Goodbye! Have a great day.

Predicted Emotion: Sad

Response: Bot: Top results for the topic 'tamoxifen':  
is there an alternative to tamoxifen? it doesn't mix with a medicine i am already on. and i'm supposed to go on it for five years. i asked r  
i don't think so. if you're premenopausal.  
i started taking tamoxifen yesterday and, to be honest, am a bit afraid because of all these horror stories i've heard. i need a few stories:  
i took tamoxifen for 7 years without significant side effects. it was much less troublesome than the ai med i'm on now.  
hi all. first post in here. i'm 10-months-post-lumpectomy with second surgery due to complications, but got clear margins. i've been on tamoxi  
i'm perimenopausal, but was still regular and even on the tamoxifen. however, my periods became even heavier and the cramping was agonizing.  
so i'm trying lupron as the last ditch before a total hysterectomy (yeah, the cramps are that bad). two shots in, most recently 6 days ago.  
is this possibly a new side effect of the lupron? i'm still on tamoxifen (and doubling the hormone blockers plus still menstruating is a giant

Related URLs:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3230769/>  
<https://lifemath.net>  
<https://www.onclive.com/view/breast-cancer-guidelines-carve-out-category-for-low-er-expression>).  
<https://breast.predict.nhs.uk/tool>(<https://breast.predict.nhs.uk/tool>)  
<https://breast.predict.nhs.uk/tool>(<https://breast.predict.nhs.uk/tool>)  
<https://breast.predict.nhs.uk/tool>(<https://breast.predict.nhs.uk/tool>)  
<https://breast.predict.nhs.uk/tool>(<https://breast.predict.nhs.uk/tool>)  
<https://breast.predict.nhs.uk/tool>(<https://breast.predict.nhs.uk/tool>)  
<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>(<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>)  
<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>(<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>)  
<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>(<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>)  
<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>(<https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900286>)  
<https://www.uptodate.com/contents/ductal-carcinoma-in-situ-treatment-and-prognosis#H935610424>  
<https://www.uptodate.com/contents/ductal-carcinoma-in-situ-treatment-and-prognosis#H935610424>  
<https://www.reddit.com/poll/omd5kx>  
<https://www.wrtv.com/news/local-news/tippecanoe-county/ovarian-cancer-drug-invented-at-purdue-approved-by-fda>  
<https://www.wrtv.com/news/local-news/tippecanoe-county/ovarian-cancer-drug-invented-at-purdue-approved-by-fda>  
[https://www.reddit.com/r/breastcancer/comments/ycqy6g/question\\_for\\_premenopausal\\_ladies\\_on\\_tamoxifen/?utm\\_source=share&utm\\_medium=ios\\_ar](https://www.reddit.com/r/breastcancer/comments/ycqy6g/question_for_premenopausal_ladies_on_tamoxifen/?utm_source=share&utm_medium=ios_ar)  
[https://www.reddit.com/r/breastcancer/comments/ycqy6g/question\\_for\\_premenopausal\\_ladies\\_on\\_tamoxifen/?utm\\_source=share&utm\\_medium=ios\\_ar](https://www.reddit.com/r/breastcancer/comments/ycqy6g/question_for_premenopausal_ladies_on_tamoxifen/?utm_source=share&utm_medium=ios_ar)  
<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2799011>  
<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2799011>

Predicted Emotion: Sad

## Challenges and Solutions:

- Chose to use TF-IDF instead of Word2Vec to better differentiate topics and improve computational efficiency, tailored to meet the unique needs of analyzing discussions on the breast cancer subreddit.
- Transitioning to the use of labeled tweet datasets for emotion analysis in breast cancer-related comments, our approach has evolved to capture the nuances of emotions more accurately. This shift highlights the challenges in selecting and preparing data for an effective Question-Answering model, emphasizing the complexity of natural language processing in diverse datasets.

- Encountered constraints in performing an exhaustive error analysis of the models, primarily due to the absence of labeled data for testing, which affected the ability to assess model performance accurately.
- Computational time is taking a lot of time because of pretrained model, this can be solved by increasing the quality of comments and using a big range processor

### **Ethics:**

### **Privacy:**

Considering the significant amount of data incorporated into our analysis, we must recognize the possible inclusion of personal or sensitive information within the dataset. This recognition necessitates a diligent approach to the management and processing of the data, with a keen emphasis on adhering to ethical research practices.

### **Robustness:**

In our context, robustness refers to the thorough evaluation of our model's performance across a spectrum of challenging conditions. This includes its resilience to simulated perturbations or data variability. Conducting such assessments is vital to confirm the dependability and thoroughness of our analysis, which is particularly important given the sensitive nature of breast cancer discourse on Reddit.

### **Conclusion:**

In conclusion, our comprehensive analysis of breast cancer discussions on social media platforms like Reddit has revealed key themes, public interests, and widespread misconceptions. This understanding is critical for healthcare providers to enhance their communication strategies and educational resources, addressing the real-world questions and emotional challenges faced by individuals. Looking ahead, the anticipation of advancements in emotion analysis through specialized datasets and the adoption of transformer-based NLP models like BERT and GPT-3 promises to deepen our understanding of these conversations. Additionally, the integration of sophisticated AI tools, including chatbots and recommendation systems, will further tailor and personalize support for breast cancer. These future developments will not only enrich conversational engagement but also extend comprehensive language support, playing a pivotal role in bridging the knowledge gap between health professionals and the public. Ultimately, this study and its future directions underscore the power of leveraging social media analytics in fostering more effective health communication and improving patient outcomes in the context of breast cancer care.