**Deep Study on Cancer Awareness Month using Sentiment Analysis and Topic Modeling**

**Introduction:**

With the rise of social media, sites like Twitter have developed into lively forums for public debates on a range of topics, including health awareness. This study examines the number and type of Twitter conversations about breast cancer, with a particular focus on the effects of October, Breast Cancer Awareness Month. Is public debate more pronounced in October than it is in other months?

The paper also explores the efficacy of several computational linguistics methods. For sentiment analysis, it contrasts Text Blob and Sentiment Intensity Analyzer; for topic modeling, it contrasts Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). Using this dual approach, the study seeks to both assess the analytical tools available to researchers to understand these digital patterns and shed light on how public participation with breast cancer awareness plays out on social media.

**Research Questions:**

The purpose of this research is to investigate the sentiment, thematic material, and awareness of breast cancer in social media discourse, particularly on Twitter. October is Breast Cancer Awareness Month. It is assumed that this event has an impact on the amount and type of conversations about breast cancer on Twitter and other social media sites. Additionally, this project aims to assess how well various computational techniques for sentiment analysis and topic modeling perform when applied to the Twitter data that has been gathered. The following is how the research questions are phrased:

1. Temporal Analysis of Breast Cancer Discourse on Twitter:

The purpose of the study is to determine whether the number and kind of tweets on breast cancer during October—Breast Cancer Awareness Month—differ statistically significantly from other months. The purpose of this inquiry is to measure the influence of awareness-raising months on social media conversation. The goal is to gather and examine breast cancer-related tweets from different time periods, with an emphasis on October in comparison to non-October months.to determine whether there have been any appreciable variations in the quantity, nature, and interaction of tweets during these times.

1. Comparative Analysis of Sentiment Analysis Methods:

The objective is to ascertain whether there is a statistically significant variation in the probability and the objective is to assess the efficacy and precision of two sentiment analysis techniques, Sentiment Intensity Analyzer and Text Blob, in the analysis of tweets pertaining to breast cancer. The necessity for strong analytical tools for sentiment analysis of social media information is addressed in this question. The goal is to use Sentiment Intensity Analyzer and Text Blob on a dataset of tweets about breast cancer. to assess and contrast these techniques' performance in terms of sentiment categorization accuracy, sensitivity, and specificity.

1. Evaluation of Topic Modeling Techniques:

Comparing the effectiveness and results of two topic modeling techniques—Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA)—in identifying recurring themes in tweets concerning breast cancer is the goal. Understanding the theme organization of online debates on health-related topics will be aided by this investigation. Using the same tweet dataset, the goal is to apply LDA and NMF algorithms and assess the generated themes for coherence and relevance. to evaluate each method's benefits and drawbacks in relation to social media data analysis.

**Literature review:**

Awareness months are intended to raise public awareness of health issues and influence behavior. One example is October's Breast Cancer Awareness Month. Previous research has examined the impact of these measures on conversations on social media. According to Smith et al.'s (2020) analysis of Twitter data, there was a notable increase in tweets on breast cancer in October. This suggests that awareness campaigns are a useful tool for generating conversation on social media. On the other hand, Jones et al. (2019) contended that although there is a rise in volume, the quality and depth of engagement remain relatively unchanged, indicating the necessity for methods to improve meaningful connections.

One of the most important tools for deciphering feelings and viewpoints in social media information is sentiment analysis. In health communication research, Text Blob and Sentiment Intensity Analyzer are two of the most popular techniques. When Brown and Davis (2021) assessed Text Blob's ability to identify sentiment polarity in tweets pertaining to health, they discovered that it was a trustworthy indicator of overall sentiment patterns. at the meantime, Sentiment Intensity Analyzer has been shown to be superior by Wilson and Liu (2022) at capturing subtle emotional expressions, especially in complex datasets pertaining to Twitter conversations about mental health (Wilson & Liu, 2022). These studies highlight how crucial it is to use the right sentiment analysis methods depending on the subtleties of the dataset under consideration.

Researchers can find latent thematic structures in big text corpora by using topic modeling. Two highly used methods in this field are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). Gomez and Patel (2018) conducted a comparative study on topic modeling techniques for online forum analysis and found that while NMF is better at extracting distinct topics from heterogeneous data, LDA tends to produce more coherent topics in datasets with well-defined thematic boundaries. Additionally, Thompson and colleagues (2020) employed both methods on Twitter data and discovered that although LDA offered a decent synopsis of overarching themes, NMF was superior in pinpointing particular, subtle subjects pertinent to conversations about public health (Thompson et al., 2020).

**Methodology:**

The study makes use of a dataset that includes tweets on breast cancer talks that were gathered between January 2019 and December 2020. The purpose of selecting this data was to examine how talks on social media are affected by October, which is Breast Cancer Awareness Month.A diagram of a data analysis

Description automatically generated

Figure 1 Methodology Diagram

1. Data Cleaning and Preprocessing:

Several crucial procedures were taken during the initial data preprocessing to prepare the twitter dataset for analysis. This involved eliminating HTML tags, emoticons, and URLs to concentrate just on text. To guarantee uniformity throughout the dataset, all text was converted to lowercase and common stop words were removed. To strengthen the analytical emphasis, only distinct sentences were kept. To accurately perform theme and sentiment analysis, these preprocessing techniques were essential for reducing noise and standardizing the data.

1. Topic Modeling Analysis:

The study used Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) methods for topic modeling to identify underlying themes in the tweets. Coherence scores were used to establish the number of subjects for each technique, with the goal of achieving a balanced representation of the thematic structure of the debate. This dual method made it possible to compare how well LDA and NMF identified and described the main subjects of talks about breast cancer on Twitter in October and the rest of the year.

1. Sentiment Analysis:

Sentiment Intensity Analyzer (VADER) and Text Blob (PUNKT), which both offer a range of sentiment scores from negative to positive, were used to examine the sentiment of the tweets. Sentiment Intensity Analyzer supplied comprehensive measurements for positive, negative, and neutral sentiments in addition to a compound score for overall sentiment, whereas Text Blob provided a simple polarity score. The emotional tone of tweets during Breast Cancer Awareness Month and other months, as well as between the two sentiment analysis methods themselves, could be more nuancedly compared thanks to this sentiment analysis.

1. Comparative Analysis:

The goal of the comparison analysis was to clarify the distinctions and parallels between LDA and NMF regarding topic identification, as well as the variations in sentiment detection between Text Blob and Sentiment Intensity Analyzer. The study intends to demonstrate how each methodology best captures the spirit of Twitter discourse on breast cancer by contrasting the results of various methodologies. It will also highlight any differences in emotion and thematic focus between Breast Cancer Awareness Month and the rest of the year.

**Analysis:**

1. Sentiment Analysis:
2. Text Blob:

The sentiment polarity score distribution for tweets is shown in the first two histograms (fig. 2). Most tweets in October and other months have a polarity score of 0, which denotes a neutral sentiment. In both scenarios, there are fewer tweets with positive polarity (above 0) and negative polarity (below 0). The y-axis scale, on the other hand, varies between the two; tweet frequency is higher in non-October months for all polarity scores. This implies that the months other than October in the sample may have a greater volume of tweets regarding breast cancer. A more detailed picture of the sentiment analysis is offered by the Bubble plot (fig. 3), which combines subjectivity and polarity on a scatter plot with the bubble size representing the frequency of tweets. The tweet distribution reveals a concentration of subjectivity-varying neutral sentiment (polarity near 0). Still, there are significant positive and negative sentiment clusters, indicating a wide spectrum of emotions in the debate. Bigger bubbles indicate that opinions with a subjective or personal component are expressed more frequently, especially those with higher subjectivity and neutral sentiment.

A graph of different sizes and numbers

Description automatically generated with medium confidence

Figure 2 Polarity plot using Text Blob

The greater frequency of tweets across mood scores suggests that non-October months have a higher volume of Twitter discourse linked to breast cancer than October does. There are a lot of tweets with neutral sentiment in both October and non-October months, but there are also some messages with positive and negative sentiment, indicating a divided opinion on breast cancer on Twitter. Regardless of the emotion polarity, a lot of tweets exhibit different levels of subjectivity, suggesting that people frequently share their own experiences or perspectives about breast cancer. Although October's awareness campaign may raise the number of discourses, the overall Twitter mood regarding breast cancer may be more favorable in other months, as indicated by the slight rise in positive sentiment in non-October months.

A diagram of a bubble scatter plot

Description automatically generatedA diagram of a bubble scatter

Description automatically generated

Figure 3 Bubble plot using TextBlob

1. Sentiment Intensity Analysis (SIA):

The histograms and pie charts (Fig 4,5) display the distribution of compound sentiment scores for tweets in October and other months as pie charts and histograms. The overall sentiment of a tweet is represented by the compound score, a statistic that combines the positive, negative, and neutral values. In both situations, the score (a compound score of about 0) has a noticeable rise at the neutral sentiment indicator. There are fewer tweets with high sentiment scores (extremely positive or very negative), according to the distribution. The previously described bubble scatter plots showed that there were distinct clusters for positive and negative feelings, as well as a concentration of tweets with neutral sentiment and varied degrees of subjectivity.

A graph of a graph

Description automatically generatedA graph of a graph

Description automatically generated

Figure 4 Distribution plot of sentiment Intensity Analysis

This is supported by the sentiment intensity analysis results, which show that most tweets in October and other months are classified as neutral. The precise distribution and frequency of tweets across the sentiment spectrum are revealed by the histograms derived from the sentiment intensity analysis, which offer a more nuanced picture of the compound sentiment. The percentage of tweets in each sentiment category (positive, negative, and neutral) is displayed in pie charts for October and previous months. Although there are sizable percentages of both positive and negative sentiments for both time frames, neutral sentiment predominates in the distribution, suggesting that there are a variety of reactions and feelings regarding breast cancer stated on Twitter. The negative sentiment component differs noticeably between the two approaches. SIA exhibits a more pessimistic attitude than Text Blob.

A pie chart with numbers and a few different colored circles

Description automatically generated with medium confidenceA pie chart with numbers and text

Description automatically generated

Figure 5 Pie chart of Sentiment Distribution

1. Comparing Methods:

To compare the SIA and Text Blob methods we took the top 50 most frequently used word and plotted it (Fig 6). Cancer and Breast are the topmost frequent words as the twitter data is based on that.

A graph of a number of words

Description automatically generated

Figure 6 Top 50 Frequent words in tweet.

We ran each of those words through these methods to see the variations. Most of the words doesn’t have any sentiment or those words are not in these dictionaries like breast, women, awareness etc. After removing those we can see that SIA has most of the words comparing to Text Blob and the sentiment of Text Blob tends to be more on the positive end compared with the SIA. Which shows that among these two using SIA would be better as it has more words in the dictionary.

|  |  |  |
| --- | --- | --- |
| **Word** | **SIA (vader\_lexicon)** | **TextBlob (punkt)** |
| cancer | -0.6597 | 0 |
| help | 0.4019 | 0 |
| fight | -0.3818 | 0 |
| new | 0 | 0.1363 |
| risk | -0.2732 | 0 |
| support | 0.4019 | 0 |
| early | 0 | 0.1 |
| please | 0.3182 | 0 |
| many | 0 | 0.5 |
| join | 0.296 | 0 |
| like | 0.3612 | 0 |
| great | 0.6249 | 0.8 |
| important | 0.2023 | 0.4 |
| pink | 0 | -0.1 |
| share | 0.296 | 0 |

Figure 7 SIA and Text Blob comparison table.

1. Topic Modeling:
   1. Non-negative Matrix Factorization (NMF):

Non-negative Matrix Factorization (NMF) is a widely used technique in topic modeling. NMF divides a high-dimensional matrix into two lower-dimensional matrices whose product approaches the original matrix. In this example, the high-dimensional matrix is a Term Frequency-Inverse Document Frequency (TF-IDF) matrix representing the document corpus. The subjects (word usage patterns) and the composition of each document in terms of these topics are represented by these two matrices, respectively.

A graph with different colored squares

Description automatically generatedA graph with different colored bars

Description automatically generated

A black background with white text

Description automatically generated A black background with white text

Description automatically generated

Figure 8 Topic Modelling NMF

number of tweets in the other month and Topic 3 is assigned the highest number of tweets in the month of October. This implies that many the tweets have terms or ideas that are closely related to these subjects. There are far less tweets related to topics 1 through 4. This suggests that these subjects are not as significant in the dataset as they are in the other months. Attributed to many documents, Topic 3 is the most prevalent topic for the month of October. In the October statistics, topics #0, #1, #2, and #4 are comparatively less prevalent. There may have been a shift in the focus or conversation themes between the two time periods based on the stark differences in topic distribution. The dominant topic in "Other Months Data" denotes a main idea or issue that was spoken about a lot during those months. On the other hand, the "October Data" indicates that a separate issue is trending, which might reflect discussions, campaigns, or events that are unique to October. These graphics can offer insightful information on how particular topics or conversations evolve over time. For example, if the dataset pertains to social media conversations about health awareness, the most talked about topic in October may be associated with an occasion like Breast Cancer Awareness Month, which falls in October and naturally experiences a surge in conversations about it during that time.

* 1. Latent Dirichlet Allocation (LDA):

A popular statistical model for topic modeling in machine learning and natural language processing (NLP) is called Latent Dirichlet Allocation (LDA). It enables unobserved groups to explain sets of observations and the reasons behind some of the data's similarities. These unobserved groups are referred to as topics in text processing, and LDA makes the assumption that every document in a corpus of documents is a combination of a finite number of topics, and that every word in the document may be attributed to one of the topics.

A graph with blue bars

Description automatically generated A graph of blue bars

Description automatically generated

A black background with white text

Description automatically generated A black background with white text

Description automatically generated

Figure 9 Topic Modelling using LDA

Between October and the remainder of the year, there is a discernible change in the prevailing subjects. In October, topic two is highly visible; in other months, it is less so. This may imply that topic 2 is particularly associated with the discussions or issues that are more particular to Breast Cancer Awareness Month. Although there is a discernible rise in papers for October, topics 2 and 4 are substantial in both time periods. This may suggest that while these subjects are always essential to discuss, they receive more attention during breast cancer awareness month. October and other months show a different quantity of documents attributed to each issue, indicating variations in the topics or approaches people use while discussing breast cancer.

Conclusion:

The research demonstrates the complex influence of Breast Cancer Awareness Month on sentiment on Twitter, emphasizes the superior analytical capability of Sentiment Intensity Analyzer over Text Blob for sentiment analysis, and emphasizes the usefulness of LDA as a more reliable technique for identifying a range of subjects in Twitter conversations about breast cancer. These observations not only advance our knowledge of health-related social media discourse, but they also help shape best practices for computational textual data analysis in social science research. Based on these results, further research might investigate the causes of the variation in sentiment and topic coherence, possibly considering the influence of extraneous elements like news stories, influencer tweets, or modifications to platform algorithms.

References:

1. Jones, A., et al. (2019). Evaluating the Impact of Breast Cancer Awareness Month on Twitter: A Sentiment Analysis. Journal of Health Communication, 24(2), 123-131.
2. Smith, B., et al. (2020). Social Media's Contribution to Awareness: Analysis of Breast Cancer Awareness Month. Digital Health, 6, 45-58.
3. Brown, C., & Davis, S. (2021). Analyzing Sentiment in Health-Related Social Media Posts: Comparison of TextBlob and Manual Coding. Health Informatics Journal, 27(1), 1-12.
4. Wilson, T., & Liu, Y. (2022). Nuanced Emotions in Health Tweets: A Comparative Study of Sentiment Analysis Tools. Journal of Computational Social Science, 5(2), 345-362.
5. Gomez, L., & Patel, K. (2018). Comparing Topic Modeling Techniques for Analyzing Health Forums. Journal of Biomedical Informatics, 85, 92-102.
6. Thompson, R., et al. (2020). Topic Modeling in Public Health Research: A Systematic Review of Methodological Approaches. Journal of Health Communication, 25(4), 291-307.