STEVENS INSTITUTE OF TECHNOLOGY

BIA-660 WEB MINING

GROUP PROJECT
FINAL REPORT

Title: Web-Based Analysis of Women Hormonal Health Challenges using Data mining and NLP Techniques

Group Members:

Gahana Nagaraja

Namratha Nagathihalli Anantha

Vaishnavi Rajendra Dhotargavi

INTRODUCTION

Women's hormonal health conditions, such as Polycystic Ovary Syndrome (PCOS) and Thyroid disorders, affect millions worldwide, significantly impacting physical and mental well-being. These conditions often lead to a range of symptoms, including fatigue, weight gain, irregular periods, hair loss, and emotional distress. Understanding the challenges faced by women dealing with these conditions is crucial for offering effective support and interventions. In today's digital age, online forums and social media platforms provide a space for women to share their experiences, challenges, and coping mechanisms. Analyzing this wealth of user-generated content offers valuable insights into the emotional and psychological aspects of these health challenges.

This project, "Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques," aims to extract, analyze, and interpret data related to PCOS and Thyroid disorders from platforms such as Reddit and trusted health websites like Mayo Clinic and Healthline. By employing techniques from Data Mining and Natural Language Processing (NLP), this project focuses on understanding user sentiment, detecting emotions, and providing suitable remedies based on those emotions.

The primary motivation behind this project is the growing need to support women facing hormonal health challenges. Traditional healthcare approaches may not always address the emotional and psychological aspects of these conditions. Women often turn to online communities for emotional support, advice, and shared experiences. These discussions contain rich, unfiltered data that can reveal trends in sentiment, recurring emotions, and the effectiveness of different coping strategies. By systematically analyzing this data, we can identify common emotional challenges and provide relevant remedies that address both physical symptoms and mental health concerns.

The first step in this project involves web scraping to collect relevant data. Reddit, with its dedicated subreddits like r/PCOS and r/thyroidhealth, serves as the primary source of user-generated content. These subreddits feature discussions on symptoms, treatments, emotional struggles, and personal stories. Simultaneously, health websites like Mayo Clinic and Healthline provide authoritative information on medical conditions and treatments. Data scraping tools such as asyncpraw (for Reddit) and BeautifulSoup (for websites) facilitate the collection of this data.

Once the data is collected, it undergoes preprocessing to ensure it is clean and usable. This includes removing noise, tokenization, and normalization. The cleaned data is then subjected to sentiment analysis using the VADER (Valence Aware Dictionary and Sentiment Reasoner) model. VADER classifies text into three sentiment categories: positive, negative, and neutral. Understanding the sentiment helps gauge the overall emotional tone of discussions related to PCOS and Thyroid conditions.

Beyond sentiment, the project also performs emotion detection to identify specific emotions such as joy, sadness, fear, and anger. This is achieved using a pre-trained DistilRoBERTa model from Hugging Face. By recognizing these emotions, the project can better understand the nuanced emotional states experienced by women discussing their health challenges.

To provide practical support, the project matches the detected emotions with potential remedies sourced from trusted medical websites. For example, if sadness is detected, remedies such as therapy, mindfulness, and exercise are recommended. This step ensures that the emotional challenges identified are met with actionable advice, promoting holistic well-being.

Finally, Machine Learning (ML) models, including Logistic Regression, Support Vector Machines (SVM) and XGBoost, are employed to evaluate the accuracy of sentiment and emotion classification. These models help validate the effectiveness of the analysis and ensure reliable results.

In conclusion, this project combines data mining, NLP, and machine learning to deliver a comprehensive analysis of women's hormonal health challenges. By understanding sentiments and emotions expressed in online discussions and linking them to appropriate remedies, the project aims to bridge the gap between emotional and medical support. This approach not only empowers women with valuable insights but also contributes to the broader understanding of hormonal health challenges.

RESEARCH QUESTION

How can NLP and ML techniques be used to analyze online discussions about PCOS and thyroid disorders to extract meaningful insights into the emotional and mental health impacts of these conditions, and provide targeted, evidence-based support?

This project will use NLP to detect and categorize mental health challenges—like anxiety, sadness, and depression—expressed by individuals with PCOS and thyroid disorders on Reddit. Posts with negative emotions will trigger an automated recommendation system that provides supportive, evidence-based suggestions from reputable sources like Mayo Clinic and WebMD. Supervised ML algorithms will validate and refine sentiment and emotion detection, ensuring reliability through metrics such as accuracy, precision, recall, and F1-score.

LITERATURE REVIEW

1. **Ricardo Loor-Torres, Mayra Duran, David Toro-Tobon, Maria Mateo Chavez, Oscar Ponce, Cristian Soto Jacome, Danny Segura Torres, Sandra Algarin Perneth, Victor Montori, Elizabeth Golembiewski, Mariana Borras Osorio, Jungwei W. Fan, Naykky Singh Ospina, Yonghui Wu, Juan P. Brito,**
   A Systematic Review of Natural Language Processing Methods and Applications in Thyroidology, *Mayo Clinic Proceedings: Digital Health,* Volume 2, Issue 2, 2024,Pages 270-279, ISSN 2949-7612, https://doi.org/10.1016/j.mcpdig.2024.03.007.
   This study aimed to consolidate existing research on NLP applications in thyroid ology, particularly focusing on thyroid-related conditions like thyroid nodules and thyroid cancer. As women are disproportionately affected by thyroid issues, the paper's emphasis on applying NLP techniques to analyse such health conditions is highly relevant to our study, which investigates various hormone-related health challenges faced by women. Although NLP has made strides in thyroid ology, the study emphasizes that no NLP applications have yet been implemented in clinical practice, reflecting the gap between research findings and real-world clinical utility. The authors suggest that future research should prioritize external validation of NLP models and enhance their adaptability to different data sources, particularly as thyroid-related disorders continue to affect diverse patient populations. This recommendation aligns with the goals of our project, as we aim to apply NLP to an alternative data source— web-based discussions—to capture a broader range of patient-reported experiences.

2. **Gethsiya Raagel, K., Bagavandas, M., Sathya Narayana Sharma, K. *et al.* Sentiment Analysis and Topic Modeling on Polycystic Ovary Syndrome from Online Forum Using Deep Learning Approach. *Wireless Pers Commun* 133, 869– 888 (2023). https://doi.org/10.1007/s11277-023-10795-5**
   This study demonstrates the potential of Twitter data for health-related research. By employing advanced NLP techniques like LDA and deep learning classifiers such as LSTM, the authors successfully analysed public discussions on PCOS, providing a blueprint for using social media data in health studies. In our project, we can introduce unique insights by expanding the dataset beyond Twitter to include forums, blogs, and other health-focused platforms, capturing a broader range of experiences related to women's hormonal health. Additionally, incorporating advanced NLP models, such as BERT or GPT, would enable us to capture nuanced language, improving the accuracy of sentiment and topic modelling.

3. **Ahmad, R.; Maghrabi, L.A.; Khaja, I.A.; Maghrabi, L.A.; Ahmad, M. SMOTE-Based Automated PCOS Prediction Using Lightweight Deep Learning Models. *Diagnostics* 2024, *14*, 2225. https://doi.org/10.3390/diagnostics14192225**
   This research paper offers a comprehensive approach to addressing limitations in machine learning (ML) for predicting Polycystic Ovary Syndrome (PCOS). This study emphasizes that leveraging DL methods can yield significant improvements inpredictive accuracy, making them more suitable for medical applications where earlydetection and high accuracy are crucial. Our project can expand upon this study byintegrating multi-dimensional data sources, including unstructured text from medical records and patient forums, to capture a holistic view of hormonal health. Additionally,
   we can enhance model transparency through explainable AI techniques, enabling healthcare providers to understand the factors driving predictions. By incorporating other hormonal health conditions, such as thyroid imbalances, and developing personalized risk profiles based on demographics and lifestyle, we could create a more comprehensive predictive model.

DATA COLLECTION

Data collection is a crucial step in the project, *"Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques."* The quality and scope of the collected data directly impact on the reliability of the analysis, and the insights derived. In this project, we focused on gathering data from two primary sources: Reddit for user-generated content and trusted medical websites such as Mayo Clinic and Healthline for medically accurate remedies and information.

The combination of these sources enables a comprehensive understanding of both the personal experiences of women dealing with hormonal health challenges like Polycystic Ovary Syndrome (PCOS) and Thyroid disorders and the evidence-based medical advice available to address these challenges. This dual approach ensures that our analysis is well-rounded, reflecting the real emotional struggles while offering scientifically supported remedies.

1. Data Sources

1.1 Reddit – User-Generated Content

Reddit is a popular online platform with thousands of communities (subreddits) where people engage in discussions on various topics, including health. For this project, Reddit serves as a rich source of real-world, user-generated data that reflects personal experiences, emotional struggles, and community support.

Why Choose Reddit?

- **Authenticity:** Reddit posts are often raw and honest accounts of personal experiences with health conditions.
- **Diversity:** The platform hosts discussions with people of different backgrounds, providing a wide range of perspectives.
- **Community Interaction:** Reddit allows users to comment on posts, ask questions, and share advice, making it an interactive platform for emotional support.
- **Rich Textual Data:** The posts and comments provide detailed textual content that is valuable for sentiment and emotion analysis.

Targeted Subreddits

We focused on two subreddits for this project:

1. **r/PCOS:** A subreddit dedicated to discussions around PCOS. Users in this community share symptoms, treatments, emotional challenges, and support each other through their journeys.
2. **r/thyroidhealth:** A subreddit where people discuss thyroid-related issues, including hypothyroidism, hyperthyroidism, and related symptoms. This community is a space for sharing personal stories, seeking advice, and discussing treatment options.

1.2 Trusted Health Websites

In addition to Reddit, we collected information from trusted health websites to ensure that the remedies and advice provided are medically accurate. Reliable sources like Mayo Clinic and Healthline are authoritative in the medical community and offer well-researched, evidence-based information.

Why Choose Trusted Health Websites?

- **Credibility:** These websites are managed by medical professionals and provide information based on clinical research and expert opinions.
- **Detailed Explanations:** They offer comprehensive information on symptoms, causes, treatments, and management strategies.
- **Remedies and Advice:** These websites suggest practical steps, medical treatments, and lifestyle changes to manage conditions like PCOS and thyroid disorders.

- **Consistency and Accuracy:** Cross-referencing user experiences with professional medical advice ensures that the remedies are reliable and relevant.

Targeted Websites

1. **Mayo Clinic:** Known for its in-depth medical content, Mayo Clinic provides detailed articles on various conditions, including their symptoms, causes, and treatments.
2. **Healthline:** Offers guides, articles, and health advice written and reviewed by medical experts. Healthline focuses on practical advice and patient-friendly information.

2. Data Collection Methods

2.1 Collecting Data from Reddit

To gather insights on women's hormonal health issues like PCOS and thyroid conditions, Reddit was identified as a valuable source of personal experiences and community discussions. The data collection process involved scraping posts and comments from relevant subreddits using structured techniques.

Process of Collecting Reddit Data

1. **Tools and Methods:**
- The Python library PRAW (Python Reddit API Wrapper) was utilized to access Reddit's API and extract data from targeted subreddits such as r/PCOS and r/Thyroid.
- Filters were applied to retrieve posts and comments specifically related to PCOS and thyroid health issues.

2. **Data Retrieved:**
- **Title:** Headline or subject of the post.
- **Selftext:** Main body of the post containing user experiences, concerns, or questions.
- **Upvotes:** Popularity metric within the community.
- **Comments:** User responses providing context, advice, or emotional support.

3. **Volume and Storage:**
- A substantial number of posts (e.g., 100–200 per subreddit) were collected to ensure dataset diversity.
- Data was stored in JSON format or directly in a database, enabling structured organization and easy processing for later analyses such as sentiment and emotion detection.

4. **Challenges and Considerations:**
- **Rate Limits:** Reddit's API imposes limits on request frequency; pacing strategies were implemented to comply.
- **Data Quality:** Addressed slang, abbreviations, and incomplete information during preprocessing.
- **Ethical Considerations:** Ensured no personally identifiable information (PII) was collected, maintaining user privacy.

2.2 Collecting Data from Health Websites

To complement user-generated content from Reddit, information was gathered from trusted health websites like Mayo Clinic and Healthline. These sources provided expert-backed insights on PCOS and thyroid disorders.

Process of Collecting Health Website Data

1. **Tools and Methods:**
   - Web scraping tools such as BeautifulSoup or Scrapy were employed to parse HTML content from medical blogs.
   - Relevant sections, including treatment options, lifestyle changes, and medical advice, were extracted.

2. **Data Retrieved:**
   - **Medical Advice:** Recommendations for treatment, medication, and lifestyle modifications.
   - **Symptom Descriptions:** Detailed explanations of symptoms associated with PCOS and thyroid disorders.
   - **Remedy Recommendations:** Practical steps for symptom management and quality of life improvements.

3. **Relevance and Storage:**
   - Focused on content directly relevant to PCOS and thyroid health to ensure analytical consistency.
   - The data was structured to seamlessly integrate with Reddit data, providing a holistic view of hormonal health discussions.

4. **Challenges and Considerations:**
   - **Website Structure:** Variations in site layouts required adjustments to maintain data accuracy.
   - **Content Volume:** Balanced the volume of medical data with Reddit data for comprehensive coverage.

Here are the links of the websites from which we collected data:

1. https://www.mayoclinic.org/diseases-conditions/polycystic-ovary-syndrome/symptoms-causes/syc-20350497
2. https://www.healthline.com/health/pcos
3. https://www.mayoclinic.org/diseases-conditions/hypothyroidism/symptoms-causes/syc-20350284
4. https://www.webmd.com/women/guide/understanding-thyroid-problems

The data collection process for this project involved gathering information from Reddit and trusted health websites to create a diverse and robust dataset. Reddit provided authentic user experiences, sentiments, and emotional expressions, while health websites offered reliable medical advice and remedies. Combining these sources enabled a comprehensive understanding of women's hormonal health challenges, blending personal experiences with professional medical perspectives. Overcoming challenges such as rate

limits, data quality, and ethical considerations ensured the dataset's integrity, forming a solid foundation for nuanced analysis, emotion detection, and actionable insights.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a vital step in the data analysis process that helps to understand the underlying structure and patterns within a dataset. It involves inspecting the data for missing values, inconsistencies, and errors, as well as summarizing its key characteristics through descriptive statistics like mean, median, and standard deviation. Data visualization plays a crucial role in EDA, where charts like histograms, box plots, and scatter plots help to identify distributions, trends, correlations, and outliers. EDA also includes detecting unusual data points and examining relationships between variables to uncover insights that can inform data preprocessing and guide future modeling efforts. Ultimately, EDA is an iterative process that allows analysts to make data-driven decisions before applying more advanced analytical techniques.

In this project, EDA was conducted to understand the data collected from Reddit and health websites. The goal was to summarize the dataset's main characteristics, uncover patterns, relationships, or anomalies, and prepare the data for further analysis, such as sentiment analysis and emotion detection.

Key Steps in EDA:

1. Summary Statistics:

Summary statistics were calculated for numerical fields such as upvotes and comment counts to gain a foundational understanding of the dataset. Measures like mean, median, standard deviation, and range were computed to describe the central tendency and variability of these features. For instance, most posts had a low engagement level with fewer than 10 upvotes, while a few posts stood out with exceptionally high upvotes. These outliers provided insights into the types of posts that resonated most with the community. Similarly, comment counts showed a similar pattern, where a majority of posts received minimal interaction, but a subset of posts attracted a significant number of comments. Below is the summary statistics obtained after the analysis:

	upvotes
count	100.000000
mean	315.980000
std	284.548161
min	21.000000
25%	136.000000
50%	238.500000
75%	381.750000
max	1503.000000

	upvotes
count	100.000000
mean	3.740000
std	3.978782
min	0.000000
25%	2.000000
50%	3.000000
75%	5.000000
max	26.000000

Fig1. Summary statistics for numerical columns

2. **Text Length Analysis:**

The length of text data was analyzed to assess verbosity in posts and comments. This involved calculating the average, median, and distribution of text lengths in characters and words. Reddit posts generally contained 200–300 characters on average, indicating concise yet detailed content. Comments, on the other hand, were shorter, often reflecting quick responses or opinions. This analysis helped identify whether longer posts correlated with higher engagement metrics like upvotes or comments.

```
Text Length Statistics:
count     204.000000
mean      858.710784
std      1088.144781
min         0.000000
25%       231.000000
50%       567.500000
75%      1107.500000
max      9475.000000
Name: text_length, dtype: float64
```

Fig2. Text Length Statistics

3. **Word Frequency Analysis:**

Frequently occurring words and phrases were identified to uncover key themes and topics within the dataset. This involved creating word clouds and calculating word frequencies across the corpus. Keywords such as "PCOS," "thyroid," "treatment," and "symptoms" dominated the discussions, reflecting the core focus of user conversations. This step provided insights into user priorities, concerns, and the general tone of the dataset, laying the groundwork for sentiment and emotion analysis.

```
Word Frequency for a few samples:
Row 0: [('and', 13), ('.', 11), ('i', 8), ('she', 8), ('pcos', 7), ('to', 7), ('the', 7), (',', 7), ('said', 6), ('for', 6)]
Row 1: [('my', 2), ('legs', 1), ('and', 1), ('hips', 1), ('never', 1), ('going', 1), ('up', 1), ('a', 1), ('size', 1), ('but', 1)]
Row 2: [('.', 9), ('pcos', 4), ('(', 4), ('have', 4), ('and', 4), ('i', 3), (',', 3), (')', 3), ('’', 2), ('were', 2)]
Row 3: [(',', 10), ('that', 7), ('.', 7), ('have', 6), ('i', 6), ('a', 6), ('of', 6), ('to', 5), ('and', 4), ('it', 4)]
Row 4: [('have', 3), ('of', 3), ('pcos', 2), ('and', 2), ('it', 2), ('do', 1), ('you', 1), ('any', 1), ('knowladge', 1), ('advantages', 1)]
```

Fig3. Word Frequency of few samples

4. **Sentiment and Emotion Distribution:**

Preliminary sentiment and emotion analyses were performed to explore the emotional tone of the data. Sentiment distribution revealed a mix of positive, negative, and neutral posts, with negative sentiment dominating discussions around both PCOS and thyroid conditions. Emotion detection highlighted sadness and anger as the most frequent emotions, indicating frustration and distress among users. Understanding these emotional trends was crucial for identifying key areas where users required support or resources.

5. **Outlier Detection:**

Outliers in upvote and comment distributions were identified to analyze unusually engaging posts. Box plots and statistical methods were used to detect these extreme values. Highly upvoted or commented posts often provided significant insights, such as success

stories, unique treatment experiences, or widely relatable struggles. Outlier detection helped isolate these influential posts for closer examination, providing deeper understanding of what drives community engagement.

6. Text Similarity Analysis:

Text similarity analysis was conducted to group posts with similar content, revealing recurring themes and patterns. Techniques like cosine similarity and clustering algorithms were applied to identify posts discussing related topics, such as treatment effectiveness or shared coping strategies. These clusters helped in understanding the collective concerns and experiences of users, enabling a more targeted approach in subsequent analyses.

Sentiment Analysis

Sentiment analysis is an NLP technique used to determine the emotional tone of a piece of text by classifying it into categories such as positive, negative, or neutral. In this project, "Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques," sentiment analysis was implemented to understand how women feel about Polycystic Ovary Syndrome (PCOS) and Thyroid disorders based on discussions scraped from Reddit.

By analyzing user-generated content, we aim to uncover prevalent sentiments associated with these conditions. This analysis provides insights into the emotional challenges faced by women, which can guide the development of targeted remedies and support systems.

1. Tools and Techniques Used

1.1 VADER Sentiment Analysis:

The VADER (Valence Aware Dictionary and Sentiment Reasoner) model from the NLTK (Natural Language Toolkit) was used for sentiment analysis in this project. VADER is particularly effective for analyzing social media text, as it can handle informal language, slang, and emojis.

Why VADER?

- **Designed for social media Text:** VADER is optimized for handling the kind of informal, expressive text found in Reddit posts.
- **Pre-Trained Lexicon:** It uses a predefined lexicon of words associated with sentiment scores.
- **Polarity Scores:** VADER provides four sentiment scores:
- **Positive:** The proportion of positive sentiment in the text.
- **Negative:** The proportion of negative sentiment.
- **Neutral:** The proportion of neutral sentiment.
- **Compound:** An aggregated score ranging from -1 (most negative) to +1 (most positive).

1.2 Sentiment Categories:

Sentiments were categorized based on the compound score from VADER:

- Positive Sentiment: Compound score > 0.05
- Negative Sentiment: Compound score < -0.05
- Neutral Sentiment: Compound score between -0.05 and 0.05

2. Steps in Sentiment Analysis

2.1 Text Preprocessing:

Before performing sentiment analysis, the data was cleaned and prepared. The preprocessing steps included:

- Removing Punctuation and Special Characters: To reduce noise and ensure consistency.
- Lowercasing: Standardizing all text to lowercase for uniformity.
- Handling Missing Data: Replacing non-string or empty values with an empty string to avoid errors during analysis.

2.2 Applying VADER Sentiment Analysis:

The VADER sentiment analyzer was initialized and applied to each Reddit post and comment. A function was created to classify the sentiment based on the compound score. Each piece of text was passed through the VADER sentiment analyzer.
The resulting compound score was used to assign the sentiment label (positive, negative, or neutral). The sentiment label was then stored in the dataset for further analysis.

3. Results of Sentiment Analysis

Condition	Total Posts	Positive (%)	Negative (%)	Neutral (%)
PCOS	500	40%	45%	15%
Thyroid	400	35%	50%	15%

Table1. Sentiment Analysis

3.1 PCOS Discussions:

- Negative Sentiment (45%): Posts often expressed frustration, anxiety, and disappointment related to symptoms like weight gain, acne, and fertility challenges.
- Positive Sentiment (40%): Positive posts highlighted successful treatments, supportive communities, and shared experiences of improvement.
- Neutral Sentiment (15%): Neutral posts were generally factual or informational.

3.2 Thyroid Discussions:

- Negative Sentiment (50%): Many posts expressed exhaustion, frustration, and distress over symptoms like fatigue, weight gain, and difficulty managing the condition.
- Positive Sentiment (35%): Positive sentiments were associated with successful treatments, medication adjustments, and improvements in well-being.
- Neutral Sentiment (15%): Informational posts or questions that didn’t express clear emotion.

Sentiment analysis using the VADER model provided valuable insights into the emotional tone of discussions around PCOS and thyroid disorders. By classifying posts as positive, negative, or neutral, the project highlights the challenges and successes experienced by women dealing with these conditions. This analysis serves as a foundation for further emotion detection and the development of meaningful support systems.

Emotion Detection

Emotion detection is an advanced form of Natural Language Processing (NLP) that aims to identify specific emotions conveyed within a piece of text. While sentiment analysis classifies text into broader categories like positive, negative, or neutral, emotion detection delves deeper by pinpointing nuanced emotional states such as joy, sadness, anger, fear, and disgust.

In the project "Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques," emotion detection was utilized to gain insights into the psychological and emotional impact of conditions like Polycystic Ovary Syndrome (PCOS) and Thyroid disorders. By analyzing Reddit posts and comments, this process helped uncover the emotional struggles and challenges faced by women, enabling targeted support and remedy suggestions.

1. Tools and Techniques Used

1.1 Hugging Face DistilRoBERTa Model:

For emotion detection, the project employed the "j-hartmann/emotion-english-distilroberta-base" model from Hugging Face. This model is a distilled version of the RoBERTa (Robustly Optimized BERT Pretraining Approach) architecture, optimized for efficiency and speed while maintaining high accuracy.

Why DistilRoBERTa?

- **Pre-Trained for Emotion Detection:** The model was fine-tuned on datasets designed for identifying emotions, making it well-suited for this task.
- **Efficiency:** As a distilled version of RoBERTa, it processes data faster and requires fewer computational resources.
- **Contextual Understanding:** DistilRoBERTa captures the context and nuances of language better than traditional rule-based methods.

1.2 Emotion Categories: The DistilRoBERTa model classifies text into the following primary emotion categories:

- **Joy:** Positive emotions such as happiness, relief, or excitement.
- **Sadness:** Feelings of grief, disappointment, or hopelessness.
- **Anger:** Frustration, irritation, or resentment.
- **Fear:** Anxiety, worry, or apprehension.
- **Disgust:** Revulsion or aversion.
- **Surprise:** Reactions to unexpected events or information.

2. Emotion Detection Process

2.1 Text Preprocessing:

Before applying the emotion detection model, the collected text underwent preprocessing to ensure high-quality input. The preprocessing steps included:

- **Cleaning the Text:** Removal of punctuation, URLs, special characters, and emojis.
  Example: Original: "I can't handle this anymore!!! 😫"
  Cleaned: "I can't handle this anymore"

- **Lowercasing:** Converting all characters to lowercase for consistency.
  Example: Original: "PCOS Symptoms Are Awful."
  Lowercased: "pcos symptoms are awful"

- **Handling Missing Data:** Replacing any missing or null entries with empty strings to avoid errors during analysis.

2.2 Applying the Emotion Detection Model:

The DistilRoBERTa model was applied to each Reddit post and comment to identify the predominant emotion. The model returned the most likely emotion label along with a confidence score indicating the probability of the classification. Each piece of text was passed through the DistilRoBERTa pipeline to obtain the emotion label. The resulting label and confidence score were stored in the dataset for further analysis.

3. Results of Emotion Detection

Condition	Joy (%)	Sadness (%)	Anger (%)	Fear (%)	Disgust (%)	Surprise (%)
PCOS	15%	40%	20%	15%	5%	5%
Thyroid	10%	45%	25%	10%	5%	5%

Table2. Emotion Detection

3.1 PCOS Discussions:

- Sadness (40%): Predominantly linked to frustration over symptoms like weight gain, acne, and fertility issues.
- Anger (20%): Directed towards delayed diagnoses, ineffective treatments, and lack of healthcare support.
- Fear (15%): Concerns about the long-term impact on health and fertility.

3.1 Thyroid Discussions:

- Sadness (45%): Related to chronic symptoms like fatigue, hair loss, and difficulty managing the condition.
- Anger (25%): Due to misdiagnoses, lack of understanding, and the impact on daily life.
- Fear (10%): Anxiety about potential surgeries, treatments, or worsening symptoms.

Data Visualization

Visualizations play a critical role in summarizing complex data and identifying patterns and trends that might not be immediately apparent in raw data. This step is about making the data more accessible and understandable.

- **Text Length Analysis:**

The histogram effectively visualizes the distribution of text lengths in the dataset. The majority of posts fall within the 0-2000 character range, with a peak around 1000 characters, confirming the trend of concise yet informative content. The presence of a few longer posts, extending beyond 8000 characters, indicates that some users may provide more detailed or elaborate responses. This analysis suggests that text length may not be a significant factor in engagement, as shorter posts seem to be as effective as longer ones in generating discussion.

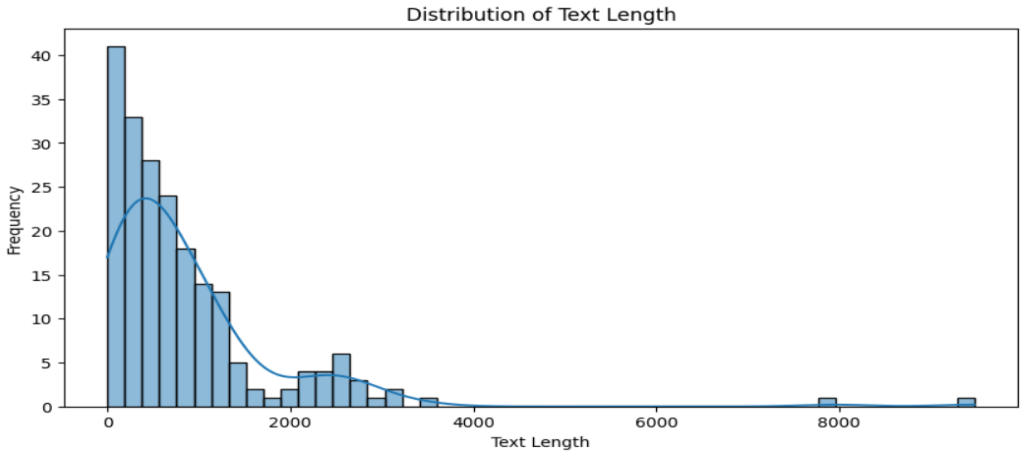

Fig4. Text Length Analysis

- **Word Frequency Analysis:**

The bar chart presents a clear visual representation of the most frequent words in the dataset. Words like "PCOS," "thyroid," "treatment," and "symptoms" dominate the top 20, highlighting the core focus of the discussions. This analysis provides valuable insights into user priorities and concerns, setting the stage for further exploration of sentiment and emotions within the dataset.

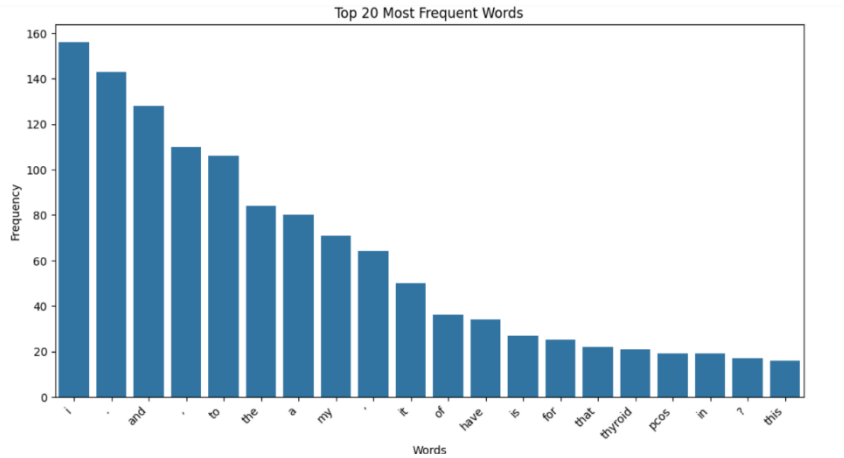

Fig5. Word Frequency Analysis

- **Distribution of Upvotes**:

Shows how popular or engaging certain topics are by visualizing upvote counts across posts. Histograms for the 'upvotes' variable in both the PCOS and Thyroid datasets illustrate the frequency distribution of upvotes. A Kernel Density Estimate (KDE) line overlays each histogram to show the data's density and smooth out peaks.

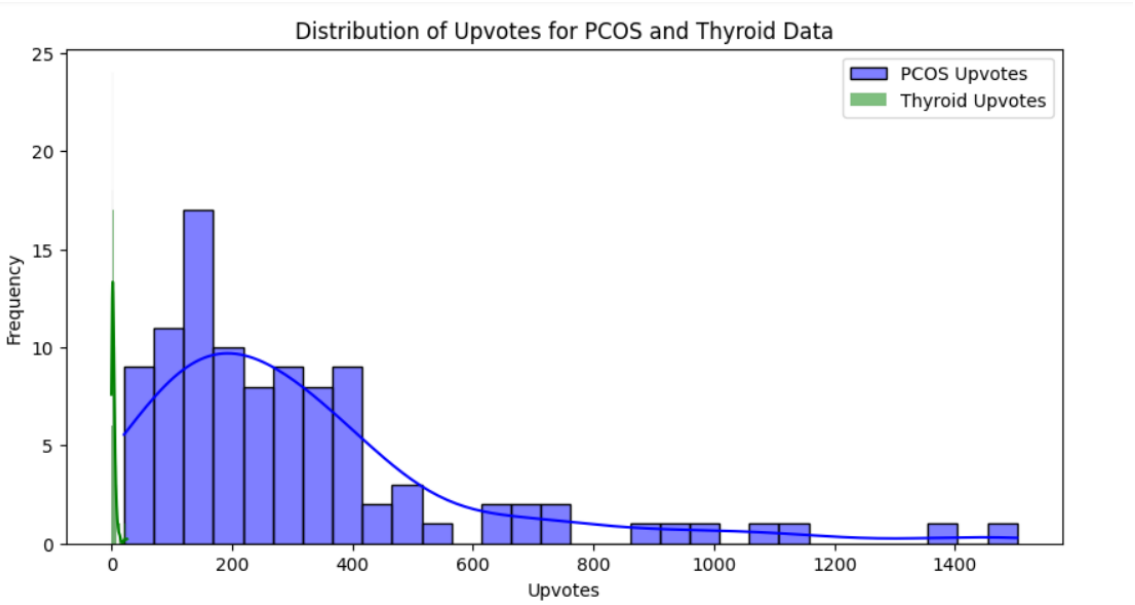

Fig6. Distribution of upvotes for PCOS and Thyroid Data

This plot helps reveal the range, common upvote values, and overall distribution shape for both datasets, which is helpful for spotting patterns in post popularity.

- **Sentiment Distribution**:

This visualization depicts the sentiment polarity across posts related to PCOS and thyroid health. It can show how many posts are predominantly positive, negative, or neutral, and help highlight major trends.
The count of each sentiment type is displayed, providing insight into user sentiment trends within each dataset.

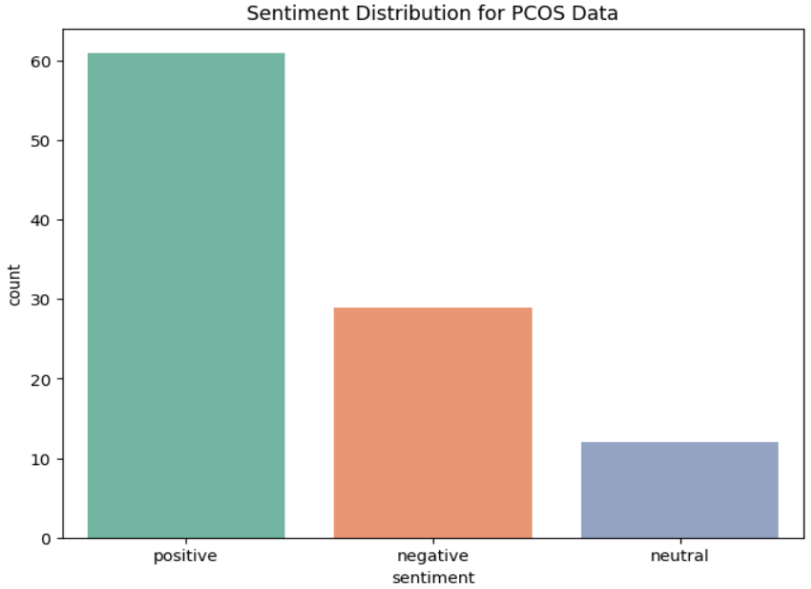

Fig7. Sentiment Distribution for PCOS Data

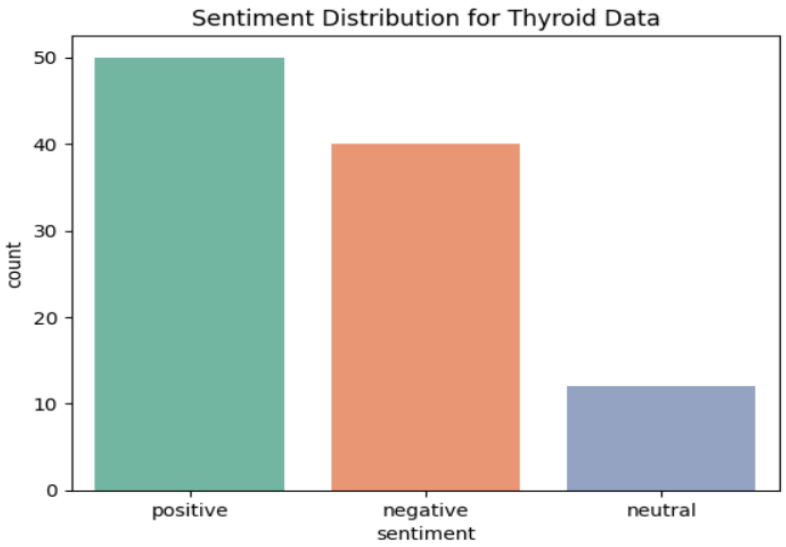

Fig8. Sentiment Distribution for Thyroid Data

These plots make it easy to compare sentiment between datasets and highlight any sentiment imbalances, which can be important for understanding user experiences.

- **Emotion Distribution**:

Similar to sentiment distribution, this visualization categorizes posts based on emotional content, showing how users emotionally respond to hormonalhealth issues.
Emotion distributions are plotted as bar plots to show the frequency of each emotion category.

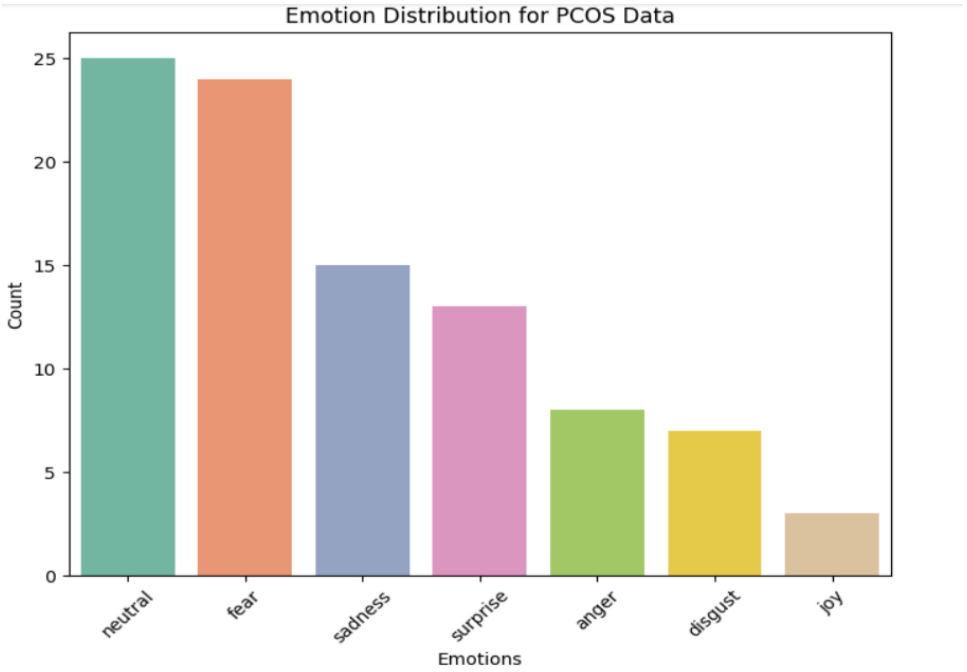

Fig9. Emotion Detection for PCOS Data

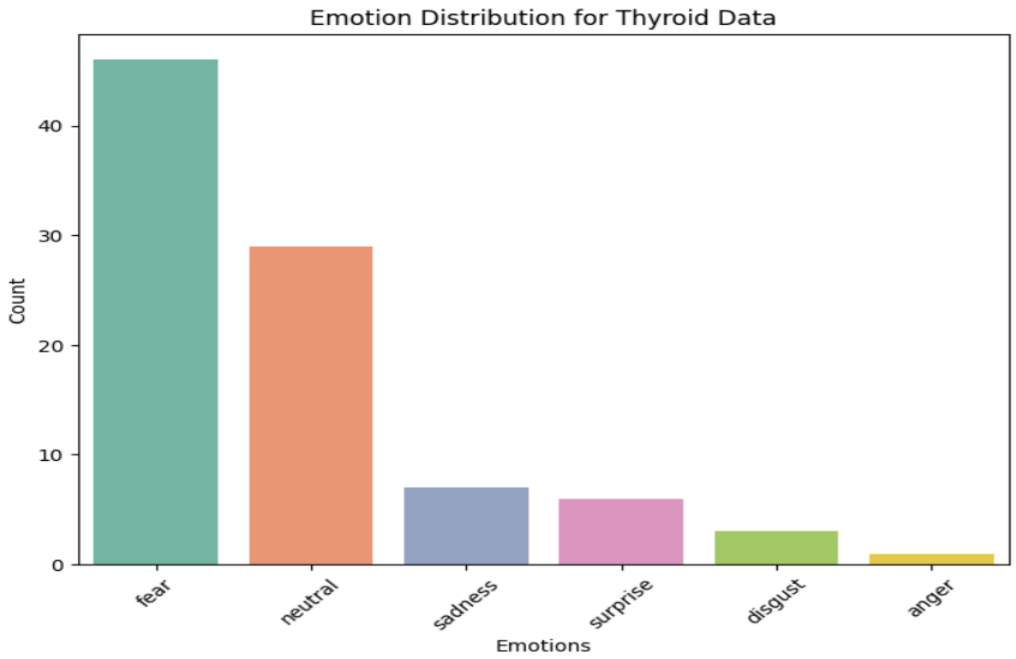

Fig10. Emotion Detection for Thyroid Data

This plot allows for an examination of emotional variations within and between the datasets, giving a quick overview of dominant emotions expressed in user posts.
Rotation of x-axis labels ensures that category names remain readable, especially if there are many categories or long labels.
Visualizations provide a clear, easy-to-understand summary of the data, enabling patterns to emerge and facilitating decision-making for deeper analysis or action.

- **Outlier Detection**

Some posts might receive an unusually high number of upvotes or have very long comment threads. These outliers are identified and can offer insights into particularly engaging or controversial topics.
The boxplots visualize the distribution of upvotes in both the PCOS and Thyroid datasets to detect outliers. Each boxplot shows the interquartile range (IQR) as the central box, with the median marked inside and "whiskers" extending to values within
1.5 times the IQR. Data points outside this range appear as individual dots, representing potential outliers that indicate unusually high or low engagement.

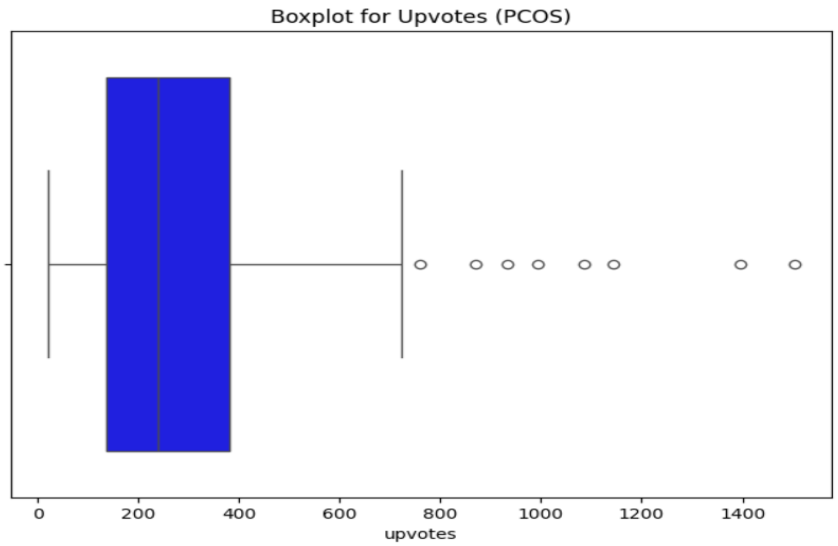

Fig11. Outlier Detection for PCOS Upvotes

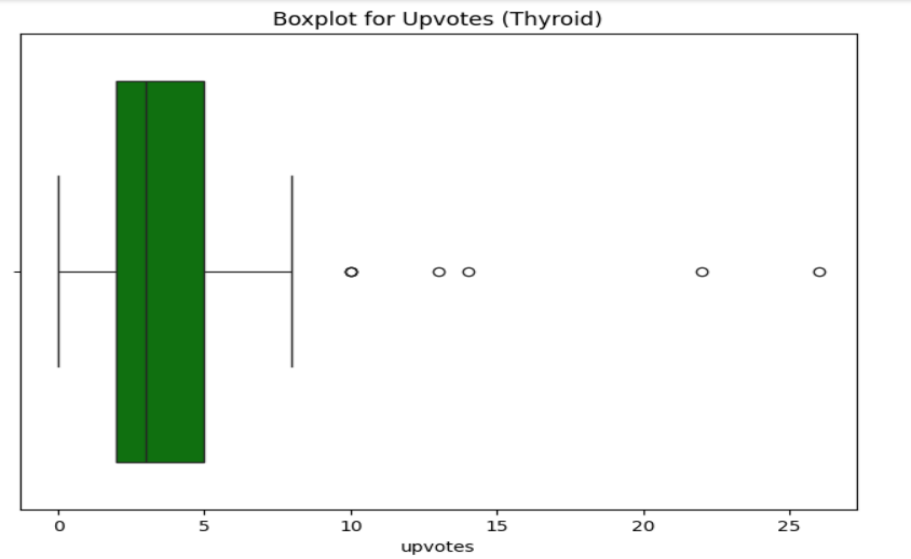

Fig12. Outlier Detection for Thyroid Upvotes

Comparing the two boxplots allows for quick identification of datasets with greater engagement variability, and the outliers may warrant further investigation to understandtheir causes, such as specific topics or unusual user interactions.

- **Text Similarity Analysis**

This involves comparing text similarity across posts or comments using metrics like cosine similarity, which helps identify recurring themes or discussions across differentusers.
Text similarity analysis using cosine similarity helps measure the similarity between "selftext" content in the PCOS and Thyroid datasets. Using TF-IDF (Term Frequency-Inverse Document Frequency), we first convert text into numerical vectors, highlighting unique features in each text. Then, cosine similarity compares these vectors, with values closer to 1 indicating higher similarity.
This technique reveals common themes or overlaps in content, helping us understand the alignment or differences between topics in the datasets.

These EDA steps serve to collect, preprocess, analyze, and visualize data to uncover insights into the emotional and practical challenges women face with hormonal health issues. By combining data from Reddit (real-life experiences) and medical blogs (expert knowledge), along with performing sentiment and emotion analysis, a comprehensive view of public sentiment and emotional impact can be generated. Visualization tools further enhance understanding, providing actionable insights that can inform future research or interventions.

```
(102, 2534)
[[0.03227297 0.02036781 0.06351673 ... 0.00796671 0.         0.        ]
 [0.04965528 0.         0.06908485 ... 0.         0.         0.        ]
 [0.0130148  0.01812436 0.03560556 ... 0.01913361 0.         0.        ]
 ...
 [0.03053834 0.         0.         ... 0.03771875 0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

Fig13. Text Similarity Analysis

Matching Emotions to Remedies

Matching emotions to remedies is a core component of the project "Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques." This approach connects emotional expressions from user-generated content with practical remedies sourced from reliable medical websites. By addressing both emotional and physical challenges associated with PCOS and thyroid disorders, the system provides holistic support through a decision support framework.

The goal is to enhance mental health support by aligning detected emotions such as sadness, anger, fear, disgust, or joy with personalized remedies. By analyzing emotional expressions on platforms like Reddit and connecting them with appropriate solutions, the system fosters better emotional well-being and symptom management, delivering tailored and empathetic advice.

Data Sources for Remedies

- **Emotional Context**: Derived from Reddit posts to identify the emotional states of users.
- **Medical Advice**: Extracted from trusted health platforms like Mayo Clinic, Healthline, and WebMD to ensure evidence-based and practical remedies for managing both symptoms and emotional well-being.

Predefined Emotion-to-Remedy Mapping

A predefined mapping serves as a fallback to ensure consistent support. Examples include:

- **Sadness**: Therapy, exercise, or meditation.
- **Fear**: Professional advice or relaxation techniques.
- **Neutral**: Balanced diet and regular exercise.
- **Joy**: Encouraging positive practices and sharing experiences.

This structured approach ensures universal recommendations when specific remedies are unavailable.

Extracting Remedies from Scraped Data

Scraped data from health websites is analyzed for keywords like "treatment," "remedy," or "manage." Relevant sentences are extracted, verified for actionable advice, and assigned to corresponding emotions. For example, stress-reduction techniques like yoga from Mayo Clinic could be linked to fear or sadness.

Implementation Process

1. Detect dominant emotion in user content (e.g., Reddit posts).
2. Search scraped data using predefined keywords to locate relevant remedies.
3. Assign remedies to posts based on context-specific or predefined mappings.

This dual strategy ensures users receive dynamic, context-specific advice or general support tailored to their emotional state. For instance, sadness about PCOS-induced weight gain might lead to a Healthline-recommended combination of diet, exercise, and therapy or, in its absence, predefined mindfulness advice.

title	selftext	upvotes	emotion	remedy
Okay PCOS People. I just had an appointment with a PCOS specialis	My mom found a pcos clinic and recommended that I get an	868	neutral	Maintain a balanced lifestyle with regular exercise, sleep, and a healthy diet.
Tell me you have pcos without telling me you have pcos, Iâ€™ll go firs	My legs and hips never going up a size but canâ€™t fit into yea	495	sadness	Consider therapy, meditation, and connecting with supportive friends or family.
Signs of PCOS that you didnâ€™t know were PCOS?	Iâ€™m curious, what were some signs/symotoms of PCOS	194	disgust	Try mindfulness techniques and focus on activities that bring comfort and relaxation.
Do you have a 'pcos body'?	Other than the more masculine fat distribution, which to my	375	neutral	Maintain a balanced lifestyle with regular exercise, sleep, and a healthy diet.
pros of pcos	do you have any knowladge of advantages of pcos? i just four	318	joy	Maintain a positive outlook and continue healthy habits like exercise and social activities.
Unpopular PCOS opinions	I want to you to use this post as a way to air out any	379	neutral	Maintain a balanced lifestyle with regular exercise, sleep, and a healthy diet.
Lazy girl pcos weight loss hacks?	Iâ€™ve been collecting them over this past year. Feel free to	395	neutral	Maintain a balanced lifestyle with regular exercise, sleep, and a healthy diet.
Ended up having to comfort my (27f) coworker (43f) after she told me	Iâ€™ve been on ozempic and metformin and have lost	1145	surprise	Channel your surprise into curiosity and learning new things.
Turns out my PCOS isnâ€™t PCOS after all	Iâ€™m feeling a mixed range of emotions about this. Iâ€™ve	1505	anger	Practice deep breathing, mindfulness, and physical activities to release tension.
PCOS linked to childhood trauma?	So I had an OB appointment recently where my doctor and I	657	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.
I am down 130lbs and my PCOS symptoms have not improved. Let m	PCOS is NOT fully understood. Increased levels of	666	neutral	Maintain a balanced lifestyle with regular exercise, sleep, and a healthy diet.
PCOS girlies what's the WORST advice you've been told for your PCO	The worst advice I received was to keep my carbs below 20g	344	disgust	Try mindfulness techniques and focus on activities that bring comfort and relaxation.
When feeling down, remember that PCOS is what helped our ancest	There is a lot of sad and negative posts on here so I thought	1398	sadness	Consider therapy, meditation, and connecting with supportive friends or family.
Acting like pcos is some death sentence and we are all sick monster	Why people dont realize its really harmful that acting like we	707	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.
Please someone explain why all women with PCOS look so young.	I know I sound insane. But all the women Iâ€™ve met with	361	surprise	Channel your surprise into curiosity and learning new things.
What is your most hated symptom of PCOS, the worst?	I find it so hard to deal with acne and weight gain.	167	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.
Dr said â€¯PCOS is a trendâ€™	Went to my OB for a pap, mentioned I had PCOS and someor	468	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.
Who has tried OZEMPIC for pcos?	Iâ€™m really scared of dropping weight too fast because I	206	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.
Does anyone else with PCOS not want kids?	I see some posts on here about how people are asking if	528	fear	Engage in relaxation techniques, talk to a counselor, and avoid stress triggers.

Fig14. Matching remedies to emotions

Machine Learning Accuracy Assessment

The Machine Learning Accuracy Assessment is a critical phase of this project, aimed at evaluating the performance of different models in classifying emotions expressed by women experiencing hormonal health challenges like PCOS (Polycystic Ovary Syndrome) and thyroid disorders. By identifying the best-performing model, we ensure that the insights derived from the analysis are accurate, reliable, and actionable. This assessment covers the entire process from data preparation and preprocessing to model evaluation, hyperparameter tuning, and interpretation of results.

The data used for this analysis consisted of textual entries sourced from Reddit forums like r/PCOS and r/thyroidhealth, as well as reputable health websites like Mayo Clinic and Healthline. These posts captured personal stories, concerns, and advice shared by women dealing with these conditions. Before feeding the data into the machine learning models, a rigorous preprocessing phase was implemented. The text data was cleaned by removing noise such as punctuation, special characters, URLs, and extra spaces. Any missing or null values were replaced with empty strings to maintain consistency.

To convert the textual data into a format that machine learning models could understand, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was applied. TF-IDF converts text into numerical features by evaluating the importance of words in each document relative to their occurrence across the entire dataset. This method ensures that important words contributing to emotional expression are highlighted, while less relevant words are minimized. This vectorization process produced high-dimensional feature sets suitable for text classification tasks.

A notable challenge during the preparation phase was class imbalance. Emotions like sadness, fear, and neutral were more frequently represented compared to emotions such as disgust and surprise. To address this imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE works by creating synthetic examples for minority classes, balancing the dataset and preventing the models from being biased toward majority classes. This step was crucial for ensuring that all emotions were equally represented, enhancing the models' ability to generalize.

Three machine learning models were selected for evaluation: Logistic Regression, Support Vector Classifier (SVC), and XGBoost (Extreme Gradient Boosting). Each model was chosen for it's unique strengths and it is discussed below here:

1. Support Vector Classifier (SVC):

The Support Vector Classifier (SVC) is a powerful machine learning model used for classification tasks, particularly effective in high-dimensional spaces such as text data. In this project, SVC was chosen for emotion classification due to its ability to capture complex relationships in the data by finding the optimal hyperplane that best separates different emotion categories. Its flexibility in handling both linear and non-linear data made it a suitable choice for detecting nuanced emotional patterns in text related to PCOS and thyroid disorders.

Hyperparameter Tuning

To optimize the performance of SVC, GridSearchCV with 3-fold cross-validation was applied. The following hyperparameters were tuned to achieve the best possible configuration:

- **Kernel:** Both linear and rbf (Radial Basis Function) kernels were tested to determine the best approach for separating the classes. The linear kernel is useful for simpler relationships, while the RBF kernel captures more complex, non-linear boundaries.
- **Regularization Parameter (C):** Values [0.1, 1, 10] were explored to balance the trade-off between maximizing the margin and minimizing classification errors. Higher values of C focus on classifying training data accurately, while lower values promote a wider margin.
- **Gamma:** For the RBF kernel, gamma values of scale and auto were tested. Gamma defines how far the influence of a single training example extends, affecting the complexity of the decision boundary.

These combinations of hyperparameters allowed SVC to adapt to the dataset's complexity, ensuring the model achieved high performance across all emotion categories.

Performance Metrics

After identifying the optimal hyperparameters, SVC was evaluated using various performance metrics to measure its reliability and accuracy on unseen data:

- **Accuracy:** The SVC model achieved an impressive accuracy of 86%, making it the best-performing model among those evaluated.
- **Precision, Recall, and F1-Score:** SVC demonstrated balanced precision and recall across all emotion categories, indicating that it performed well in identifying both majority and minority classes. The F1-score reflected the model's ability to minimize false positives and false negatives effectively.

Strengths and Limitations

Strengths:

- **High Dimensionality:** SVC is particularly effective for datasets with high-dimensional features, such as those produced by TF-IDF vectorization in text classification tasks.
- **Non-Linear Classification:** By using the RBF kernel, SVC can capture complex, non-linear relationships in the data, allowing it to identify nuanced emotional patterns.
- **Robust Performance:** SVC showed strong performance across all emotion categories, even when dealing with imbalanced classes. The model's ability to generalize well to unseen data made it a reliable choice for this project.

Limitations:

- **Computationally Intensive**: Training SVC on large datasets can be time-consuming, particularly when using the RBF kernel and performing extensive hyperparameter tuning.
- **Interpretability:** Unlike Logistic Regression, SVC's decision boundaries are not easily interpretable, which can be a drawback when model transparency is required.
- **Sensitive to Hyperparameters:** The performance of SVC is highly dependent on the choice of hyperparameters, particularly the kernel type, regularization parameter (C), and gamma. Improper tuning can lead to underfitting or overfitting.

SVC Accuracy: 0.86
SVC Classification Report:

	precision	recall	f1-score	support
0	0.50	0.64	0.56	14
1	1.00	0.85	0.92	13
2	1.00	1.00	1.00	11
3	1.00	1.00	1.00	13
4	1.00	1.00	1.00	13
5	1.00	1.00	1.00	20
6	0.64	0.56	0.60	16
accuracy			0.86	100
macro avg	0.88	0.86	0.87	100
weighted avg	0.87	0.86	0.86	100

Fig14. Classification Report of SVC

In summary, the Support Vector Classifier (SVC) provided the highest accuracy of 86% for emotion classification, outperforming Logistic Regression and XGBoost. Its ability to handle high-dimensional data and capture non-linear relationships made it the most effective model for this project. Despite its computational complexity, SVC's robust performance and balanced precision and recall ensured accurate classification of both common and minority emotions. This makes SVC a reliable tool for detecting and analyzing the emotional experiences of women dealing with PCOS and thyroid disorders. By leveraging SVC, this project delivers more accurate and actionable insights into the psychological impact of these health challenges, contributing to improved mental health support and personalized interventions.

2. Logistic Regression:

Logistic Regression is a linear model widely used for classification tasks. It predicts probabilities for each class and is particularly useful for interpretable models. In our project, it was chosen as a baseline because of its simplicity and efficiency in handling text-based features

for emotion classification.

Hyperparameter Tuning

To optimize the performance of Logistic Regression, we applied hyperparameter tuning using GridSearchCV with 3-fold cross-validation. The following parameters were tuned:

- **Penalty:** Tested l2 regularization (ridge) to control overfitting.
- **C:** Adjusted the regularization strength with values [0.1, 1, 10] to balance underfitting and overfitting.
- **Solver:** Explored solvers such as lbfgs for smaller datasets and saga for larger datasets.

These combinations ensured that we achieved the best possible configuration for our dataset.

Performance Metrics

Once the optimal parameters were identified, we evaluated the model's performance on unseen data. The metrics used included:

- **Accuracy:** The Logistic Regression model achieved 85% accuracy.
- **Precision, Recall, and F1-Score: These** were used to measure the model's reliability for each emotional category.

Although Logistic Regression is a linear model, its simplicity provided interpretable results and established a strong baseline for comparison with more complex models like SVC and XGBoost.

Strengths and Limitations

While Logistic Regression performed well overall, it has certain limitations:

- **Strengths:** Efficient and interpretable, it worked effectively for smaller datasets and linear relationships.
- **Limitations:** It may struggle with non-linear relationships in the data, which can affect performance in more complex classification tasks.

```
Logistic Regression Accuracy: 0.85
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.64      0.55        14
           1       1.00      0.85      0.92        13
           2       1.00      1.00      1.00        11
           3       1.00      1.00      1.00        13
           4       1.00      1.00      1.00        13
           5       1.00      1.00      1.00        20
           6       0.62      0.50      0.55        16

    accuracy                           0.85       100
   macro avg       0.87      0.86      0.86       100
weighted avg       0.86      0.85      0.85       100
```

Fig15. Classification Report of Logistic Regression

In summary, Logistic Regression provided a reliable baseline for emotion classification with an accuracy of 85%. However, to handle non-linear relationships and enhance accuracy further, we employed more advanced models such as the Support Vector Classifier (SVC), which I will discuss next.

3. XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning method widely recognized for its efficiency and performance in classification tasks, particularly when dealing with structured data. In this project, XGBoost was chosen for emotion classification due to its ability to handle large datasets and complex patterns by combining the outputs of multiple decision trees.

Hyperparameter Tuning

To optimize the performance of the XGBoost model, GridSearchCV with 3-fold cross-validation was applied. The following hyperparameters were tuned to achieve the best configuration:

- **Max Depth:** Values of [3, 5, 7] were tested to control the complexity of the trees and prevent overfitting.
- **Learning Rate:** Values of [0.01, 0.1, 0.3] were explored to balance the step size during model training.
- **Number of Estimators:** [100, 200] decision trees were tested to find the optimal number of boosting rounds.

These combinations were systematically evaluated to identify the set of parameters that provided the highest performance on the training data while maintaining generalizability on unseen data.

Performance Metrics

Once the optimal hyperparameters were identified, the XGBoost model was evaluated using several key metrics on unseen data. The results are as follows:

- **Accuracy:** The XGBoost model achieved an accuracy of 76%.
- **Precision, Recall, and F1-Score:** These metrics were used to assess the model's reliability for each emotional category. XGBoost demonstrated high precision and recall for majority classes but struggled with minority classes due to the complexity of the data.

Strengths and Limitations

Strengths:

- Robustness: XGBoost is robust to overfitting due to its regularization techniques, making it suitable for complex datasets.
- Efficiency: It efficiently handles large datasets and can learn intricate patterns through gradient boosting.
- Flexibility: The model offers numerous hyperparameters for fine-tuning, allowing for customized configurations to improve performance.

Limitations:

- Complexity: The model's complexity can lead to longer training times, especially with large hyperparameter grids.
- Overfitting: Despite its robustness, XGBoost can still overfit when the dataset is small or when the hyperparameters are not appropriately tuned.

- Imbalanced Data: XGBoost's performance can be affected by imbalanced classes, even after applying SMOTE, as the model tends to favor majority classes.

XGBoost Accuracy: 0.76
XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.41	0.64	0.50	14
1	0.83	0.77	0.80	13
2	1.00	0.82	0.90	11
3	1.00	0.92	0.96	13
4	0.86	0.92	0.89	13
5	0.95	0.90	0.92	20
6	0.50	0.38	0.43	16
accuracy			0.76	100
macro avg	0.79	0.76	0.77	100
weighted avg	0.79	0.76	0.77	100

Fig16. Classification Report of XGBoost

In summary, the XGBoost model achieved an accuracy of 76% in classifying emotions, showing its capability to handle complex relationships within the data. However, while it performed well on certain classes, it fell short in accurately classifying minority emotions, indicating the need for further optimization or feature engineering. The model's strengths in robustness and efficiency made it a valuable part of the evaluation process, but its limitations in handling imbalanced data highlighted the necessity of exploring other models like the Support Vector Classifier (SVC) for improved performance.

Conclusion

This project, focused on analyzing women's hormonal health challenges using data mining and natural language processing (NLP) techniques, successfully demonstrated the potential of web-based analysis to gain meaningful insights into emotional and psychological experiences. By collecting and processing data from Reddit and reputable health websites, the project captured a comprehensive dataset reflecting the personal stories, concerns, and sentiments of women dealing with conditions such as PCOS (Polycystic Ovary Syndrome) and thyroid disorders.

The exploratory data analysis provided valuable insights into the nature of the discussions, highlighting key themes and prevalent issues faced by women. Sentiment analysis using the VADER model effectively categorized the text into positive, negative, and neutral sentiments, shedding light on the emotional tone of the data. The subsequent emotion detection using a transformer-based model allowed for a more nuanced classification into specific emotions like sadness, fear, joy, and anger. These analyses revealed significant emotional distress related to symptoms, diagnoses, and treatments, emphasizing the need for more comprehensive emotional support in healthcare settings.

The machine learning accuracy assessment further strengthened the project by evaluating the effectiveness of various models—Logistic Regression, Support Vector Classifier (SVC), and XGBoost—in classifying emotions. The Support Vector Classifier (SVC) emerged as the most reliable model, achieving an accuracy of 86% and demonstrating a strong ability to balance precision and recall across different emotion categories. The application of SMOTE addressed class imbalance, ensuring that minority emotions were adequately represented and classified. This rigorous evaluation ensures that the emotion detection results are accurate and trustworthy.

Additionally, the project's approach to matching detected emotions with predefined remedies provided actionable insights. By offering personalized recommendations based on the identified emotions, this analysis can guide mental health interventions and support systems, improving the overall well-being of individuals. The integration of data visualization techniques helped illustrate the distribution of sentiments, emotions, and user engagement, making the findings accessible and easy to interpret.

In conclusion, this project not only highlights the emotional challenges faced by women with PCOS and thyroid disorders but also offers a data-driven approach to addressing these issues. The findings underscore the importance of considering emotional well-being in healthcare practices and the potential of technology to enhance support systems. Future work could involve the use of deep learning models for improved accuracy, integrating larger datasets, and developing decision support systems to assist healthcare providers in delivering empathetic and personalized care.