STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

CS- 513 - Knowledge Discovery and Data Mining

Final Project on:
Web-Based Analysis of Women Hormonal Health Challenges using Data Mining and NLP Techniques

Instructor: Dr. Jingyi Sun

Bhaskara Sai Vamsi Krishna Padala

CONTENT

INTRODUCTION

- Our 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**.
- The first step in our project involves **web scraping** to collect relevant data. Reddit, with its dedicated subreddits like **r/PCOS** and **r/thyroidhealth.** Data scraping tools such as **asyncpraw** (for Reddit) and **BeautifulSoup** (for websites) facilitate the collection of the data from other websites.
- Once the data is collected, it undergoes **preprocessing** to ensure it is clean and usable. The cleaned data is then subjected to **sentiment analysis** using the **VADER (Valence Aware Dictionary and sEntiment Reasoner)** model.
- 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.

- To provide practical support, the project matches the detected emotions with potential **remedies** sourced from trusted medical websites.
- Finally, **Machine Learning (ML)** models, including **Support Vector Classifier (SVC), XGBoost and Logistic Regression** are employed to evaluate the accuracy of sentiment and emotion classification. These models help validate the effectiveness of the analysis and ensure reliable results.

Research Statement: *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 the sentiment and emotion detection, ensuring reliability through metrics such as accuracy, precision, recall, and F1-score.*

LITERATURE SURVEY

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.

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.

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 .

DATA COLLECTION

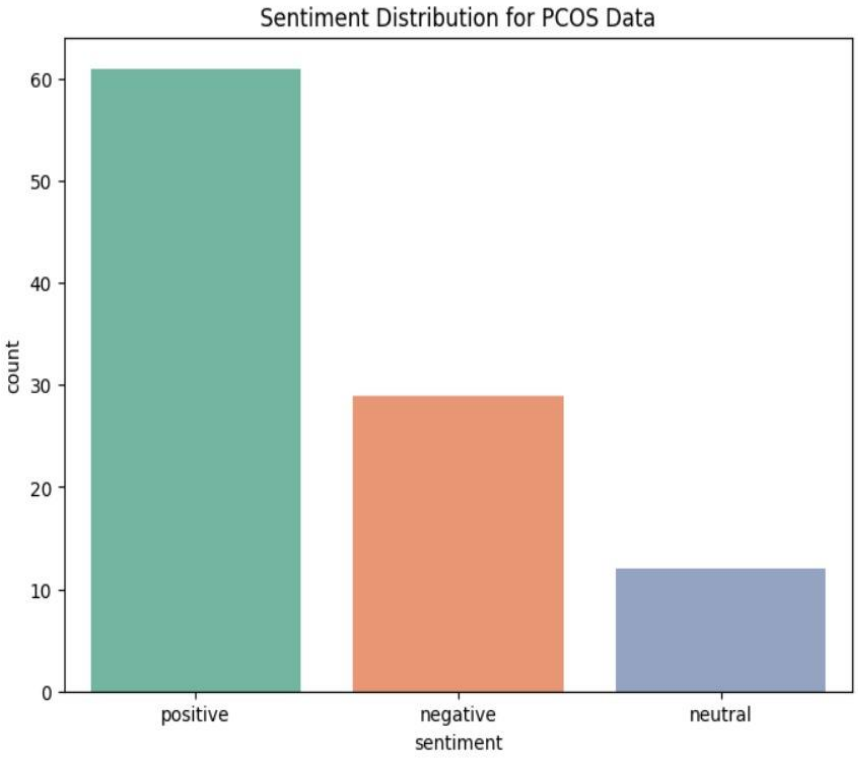
- **User-Generated Content from Reddit:** The well-rounded data collected from Reddit includes the post title and content, which allowed us to understand the primary topics and themes within each community. Additionally, we recorded the upvote count for each post as a measure of community engagement and the perceived relevance of each discussion.
- **Web Scraping from Reddit Data:** This step is crucial for collecting raw data from online forums, particularly those focused on women's hormonal health issues like PCOS and thyroid conditions. The Python library PRAW (Python Reddit API Wrapper) is utilized to access Reddit's API and collect data. Specifically, it is used to retrieve data from targeted subreddits such as r/PCOS and r/Thyroid. The collected data is stored in JSON format or directly in a database, which simplifies the organization, retrieval, and processing for future analyses like sentiment and emotion detection.

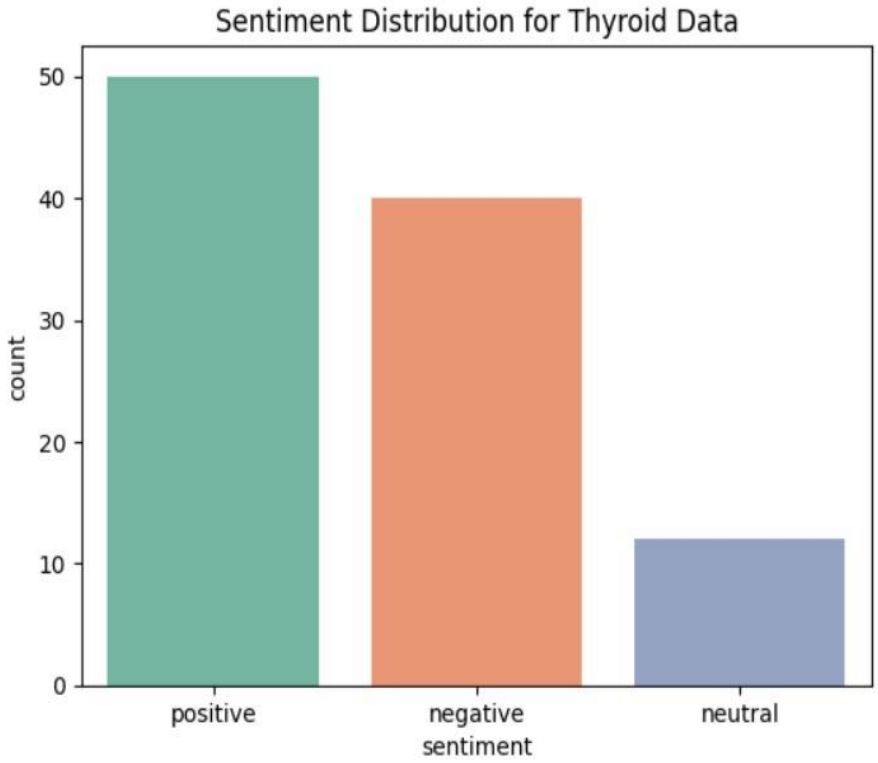
- **Supplementary Health Information from Medical Blogs:** To provide a factual foundation alongside the community-based insights from Reddit, we gathered supplementary health information, which contextualizes many of the health concerns expressed in the Reddit discussions.
- **Web Scraping from Medical Blogs:** This step complements the Reddit data by scraping expert-backed content from trusted medical sources like Mayo Clinic and Healthline. Web scraping tools such as BeautifulSoup or Scrapy are used to parse HTML content from these websites. The libraries allow us to extract relevant sections such as treatment options, lifestyle changes, or medical advice specifically related to PCOS and thyroid issues.

EXPLORATORY DATA ANALYSIS

1. Sentiment Analysis:

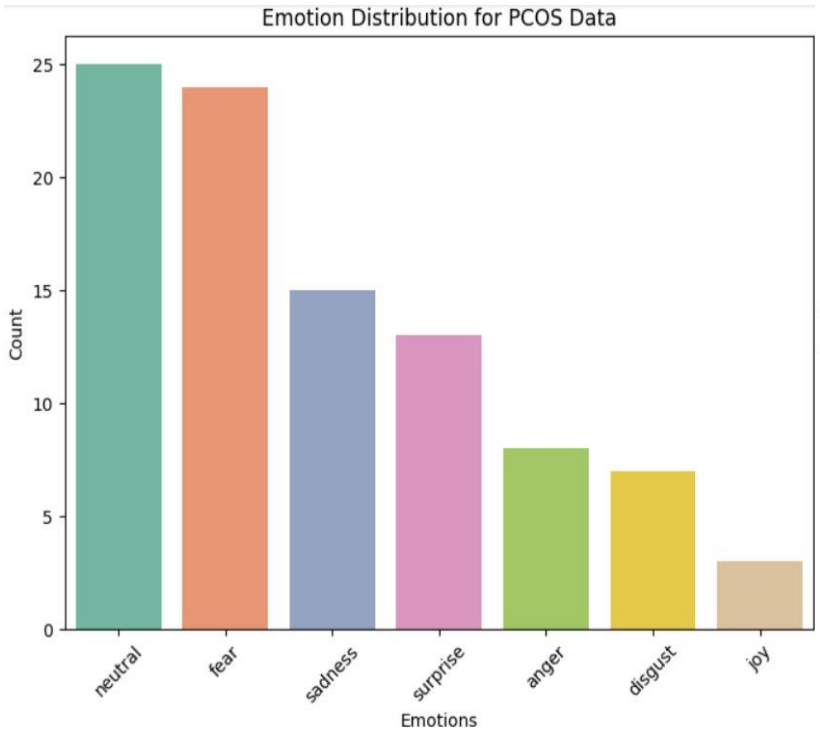
- The VADER (Valence Aware Dictionary and sentiment Reasoner) sentiment analysis tool is used in our project. VADER is well-suited for analyzing text data from social media and is capable of handling informal language, which is often present in Reddit posts.
- Each Reddit post and its comments are analyzed to determine whether the sentiment expressed is positive, negative, or neutral. This is done by evaluating the text's words and their associated polarity. By aggregating these results, we can uncover general trends in user sentiment towards specific topics within the hormonal health realm.
- This helps in identifying the general mood of discussions, allowing for a deeper understanding of public opinion on different aspects of PCOS and thyroid related health issues.

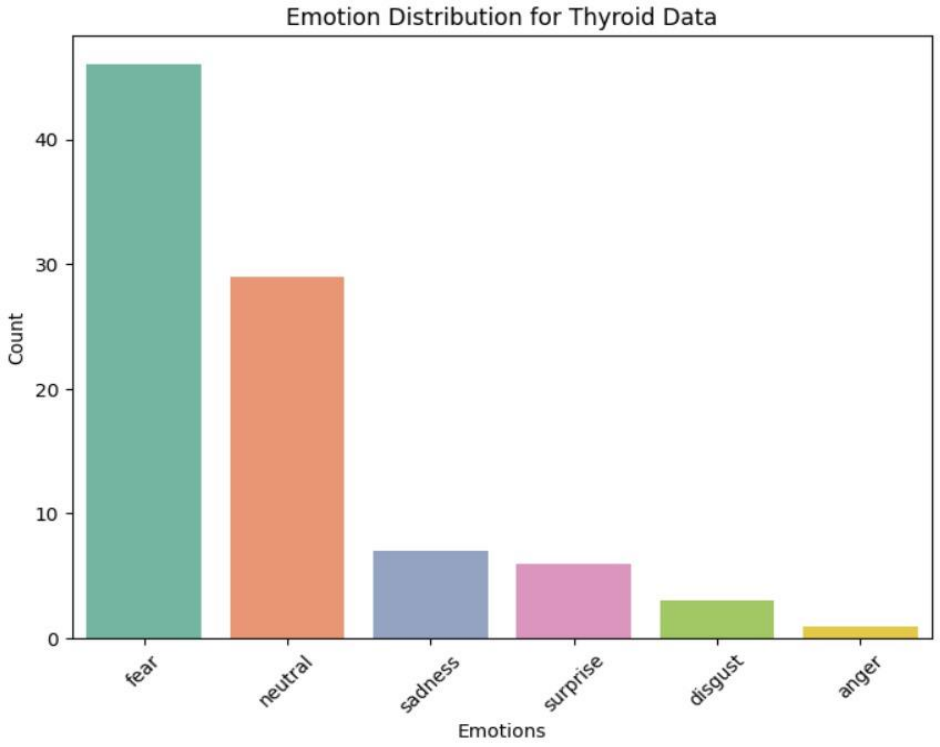

Sentiment Distribution for PCOS Data
count
0
10
20
30
40
50
60
positive
negative
neutral
sentiment


Sentiment Distribution for Thyroid Data
count
0
10
20
30
40
50
positive
negative
neutral
sentiment

2. Emotion Detection:

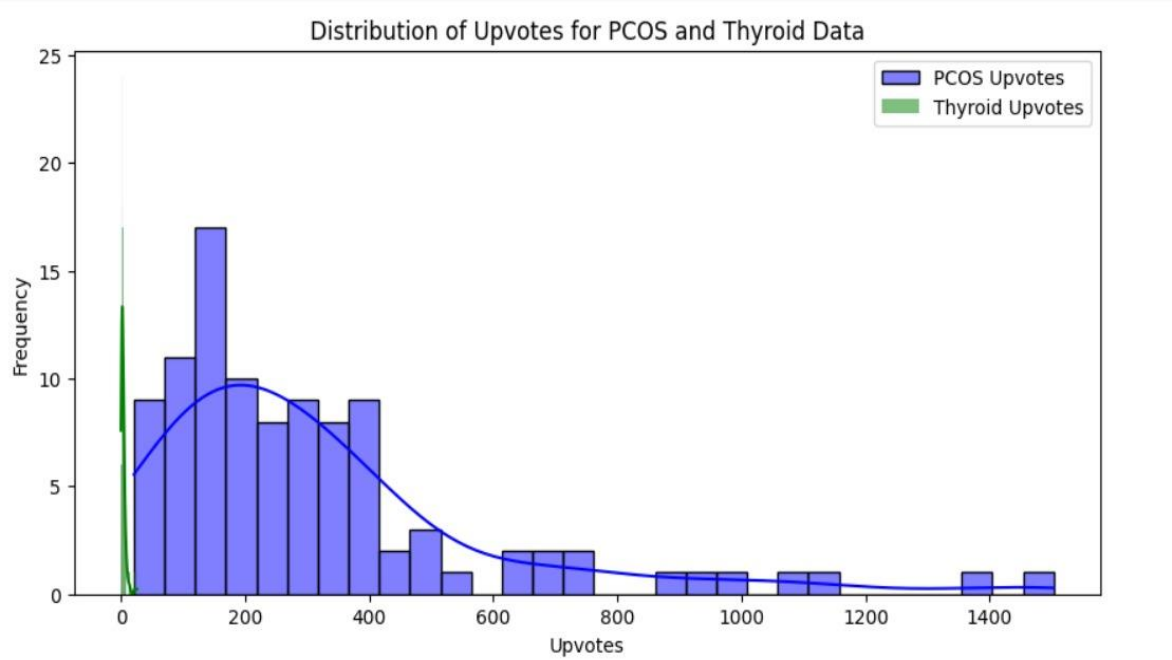
- Pre-trained models from the transformer library (such as BERT based models) can be applied for emotion detection. These models can detect a wide range of emotions such as anxiety, frustration, anger, happiness, and sadness from textual data.
- The emotion detection models classify posts into specific emotional categories. For example, a post about struggling with thyroid-related symptoms might be categorized as expressing frustration or anxiety, while another post may express relief or joy after discovering a successful treatment.
- This granularity helps in understanding not only the general sentiment but also the psychological and emotional impact of hormonal health issues on individuals, allowing for more tailored interventions or solutions.


Emotion Distribution for PCOS Data
Count
Emotions
neutral
fear
sadness
surprise
anger
disgust
joy
0
5
10
15
20
25


Emotion Distribution for Thyroid Data
Count
Emotions
fear
neutral
sadness
surprise
disgust
anger
0
10
20
30
40

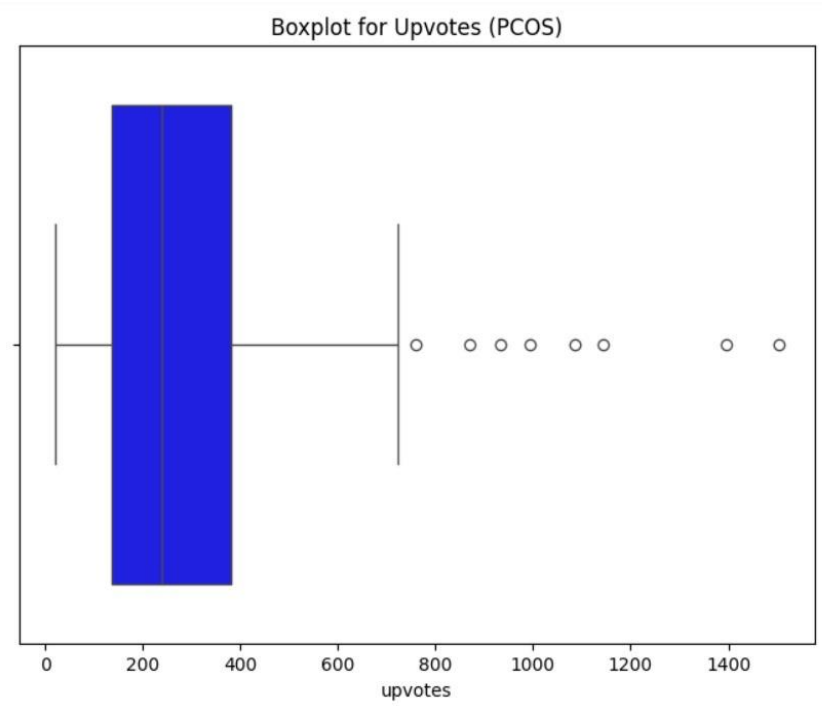
3. Data Visualization:

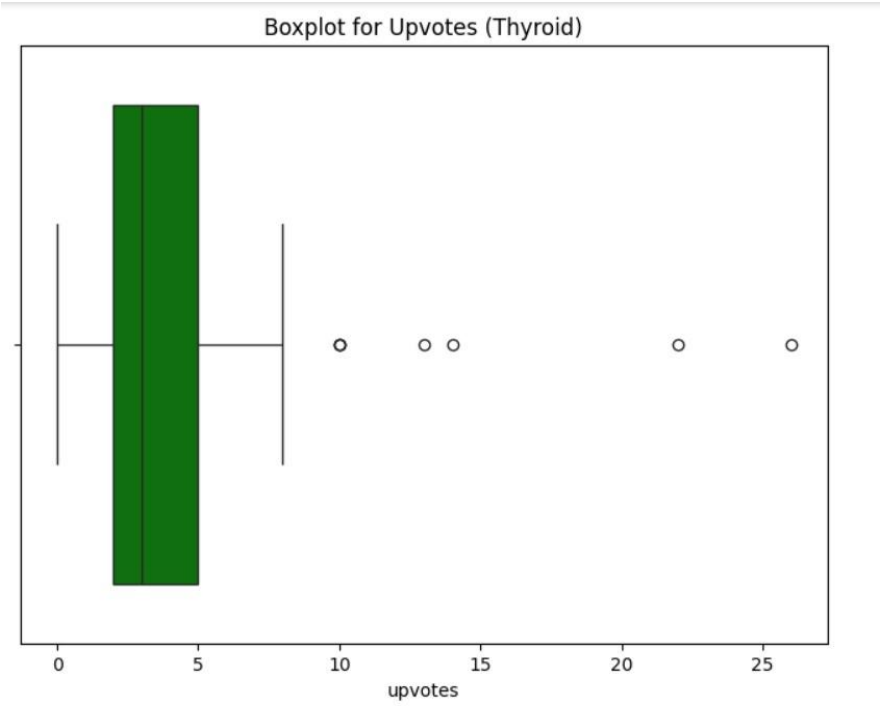
- **Distribution of Upvotes:** Shows how popular or engaging certain topics are by visualizing upvote counts across posts.



- **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.





- **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 different users.

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.

MATCHING EMOTIONS TO REMEDIES

- Matching emotions to remedies is an essential component that aims to provide actionable advice and support by aligning the emotions expressed in user-generated content with practical remedies sourced from trusted medical websites.
- The primary objective of matching emotions to remedies is to enhance mental health support and provide personalized advice.
- A dictionary, **emotion_remedies**, pairs emotions such as joy, sadness, and anger with relevant coping strategies, like maintaining healthy habits, practicing mindfulness, or engaging in relaxation techniques.
- The **match_remedy** function retrieves these remedies based on the detected emotion in the dataset. If an emotion is not recognized, the function returns a default message indicating no remedy is available for that emotion.
- The remedy-matching process is applied to the **emotion** column of the **data** DataFrame, creating a new column, **remedy**, which stores the corresponding suggestions. For each row, the function fetches a remedy tailored to the detected emotion or provides a fallback message for undefined emotions.

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.

ML MODELS ACCURACY ASSESSMENT

1. XGBOOST CLASSIFIER

XGBoost, a gradient boosting model, was applied for robust and efficient multi-class classification, leveraging its ensemble learning capabilities.

Hyperparameter Tuning:

- max_depth: Controlled tree depth ([3, 5, 7]) to prevent overfitting.
- learning_rate: Fine-tuned learning rates ([0.01, 0.1, 0.3]) to optimize weight updates.
- n_estimators: Adjusted the number of boosting rounds ([100, 200]) to enhance predictive power.

Training: GridSearchCV applied hyperparameter tuning with 3-fold cross-validation.

Evaluation: The model was assessed on X_test, with detailed metrics to evaluate its performance. XGBoost demonstrated high accuracy and flexibility, particularly in handling imbalanced and large datasets.

Accuracy: XGBoost Classifier achieved an accuracy of 76%.

2. LOGISTIC REGRESSION

Logistic Regression was used as a baseline linear model for multi-class classification to predict emotions based on textual features.

Hyperparameter Tuning:

- penalty: Tested l2 regularization (ridge) and no regularization (none) to control overfitting.
- C: Adjusted regularization strength with values [0.1, 1, 10].
- solver: Optimized using lbfgs (efficient for smaller datasets) and saga (supports larger datasets).

Training: GridSearchCV was applied to find the best combination of hyperparameters over 3-fold cross-validation.

Evaluation: The optimized model was tested on unseen data (X_test), and its accuracy and classification report (precision, recall, F1-score) were printed. Logistic Regression's simplicity provided interpretable results but may not handle non-linear relationships effectively.

Accuracy: Logistic Regression achieved an accuracy of 85%.

3. SUPPORT VECTOR CLASSIFIER (SVC)

SVC was utilized to model the classification task with the ability to handle both linear and non-linear decision boundaries using kernels.

Hyperparameter Tuning:

- kernel: Explored linear (for simpler decision boundaries) and rbf (for non-linear separations).
- C: Regularization parameter ([0.1, 1, 10]) was tuned to balance margin maximization and classification accuracy.
- gamma: Adjusted the influence of individual training samples in non-linear kernels with scale and auto.

Training: GridSearchCV was used to perform exhaustive hyperparameter search over 3-fold cross-validation.

Evaluation: The best model was validated on X_test, and metrics were logged. SVC provided flexibility for complex relationships but can be computationally intensive for large datasets.

Accuracy: SVC achieved an accuracy of 86%.

CONCLUSION

Our project successfully analyzed and categorized emotional responses from online discussions related to PCOS and thyroid disorders using data mining and NLP techniques. Sentiment and emotion detection were key components, with sentiment analyzed via VADER and emotions identified using the pre-trained DistilRoBERTa model. Matching detected emotions to remedies provided personalized and actionable advice, demonstrating the potential for real-world application in mental health support.

Three supervised machine learning models were employed to validate the emotion detection process: Logistic Regression, Support Vector Classifier (SVC), and XGBoost. Among these, SVC achieved the highest accuracy at 86%, closely followed by Logistic Regression at 85%, both showcasing strong performance for the classification task. XGBoost, while slightly less accurate at 76%, demonstrated robustness in handling class imbalances and large datasets. The combination of NLP techniques and machine learning ensured reliable results, supporting the project's goal of leveraging web-based data to enhance understanding and support for women's hormonal health challenges.

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu

THANK YOU