



NLP-based app focused on Health Awareness in Rural Rajasthan,  
especially with Marwari text-to-speech (TTS)

Capstone Project

CS-IT Department

Teena Sharma

# PROBLEM STATEMENT

In many rural areas of Rajasthan, women lack access to health information in their local Marwari language. Language and literacy barriers prevent them from understanding health advice. The "Hindi-to-Marwari Health Awareness Web Application/App" solves this problem by translating Hindi health queries into Marwari text and audio. This user-friendly platform empowers rural women by providing health information in their native dialect, improving awareness, preventive care, and informed health decisions.

# DATASET

1.Total Sentences: 100 health-related sentences

- Language Pairs Breakdown:

- Hindi Sentences: Collected from health awareness materials, government posters, local clinics

2. Marwari Translations: Verified by native speakers, community health workers, online Marwari resource

3. Topics Covered:

- Hygiene Practices (e.g., handwashing, clean water)

- Menstrual Health (e.g., menstrual hygiene tips)

- Disease Prevention (e.g., vaccination, mosquito control)

# Objective

1. To develop a user-friendly web application that translates Hindi health-related content into Marwari audio.
2. To integrate a Marwari text-to-speech system for generating clear and understandable audio output.
3. To build a database for securely storing translation history, including input text, translated text, and audio file paths.
4. To provide an easy-to-use interface that enables non-technical rural users to access health information without difficulty.
5. To improve accessibility of health awareness content by combining translation and audio playback in a single platform.





# Methodology

## 1 Need

Many rural women in Rajasthan do not have access to health information in their local Marwari language. There is a need for a simple platform that can translate health-related messages into Marwari text and audio to help them understand better.

## 2 Objective

To build a user-friendly web app that translates Hindi health questions into Marwari text and speech, using a machine learning model and text-to-speech system, so that even non-literate users can access health information.

## 3 Approach

We created a small Hindi-Marwari dataset, trained a Logistic Regression model for translation, and used Google Text-to-Speech (gTTS) for audio output. A web interface was built using Streamlit to allow users to enter Hindi sentences and get Marwari responses in both text and audio formats.

## 4 Challenges

The main challenge was creating accurate translations with a small dataset and integrating Marwari audio using tools not directly made for Marwari. It was also important to make the app simple enough for rural users.

## 5 Future Scope

In the future, more sentences can be added to the dataset to improve translation quality. A mobile version of the app can also be developed to increase accessibility for women using smartphones.

# Methodology

## 1. Dataset Collection & Preparation

- Collected 100+ Hindi health sentences covering hygiene, disease prevention, menstrual health.
- Translated each sentence into Marwari with help of native speakers and online resources.
- Organized the data in an Excel file and converted to CSV for model use.

## 2. Model Selection & Training

- Used **Logistic Regression** to build the translation model.
- Vectorized Hindi sentences using **CountVectorizer**.
- Trained the model to map Hindi input to Marwari output.
- Saved the trained model as **translation\_model.pkl** for reuse.

hindi question	marwari answer	Category
मैं ठीक महसूस नहीं कर रहा	हूँ ठीक महसूस नीं कर रियो	General
मुझे खांसी है	म्हाने खांसी है	General
मुझे सांस लेने में दिक्कत है	म्हाने सांस लेवण में तकलीफ है	General
आपको किस चीज़ से एलर्जी है?	थाने कोणसी चीज़ सूँ एलर्जी है?	General
क्या आप गर्भवती हैं?	थूँ पेट सूँ हो?	General
मुझे सिरदर्द है	म्हाने मथा दुखे है	General
मुझे उल्टी आ रही है	म्हाने उल्टी आ री है	General
आपको डॉक्टर से मिलना चाहिए	थाने डाक्टर सूँ मिलणो चाही	General
यहाँ कोई क्लिनिक है?	ईया कोई दवाखानो है?	General
बुखार आने पर क्या करें?	बुखार आवे तो आराम करो अर डॉक्टर सूँ संपर्क करो।	prevent disease
मासिक धर्म के दौरान सफाई क्यों जरूरी है?	मासिक धर्म में सफाई राखण जरूरी है जिकां शरीर साफ रहे और इंफेक्शन न होवे।	mensuration hygiene
हाथ धोना कब जरूरी होता है?	हाथ धोना हर बार जरूरी है जद खाणे पीण लागो, टॉयलेट सूँ पाछो आओ या बाहर सूँ आओ।	hygiene
गंदा पानी पीने से कौन-कौन सी बीमारियाँ होती हैं?	गंदो पानी पीण सूँ पेट खराब होवे, टाइफाइड आ साके, या पीलिया।	disease prevention
मच्छर से बचाव कैसे करें?	मच्छर सूँ बचण खातर मच्छरदानी काम में लेवो अर आसपास पानी इकठ्ठो मत होण देवो।	disease prevention
क्यों हमें साबुन से नहाना चाहिए?	साबण सूँ नहाणे सूँ शरीर साफ रहवे अर बीमारियां दूर रहें।	hygiene
क्या रोज़ दाँत साफ़ करना जरूरी है?	हां, रोज़ दाँत साफ़ करवा जरूरी है जिकां दाँत मजबूत रहें अर कीड़े ना लागे।	hygiene
साफ पानी कैसे स्टोर करें?	साफ पानी ढक्कण वाली बाल्टी में राखो अर रोज ढक्कण लगावो।	hygiene
गर्भावस्था में क्या सावधानी रखनी चाहिए?	गर्भावस्था में साफ-सफाई राखणी, पोष्टिक खाना खाणो अर डॉक्टर की सलाह लेणी।	mensuration health
बच्चों को दस्त क्यों लगते हैं?	बच्चा गंदगी में खेले या गंदा पानी पीवे तो दस्त लाग सके।	disease prevention
कचरा कहाँ फेंकना चाहिए?	कचरा हमेशा डस्टबिन में फेंको, इधर-उधर मत फेंको।	hygiene

```
[ ] import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
import pickle

# --- STEP 1: Load Dataset ---
def load_dataset(filepath='/content/hindi_marwari.csv'):
    df = pd.read_excel('/content/Hindi_Marwari.xlsx')
    df.dropna(inplace=True)
    return df

# --- STEP 2: Prepare Data ---
def prepare_data(df):
    # Check the actual column names in your DataFrame
    # Print df.columns to see the available columns
    print(df.columns)

    # Replace 'Hindi_Text' and 'Marwari_Text' with the correct column names
    # These are placeholders and might need adjustment based on your CSV file
    X = df[df.columns[0]] # Accessing using column index instead of column name
    y = df[df.columns[1]] # Accessing using column index instead of column name
    return X, y

# --- STEP 3: Load + Prepare ---
df = load_dataset()
X, y = prepare_data(df)

# --- STEP 4: Vectorize ---
vectorizer = CountVectorizer()
X_vectorized = vectorizer.fit_transform(X)

# --- STEP 5: Train the model ---
model = MultinomialNB()
model.fit(X_vectorized, y)

# --- STEP 6: Save the trained model and vectorizer ---
with open('/content/translation_model.pkl', 'wb') as f:
    pickle.dump((model, vectorizer), f)

print("✅ Model trained and saved as 'translation_model.pkl'!")

Index(['hindi', 'marwari'], dtype='object')
✅ Model trained and saved as 'translation_model.pkl'!
```



### 3) Text-to-Speech Integration

- Used **Google Text-to-Speech (gTTS)** to convert Marwari text into audio files.
- Saved audio files in a separate folder with unique filenames for each translation.
- Integrated the audio playback feature inside the web app so users can listen easily.
- Ensured audio files are automatically linked to their corresponding translation in history.

```
import pandas as pd
from gtts import gTTS
import os

# Convert predicted text to audio
tts = gTTS(text=predicted_marwari, lang='hi') # For demo, using Hindi TTS voice
audio_path = "/content/marwari_audio.mp3"
tts.save(audio_path)










# Play audio (if running locally, this works)
from IPython.display import Audio
Audio(audio_path)

import pandas as pd
from gtts import gTTS
import os

def save_to_csv(hindi, marwari, audio_path):
    file = 'dataset/transaction_history.csv'
    df = pd.DataFrame([[hindi, marwari, audio_path]], columns=["Hindi", "Marwari", "Audio"])

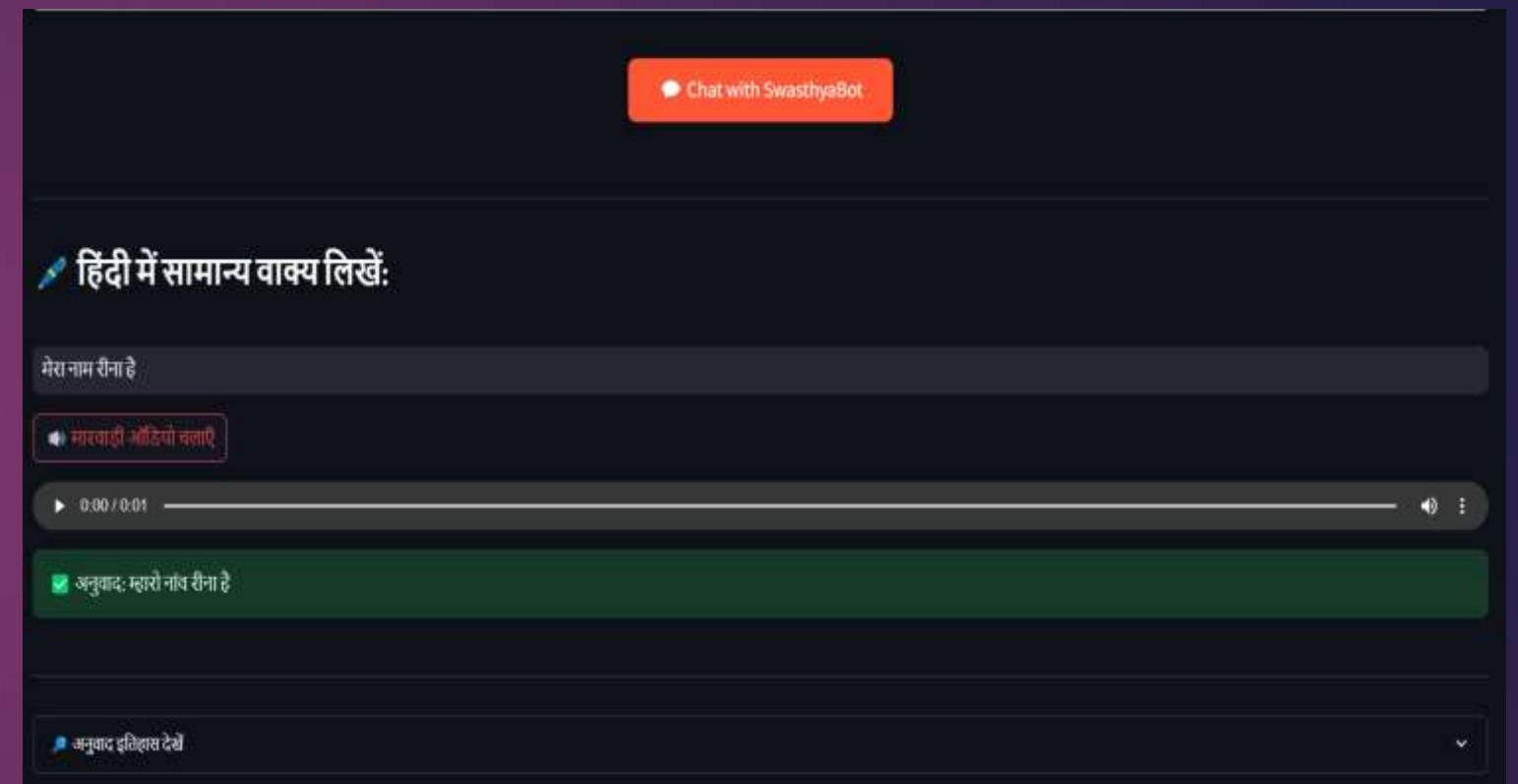
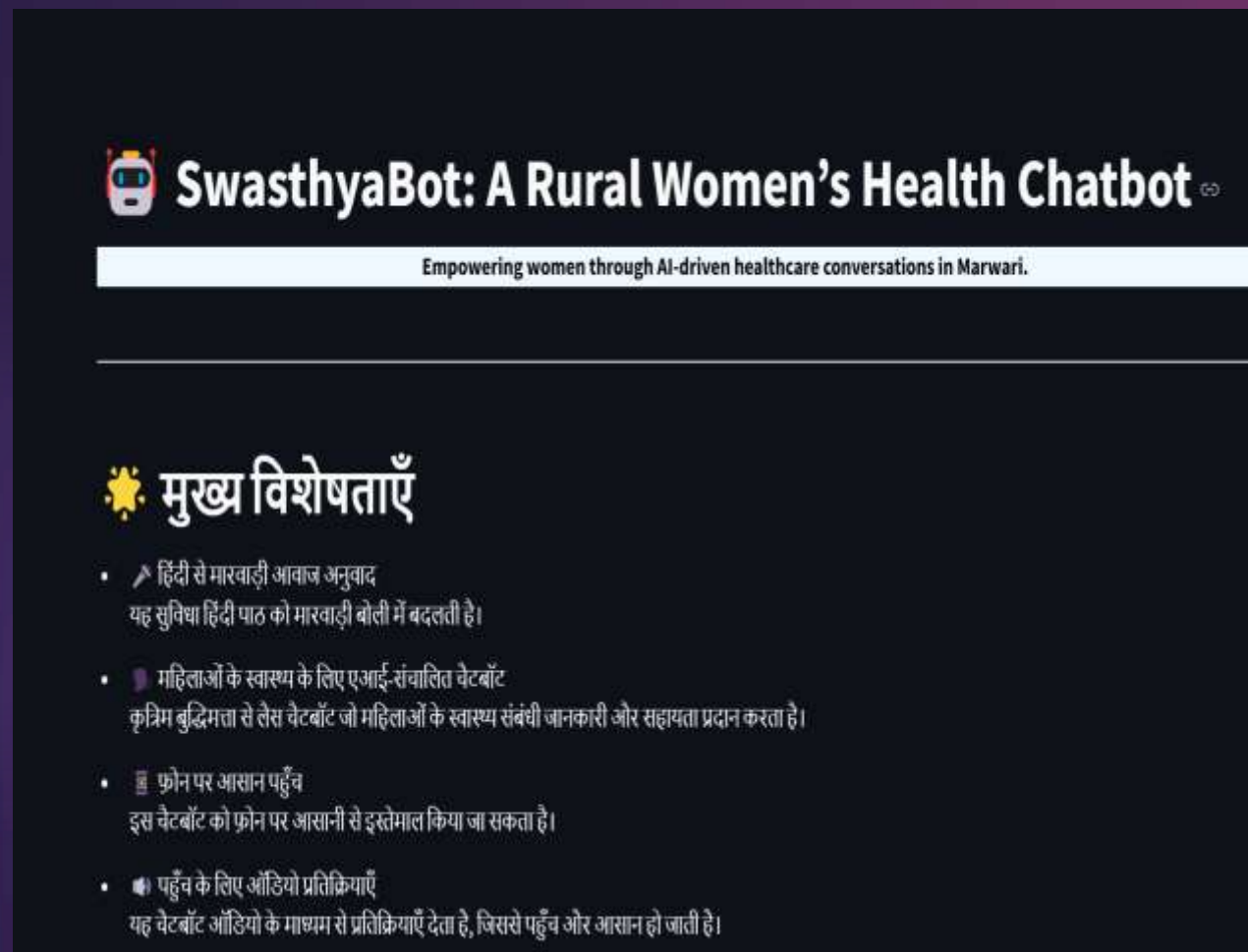
    if os.path.exists(file):
        df.to_csv(file, mode='a', header=False, index=False)
    else:
        df.to_csv(file, index=False)

    print("✅ Data saved to CSV.")
```

Name	#	Title	Contributing artists
 marwari_audio			
 ई_दवई_कनन_वर_लन_ह			
 गद_पन_पण_स_पट_खरब_...			
 थ_कस_ह			
 थर_उमर_कनन_ह_20250...			
 थर_उमर_कनन_ह_20250...			
 थर_नव_कइ_ह_20250505...			
 धनयवद_202505051525...			
 बखर_आव_त_आरम_कर_...			

## 4. Web App Development

- Developed web interface using **Streamlit** for easy user interaction.
- Added input box for Hindi sentences and buttons to display Marwari translation + play audio.
- Displayed translation history inside app for user reference.





## 🌸 SwasthyaBot 🌸

हमारा SwasthyaBot एक AI-पावर्ड हेल्थ ट्रेससेटर चैटबॉट है, जो ग्रामीण महिलाओं के लिए स्वास्थ्य शिक्षा को सरल और प्रभावी तरीके से प्रस्तुत करता है। वहीं आप हिंदी में पूछकर मारवाड़ी में स्वास्थ्य संबंधी सलाह और जानकारी प्राप्त कर सकते हैं।



### 👤 स्वस्थबॉट की विशेषताएँ

- ग्रामीण भारतीय महिलाओं के लिए सांस्कृतिक रूप से प्रासंगिक 🇮🇳
- आवाज समर्पण के साथ उपयोग में आसान 🗣️
- स्वास्थ्य सुझाव प्रदान करता है, सवालों के जवाब देता है और मार्गदर्शन देता है 📋

### 🔧 कैसे इस्तेमाल करें?

1. अपना स्वास्थ्य प्रश्न हिंदी में लिखिए 📝
2. बॉट इसे मारवाड़ी में अनुवाद करता है 🗣️
3. एआई संसाधित करता है और स्वास्थ्य सलाह के साथ उत्तर देता है 📋
4. आप ऑडियो प्रारूप में प्रतिक्रिया सुनते हैं 🎧

### 🌟 मुख्य विशेषताएँ

#### 1. सामान्य और विशिष्ट स्वास्थ्य प्रश्नों का मारवाड़ी में ऑडियो उत्तर

आपका चैटबॉट उपयोगकर्ताओं द्वारा पूछे गए सामान्य स्वास्थ्य संबंधी प्रश्नों और मासिक धर्म स्वच्छता व बीमारियों की रोकथाम जैसे विशिष्ट क्षेत्रों से संबंधित हिंदी पाठ को मारवाड़ी ऑडियो में उत्तर देगा। इससे उपयोगकर्ताओं को अपनी भाषा में जानकारी आसानी से समझने में मदद मिलेगी।

#### 2. सुलभ और श्वस्य स्वास्थ्य मार्गदर्शन

यह चैटबॉट स्वास्थ्य संबंधी जानकारी को सुनने योग्य प्रारूप में उपलब्ध कराएगा, जिससे उन लोगों के लिए भी यह सुलभ होगा जो पढ़ना पसंद नहीं करते या जिन्हें पढ़ने में कठिनाई होती है। मारवाड़ी ऑडियो प्रारूप जानकारी को अधिक व्यक्तिगत और समझने में आसान बना सकता है।

#### 3. विशिष्ट स्वास्थ्य विषयों पर स्थानीय भाषा में जागरूकता

आपका मॉडल मासिक धर्म स्वच्छता और बीमारियों की रोकथाम जैसे महत्वपूर्ण स्वास्थ्य विषयों पर मारवाड़ी भाषा में जानकारी प्रदान करके स्थानीय समुदाय में जागरूकता बढ़ाने में महत्वपूर्ण भूमिका निभाएगा। यह उन लोगों तक पहुंचने में मदद करेगा जो हिंदी में सहज नहीं हैं, जिससे स्वास्थ्य संबंधी महत्वपूर्ण संदेश प्रभावी ढंग से संप्रेषित किए जा सकेंगे।

## 5) Model Integration & Testing

- Integrate the trained model with the web application.
- Test the app to ensure correct predictions and clear explanations.
- Fix any issues related to model accuracy, UI design, or performance.

[illegible]

## FUTURE DEVELOPMENT

### 1 Mission 01

- **Mobile App:** Transform the current web application into a mobile app for easier access on smartphones by rural women.

### 2 Mission 02

- **Voice Input:** Add a voice input feature so users can ask health questions by speaking instead of typing.

### 3 Mission 03

- **Expanded Dataset:** Increase the number of Hindi-Marwari health sentences to improve translation quality and cover more health topics.

# CONCLUSION

The Hindi-to-Marwari Health Awareness Web Application successfully delivers health information in the Marwari language through both text and audio. By bridging language and literacy barriers, it empowers rural women to access crucial health knowledge in their native dialect, leading to better awareness and healthier choices. The project demonstrates how technology can be used to support community health education, promote preventive care, and help rural women make informed decisions about their well-being. It shows the potential of combining machine learning, text-to-speech, and web development to create solutions for social good, making health information more inclusive and accessible for underserved populations.



# References

- AI4Bharat's Indic-TTS: An open-source initiative providing state-of-the-art TTS models for Indian languages, including Rajasthani.  
[github.com](https://github.com/aishwaryap99/indic-tts)
- "Multilingual Text-to-Speech Models for Indic Languages": An article discussing the development of TTS models for multiple Indic languages.  
[analyticsvidhya.com](https://analyticsvidhya.com)
- "Deep Learning Based TTS-STT Model with Transliteration for Indic Languages": A research paper introducing an end-to-end TTS model using the ESPnet toolkit.  
[ijraset.com](https://ijraset.com)
- "Improving health outcomes for under-served communities in rural Rajasthan": An article discussing initiatives aimed at increasing community awareness of health information in rural Rajasthan.  
[biovoicenews.com](https://biovoicenews.com)
- "mHealth App Improves Maternal Health Awareness, Behavior in Rural India": A study on the impact of a maternal health mobile app on pregnant women's health awareness and behaviors.  
[ajmc.com](https://ajmc.com)