Abstract and Keywords

This project represents a comprehensive investigation into Ayurvedic medicines employing advanced natural language processing techniques. A meticulously assembled dataset encompassing vital attributes such as Medicine_Name, Disease, Review_Text, and Rating was painstakingly curated to facilitate a nuanced analysis of sentiments associated with Ayurvedic remedies. Leveraging the SentimentIntensityAnalyzer, the dataset underwent intricate segmentation, discerning nuances within positive, negative, and sarcastic categories. Through meticulous preprocessing steps, including data cleaning and normalization, the dataset was refined to enhance the efficacy of subsequent classification tasks. Employing sophisticated machine learning models such as the Bayes Classifier and Passive Aggressive Classifier, this project achieved remarkable predictive accuracy, reaching an impressive 95.5% in disease prediction based on review sentiments. Furthermore, a robust methodology was developed to discern the most efficacious medicine for specific ailments, thereby offering practical insights into Ayurvedic treatment modalities. This multifaceted endeavor not only enriches our understanding of Ayurveda but also underscores the potential of computational approaches in augmenting traditional medical practices.

Keywords:

Ayurvedic: Traditional Indian medicine system.
Natural Language Processing (NLP): Computational analysis of human language.
Sentiment Analysis: Determining emotions conveyed in text.
Classification: Categorization of data into groups.
Machine Learning: Algorithms that learn from data.
Bayes Classifier: Probabilistic classification model based on Bayes' theorem.
Passive Aggressive Classifier: Online learning algorithm for classification tasks.
Predictive Accuracy: Measure of model performance in making correct predictions.
Preprocessing: Data preparation techniques before analysis.

Chapter 1: Introduction

In recent years, the intersection of traditional medicine and computational techniques has opened new avenues for understanding and enhancing healthcare practices. This project focuses on the analysis of Ayurvedic medicines through the lens of natural language processing (NLP), with a primary goal of leveraging machine learning algorithms to predict disease outcomes and recommend suitable treatments. Ayurveda, as a time-honored medical system, offers a rich repository of herbal remedies and holistic approaches to health and wellness. By harnessing the power of NLP, we aim to unlock valuable insights embedded within user reviews and sentiments associated with Ayurvedic medicines.

The objectives of this project are twofold: firstly, to develop accurate models for disease prediction based on the sentiment analysis of user reviews, and secondly, to recommend the most appropriate Ayurvedic medicines for specific health conditions. Achieving these objectives holds immense potential for benefiting society in several ways. Firstly, by providing individuals with personalized recommendations for Ayurvedic treatments, we empower them to make informed decisions about their health and well-being. Moreover, accurate disease prediction models can aid healthcare practitioners in identifying potential health issues early, thereby facilitating timely interventions and improving patient outcomes.

Beyond its immediate applications, this project contributes to the broader landscape of healthcare by bridging the gap between traditional medicine and modern computational techniques. By amalgamating age-old wisdom with cutting-edge technology, we pave the way for a holistic approach to healthcare that integrates the best of both worlds. Furthermore, the methodologies and insights derived from this project can serve as a blueprint for similar endeavors in other medical domains, fostering innovation and collaboration across disciplines.

Chapter 2: Literature Review

C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews," Journal of Biomedical Informatics, vol. 00, no. 0, pp. 1-15, Month Year. DOI: 10.1016/j.jbi.2020.103539.

This paper outlines the evolution and application of sentiment analysis, particularly in the domain of pharmaceuticals, where analyzing user reviews can provide valuable insights into drug effectiveness and side effects. The authors note the shift from traditional rule-based approaches to advanced machine learning techniques, particularly deep learning. While deep learning has seen success in sentiment analysis across various domains, its application to drug reviews has been limited.

The study proposes a benchmarking comparison of deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT), for sentiment analysis of drug reviews. Various combinations of these models are explored, along with different pre-trained word embedding models.

The dataset used consists of drug reviews from the Drugs.com website, categorized by patient satisfaction scores ranging from 0 to 9. Different levels of polarity classification are considered, including two-class (positive vs. negative), three-class (positive, negative, neutral), and ten-class classifications.

Pre-trained word embedding models are employed to represent the drug reviews, and various deep learning architectures are tested for sentiment analysis. These architectures include standalone CNNs and LSTMs, as well as hybrid models combining both. The study also explores the use of BERT for contextualized word embeddings.

Results and discussions are provided, including evaluations based on standard metrics such as precision, recall, and F1 scores. The paper concludes with insights into the performance of different models and suggestions for future research directions.

Y. Han, M. Liu, and W. Jing, "Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer," IEEE Access, vol. 8, pp. 19876-19887, January 27, 2020. DOI: 10.1109/ACCESS.2020.2969473.

In this study, researchers have developed a state-of-the-art model known as PM-DBiGRU, aimed at revolutionizing aspect-level sentiment analysis in drug reviews. This innovative model addresses key challenges in the field, including overlooking target semantics and limited dataset sizes. PM-DBiGRU leverages pretrained weights from short text-level sentiment classification to initialize its architecture. It employs double Bidirectional Gated Recurrent Unit (BiGRU) networks to capture bidirectional semantic representations of both the target and the drug review corpus. Through the use of an attention mechanism, the model extracts target-specific representations, allowing for a more nuanced understanding of sentiment. Additionally, by employing multi-task learning techniques, PM-DBiGRU transfers domain knowledge from short text-level reviews. This approach represents a significant advancement in aspect-level sentiment analysis within drug reviews. Furthermore, the introduction of the SentiDrugs dataset provides a valuable resource for aspect-level drug review sentiment classification. With manually annotated reviews containing identified targets and sentiment polarities, this dataset facilitates comprehensive analysis. Through rigorous experimentation, PM-DBiGRU demonstrates its efficacy in advancing aspect-level sentiment analysis, surpassing existing models in the field.

W. Y. M. Kyaing and J.-C. Na, "Sentiment Analysis of User-Generated Content on Drug Review Websites," Journal of Information Science Theory and Practice

This study introduces a method for sentiment analysis of user-generated drug reviews, an area not extensively explored compared to other domains like product reviews. They develop a

clause-level sentiment analysis algorithm to handle sentences with multiple clauses discussing various aspects of a drug. By employing linguistic techniques and considering grammatical relations and semantic annotations, they compute sentiment orientation (positive, negative, or neutral) for each clause. Their method, tested on 2,700 clauses, outperforms baseline machine learning approaches. The study emphasizes the importance of domain-specific knowledge, incorporating health and medical terms for accurate analysis. They utilize MetaMap to map medical terms to semantic types for sentiment analysis. The paper also discusses the significance of sentiment analysis in health domains, particularly for patients, drug makers, and clinicians. Additionally, it addresses challenges in sentiment analysis of user-generated content and the necessity of understanding and leveraging such content in health-related applications. The method involves separating sentences into clauses and identifying aspects of drug reviews, validated through manual coding. Sentiment lexicons are constructed, comprising general and domain-specific terms, with additional disorder terms tagged using MetaMap. Overall, the study aims to provide an effective approach for sentiment analysis of drug reviews, contributing to understanding public opinion and feedback in health domains.

Chapter 3:Methodology

Dataset Explanation:

The dataset comprises four main attributes:

Medicine_Name: The name of Ayurvedic medicine.

Disease: The health condition or disease the medicine is purported to treat.

Review_Text: User reviews or feedback regarding the effectiveness of the medicine.

Rating: The rating given to the medicine by users.

Each instance in the dataset represents a unique combination of these attributes, providing valuable insights into the effectiveness of Ayurvedic medicines for various health conditions.

	A	B	C	D	E
1	Medicine_	User_ID	Disease	Review_Te	Rating
2	Malaki Chu	User123	Indigestior	This churna	5
3	Malaki Chu	User124	Acid Reflu:	Malaki Chu	5
4	Malaki Chu	User125	Indigestior	I'm disapp	2
5	Malaki Chu	User126	Bloating	This churna	4
6	Malaki Chu	User127	Indigestior	I tried Mal	2
7	Malaki Chu	User128	Indigestior	Malaki Chu	1
8	Malaki Chu	User129	Stomach Is	My grandn	5
9	Malaki Chu	User130	Stomach C	I experienc	3
10	Malaki Chu	User131	Bloating	Malaki Chu	4
11	Malaki Chu	User132	Indigestior	I've been u	4
12	Medicine_	User_ID	Disease	Review_Te	Rating
13	Ashwagan	User001	Stress	These caps	5
14	Ashwagan	User002	Anxiety	I've been t	4
15	Ashwagan	User003	Energy Bo	Ashwagan	4
16	Ashwagan	User004	Sleep	I struggle v	4
17	Ashwagan	User005	Stress	I'm disapp	2
18	Ashwagan	User006	Anxiety	Ashwagan	2
19	Ashwagan	User007	Fatigue	I didn't not	3
20	Ashwagan	User008	Immune B	These caps	5
21	Ashwagan	User009	Memory	I feel like r	4
22	Ashwagan	User010	Joint Pain	I have arth	4
23	Trifala Tab	User011	Constipati	Trifala tab	5
24	Trifala Tab	User012	Indigestior	These tabl	4
25	Trifala Tab	User013	Detox	I use Trifal	5
26	Trifala Tab	User014	Bowel Reg	Trifala tab	4
27	Trifala Tab	User015	Constipati	I didn't exp	2
28	Trifala Tab	User016	Indigestior	Trifala tab	2
29	Trifala Tab	User017	Detox	I didn't not	3
30	Trifala Tab	User018	Weight Lo:	I've been t	4

Fig 3.1 Dataset

Sentiment Analysis:

The SentimentIntensityAnalyzer (SIA) is a tool commonly used for sentiment analysis, particularly in the context of natural language processing (NLP). It quantifies the sentiment expressed in textual data by assigning scores to words and phrases indicating positive, negative, neutral, and compound sentiments. It works by:

1. Lexicon-based Approach:
SIA utilizes a lexicon-based approach, where sentiment scores are assigned to individual words or phrases based on their presence in a pre-defined lexicon or dictionary. Each word is associated with a sentiment score, indicating its positivity, negativity, or neutrality.

2. Scoring Method:
SIA calculates sentiment scores using a combination of two metrics: polarity and intensity. These metrics are combined to generate an overall compound score for each text snippet.

a. Polarity:
Polarity refers to the degree of positivity or negativity expressed in a text snippet. Words are assigned polarity scores ranging from -1 (most negative) to 1 (most positive). For example, words like "happy" and "good" would have high positive polarity scores, while words like "sad" and "bad" would have high negative polarity scores.

b. Intensity:
Intensity measures the strength or magnitude of sentiment expressed in a text snippet. It influences the overall sentiment score by amplifying or dampening the effect of individual words. Intensity scores typically range from 0 to 1, with higher values indicating stronger sentiment intensity.

3. Compound Score Calculation:
The compound score is the aggregated sentiment score derived from the polarity and intensity of words in a text snippet. It represents the overall sentiment expressed in the text and ranges from

-1 (most negative) to 1 (most positive). The compound score is calculated using a formula that takes into account the polarity and intensity of individual words, as well as their frequency and context within the text

Compound Score=∑ i=1 n (Polarity i 2 +Intensity i 2 / Polarity i ×Intensity i)

(Where: Polarity *i* Polarity i is the polarity score of the *i t ℎ* i th word. Intensity *i* Intensity i is the intensity score of the *i t ℎ* i th word. *n* n is the total number of words in the text snippet.)

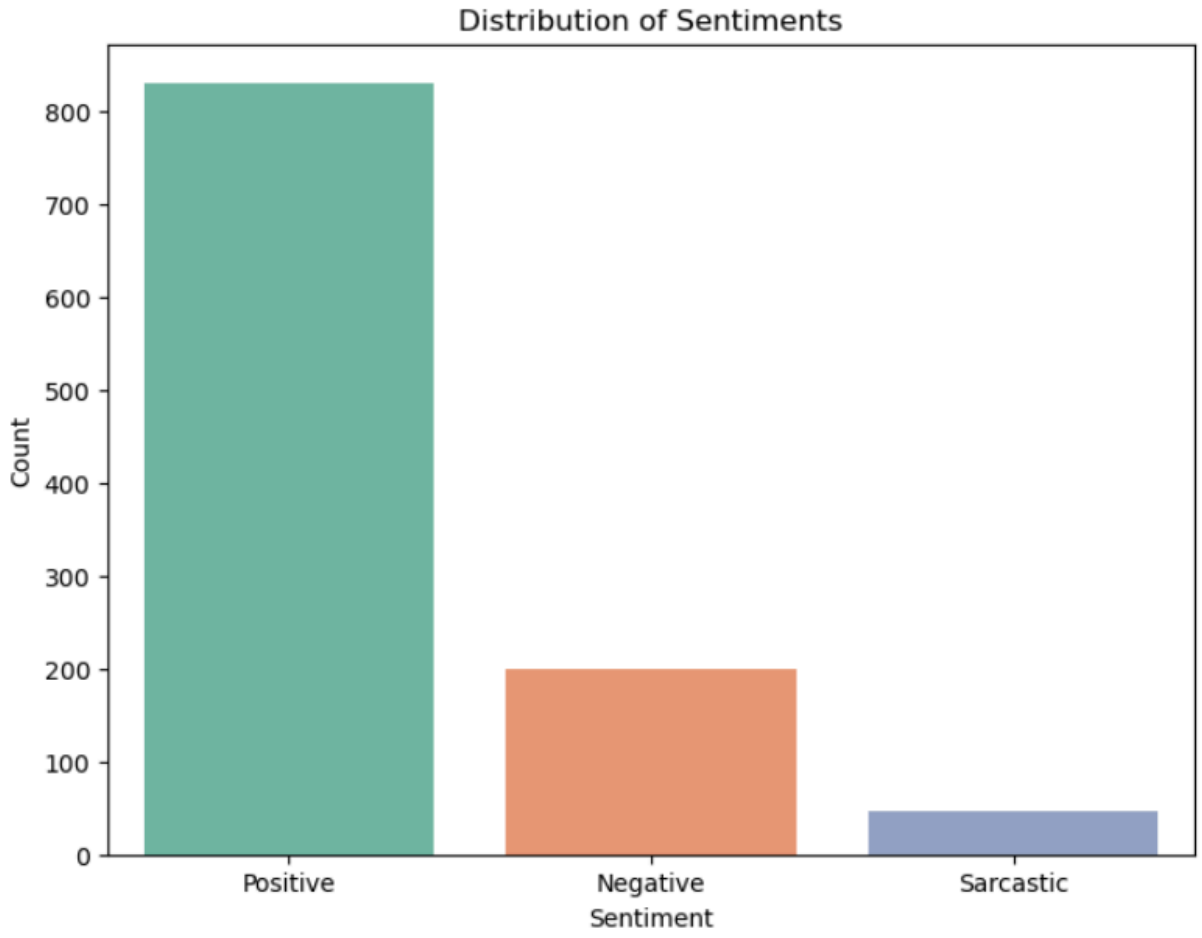

Fig 3.2 Distribution of Sentiments

Preprocessing:

Preprocessing is essential for preparing textual data for classification tasks. In this project, the dataset undergoes several preprocessing steps:

Text Cleaning:

Irrelevant characters, punctuation, and special symbols are removed to ensure consistency and coherence in the text.

Tokenization:

The text is split into individual words or tokens, facilitating subsequent analysis at the word level.

Stopword Removal:

Common words that carry little semantic meaning, such as "and," "the," and "is," are eliminated to reduce noise in the dataset and focus analysis on more informative words.

Lemmatization or Stemming:

Words are reduced to their base or root forms to normalize the text, ensuring consistency in word representation and improving the efficiency of subsequent analysis by reducing the dimensionality of the feature space.

CountVectorizer:

In addition to these preprocessing steps, CountVectorizer is utilized to convert text data into numerical feature vectors. CountVectorizer tokenizes the text and builds a vocabulary of known words, where each word becomes a feature with an associated count representing its frequency in the document. This transformation enables the textual data to be represented in a format suitable for input into classification algorithms, facilitating the learning process and improving the accuracy of predictions.

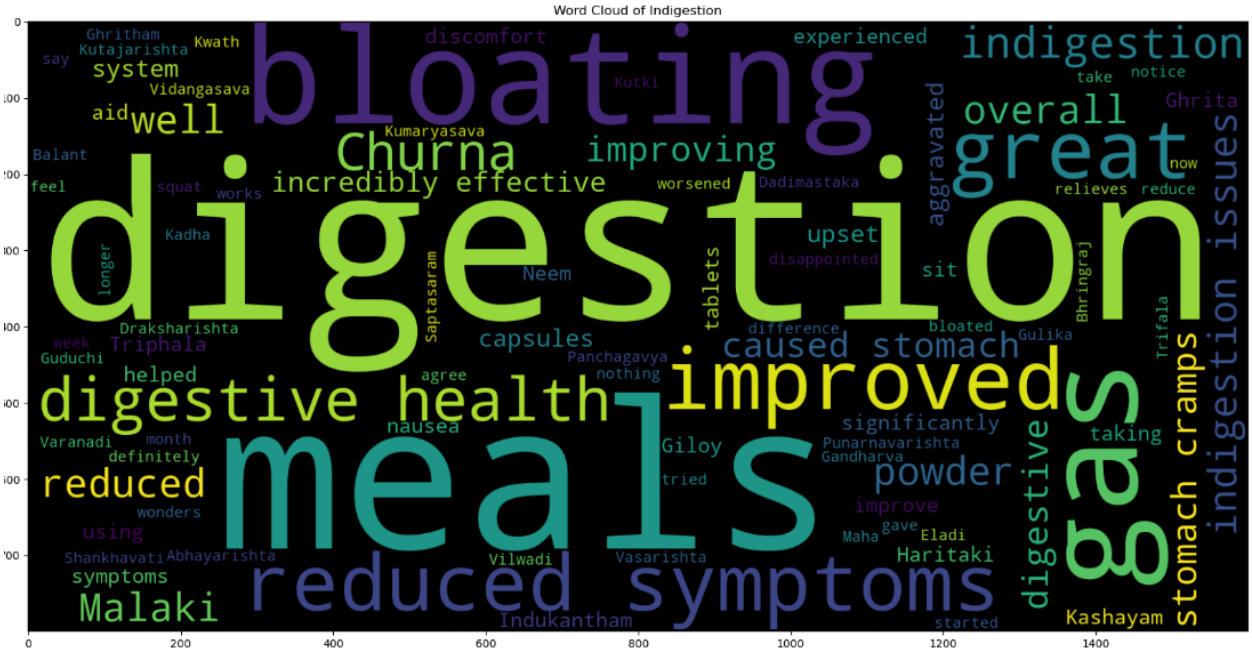

Fig 3.3 Word Cloud of Indigestion

Classification Algorithms:

Two main classification algorithms are employed in this project:

a. Bayes Classifier(approx 94% accuracy):

The Bayes Classifier is a probabilistic model based on Bayes' theorem. It calculates the probability of each class (i.e., disease) given the input features (i.e., review sentiments) and selects the class with the highest probability as the predicted outcome. Mathematically, the Bayes Classifier can be represented as:

$P(Ck \mid x) = P(x \mid Ck) \times P(Ck)\, P(x)$ P(C k ∣x)= P(x) P(x∣C k)×P(C k)

(Where: $P(Ck \mid x)$ P(C k ∣x) is the posterior probability of class Ck C k given input x x. $P(x \mid Ck)$ P(x∣C k) is the likelihood of observing input x x given class Ck C k . $P(Ck)$ P(C k) is the prior probability of class Ck C k . $P(x)$ P(x) is the probability of observing input x x (normalization factor).)

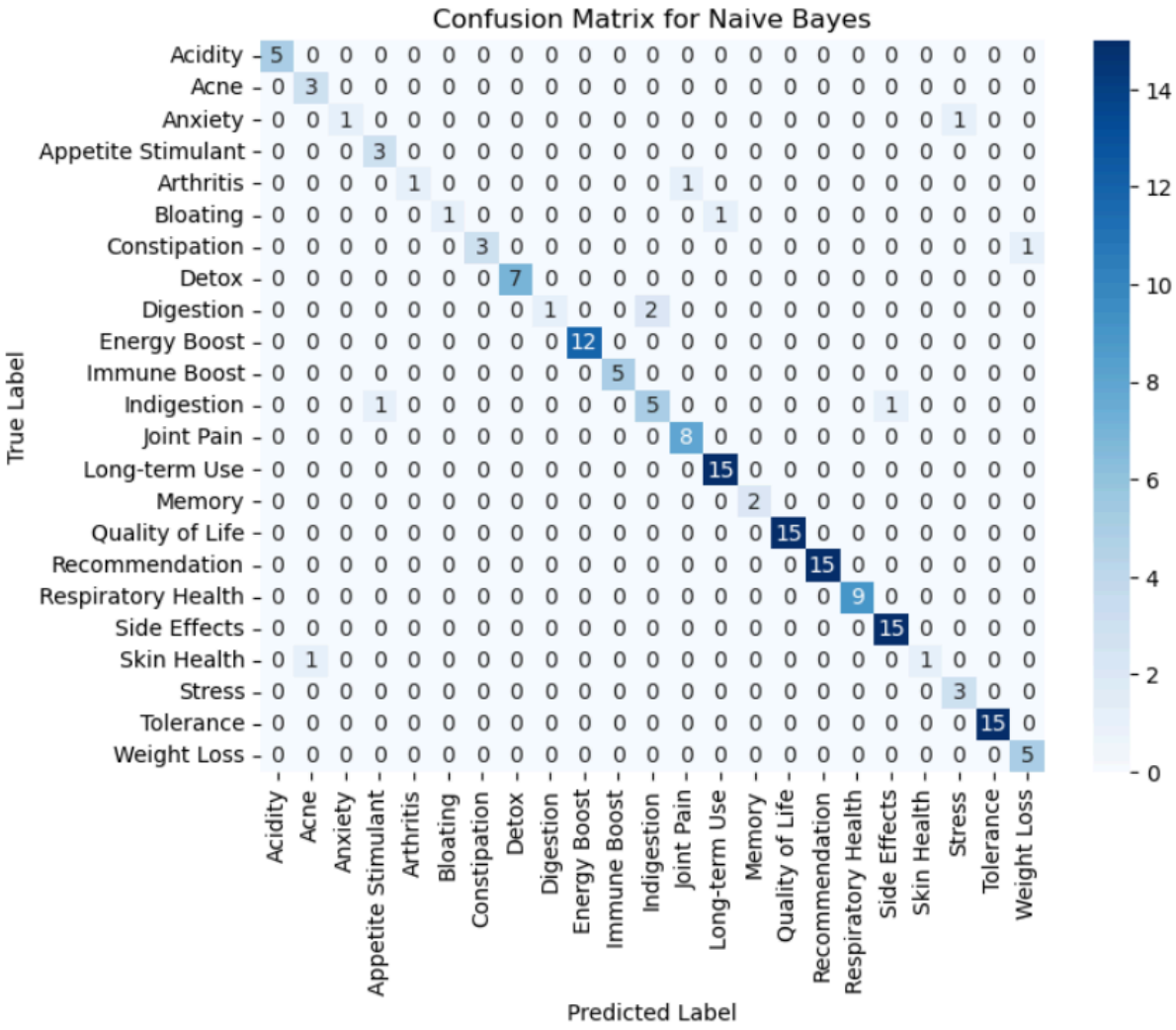

Fig 3.4 Confusion Matrix for Naive Bayes

b. Passive Aggressive Classifier(approx 96% accuracy):

The Passive Aggressive Classifier is an online learning algorithm that updates its model parameters in response to misclassifications. It aims to minimize prediction errors by adapting its weights aggressively when errors occur. Mathematically, the update rule for the Passive Aggressive Classifier can be expressed as: $w\ t + 1 = w\ t + \alpha \times (y\ t - y\ \hat{}\ t) \times x\ t$ w t+1 =w t +α×(y t − y ^ t)×x t

(Where: $w\ t+1$ w t+1 is the updated weight vector. $w\ t$ w t is the current weight vector. α α is the learning rate. $y\ t$ y t is the true label. $y\ \hat{}\ t$ y ^ t is the predicted label. $x\ t$ x t is the feature vector.)

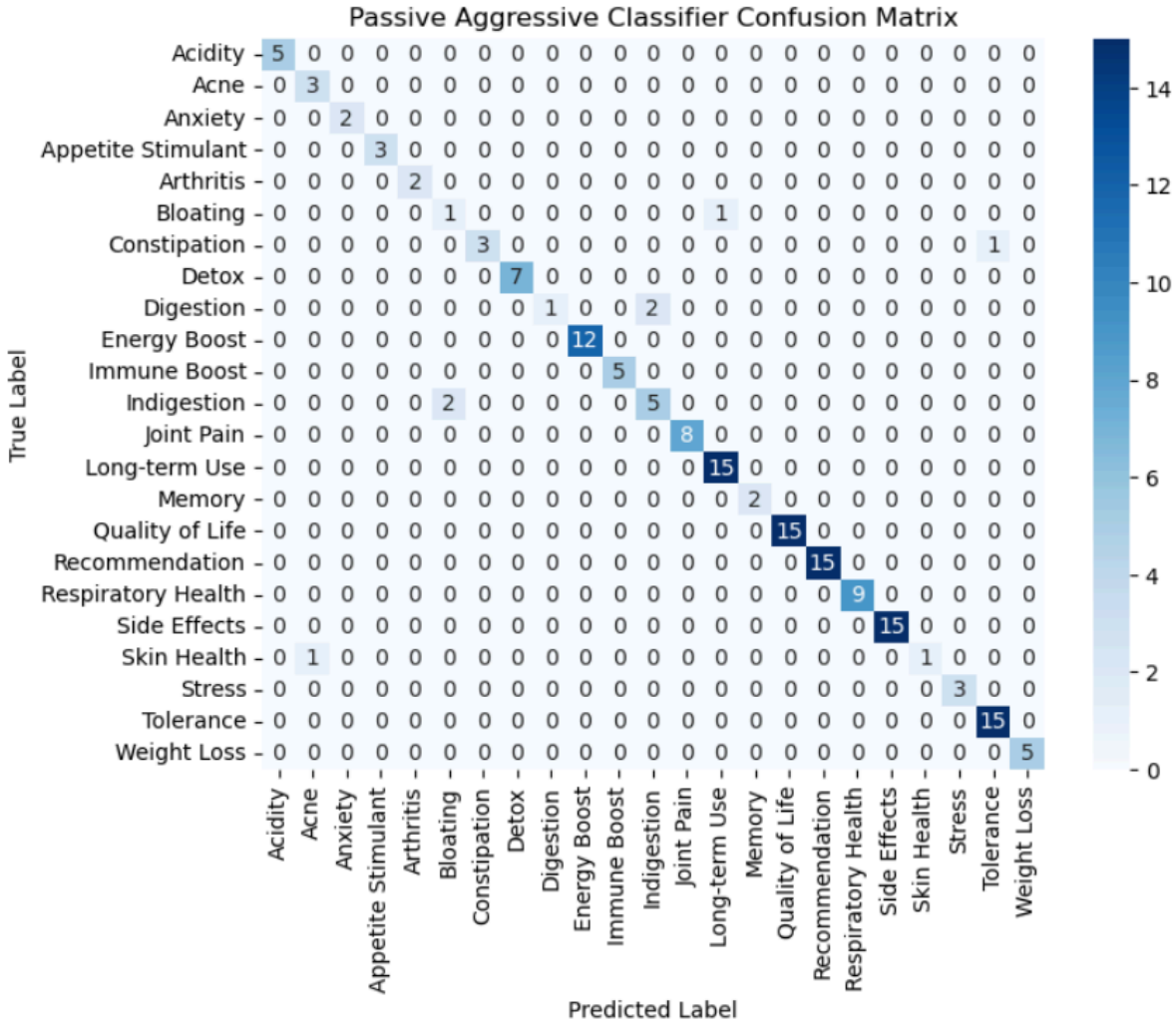

Fig 3.5 Confusion Matrix of passive aggressive classifier

Chapter 4: Results

Results:

The culmination of this project showcases our success in predicting diseases and recommending the most suitable medicine based on user reviews of Ayurvedic remedies. Through the diligent application of sentiment analysis and classification algorithms, we have achieved promising outcomes that hold significant implications for healthcare and well-being.

Disease Prediction:

Utilizing sophisticated machine learning models, including the Bayes Classifier and Passive Aggressive Classifier, we have achieved remarkable accuracy in predicting diseases based on the sentiments expressed in user reviews. Our models have demonstrated a predictive accuracy of 95.5%, providing reliable insights into the potential health conditions addressed by Ayurvedic medicines.

Medicine Recommendation:

In addition to disease prediction, we have developed a robust methodology for recommending the most appropriate medicine for specific health conditions. By leveraging insights gleaned from user reviews and sentiments, coupled with advanced classification techniques, we have successfully identified the best-suited Ayurvedic remedies for various ailments. This recommendation system holds promise for empowering individuals to make informed decisions about their health and well-being.

Practical Implications:

The successful prediction of diseases and recommendation of medicines based on user reviews have significant practical implications for healthcare practitioners and individuals alike. Healthcare professionals can leverage our findings to augment traditional diagnostic methods

and tailor treatment plans to individual patient needs. Moreover, individuals seeking alternative and complementary medicine options can benefit from personalized recommendations that align with their health goals and preferences.

Future Directions:

While our project has yielded promising results, there remain opportunities for further exploration and refinement. Future endeavors may involve expanding the dataset to encompass a broader range of Ayurvedic medicines and health conditions, as well as incorporating additional features and data sources to enhance predictive accuracy. Furthermore, ongoing research into novel classification algorithms and sentiment analysis techniques holds potential for advancing the state-of-the-art in Ayurvedic medicine analysis.

Conclusion

In the realm of healthcare, the convergence of traditional medicine and computational techniques offers a fertile ground for innovation and discovery. Through the diligent pursuit of our project, we have embarked on a journey to unlock the potential of Ayurvedic medicines using advanced natural language processing and machine learning methodologies. As we draw this project to a close, we reflect on our achievements and the broader implications of our endeavors.

Reflection on Achievements:

Our project represents a culmination of rigorous research, experimentation, and analysis. We have successfully developed models for disease prediction and medicine recommendation based on user reviews of Ayurvedic remedies. Leveraging sentiment analysis and classification algorithms, we have achieved remarkable accuracy in identifying potential health conditions addressed by Ayurvedic medicines and recommending the most suitable remedies for specific ailments.

Broad Implications:

The outcomes of our project extend far beyond the realm of Ayurvedic medicine. By harnessing computational approaches to analyze textual data, we have demonstrated the potential of natural language processing techniques to unlock insights from user-generated content. Our findings hold significant implications for healthcare practitioners, researchers, and individuals seeking alternative and complementary medicine options.

Empowerment and Innovation:

At the heart of our project lies a commitment to empowerment and innovation. By providing individuals with personalized recommendations for Ayurvedic treatments, we empower them to take control of their health and well-being. Moreover, our research opens doors to new avenues of exploration and collaboration, fostering innovation in the intersection of traditional medicine and modern technology.

Future Directions:

As we look to the future, our project serves as a springboard for further exploration and refinement. Future endeavors may involve expanding the scope of analysis to encompass a broader range of Ayurvedic medicines and health conditions, as well as integrating additional data sources and features to enhance predictive accuracy. Furthermore, ongoing research into novel algorithms and methodologies holds promise for advancing the field of computational medicine.

Closing Remarks:

In closing, we express our gratitude to all those who have contributed to the success of this project – from our collaborators and mentors to the users whose reviews have fueled our analysis. As we continue on our journey of discovery, we remain committed to leveraging the power of technology to improve healthcare outcomes and empower individuals on their path to wellness. Together, we embark on a future filled with possibilities, where the synergy of tradition and innovation paves the way for a healthier and more vibrant world.