DRUGREVIEW ANALYSIS USING MACHINE LEARNING

Mrs.GomathiM[1], Abishek.E[2]Arulamuthan.S[2]and JoelRockson.G[2]
Assistant Professor[1], Final year[2]
Department of Information Technology
St.Joseph College of Engineering
Sriperumbudur , Chennai-602 117

Abstract—**This paper introduces a robust, AI-powered framework for the automated analysis of drug reviews, aimed at extracting actionable insights into patient sentiment, therapeutic efficacy, and adverse effects. The system integrates Natural Language Processing (NLP), supervised Machine Learning (ML), and Generative AI to perform multi-faceted analysis, including sentiment classification, rating prediction, thematic topic modeling, and abstractive summarization. Built using Python, Streamlit, and MongoDB, the framework leverages established models such as Support Vector Machine (SVM), Linear Regression, Random Forest, XGBoost, and Latent Dirichlet Allocation (LDA). A notable enhancement is the incorporation of Generative AI (Gemini Pro), which enables personalized and concise summarization of patient feedback, enhancing interpretability and user engagement.Designed for accessibility and scalability, the application provides an interactive dashboard for real-time analysis, catering to patients, healthcare practitioners, and pharmaceutical stakeholders. By transforming unstructured textual reviews into structured insights, this system offers a meaningful contribution to data-driven pharmacovigilance, patient-centric care, and informed clinical decision-making.**

I.INTRODUCTION

In the modern digital era, online drug reviews have emerged as a vital source of information, offering deep insights into medication effectiveness and patient experiences. The widespread availability of such reviews allows patients and healthcare providers to better understand the real-world performance of pharmaceuticals beyond clinical trials. However, the sheer volume and unstructured nature of this text data present challenges in extracting meaningful insights. To address this, the present study proposes an advanced system that employs Artificial Intelligence (AI), Machine Learning (ML), and Generative AI techniques to automate the analysis of drug reviews. By applying methods such as sentiment classification, regression analysis, topic modeling, and text summarization, the system aims to classify reviews aspositive, neutral, or negative, predict ratings, and identify key patient concerns such as side effects and drug efficacy. This analysis contributes to improved drug evaluation and supports data-driven healthcare decision-making.

II.EXISTING AND PROPOSING SYSTEM

Existing drug review analysis systems mostly rely on machine learning and deep learning techniques. Traditional models like Naïve Bayes, Support Vector Machines (SVM), and Random Forest use methods such as TF-IDF and Word2Vec to classify sentiments in reviews. However, these models struggle to understand the deeper context of medical terms and complex sentences. Deep learning models, such as LSTM and CNN, offer better performance by capturing sequential patterns in text, but they require large datasets and high computational power, making them expensive to train and deploy. Transformer-based models like BERT, BioBERT, and SciBERT handle medical context effectively and deliver state-of-the-art results, but they too are resource-intensive and require fine-tuning. Aspect-Based Sentiment Analysis (ABSA) models go further by detecting sentiments on specific aspects like drug effectiveness and side effects, but they are harder to scale for large datasets.

To overcome these challenges, the proposed system integrates AI, ML, and Generative AI technologies for a more efficient and scalable solution. It features

sentiment analysis using SVM, rating prediction through regression models, and topic modeling with LDA to identify key themes in reviews. Additionally, it includes a Generative AI module (Gemini Pro) that summarizes long reviews, making insights more readable. An interactive Streamlit web app allows users—patients, doctors, and pharmaceutical companies—to analyze reviews in real time, filter results, and gain meaningful insights without technical complexity. Thus, the proposed system offers a comprehensive, cost-effective, and user-friendly solution for modern drug review analysis.

In addition to addressing performance and scalability, the proposed system emphasizes accessibility and real-time analysis. Unlike heavy deep learning models that require specialized hardware, this system is designed to run efficiently on standard computing resources. By leveraging open-source libraries and deploying through cloud platforms like Streamlit Cloud or AWS, it offers a low-cost solution that can be accessed by users from any location. The inclusion of real-time sentiment prediction and rating estimation allows healthcare professionals and patients to instantly evaluate drug reviews and make informed decisions. Furthermore, by summarizing user feedback through Generative AI, the system simplifies complex and lengthy reviews, making drug evaluation quicker and more actionable for non-technical users.

III.SYSTEM STUDY

A. Technical Feasibility

The technical foundation of this system is robust and feasible with current technologies. It is implemented using Python 3.8+, leveraging popular machine learning libraries such as Scikit-learn, NLTK, and TextBlob. Streamlit is used to build the web-based user interface, providing an interactive and responsive platform for users. The system is compatible with major operating systems, including Windows, macOS, and Linux, and can be deployed on cloud platforms like Streamlit Cloud, AWS, or Heroku. These choices ensure that the system is technically sound and easy to deploy.

B. Economic Feasibility

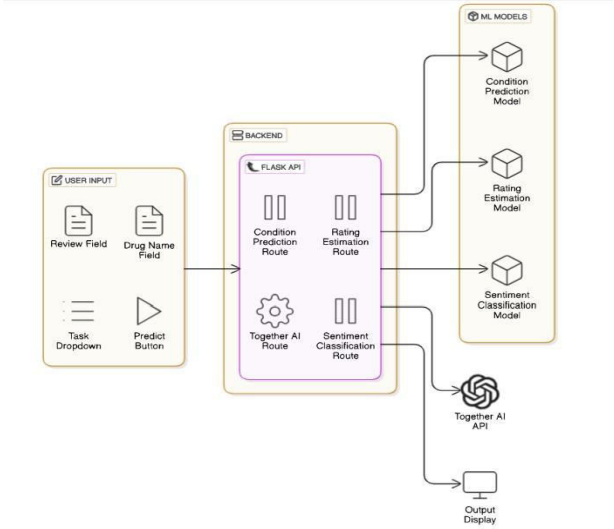
From an economic perspective, the system offers a cost-effective solution for drug review analysis. It is built using open-source tools, eliminating the need for expensive software licenses. Deployment on affordable cloud services minimizes infrastructure costs, and the system's architecture does not require high-end hardware, making it accessible to small and medium-sized healthcare institutions. By automating review analysis, the system reduces manual labor costs and improves efficiency.

C. Operational Feasibility

Operationally, the system is designed to be user-friendly and efficient. The Streamlit-based interface allows users with minimal technical knowledge to interact with the system easily. It supports real-time sentiment and rating predictions, enabling healthcare providers, patients, and pharmaceutical companies to gain instant insights. The modular design ensures that the system can be maintained and upgraded with minimal disruption to operations, supporting its practical deployment in real-world settings.

IV.ARCHITECTURE DIAGRAM

The system architecture is composed of multiple interconnected modules, each performing a specific function in the analysis pipeline. Data flows from the collection and preprocessing stages through feature engineering and model training, culminating in deployment on a Streamlit web application. The architecture supports real-time user input, allowing dynamic sentiment and rating analysis. A database component, such as MongoDB, can be integrated for storing user inputs and results, enabling longitudinal analysis and real-time monitoring.



.

V.MODULES

1. Data Collection and Preprocessing
2. Sentiment Analysis
3. Rating Prediction
4. Topic Modelling
5. Text Summarization
6. Visualization and Reporting

A. Data Collection and Preprocessing

This module gathers drug review data from publicly available datasets such as Drugs.com. It prepares the raw data by cleaning the text, which includes removing special characters, stopwords, and converting text to lowercase. Missing values are handled appropriately to ensure data quality. Features are extracted using TF-IDF vectorization and label encoding, transforming both text and categorical variables into machine-readable formats for further analysis.

B. Sentiment Analysis

In this module, each drug review is classified as positive, neutral, or negative using a Support Vector Machine (SVM) classifier. SVM is chosen for its robustness and high accuracy in text classification, especially when combined with TF-IDF features. This helps users quickly understand the overall sentiment trend associated with each drug.

C. Rating Prediction

This module uses regression models such as Linear Regression, Random Forest Regression, and XGBoost to predict the numerical rating of a drug based on the review text. By estimating a structured rating, this module offers a data-driven method for assessing drug effectiveness from patient feedback.

D. Topic Modelling

Using the Latent Dirichlet Allocation (LDA) algorithm, this module identifies hidden themes and recurring patterns within the reviews. It helps extract key topics such as effectiveness, side effects, and cost, providing deeper insights beyond simple sentiment classification.

E. Text Summarization

This module integrates Generative AI, specifically the Gemini Pro model, to summarize lengthy and detailed drug reviews into concise and easily readable summaries. This makes the insights more accessible, allowing users to quickly grasp key points without reading through extensive text

Visualization and Reporting

The final module presents the processed insights through an interactive web application developed with Streamlit. It allows users to filter reviews by drug names, view sentiment distributions, access predicted ratings, and explore summarized reviews in real time. The app also supports user input, enabling instant analysis of new reviews. give subheading only

VI.SCOPE OF FUTURE DEVELOPMENT

The proposed system offers significant potential for further enhancement and expansion. One major area of future development is the integration of real-time drug monitoring capabilities. By connecting the system to online health forums, pharmaceutical websites, and social media platforms, it can automatically collect and analyze newly posted reviews, allowing healthcare providers and regulatory authorities to detect emerging concerns or adverse effects in a timely manner. Additionally, expanding the system's capability to handle multilingual reviews will make it more inclusive and globally applicable. By incorporating multilingual natural language processing (NLP) models, the system can analyze feedback written in various languages, enabling it to serve a wider patient population beyond English-speaking users.

Moreover, the adoption of more advanced AI models, such as transformer-based architectures like BERT, BioBERT, or specialized healthcare language models, can further improve the system's accuracy in sentiment classification, rating prediction, and topic modeling. These models will enable better handling of complex medical terms and nuanced patient feedback. Another area of growth lies in integrating Explainable AI (XAI) techniques to enhance the interpretability of the system's predictions, helping users and healthcare professionals understand the reasoning behind specific sentiment classifications or rating estimates. Finally, developing a mobile

application version of the system will broaden accessibility, allowing patients and doctors to analyze drug reviews and receive personalized insights directly on their smartphones, thus extending the reach and practical impact of the solution.

VII.CONCLUSION

The proposed Drug Review Sentiment Analyzer presents an effective and comprehensive solution for automating the analysis of patient feedback on medications. By leveraging a combination of machine learning, natural language processing, and generative AI techniques, the system successfully classifies drug reviews into positive, neutral, and negative sentiments, predicts numerical ratings, and extracts key themes such as drug effectiveness and side effects. The integration of Generative AI for text summarization further enhances the system's usability by converting lengthy and complex reviews into concise summaries, making the insights more accessible to patients, doctors, and pharmaceutical companies. Additionally, the interactive Streamlit web application provides real-time analysis, allowing users to filter reviews, view sentiment trends, and submit their own feedback for instant evaluation. Overall, this system empowers users to make informed healthcare decisions based on collective patient experiences. Future enhancements, such as real-time monitoring, multilingual support, advanced AI models, and mobile app development, will further strengthen the system's accuracy, scalability, and global impact in healthcare analytics.

VIII.REFERENCES

[1] Drugs.com, "Medication Reviews Dataset," [Online]. Available: https://www.drugs.com. [Accessed: May 8, 2025].

[2] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[3] Scikit-learn, "Machine Learning in Python," [Online]. Available: https://scikit-learn.org. [Accessed: May 8, 2025].

[4] NLTK, "Natural Language Toolkit," [Online]. Available: https://www.nltk.org. [Accessed: May 8, 2025].

[5] Streamlit, "Streamlit Documentation," [Online]. Available: https://docs.streamlit.io. [Accessed: May 8, 2025].

[6] R. Vijayaraghavan and A. Basu, "Sentiment Analysis Using Machine Learning," *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2003.11643.

[7] A. Nair, P. Patel, and R. Jain, "Transformer Models in Healthcare," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2404.13057.

[8] F. Gräßer, P. Kallumadi, A. Malberg, and S. Zaunseder, "Aspect-Based Sentiment Analysis of Drug Reviews Using Hybrid Deep Learning," *ResearchGate*, 2018. [Online]. Available: https://www.researchgate.net/publication/324736492.