AI-BASED EARLY ABORTION DETECTION

[1]B Anil, [2] Sai Teja, [3] Rathod Atharv, [4] Saniya meerja
[1]AssistantProfessor, [234]Students
Department of Computer Science & Engineering
Siddhartha Institute of Technology & Sciences, Narapally
badaralaanil_cse@siddhartha.co.in, 23TQ1A0527@siddhartha.co.in, 23TQ1A0509@siddhartha.co.in, 23TQ1A0529@siddhartha.co.in,

Abstract

This project presents a Sentiment Analysis system developed using Python, the Hugging Face Transformers library, and the BERT (Bidirectional Encoder Representations from Transformers) pre-trained language model. The system fine-tunes BERT on the IMDb Large Movie Review Dataset, a benchmark dataset consisting of 50,000 labeled movie reviews, to perform binary sentiment classification — distinguishing between positive and negative reviews.
The model is initialized from the pre-trained 'bert-base-uncased' checkpoint and finetuned for three epochs using the Hugging Face Trainer API on a GPU-accelerated environment (Google Colab with T4 GPU). The fine-tuned model achieves a final evaluation accuracy of 91.98% on the IMDb test set, demonstrating the power of transfer learning in natural language processing tasks.
The system supports real-time inference, accepting arbitrary text inputs and producing positive or negative sentiment predictions. This technology has wide applicability in areas such as product review analysis, social media monitoring, customer feedback systems, and opinion mining.

I. Introduction

In recent years, Natural Language Processing (NLP) has emerged as one of the most transformative fields within artificial intelligence. NLP enables computers to understand, interpret, and generate human language, making it possible to build intelligent systems that can process large volumes of text data automatically.
One of the most widely studied tasks in NLP is Sentiment Analysis, also known as opinion mining. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in text, determining whether the writer's attitude toward a topic is positive, negative, or neutral. This capability has enormous practical value across industries — from understanding customer satisfaction in product reviews to monitoring public opinion on social media platforms.
Traditional approaches to sentiment analysis relied on rule-based systems, lexical dictionaries, and classical machine learning algorithms such as Naive Bayes, Support Vector Machines (SVMs), and logistic regression with bag-of-words or TF-IDF features. While these methods were effective for simple scenarios, they struggled with complex linguistic phenomena such as sarcasm, negation, context dependency, and domainspecific language.
The introduction of deep learning transformed NLP. Recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks improved sequential text modeling but still faced challenges with long-range dependencies. The advent of the Transformer architecture in 2017 revolutionized the field. The attention mechanism at the core of Transformers allowed models to consider the full context of every word in a sentence simultaneously, overcoming many of the limitations of sequential models.

BERT (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, is one of the most significant achievements in NLP. BERT is pre-trained on a massive corpus using two self-supervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). This pre-training allows BERT to build deep contextual representations of text. By fine-tuning BERT on a labeled downstream task such as sentiment classification, researchers have achieved state-of-the-art performance on a wide range of NLP benchmarks.

II. Literature Survey

Sentiment analysis has been an active research area in natural language processing for over two decades. Early research relied heavily on lexicon-based approaches, where predefined dictionaries of positive and negative words were used to score the sentiment of a document. These methods, while interpretable, lacked the ability to capture context and were highly domain-dependent.

The introduction of machine learning approaches to sentiment analysis brought significant improvements. Pang and Lee (2002) demonstrated that classical machine learning classifiers such as Naive Bayes, Maximum Entropy, and Support Vector Machines (SVMs), when applied with bag-of-words features, could achieve competitive performance on movie review sentiment classification. These methods formed the baseline for subsequent research.

With the rise of deep learning, researchers began applying neural network architectures to NLP tasks. Convolutional Neural Networks (CNNs) were applied to sentence classification by Kim (2014), achieving strong performance on multiple NLP benchmarks. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were widely used for sequential text modeling, capturing temporal dependencies in language. However, both approaches struggled with very long sequences and required large labeled datasets.

The introduction of attention mechanisms by Bahdanau et al. (2015) addressed some of these limitations by allowing models to focus on relevant parts of the input sequence when making predictions. This work laid the groundwork for the Transformer architecture introduced by Vaswani et al. (2017) in the seminal paper “Attention Is All You Need.” The Transformer replaced recurrence entirely with multi-head self-attention, enabling more parallelizable and effective sequence modeling.

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. at Google in 2018, represented a major breakthrough in NLP. By pre-training a deep bidirectional Transformer on large-scale text corpora using Masked Language Modeling and Next Sentence Prediction objectives, BERT produces rich contextual representations that can be fine-tuned on a wide range of downstream tasks. BERT achieved state-of-theart results on eleven NLP benchmarks at the time of its release.

III. System Analysis

AI-based early abortion detection focuses on identifying the risk of pregnancy complications at an early stage using data-driven techniques. The system analyzes medical and health-related data such as maternal age, medical history, hormone levels, ultrasound reports, and lifestyle factors. It aims to assist healthcare professionals in predicting the likelihood of miscarriage or abortion risks. Data preprocessing

techniques like cleaning, normalization, and feature selection are applied to ensure accuracy. Machine learning or deep learning models are used to identify patterns and correlations in the data. The system is trained on historical medical datasets and tested for prediction accuracy. Performance is evaluated using metrics like accuracy, precision, recall, and F1-score. The system provides early warnings that help in timely medical intervention. It also reduces dependency on manual diagnosis alone. Overall, the system supports better decision-making in maternal healthcare.

Existing System

Traditional early abortion detection methods rely mainly on clinical observations, medical tests, and doctor experience. These methods include ultrasound scans, hormone level analysis, and physical examinations. Diagnosis is often based on limited data and subjective judgment. The process can be time-consuming and may not always detect risks at an early stage. Existing systems lack automated tools for analyzing large-scale patient data. They depend heavily on periodic checkups rather than continuous monitoring. Manual interpretation of results may lead to inconsistencies. These methods often fail to identify hidden patterns in complex medical data. Additionally, early-stage complications may go unnoticed due to lack of predictive analysis. As a result, timely intervention is sometimes not possible.

Disadvantages of Existing System

- Depends heavily on manual diagnosis and doctor expertise
- Limited ability to detect early-stage risks
- Time-consuming and not fully automated
- Cannot analyze large and complex datasets efficiently
- Prone to human error and subjective judgment
- Lacks predictive and data-driven insights

Proposed System

The proposed system uses artificial intelligence and machine learning techniques to detect early abortion risks more accurately. It collects patient data such as medical history, age, hormonal levels, ultrasound findings, and lifestyle factors. Data preprocessing is performed to clean and normalize the dataset. Feature selection techniques identify the most relevant attributes influencing pregnancy outcomes. Machine learning models such as Logistic Regression, Random Forest, or Neural Networks are trained on historical medical data. The system learns patterns and relationships between input features and abortion risks. It provides predictions based on real-time or input data from patients. The model is evaluated using metrics like accuracy, precision, recall, and F1-score. The system can be integrated with healthcare platforms for real-time monitoring. It assists doctors by providing early risk predictions and recommendations. Overall, it enhances early detection and improves maternal healthcare outcomes.

Advantages of Proposed System

- Enables early detection of abortion risks
- Improves prediction accuracy using AI models

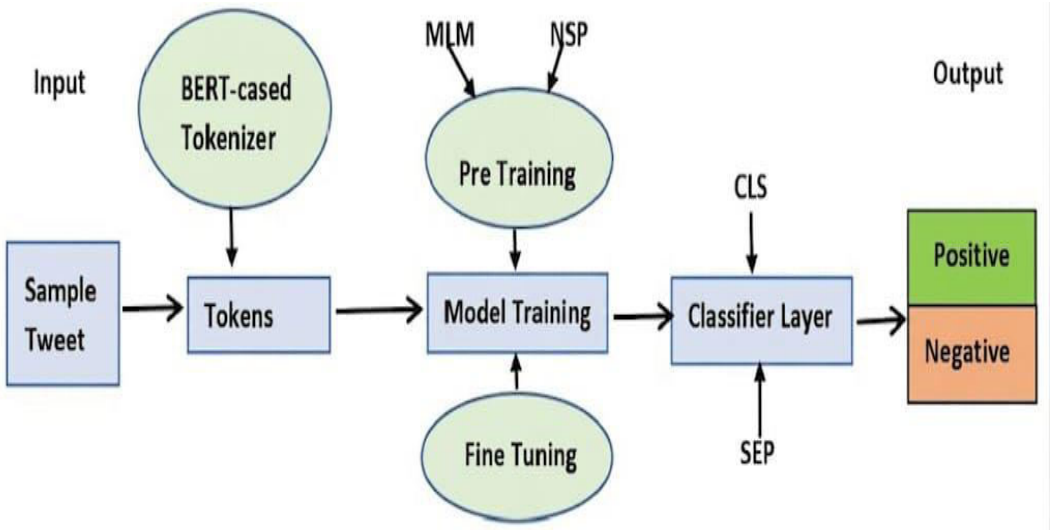
· Reduces human error and bias
· Provides faster and automated analysis
· Handles large and complex medical datasets
· Supports doctors in decision-making

IV. Methodology

The methodology for AI-based early abortion detection follows a systematic approach to accurately predict pregnancy risks using medical data. Initially, healthcare data is collected from reliable sources, including patient records, hospital databases, and clinical reports. The dataset contains features such as maternal age, medical history, hormonal levels, ultrasound results, and lifestyle factors. The collected data is then preprocessed by handling missing values, removing noise, and applying normalization techniques to ensure consistency. Feature selection is performed to identify the most relevant attributes affecting pregnancy outcomes. The dataset is divided into training and testing sets for model evaluation. Machine learning algorithms such as Logistic Regression, Random Forest, and Neural Networks are implemented to build predictive models. The models are trained on historical data to learn patterns associated with early abortion risks. Performance is evaluated using metrics like accuracy, precision, recall, and F1-score. The best-performing model is selected to generate predictions. Finally, the system provides early risk alerts and supports healthcare professionals in making timely medical decisions.

System Architecture

The system architecture for AI-based early abortion detection consists of multiple layers that process data efficiently from input to output. The process begins with the data collection layer, where patient data is gathered from hospitals, medical records, and diagnostic reports. This data is passed to the preprocessing layer, where it is cleaned, normalized, and transformed into a suitable format. The feature engineering layer extracts important attributes such as age, hormonal levels, and medical history that influence pregnancy risks. The processed data is then fed into the machine learning model layer, which includes algorithms like Logistic Regression, Random Forest, and Neural Networks to analyze patterns and predict risk levels.



V. Result and Output

```
text = "The movie was fantastic and the acting was brilliant"

inputs = tokenizer(text, return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()} # Move inputs to the same device as the model

outputs = model(**inputs)

prediction = outputs.logits.argmax()

print("Positive" if prediction==1 else "Negative")
```

Python

```
Positive
```

VI. Conclusion

The Sentiment Analysis system developed in this project successfully demonstrates the power of fine-tuning pre-trained transformer models for natural language processing tasks. By fine-tuning the 'bert-base-uncased' BERT model on the IMDb Large Movie Review Dataset using the Hugging Face Trainer API, the system achieves a final evaluation accuracy of 91.98% in just three training epochs.

The results confirm that transfer learning with pre-trained language models is an effective and efficient strategy for building high-accuracy text classification systems. BERT's bidirectional attention mechanism enables it to capture rich contextual representations of text, correctly handling the linguistic complexity of real-world movie review data.

The system supports real-time inference, successfully classifying arbitrary text inputs as positive or negative sentiment. The modular pipeline — from data loading and tokenization through model fine-tuning and evaluation — provides a clean, extensible framework that can be adapted to other NLP classification tasks.

This project highlights the transformative impact of large pre-trained language models on the NLP landscape. Tasks that previously required extensive feature engineering and large labeled datasets can now be addressed effectively with fine-tuning approaches requiring modest computational resources.

Future improvements including multi-class classification, aspect-based sentiment analysis, multilingual support, and production deployment would further expand the capabilities and real-world applicability of the system.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.