Medical Text Analysis System

Nishanth Hanmanthureddygari, Revanth Kumar Chekuri, Satheesh Meadi, Srilakshmi Pyarsani
Department of Data Science
University of Maryland Baltimore County
DATA 690 - Introduction to Natural Language Processing
Prof: Dr Antony Diana
Fall 2024

***Abstract* - This paper introduces the Medical Text Analysis System, an advanced platform for analyzing medical text through summarization, translation, sentiment analysis, Named Entity Recognition (NER), and an interactive Q&A chatbot. Its modular architecture includes three primary layers: an Input Layer for diverse data formats, a Core Processing Layer powered by pre-trained NLP models, and an Integration Layer that provides a user-friendly interface via Streamlit. This framework delivers a unified, scalable, and customizable solution, overcoming traditional text analysis limitations. It automates the extraction of insights from complex clinical notes, patient feedback, and research abstracts, reducing time and effort while improving data accessibility and decision-making. Advanced features such as multilingual support, dynamic visualizations, and interactive chatbot functionality enhance its versatility. MTAS provides a powerful tool for healthcare professionals and researchers, enabling evidence-based decision support, improving research efficiency, and fostering better patient outcomes through actionable insights and comprehensive data analysis.**

Keywords: *Medical Text Analysis System, Biomedical NLP, Named Entity Recognition, Sentiment Analysis, Text Translation.*

1.INTRODUCTION

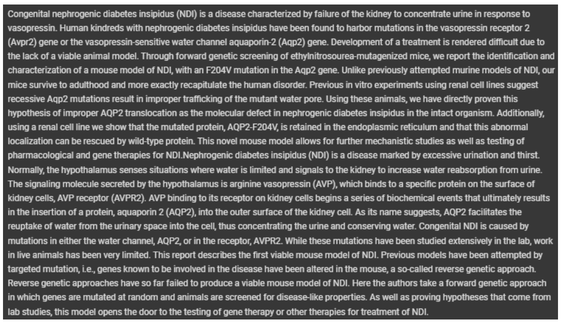
The healthcare field generates vast amounts of data from various sources, including medical documents like clinical notes and research papers as well as patient records. Analyzing this data traditionally involved manual review and basic keyword-based tools which had limitations in scalability and accuracy[1]. The complex and specialized nature of documents makes it challenging to extract information. Although current solutions handle tasks well individually, they often cannot fully integrate for comprehensive text analysis of text data. This division creates a gap in the processing of healthcare information that could affect decision-making outcomes as well as the efficiency of research and the quality of patient care provided to patients in need of medical attention and treatment services. To tackle these obstacles effectively and efficiently that have been hindering progress in the healthcare sector for quite some time now we introduce a Medical Text Analysis System that makes use of PubMedBERT specialized language comprehension capabilities. This system combines Natural Language Processing (NLP) methods to carry out analytical tasks such as condensing text information into summaries, identifying named entities mentioned in the text, analyzing sentiments conveyed within the text content, and translating content across different languages, for broader accessibility and understanding. This thorough strategy addresses the drawbacks of techniques by enhancing data retrieval and supporting healthcare decisions based on evidence.

2. LITERATURE REVIEW

Biomedical text summarization has evolved from basic computational methods using term frequency and sentence position to advanced deep learning approaches. Early systems struggled with the complexity of biomedical texts [1]. Domain-specific tools like UMLS (Unified Medical Language System) improved context understanding but required extensive maintenance [2]. Models such as BERT and BioBERT enabled dynamic, context-aware representations, are not significantly enhancing accurate data. Lastly, the model's performance may degrade when handling multi-document or query-based tasks, requiring further research to enhance adaptability and scalability[4].

3.DATASET DESCRIPTION

The dataset is sourced from the JohnSnowLabs GitHub repository. It comprises up to 1,000 text files, containing abstracts from medical research articles, with each file focusing on a specific medical condition or study. A sample abstract discusses congenital nephrogenic diabetes insipidus (NDI), exploring its genetic causes, mechanisms, and potential therapies. The data highlights mutations in the vasopressin receptor 2 (AVPR2) and aquaporin-2 (AQP2) genes as key contributors to the condition. Such files provide rich biomedical insights, making the dataset ideal for tasks like summarization, named entity recognition, and sentiment analysis in medical NLP research.



Congenital nephrogenic diabetes insipidus (NDI) is a disease characterized by failure of the kidney to concentrate urine in response to vasopressin. Human kindreds with nephrogenic diabetes insipidus have been found to harbor mutations in the vasopressin receptor 2 (Avpr2) gene or the vasopressin-sensitive water channel aquaporin-2 (Aqp2) gene. Development of a treatment is rendered difficult due to the lack of a viable animal model. Through forward genetic screening of ethylnitrosourea-mutagenized mice, we report the identification and characterization of a mouse model of NDI, with an F204V mutation in the Aqp2 gene. Unlike previously attempted murine models of NDI, our mice survive to adulthood and more exactly recapitulate the human disorder. Previous in vitro experiments using renal cell lines suggest recessive Aqp2 mutations result in improper trafficking of the mutant water pore. Using these animals, we have directly proven this hypothesis of improper AQP2 translocation as the molecular defect in nephrogenic diabetes insipidus in the intact organism. Additionally, using a renal cell line we show that the mutated protein, AQP2-F204V, is retained in the endoplasmic reticulum and that this abnormal localization can be rescued by wild-type protein. This novel mouse model allows for further mechanistic studies as well as testing of pharmacological and gene therapies for NDI.Nephrogenic diabetes insipidus (NDI) is a disease marked by excessive urination and thirst. Normally, the hypothalamus senses situations where water is limited and signals to the kidney to increase water reabsorption from urine. The signaling molecule secreted by the hypothalamus is arginine vasopressin (AVP), which binds to a specific protein on the surface of kidney cells, AVP receptor (AVPR2). AVP binding to its receptor on kidney cells begins a series of biochemical events that ultimately results in the insertion of a protein, aquaporin 2 (AQP2), into the outer surface of the kidney cell. As its name suggests, AQP2 facilitates the reuptake of water from the urinary space into the cell, thus concentrating the urine and conserving water. Congenital NDI is caused by mutations in either the water channel, AQP2, or in the receptor, AVPR2. While these mutations have been studied extensively in the lab, work in live animals has been very limited. This report describes the first viable mouse model of NDI. Previous models have been attempted by targeted mutation, i.e., genes known to be involved in the disease have been altered in the mouse, a so-called reverse genetic approach. Reverse genetic approaches have so far failed to produce a viable mouse model of NDI. Here the authors take a forward genetic approach in which genes are mutated at random and animals are screened for disease-like properties. As well as proving hypotheses that come from lab studies, this model opens the door to the testing of gene therapy or other therapies for treatment of NDI.

Sample Text Data From Zip File

4. SYSTEM ARCHITECTURE

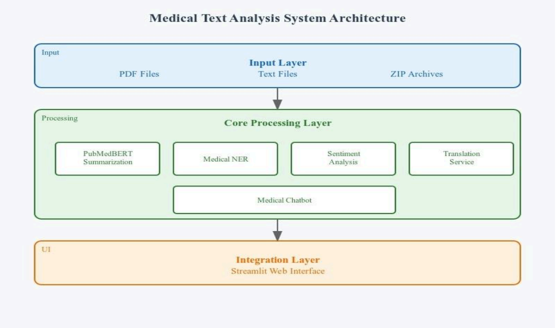


***Fig 1.** Block Diagram of the system*

- The Medical Text Analysis System is structured to ensure seamless processing, analysis, and presentation of insights from medical text data.
- The architecture consists of three primary layers: Input Layer, Core Processing Layer, and Integration Layer. Each layer serves distinct functionalities while working to deliver an end-to-end solution.

Input Layer

The input layer is designed to handle various data formats such as PDF files, text files, and ZIP archives ensuring compatibility with diverse medical document types. It serves as a unified entry point for ingesting raw data into the system.

Core Processing Layer

The core of our system architecture comprises an integrated suite of Natural Language Processing modules, each specialized for medical domain applications [11]. The system incorporates five key components:

- **PubMedBERT Summarization:** Implements a transformer-based architecture to extract crucial information from complex medical documents while maintaining clinical accuracy through domain-specific training.
- **Medical Named Entity Recognition (NER)**: Systematically identifies and labels medical entities (diseases, medications, procedures, biological terms) within unstructured clinical narratives, enabling structured data extraction.
- **Sentiment Analysis**: Evaluates medical text through sentiment classification, visualizing results with a color-coded graph (green for positive, red for negative) and providing confidence scores to help healthcare providers assess patient feedback and clinical communications.
- **Translation Service**: Facilitates multilingual support by translating medical text into user-preferred languages.
- **Medical Chatbot:** Offers an intuitive conversational interface for navigating system features and medical queries, leveraging PubMedBERT to ensure accurate and contextual responses.

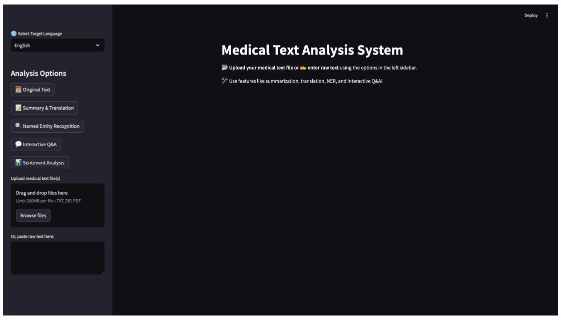
These components are built upon pre-trained models and frameworks like PubMedBERT and Hugging Face Transformers, specifically optimized for medical applications through domain-specific fine-tuning.

Integration Layer

The Integration Layer connects the backend processing with the user-interface. It employs the **Streamlit Web Interface**, providing an intuitive, browser-based platform for healthcare professionals. This layer ensures that insights generated by the Core Processing Layer are accessible in a user-friendly format, enabling smooth visualization and interaction.

5. PRODUCT DEMONSTRATION

5.1 *User Interface Overview:*



Medical Text Analysis System

📁 **Upload your medical text file** or ✍️ **enter raw text** using the options in the left sidebar.

🛠️ Use features like summarization, translation, NER, and interactive Q&A!

***Fig 2.** User Interface of the Model*

The sidebar provides a range of analysis options, such as Summary, Translation, Named Entity Recognition (NER), Sentiment Analysis, Interactive Q&A, and viewing the Original Text. Users can upload raw information directly into the system. The central design prioritizes efficiency and adaptability, making complex text analysis accessible to non-technical users.

5.2 Feature Demonstration

Text Input & Summarization:

This interface seamlessly integrates with the underlying PubMedBERT model, which processes the input through its transformer-based architecture. The system displays the original text. The implementation handles document processing through a pipeline that includes text preprocessing, embedding generation, and feature extraction. Additionally, it ensures efficient management of computational resources, enabling scalability for processing large datasets.

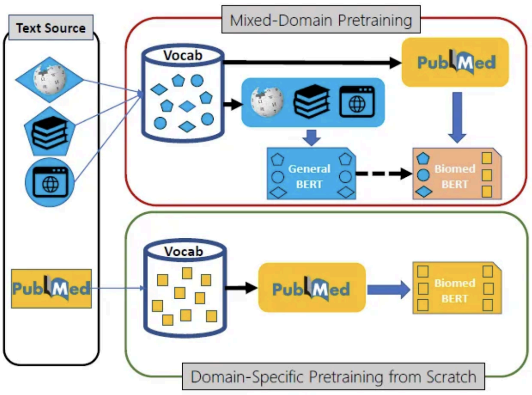
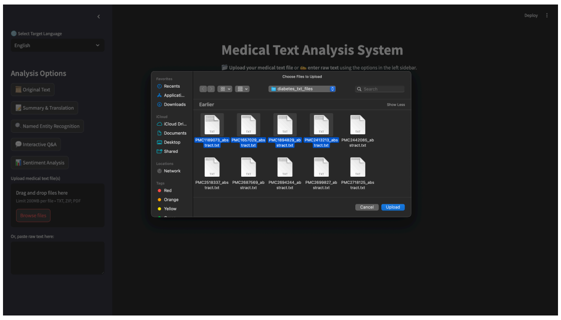


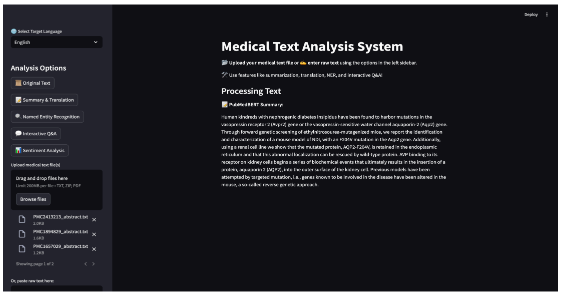
***Fig 3**. PubMedBERT Architecture (Source: [1])*

Users can input medical text through direct file upload (supporting formats like TXT, ZIP, and PDF) or by pasting raw text to the system.


Medical Text Analysis System

📂 Upload your medical text file or ✍️ enter raw text using the options in the left sidebar.

***Fig 4.** Uploading multiple text files.*


Medical Text Analysis System

📂 **Upload your medical text file** or ✍️ **enter raw text** using the options in the left sidebar.

🛠️ Use features like summarization, translation, NER, and interactive Q&A!

Processing Text

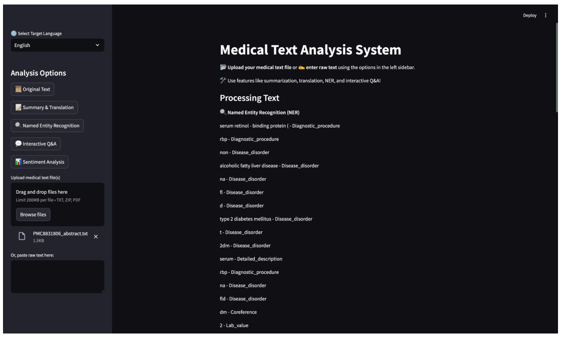
📝 **PubMedBERT Summary:**

Human kindreds with nephrogenic diabetes insipidus have been found to harbor mutations in the vasopressin receptor 2 (Avpr2) gene or the vasopressin-sensitive water channel aquaporin-2 (Aqp2) gene. Through forward genetic screening of ethylnitrosourea-mutagenized mice, we report the identification and characterization of a mouse model of NDI, with an F204V mutation in the Aqp2 gene. Additionally, using a renal cell line we show that the mutated protein, AQP2-F204V, is retained in the endoplasmic reticulum and that this abnormal localization can be rescued by wild-type protein. AVP binding to its receptor on kidney cells begins a series of biochemical events that ultimately results in the insertion of a protein, aquaporin 2 (AQP2), into the outer surface of the kidney cell. Previous models have been attempted by targeted mutation, i.e., genes known to be involved in the disease have been altered in the mouse, a so-called reverse genetic approach.

***Fig 5**. Text Summarization*

Named Entity Recognition(NER)

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique used to extract and identify essential information from text, categorizing it into predefined entities, In medical text analysis, NER is used to identify and classify key entities within unstructured text [5], such as diseases, procedures, medications, and clinical findings[10]. In our application, we use the "d4data/biomedical-ner-all," a transformer-based model specifically optimized for biomedical text, to accurately extract relevant medical entities. We initialized the NER pipeline using Hugging Face's pipeline function, loading the model for smooth integration.


Medical Text Analysis System

📂 **Upload your medical text file** or ✍️ **enter raw text** using the options in the left sidebar.

🛠️ Use features like summarization, translation, NER, and interactive Q&A!

Processing Text

🔍 **Named Entity Recognition (NER)**

serum retinol - binding protein (- Diagnostic_procedure

rbp - Diagnostic_procedure

non - Disease_disorder

alcoholic fatty liver disease - Disease_disorder

na - Disease_disorder

fl - Disease_disorder

d - Disease_disorder

type 2 diabetes mellitus - Disease_disorder

t - Disease_disorder

2dm - Disease_disorder

serum - Detailed_description

rbp - Diagnostic_procedure

na - Disease_disorder

fld - Disease_disorder

dm - Coreference

2 - Lab_value

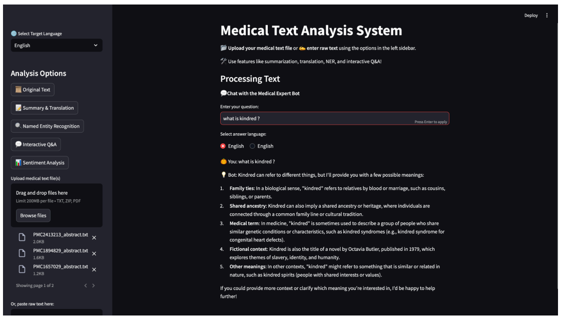
***Fig 6.** NER(Name Entity Recognition) Image.*

As shown in *Fig 6,* when processing biomedical text, the NER module efficiently extracts and categorizes key entities.

Interactive Medical QA(Chatbot):

The system employs a locally deployed Ollama server [12] integrated with an OpenAI-compatible API to power its interactive medical chatbot. A custom medical chatbot class enables real-time interaction and context-aware responses tailored to user queries. The system employs context awareness by storing and reusing the last three interactions in its conversation history, ensuring continuity and precision in multi-turn conversations. Each response is enriched by leveraging both the question and the provided medical context. chatbot uses the Llama 2 model, deployed through Ollama's framework for high-quality language understanding and generation [6,7]. By supporting input for both user queries, the chatbot adjusts its responses to align with the specific requirements of medical professionals or researchers.

Users can reset the interaction history using a clear history feature, enabling flexibility in managing new, independent conversations without retaining prior data. Chatbot QA can be used in clinical decision support, Medical Education, and Research Assistance. The use of a local Ollama server ensures data privacy and security, a critical requirement in healthcare settings where sensitive information may be involved. The architecture allows easy extension to support additional features, such as multilingual Q&A or integration with other NLP tasks like summarization and named entity recognition.



PMC2413213_abstract.txt 2.0KB
PMC1894829_abstract.txt 1.6KB
PMC1657029_abstract.txt 1.2KB

Medical Text Analysis System

📁 **Upload your medical text file** or ✍ **enter raw text** using the options in the left sidebar.

🛠 Use features like summarization, translation, NER, and interactive Q&A!

Processing Text

💬**Chat with the Medical Expert Bot**

Enter your question: what is kindred ?

Select answer language: ● English ○ English

🎃 You: what is kindred ?

💡 Bot: Kindred can refer to different things, but I'll provide you with a few possible meanings:

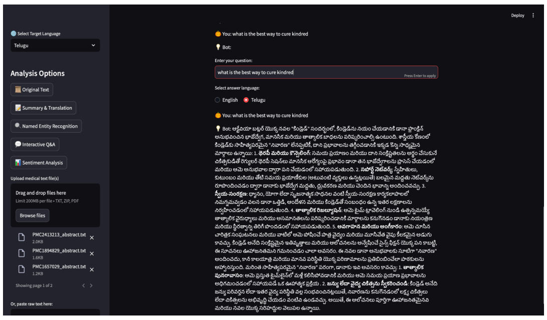
1. **Family ties**: In a biological sense, "kindred" refers to relatives by blood or marriage, such as cousins, siblings, or parents.
2. **Shared ancestry**: Kindred can also imply a shared ancestry or heritage, where individuals are connected through a common family line or cultural tradition.
3. **Medical term**: In medicine, "kindred" is sometimes used to describe a group of people who share similar genetic conditions or characteristics, such as kindred syndromes (e.g., kindred syndrome for congenital heart defects).
4. **Fictional context**: Kindred is also the title of a novel by Octavia Butler, published in 1979, which explores themes of slavery, identity, and humanity.
5. **Other meanings**: In other contexts, "kindred" might refer to something that is similar or related in nature, such as kindred spirits (people with shared interests or values).

If you could provide more context or clarify which meaning you're interested in, I'd be happy to help further!

Fig 7. Q&A Session

Multi-lingual support (Translation)

The system's translation module accurately translates medical content into multiple languages. This improves access for non-English-speaking healthcare practitioners and researchers. The interactive interface enables users to upload raw text or medical documents and select a destination language. It has real-time translation and other NLP functions including summarization and named entity recognition. The Google Translation API v2 is used to offer accurate and dependable translations of medical documents[8].



PMC2413213_abstract.txt 2.0KB
PMC1894829_abstract.txt 1.6KB
PMC1657029_abstract.txt 1.2KB

🎃 You: what is the best way to cure kindred

💡 Bot:

Enter your question: what is the best way to cure kindred

Select answer language: ○ English ● Telugu

🎃 You: what is the best way to cure kindred

💡 Bot: ఆక్టేవియా బట్లర్ యొక్క నవల "కింద్రెడ్" సందర్భంలో, కింద్రెడ్‌ను నయం చేయడానికి డానా ఫ్రాంక్లిన్ అనుభవించిన భావోద్వేగ, మానసిక మరియు తాత్కాలిక బాధలను పరిష్కరించాల్సి ఉంటుంది. శాస్త్రీయ కోణంలో కింద్రెడ్‌కు సాహిత్యపరమైన "నివారణ" లేనప్పటికీ, దాని ప్రభావాలను తగ్గించడానికి ఇక్కడ కొన్ని సాధ్యమైన మార్గాలు ఉన్నాయి: 1. **థెరపీ మరియు కౌన్సెలింగ్**: సమయ ప్రయాణం మరియు దాని సంక్లిష్టతలను అర్థం చేసుకునే చికిత్సకుడితో రెగ్యులర్ థెరపీ సెషన్‌లు మానసిక ఆరోగ్యంపై ప్రభావం డానా తన భావోద్వేగాలను ప్రాసెస్ చేయడంలో మరియు ఆమె అనుభవాల ద్వారా పని చేయడంలో సహాయపడుతుంది. 2. **సపోర్ట్ నెట్‌వర్క్**: స్నేహితులు, కుటుంబం మరియు తోటి సమయ ప్రయాణికుల (అటువంటి వ్యక్తులు ఉన్నట్లయితే) బలమైన మద్దతు నెట్‌వర్క్‌ను రూపొందించడం ద్వారా డానాకు భావోద్వేగ మద్దతు, ధ్రువీకరణ మరియు చెందిన భావాన్ని అందించవచ్చు. 3. **స్వీయ-సంరక్షణ**: ధ్యానం, యోగా లేదా సృజనాత్మక సాధనల వంటి స్వీయ-సంరక్షణ కార్యకలాపాలలో నిమగ్నమవ్వడం వలన డానా ఒత్తిడి, ఆందోళన మరియు కింద్రెడ్‌తో సంబంధం ఉన్న ఇతర లక్షణాలను నిర్వహించడంలో సహాయపడుతుంది. 4. **తాత్కాలిక రిజల్యూషన్**: ఆమె టైమ్ ట్రావెలింగ్ నుండి ఉత్పన్నమయ్యే తాత్కాలిక వైరుధ్యాలు మరియు అసమానతలను పరిష్కరించడానికి మార్గాలను కనుగొనడం డానాకు నియంత్రణ మరియు స్థిరత్వాన్ని తిరిగి పొందడంలో సహాయపడుతుంది. 5. **అవగాహన మరియు అంగీకారం**: ఆమె చూసిన చారిత్రక సంఘటనలు మరియు వాటిలో ఆమె పోషించే పాత్ర వైద్యం మరియు మూసివేత వైపు కీలకమైన అడుగు కావచ్చు. కింద్రెడ్ అనేది సంక్లిష్టమైన ఇతివృత్తాలు మరియు ఆలోచనలను అన్వేషించే సైన్స్ ఫిక్షన్ యొక్క పని కాబట్టి, ఈ సూచనలు ఊహాజనితమని గమనించడం చాలా అవసరం. ఈ నవల డానా అనుభవాలకు సూటిగా "నివారణ" అందించదు, కానీ కాలయాత్ర మరియు మానవ పరిస్థితి యొక్క పరిణామాలను ప్రతిబింబించేలా పాఠకులను ఆహ్వానిస్తుంది. మరింత సాహిత్యపరమైన "నివారణ" పరంగా, డానాకు ఇవి అవసరం కావచ్చు: 1. **తాత్కాలిక పునరావాసం**: ఆమె ప్రస్తుత టైమ్‌లైన్‌లో మళ్లీ కలిసిపోవడానికి మరియు ఆమె సమయ ప్రయాణ ప్రభావాలను అధిగమించడంలో సహాయపడే ఒక ఊహాత్మక ప్రక్రియ . 2. **జన్యు లేదా వైద్య చికిత్సను స్వీకరించండి**: కింద్రెడ్ అనేది జన్యు పరివర్తన లేదా ఇతర వైద్య పరిస్థితి వల్ల సంభవించినట్లయితే, నివారణను కనుగొనడంలో లక్ష్య చికిత్సలు లేదా చికిత్సలను అభివృద్ధి చేయడం వంటివి ఉండవచ్చు. అయితే, ఈ ఆలోచనలు పూర్తిగా ఊహాజనితమైనవి మరియు నవల యొక్క సరిహద్దుల వెలుపల ఉన్నాయి.

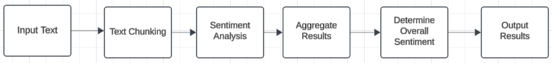
***Fig 8**.Text Translation in Multilingual*

The API dynamically translates the content into the specified target language (e.g., Telugu, as shown in the example). A translation cache is implemented to optimize performance and minimize repeated API calls. The system stores previously translated text for reuse, reducing latency and computational costs. The system includes comprehensive error handling for failed API requests, ensuring users receive fallback messages when translation issues occur. The error messages are user-friendly and indicate potential issues like API limitations or input constraints. The module supports a wide range of languages, enabling global accessibility in multilingual healthcare environments. This adaptability makes the system valuable for international collaboration and research dissemination.

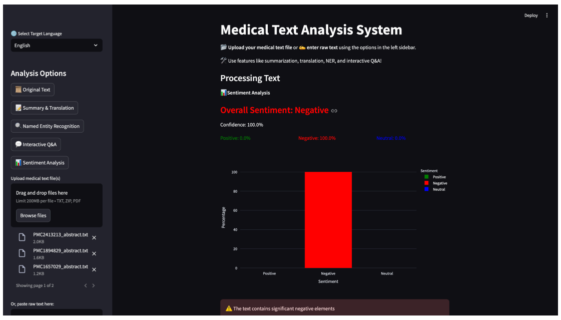
6. ADVANCED FEATURES:

Sentimental Analysis Visualization **:**

The system utilizes the Hugging Face Transformers pipeline for sentiment analysis by implementing a pre-trained model optimized for general sentiment classification. To handle variability in text input, it validates and excludes empty or null entries. For long texts, it segments them into smaller chunks of up to 1,000 characters to ensure compatibility with the model's input constraints. Each chunk is analyzed independently, and the results are categorized into positive, negative, and neutral sentiments, which are aggregated for a comprehensive evaluation.


Input Text → Text Chunking → Sentiment Analysis → Aggregate Results → Determine Overall Sentiment → Output Results

Sentiment percentages are calculated for each category relative to the total chunks, determining the dominant sentiment and its confidence score. The results are visually represented with intuitive labels and color codes. This approach efficiently analyzes lengthy and complex texts, such as medical abstracts, enabling the identification of sentiment distribution across sections. It is particularly suited for extracting insights from the medical data.



PMC2413213_abstract.txt 2.0KB
PMC1894829_abstract.txt 1.6KB
PMC1657029_abstract.txt 1.2KB

Medical Text Analysis System

📁 **Upload your medical text file** or ✍ **enter raw text** using the options in the left sidebar.

🛠 Use features like summarization, translation, NER, and interactive Q&A!

Processing Text

📊**Sentiment Analysis**

Overall Sentiment: Negative

Confidence: 100.0%

Positive: 0.0% Negative: 100.0% Neutral: 0.0%

⚠ The text contains significant negative elements

***Fig 9**. Sentimental Analysis of Text Data.*

7. CHALLENGES AND LIMITATIONS:

- Biomedical NER models, such as BioBERT, ClinicalBERT might generate a false positive by tagging common terms like "non" as entities due to overgeneralization during training or biases in the dataset.
- Initially, we have utilized the ChatGPT-3.5 model for the Q&A application. However, challenges raised due to credit limitations, which is a significant drawback for long-term sustainability. After exploring alternative solutions, the Olama system was selected as the final model. This decision was driven by its cost-free availability and user-friendly interface, making it a practical choice for implementing the Q&A session effectively.

8. FUTURE ADVANCEMENTS:

Text-to-Speech (TTS) Integration:

The system could integrate state-of-the-art text-to-speech models such as Google's WaveNet, Amazon Polly, or Meta's TTS models. These technologies can convert analyzed text, summaries, or sentiment outcomes into natural-sounding speech. TTS would enable accessibility for visually impaired users, providing a more inclusive system for medical, educational, or general use cases. Automatically generate spoken summaries of research articles or medical reports for researchers and clinicians on the go. Real-time conversion of sentiment analysis results into audio, enhancing user experience in interactive dashboards.

Avatar-Based Communication:

The inclusion of a virtual avatar (powered by animation frameworks like Unity or Unreal Engine) can make the system more interactive and engaging. The avatar could: "Speak" is the summarized content or sentiment analysis results using TTS, mimicking human facial expressions and gestures for enhanced comprehension. Be tailored for specific domains, such as a medical professional avatar explaining healthcare reports or a teacher avatar summarizing educational content. Extend TTS and avatar communication to support multilingual functionality, allowing users to select their preferred language for text summaries or sentiment feedback. Coupled with translation features, the system could cater to a global audience, improving accessibility in non-English-speaking regions. For user experience, these features can be integrated into IoT devices such as smart speakers (e.g., Amazon Echo, Google Nest) or wearable AR/VR devices.

9. CONCLUSION:

In conclusion, the developed framework leverages advanced Natural Language Processing (NLP) techniques to deliver robust functionalities for text summarization and sentiment analysis. The integration of PubMedBERT provides domain-specific embeddings for effective summarization, highlighting key sentences with high relevance, while the sentiment analysis pipeline accurately classifies text into positive, negative, or neutral sentiments with a breakdown of sentiment scores. By combining pre-trained models with efficient preprocessing and chunking strategies, the system ensures scalability for handling large or complex texts. Future potential for enhancements, such as real-time interactivity and accessibility features, further underscores its value for applications in research, healthcare, and education. This framework demonstrates the powerful synergy of pre-trained language models and innovative feature engineering to address real-world challenges in text analysis.

10. REFERENCES:

[1] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, "Clinical context–aware biomedical text summarization using deep neural network: model development and validation," J. Med. Internet Res., 2020.

[2] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," Nucleic Acids Res., 2004.

[3] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," Knowl. Inf. Syst., 2023.

[4] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Artif. Intell. Med., 2019.

[5] O. S. Chirila, C. B. Chirila, and L. Stoicu-Tivadar, "Named entity recognition and classification for medical prospectuses," Health Informatics Vision: From Data via Information to Knowledge, IOS Press, 2019.

[6] F. Liu, Z. Kang, and X. Han, "Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models," arXiv preprint, 2024.

[7] I. A. Pap and S. Oniga, "eHealth Assistant AI Chatbot Using a Large Language Model to Provide Personalized Answers through Secure Decentralized Communication," Sensors, 2024.

[8] P. Yellamma et al., "Automatic and Multilingual Speech Recognition and Translation by using Google Cloud API," Int. Conf. Mobile Computing and Sustainable Informatics, 2024.

[9] S. Pokhrel, S. Ganesan, T. Akther, and L. Karunarathne, "Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit," J. Inf. Technol. Digit. World, 2024.

[10] Google Cloud, "A Comprehensive Guide to Named Entity Recognition (NER)," Google Cloud Documentation, 2023.

[11] Y. Gu et al., "PubMedBERT: Domain-Specific Language Model Pre Training for Biomedical Natural Language Processing," ACM Trans. Compute. Healthcare, 2022.

[12] A. Maurya, "Ollama: A Deep Dive into Running Large Language Models Locally (PART-1)," Medium, 2023.

[13] JohnSnowLabs. (n.d.). spark-nlp-workshop/healthcare-nlp/data/diabetes_txt_files.zip at master · JohnSnowLabs/spark-nlp-workshop. GitHub.