Comparative analysis of diabetes diagnosis: WE-LSTM networks and WizardLM-powered DiabeTalk chatbot

Domenico Rossi
Department of Computer Science
University of Salerno
Fisciano, Italy
dorossi@unisa.it

Alessia Auriemma Citarella
Department of Computer Science
University of Salerno
Fisciano, Italy
aauriemmacitarella@unisa.it

Fabiola De Marco
Department of Computer Science
University of Salerno
Fisciano, Italy
fdemarco@unisa.it

Luigi Di Biasi
Department of Computer Science
University of Salerno
Fisciano, Italy
ldibiasi@unisa.it

Genoveffa Tortora
Department of Computer Science
University of Salerno
Fisciano, Italy
tortora@unisa.it

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10821742

***Abstract*—Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels due to insufficient insulin production or insulin resistance. It primarily manifests in two forms: Type 1 diabetes, an autoimmune condition typically diagnosed in younger individuals, and Type 2 diabetes, which is more prevalent and often linked to lifestyle factors such as obesity and inactivity. This study evaluates the performance of Long Short-Term Memory networks in diagnosing the two types of diabetes from Italian medical text across four progressively refined pre-processing scenarios. Each scenario incrementally builds on the previous one to enhance text cleaning and data preparation, allowing for a more refined and effective data processing pipeline. In parallel, this study introduces *DiabeTalk*, a chatbot developed on the WizardLM model, designed to provide specialized advice and support for diabetes diagnosis. While the WE-long short term memory models were fine-tuned with clinical data, DiabeTalk was tested without prior training on clinical diaries, allowing us to evaluate its performance in a real-world context. The results indicate that, despite the lack of domain-specific pre-training, DiabeTalk effectively employs natural language understanding and decision-making algorithms to predict diabetes type and respond to user inquiries. However, the testing revealed limitations in accuracy (*77.56% versus* 97.80%), with the chatbot achieving a lower performance than the WE-long short term memory model, which was applied to minimally pre-processed raw data. The findings underscore the importance of training large language models on relevant clinical datasets to enhance their response capabilities.**

***Index Terms*—Type 1 and Type 2 diabetes, Long Short-Term Memory, chatbot, Natural Language Processing, WizardLM**

This study was carried out within the Project PNC 0000001 D3 4 Health, - CUP B83C22006120001, The National Plan for Complementary Investments to the NRRP , Funded by the European Union - NextGenerationEU (Progetto PNC 0000001 D3 4 Health, - CUP B83C22006120001, Piano nazionale per gli investimenti complementari al PNRR, finanziato dall'Unione europea – NextGenerationEU).

I. Introduction

Artificial Intelligence (AI) has revolutionized various industries, providing innovative solutions to complex problems through data-driven insights, automation, and enhanced decision-making processes, especially in the healthcare field [1]. In recent years, one of the most transformative applications of AI has been in Natural Language Processing (NLP), a subset of AI, which plays a pivotal role in enabling machines to understand, interpret, and respond to human language. NLP techniques, such as chatbots, are widely used in AI-powered systems to process and analyze large volumes of unstructured text data [2]. A key focus in NLP is extracting data from unstructured clinical records, which supports decision-making and helps create structured representations of clinical notes. Through these capabilities, NLP not only enables chatbots to have fluent and context-aware conversations but also opens new possibilities for extracting valuable insights from diverse data sources [3].

For instance, in the healthcare sector, NLP algorithms can analyze electronic health records (EHRs) to identify patient symptoms, diagnoses, and treatment patterns. This information can then support clinical decision-making, detect potential health trends, or identify early signs of disease [4].

Large Language Models (LLMs) have emerged as one of the most impactful advancements in the field of NLP to process, generate, and understand human language. LLMs are deep learning models trained on vast amounts of text data, enabling them to capture the intricacies of language, context, and meaning [5]. These models have been at the forefront of conversational AI, supporting applications like chatbots, automated customer service, and medical diagnosis systems [4], [6]. In particular, chatbots use NLP algorithms, allowing

Authorized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions apply.
979-8-3503-8622-6/24/$31.00 ©2024 IEEE

them to understand user intent, respond accurately, and learn over time through interaction [7].

Building on this advancement, symbiotic AI—a collaborative approach where humans and AI systems work together in a mutually beneficial relationship—has further enhanced the capabilities of chatbots. Symbiotic AI combines intuition, creativity, and emotional intelligence with AI's data processing and computational speed, resulting in more adaptive and responsive systems [8]. This fusion allows AI chatbots to offer not only factual responses but also emotionally intelligent and contextually aware interactions [9].

This work presents DiabeTalk, a chatbot designed with WizardLM to assist in the diagnosis of DMT1 and DMT2, in comparison with WE-LSTM network. Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from insulin deficiency or insulin resistance. The condition is primarily classified into two major types: Type 1 diabetes (DMT1) [10], which is autoimmune and typically diagnosed in younger individuals, and Type 2 diabetes (DMT2) [11], a more prevalent form that occurs due to insulin resistance and is often linked to obesity and lifestyle factors. By integrating WizardLM for real-time, conversational diagnostics, DiabeTalk offers personalized insights based on symptoms, lifestyle factors, and medical history.

The primary objective of this study is to assess how effectively the knowledge base of WizardLM can respond to medical questions. To achieve this, we trained a WE-LSTM network using a private dataset of clinical data from diabetic patients and evaluated its performance. Simultaneously, we provided the same dataset to WizardLM, without any fine-tuning specific to the dataset, to analyze its performance and compare the outcomes with those of the WE-LSTM network. The results indicate that the WE-LSTM significantly outperforms the chatbot, although the chatbot still achieves acceptable results despite not being pre-trained on the clinical data used.

This study introduces **DiabeTalk**, a chatbot designed using **WizardLM** to assist in the diagnosis of Type 1 (DMT1) and Type 2 diabetes (DMT2), enabling real-time, conversational diagnostics.

- It leverages **Natural Language Processing (NLP)** techniques to support decision-making through the analysis of unstructured clinical records, facilitating the creation of structured representations of clinical notes.
- The study highlights the use of **Large Language Models (LLMs)** for understanding human language in medical contexts, supporting applications such as medical diagnosis systems.
- By integrating **symbiotic AI**, the study combines human intuition with AI's data processing capabilities, enhancing chatbot interactions with emotional intelligence and context-awareness.
- The effectiveness of **WizardLM's** responses to medical questions was assessed and compared against a WE-LSTM network trained on clinical data from diabetic patients.
- The study found that while **WE-LSTM** significantly outperformed the chatbot, WizardLM achieved acceptable results even without fine-tuning on the specific clinical dataset.

The article's structure is as follows: Section II offers a brief overview of NLP and chatbots in the context of diabetes. Section III introduces the methodology and the proposed chatbot DiabeTalk. Section IV details the experimental results, followed by the discussion and the related conclusions (Section V).

II. Related Works

A. NLP approach & chatbots in diabetes

NLP plays a crucial role in analyzing EHRs. By processing unstructured data like clinical notes, lab reports, and physician observations, NLP helps identify relevant information such as symptoms, treatments, and risk factors for disease management [12], [13], [14]. Not all diabetic patient datasets provide access to textual data for analysis. While many datasets focus on structured information like lab results, diagnoses, and medication lists, they may lack unstructured text such as physician notes, patient reports, or clinical observations [15]. This limitation restricts the ability to explore deeper insights that could be gained from analyzing narrative data, which can offer context, reveal subtle symptoms, or document lifestyle factors critical to diabetes management. Furthermore, since diabetes frequently occurs alongside other health issues such as cardiovascular disease, hypertension, or obesity, these datasets may encompass data related to multiple pathologies. In [16], the study explored using NLP to extract structured data from unstructured clinical notes in electronic healthcare records for diabetes patients. By developing deep learning models based on WE-LSTMs, the research aimed to automatically identify hospitalizations related to cardiovascular disease (CVD) from routine visit texts, considering four-time windows: infinite, 24 months, 12 months, and 6 months. Results showed that the NLP approach performed well for the infinite and 24-month windows, allowing for effective updates to medical records with minimal clinician input. In [17], the authors analyzed the discharge notes of over 40,000 ICU patients from Beth Israel Hospital over 11 years, focusing on type 2 diabetic patients. Using NLP, they extracted indicators of compliance with diet, exercise, and medication to detect ineffective self-management in diabetes. Then, they employed a three-step methodology: first, the authors identified diabetic patients using the keywords provided by the American Diabetes Association (ADA), while considering both those with current diagnoses and those with a history of diabetes. They pre-processed the data with tokenization, removing non-ASCII characters, eliminating stop words and punctuation, and lemmatizing the text. Second, they categorized patients based on their A1C levels, using normal A1C to indicate effective diabetes self-management (DSM). Those with abnormal A1C used BMI to assess diet, exercise, and medication compliance. Lastly, they conducted a bi-gram analysis to identify key phrases related to six compliance categories ((patient history, alcohol intake, medication

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

intake, diet, exercise, and general diabetes signs), which could inform patient feedback. Results from 471 patients showed patterns in BMI groups (normal, overweight, and obese) and compliance categories, suggesting that BMI trends can provide insights into DSM. In [18], the authors predicted the future risk of T2D by allowing neural models to analyze concepts extracted from the temporal sequence of patient visits. They used a normalized concept-based approach, extracting clinical concepts from the text—such as medications, diseases, symptoms, procedures, and anatomical sites—and mapping them to unique identifiers (CUIs) within the Unified Medical Language System (UMLS) meta thesaurus. This is achieved using the NLP tool cTAKES. The CNN-WE-LSTM model using cui2vec achieved the highest performance, with a macro average F1 score of 74.15%. Finally, they explore the driving factors behind model decision using explainable AI techniques and their analysis showed that highly predictive features are aligned with established medical knowledge, fostering trust in the models.

Chatbots have emerged as a valuable tool in disease prediction and management by providing real-time, personalized support to patients. These chatbots can use NLP and AI algorithms to engage users in interactive conversations, gathering essential information about their health behaviors, symptoms, and medical history. By analyzing this data, chatbots can identify risk factors for diabetes, offer personalized advice on lifestyle modifications, and remind patients to adhere to medication regimens. Chatbots in the literature for diabetes focus mainly on disease management, rather than classification. In [19], the authors presented an innovative question-answering and user-friendly system for chronic medical conditions based on large language models (LLMs). The system leverages a comprehensive medical knowledge corpus to intelligently create a chatbot to diagnose common chronic diseases. Empirical results from the CUQ test indicate that the system is effective and user-friendly. In [20], the study evaluates WizardLM's accuracy and clarity in providing nutritional management for T2DM and Metabolic Syndrome (MetS), based on the Academy of Nutrition and Dietetics guidelines. Using 63 prompts, two dietitians assessed WizardLM's responses. While the chatbot performed well, gaps were identified in weight loss recommendations, energy deficits, anthropometric assessments, and specific nutrients. Limitations were also noted in physical activity guidance and the nutrition care process, particularly in diagnostic statements and evaluation. Despite these gaps, WizardLM's output was rated as good or excellent for clarity.

III. Methods

In Fig. 1, we graphically depicted the entire workflow of the process. The goal of this experiment is to evaluate the performance of WE-LSTM networks in comparison with DiabeTalk, applied to the dataset in four specific scenarios, each built incrementally from the pre-processing of the previous scenarios. This progressive approach allows for a gradual improvement in data processing, refining the models as new layers of complexity and features are incorporated, enabling a more in-depth and detailed assessment of the capabilities of the WE-LSTM network in different contexts:

- *Scenario 1* with minimal pre-processing: only basic cleaning is applied to the text data, converting all text to lowercase and removing punctuation. This scenario evaluates WE-LSTM performance on raw, natural, medical Italian text;
- *Scenario 2* advanced cleaning: in addition to the basic steps, stop words are removed, and lemmatization is applied (reducing words to their root forms);
- *Scenario 3* additional word length filtering: this extends Scenario 2 by also removing very short and very long words. The aim is to further refine the text by filtering out irrelevant words, assessing whether this additional cleaning improves the model's performance;
- *Scenario 4*: refinement of the dataset with the sum of the previous scenarios.

Then, the pre-processed data for each scenario are passed into an WE-LSTM model, trained to predict the diagnosis of the two types of diabetes, DMT1 and DMT2. In parallel, the chatbot DiabeTalk acts as a conversational agent to interact with users and assist in diagnosing diabetes. By leveraging real-time user inputs such as symptoms, lifestyle factors, and medical history, the chatbot dynamically interprets the information through contextual natural language understanding (NLU) and applies decision-making algorithms to estimate whether the user will likely have DMT1 or DMT2. Following, we compared the accuracy of responses between the WE-LSTM network and the chatbot to assess the performance of both approaches.

A. Dataset & pre-processing

The used private dataset, named here as *DiabetesDB*, represents a retrospective observational Italian dataset that includes information collected from diabetic patients, capturing both their clinical conditions and disease progression over time. It is designed to support the development of predictive models that can anticipate disease outcomes, primarily focusing on DMT1 and DMT2. Some records or observations are incomplete, particularly with missing information in the "Clinical Diary" column. This missingness poses a challenge for predictive analysis since key details about the patient's condition or treatment over time are absent. A single observation (record) is associated with two distinct outcomes in certain cases. This occurs when the same diary entry (clinical record) is linked to multiple labels, leading to ambiguity.

The first pre-processing step removes all empty rows. In the second step, any row where the diabetes column contains values other than DMT1 or DMT2 is removed from the dataset. After cleaning the diabetes column, the labels are transformed into a categorical data type. Categorical variables are more efficient for storing and processing information, especially when the variable represents a fixed set of categories (like DMT1 and DMT2). This step is sometimes called "label lemmatization", ensuring that the labels are standardized and ready for analysis.

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

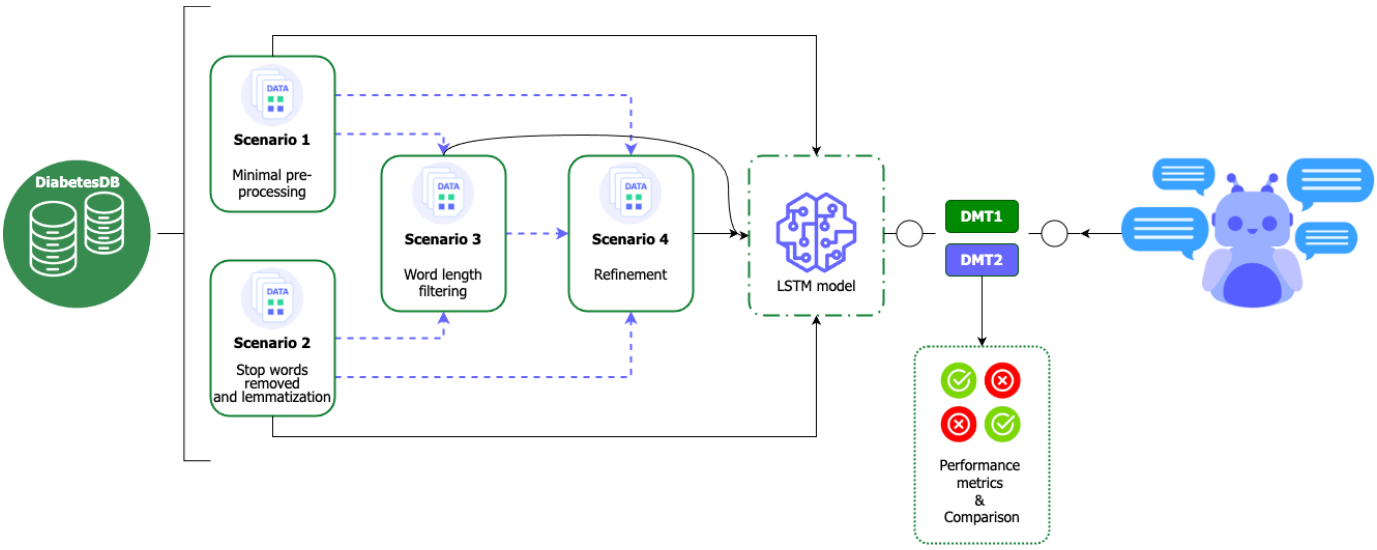

Fig. 1. Workflow of the process

B. WE-LSTM network

LSTM is a Recurrent Neural Network (RNN) type designed to capture long-term dependencies in sequence data, making them effective for tasks like NLP, time series forecasting, and sequence classification. They maintain an internal state that allows them to remember important information over long sequences, avoiding issues like vanishing gradients [21]. In this specific case, the WE-LSTM is used within a pipeline that takes sequences of word embeddings (likely text data) as input and predicts a class (likely a label for each sequence), using a set of hyperparameters tuned for experimentation.

The described WE-LSTM network architecture is used in the four different experimental scenarios that vary based on the level of text pre-processing. These scenarios are intended to evaluate how different text refinement levels impact the WE-LSTM's performance in classifying medical text in Italian.

The core architecture of the network remains constant across all scenarios:

- *Sequence Input Layer*: it receives input in the form of sequences of text data. The sequences are processed word by word;
- *Word Embedding Layer*: it maps words to dense vectors of size embeddingDimension, capturing semantic similarities between words;
- *WE-LSTM Layer*: it processes the sequence using an WE-LSTM with numHiddenUnits. The output is the final hidden state after processing the entire sequence, representing a summary of the input;
- *Fully Connected Layer*:it transforms the WE-LSTM output into predictions across numClasses, which represent the classification categories;
- *Softmax Layer:*: it converts raw predictions into probabilities for each class.

The training process uses the Adam optimizer, with specific hyperparameters like mini-batch size, gradient threshold, and validation frequency set dynamically based on the experiment. The GradientThreshold is set to 2 to avoid exploding gradients, which can destabilize training. The network is trained on a CPU environment with data shuffling at each epoch. The network performance is stored with a function that monitors how well the model generalizes across training, validation, and test sets with a given data split (70% training, 20% validation, 10% testing).

C. DiabeTalk design

DiabeTalk is a conversational AI chatbot designed to diagnose diabetes, types DMT1 and DMT2, with user-friendly web technologies.

To enhance DiabeTalk's capabilities, we integrate WizardLM, a cutting-edge LLM designed for natural language understanding and conversational tasks [22]. WizardLM creates a hybrid system that balances general-purpose understanding with specialized domain expertise. The integration ensures that WizardLM is fine-tuned to specialize in conversational flow management, maintaining consistency in long, context-aware conversations related to diabetes. WizarlLM focuses on processing medical content, leveraging its transformer architecture to interpret intricate diabetes-related information and produce accurate classifications. This approach allows DiabeTalk to respond effectively to both common and complex diabetes-related inquiries, ensuring high-quality, medically accurate interactions.

The front-end interface of DiabeTalk is designed to be user-friendly and intuitive, enabling seamless interaction with the chatbot (see Fig. 2). HTML (HyperText Markup Language) provides the structural foundation of the chatbot interface, organizing the content and elements like text boxes, buttons, and chat windows. CSS (Cascading Style Sheets) styles these elements, ensuring the interface is visually appealing, professional, and consistent. JavaScript is used to add interactive functionality to the chatbot. AJAX (Asynchronous JavaScript

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

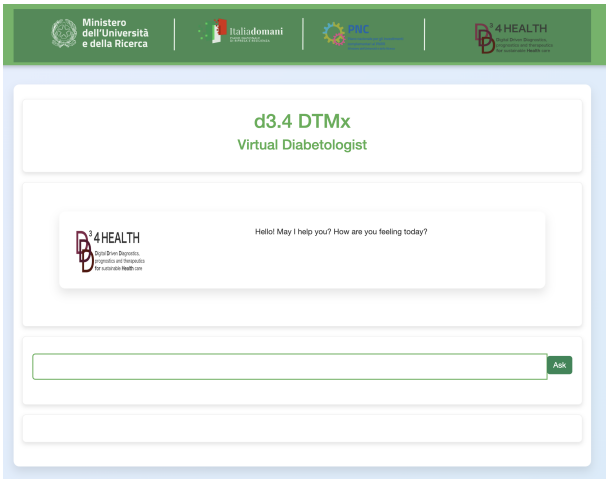

Fig. 2. Interface of the DiabeTalk

and XML) enables the interface to communicate with the back-end server in real-time without reloading the page. This results in a fluid, uninterrupted user experience where conversations happen instantaneously, as data is exchanged in the background. jQuery, a JavaScript library, is implemented to simplify the development process, particularly for managing dynamic content updates and handling user inputs efficiently.

D. Evaluation metrics

In evaluating the performance of classification models, especially in medical diagnostics, it is crucial to use a set of well-defined metrics that provide insights into the accuracy, sensitivity, specificity, and false negative rate of the models. These metrics (see Equations 1-4) can be calculated using the fundamental classification outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$ACC = \frac{TP + TN}{TN + FP + FN + TP} \tag{1}$$

$$SN = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$FNR = \frac{FN}{FN + TP} \tag{4}$$

Accuracy (ACC) is a measure of how often the classifier correctly identifies both positive and negative instances. It represents the proportion of correctly classified samples (both true positives and true negatives) out of the total samples. Sensitivity (SN), also known as Recall, measures the proportion of actual positive cases correctly identified by the model. Specificity (SP) measures the proportion of actual negative cases correctly identified by the model. It evaluates the model's ability to avoid false positives and is particularly relevant when FP needs to be minimized. The False Negative Rate (FNR) is the complement of sensitivity. It measures the proportion of positive instances that are incorrectly classified as negative. A high FNR can be critical in medical diagnostics, as it indicates the model is missing cases that should be diagnosed as positive.

IV. Results

A. Data analysis

In the data analysis phase, we employed two visual representation techniques to enhance our understanding of the dataset. This step is crucial for uncovering patterns, trends, and insights that can inform subsequent modeling and decision-making processes. Histograms represent the distribution of numerical data, particularly in our analysis of document characteristics. In this case, we focus on two key variables: the x-axis, which represents the length of documents, and the y-axis, which indicates the number of documents within each length range. Word clouds are visual representations of text data where the size of each word reflects its frequency or importance in the dataset [23]. Larger words appear more often, while smaller ones are less common. The words are usually arranged randomly, creating a cloud-like effect. They offer a quick, visual summary of the key terms in a text, making it easy to identify them.

In Fig.3, we depicted the histograms for the four scenarios.

In particular, Scenarios 1 and 2 show a distribution where most chunks fall under the 200-length threshold, suggesting that the content being processed in these experiments contains smaller units. This could imply a higher frequency of shorter segments, likely influenced by the nature of the dataset or the pre-processing steps applied. On the other hand, Scenario 4 demonstrates a significant reduction in chunk size following the exclusion of shortWords. Removing these short words appears to streamline the chunking process, leading to a more compact and consistent chunk size distribution, with no chunk exceeding the 100-length mark. This suggests that shortWords play a substantial role in inflating chunk sizes in the previous experiments, and their removal brings about a more controlled data segmentation.

In Fig. 4, word clouds for all the scenarios are reported. In contrast to a raw dataset (see Scenario 1, Fig.4(a)), which might display high-frequency but less informative words, the word cloud of the refined dataset reports frequent terms with a higher semantic content at the medical level. This refinement likely stems from the pre-processing steps, such as removing common stop words, irrelevant terms, or noise, which allows medically significant terminology to emerge more clearly. As a result, the word cloud is no longer cluttered with generic or unrelated terms but instead highlights domain-specific language crucial for medical analysis.

B. Classification results

In this section, we present the performance outcomes of the proposed model based on key classification metrics. The results are derived from testing the four proposed scenarios.

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

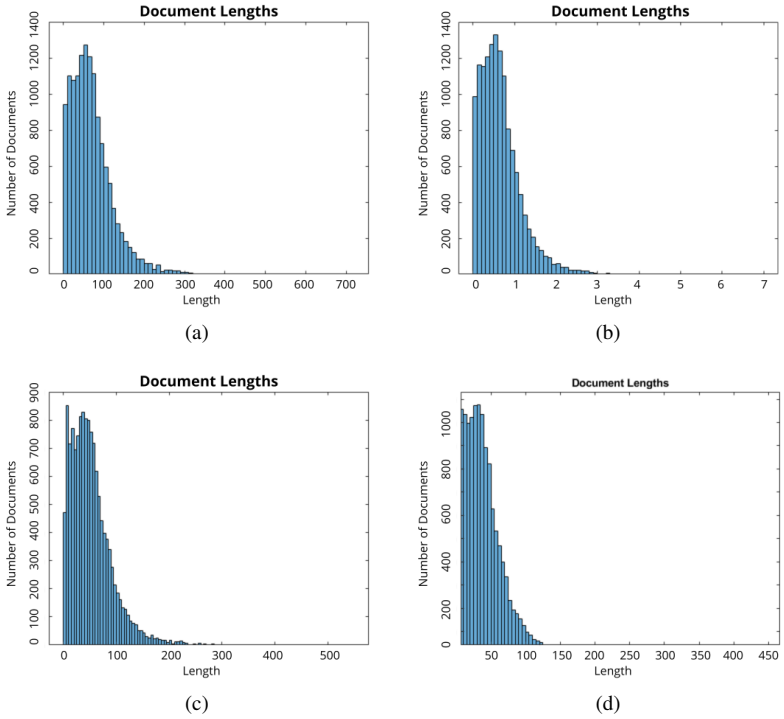

Fig. 3. Histogram for Scenario 1 (a), Scenario 2 (b), Scenario 3 (c) and Scenario 4 (d)

Fig. 4. Word clouds for Scenario 1 (a), Scenario 2 (b), Scenario 3 (c) and Scenario 4 (d)

Scenario	ACC (%)	SN (%)	SP (%)	FNR
Scenario 1	97.80	98.60	95.90	0.01
Scenario 2	96.52	97.40	93.70	0.03
Scenario 3	97.26	98.5	93.5	0.02
Scenario 4	97.08	97.3	96.5	0.03

TABLE I
PERFORMANCE METRICS FOR ALL THE SCENARIOS

Table I presents the performance metrics of the WE-LSTM model across four different text preprocessing scenarios. Each scenario applies varying levels of text cleaning, from minimal pre-processing to advanced cleaning and filtering. In *Scenario 1*, where only basic text cleaning (lowercasing and punctuation removal) is applied, the model achieves the highest ACC (97.80%) and SN (98.60%). The SP (95.90%) is also strong, but slightly lower compared to SN, indicating that while the model is very good at identifying true positives (SN), it occasionally misclassifies negative instances (FP). The FNR is extremely low (0.01), suggesting the model almost never misses a true positive. However, this performance might be influenced by overfitting to irrelevant patterns due to minimal cleaning. With *Scenario 2*, the application of stop word removal and lemmatization results in a slight decrease in ACC (96.52%) and SN (97.40%) compared to Scenario 1. However, SP (93.70%) also drops, indicating that the ability of the model to correctly identify negative cases is worse. The FNR (0.03) has increased slightly, implying that the model is now more likely to miss positive cases after this additional cleaning. *Scenario 3* introduces further refinement by filtering words based on length. This results in an improvement in SN (98.50%), close to the level observed in Scenario 1, and a moderate ACC (97.26%), slightly better than Scenario 2. However, SP (93.50%) remains lower, similar to Scenario 2, indicating the model still struggles with false positives. The FNR (0.02) improves compared to Scenario 2, suggesting the additional filtering has helped the model better identify true positives

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

without losing too much valuable information. In *Scenario 4*, which combines all pre-processing steps from the previous scenarios, ACC (97.08%) and SN (97.30%) are slightly lower than in Scenarios 1 and 3, but SP (96.50%)** see a significant increase compared to the previous scenarios. This indicates a more balanced performance between identifying true positives and true negatives. The FNR (0.03) is comparable to Scenario 2, meaning the comprehensive cleaning approach still misses many true positives. The results suggest that the optimal pre-processing approach depends on the specific emphasis of the task— whether minimizing FP (as seen in Scenario 4) or maximizing TP (as in Scenario 1).

C. DiabeTalk results and comparison

In Table II, we reported the results and comparison between the classification performed by the best WE-LSTM model and, on the other hand, by the approach with DiabeTalk. Fig. 5 shows an example of input data and answers received from the chatbot.

Methods	ACC (%)	SN (%)	SP (%)	FNR
WE-LSTM	**97.80**	**98.60**	**95.90**	**0.01**
DiabeTalk	77.56	87.2	57.80	0.13

TABLE II
PERFORMANCE METRICS FOR ALL THE SCENARIOS

In Scenario 1, the WE-LSTM model was applied to raw, minimally pre-processed data. It achieved the highest performance, with an ACC of 97.80%, SN of 98.60%, and SP of 95.90%. The FNR was lower (0.01). These results indicate that the model performed exceptionally well in identifying both positive and negative cases with minimal pre-processing, highlighting its robustness in handling raw medical text data without extensive cleaning or transformation. DiabeTalk, launched without pre-training on clinical diaries, was evaluated to assess its performance without any fine-tuning on domain-specific data. The results show that DiabeTalk achieved a significantly lower ACC of 77.56%, SN of 87.2%, and SP of 57.80%, with a higher FNR of 0.13. These outcomes reflect the limitations of using a chatbot without pre-training on specific clinical data. While DiabeTalk demonstrated a reasonable ability to identify TP cases (as seen from its sensitivity), its specificity was much lower, indicating a higher number of false positives. This suggests that without domain-specific fine-tuning, the chatbot struggles with distinguishing between positive and negative cases accurately, potentially leading to a higher rate of incorrect classifications in real-world medical scenarios. Consequently, Scenario 1, based on WE-LSTM model, results in the best-case scenario in terms of accuracy, sensitivity, and overall balance in classification.

V. CONCLUSION

This study compared WE-LSTM networks with the DiabeTalk chatbot powered by WizardLM in diagnosing diabetes mellitus types 1 and 2. The evaluation included four scenarios that employed refined text pre-processing techniques to assess their influence on model performance. The results showed that WE-LSTM networks, particularly in Scenario 1, which utilized raw data, achieved the Future research should prioritize several key areas to enhance the performance and applicability of the DiabeTalk chatbot and similar AI-driven diagnostic tools. First, the integration of LLMs that have shown promise in understanding complex medical language could significantly improve the diagnostic accuracy of the chatbot. By fine-tuning these models on comprehensive clinical datasets, DiabeTalk can better capture the medical terminology and patient presentations, allowing us to obtain more reliable and contextually appropriate responses. Moreover, expanding the chatbot's knowledge base to include a wider range of chronic diseases beyond diabetes would enhance its utility as a multifaceted diagnostic tool. Another vital area for future exploration is the enhancement of model interpretability by applying explainable AI techniques. Developing transparent models to elucidate the reasoning behind their diagnostic suggestions is very important in the biomedical field. Lastly, it would be beneficial to explore the integration of DiabeTalk with existing EHR systems, enabling the chatbot to provide more personalized and informed diagnostic insights based on the patient's medical history.

REFERENCES

[1] S. Aminizadeh, A. Heidari, M. Dehghan, S. Toumaj, M. Rezaei, N. J. Navimipour, F. Stroppa, and M. Unal, "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," *Artificial Intelligence in Medicine*, vol. 149, p. 102779, 2024.

[2] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[3] R. Srinivasan, M. Kavitha, and S. Uma, "Natural language processing: Concepts and applications using chatbot," in *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2023, pp. 331–337.

[4] G. K. Thakur, A. Thakur, N. Khan, and H. Anush, "The role of natural language processing in medical data analysis and healthcare automation," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, vol. 1. IEEE, 2024, pp. 1–5.

[5] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.

[6] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature medicine*, vol. 30, no. 9, pp. 2613–2622, 2024.

[7] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *IFIP international conference on artificial intelligence applications and innovations*. Springer, 2020, pp. 373–383.

[8] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making," *Business horizons*, vol. 61, no. 4, pp. 577–586, 2018.

[9] Y. Yu, R. Guan, J. Ma, Z. Jiang, and J. Huang, "When and who? conversation transition based on bot-agent symbiosis learning network," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4056–4066.

[10] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L. M. Jacobsen, D. A. Schatz, and Å. Lernmark, "Type 1 diabetes mellitus," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–17, 2017.

[11] R. A. DeFronzo, E. Ferrannini, L. Groop, R. R. Henry, W. H. Herman, J. J. Holst, F. B. Hu, C. R. Kahn, I. Raz, G. I. Shulman *et al.*, "Type 2 diabetes mellitus," *Nature reviews Disease primers*, vol. 1, no. 1, pp. 1–22, 2015.

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app

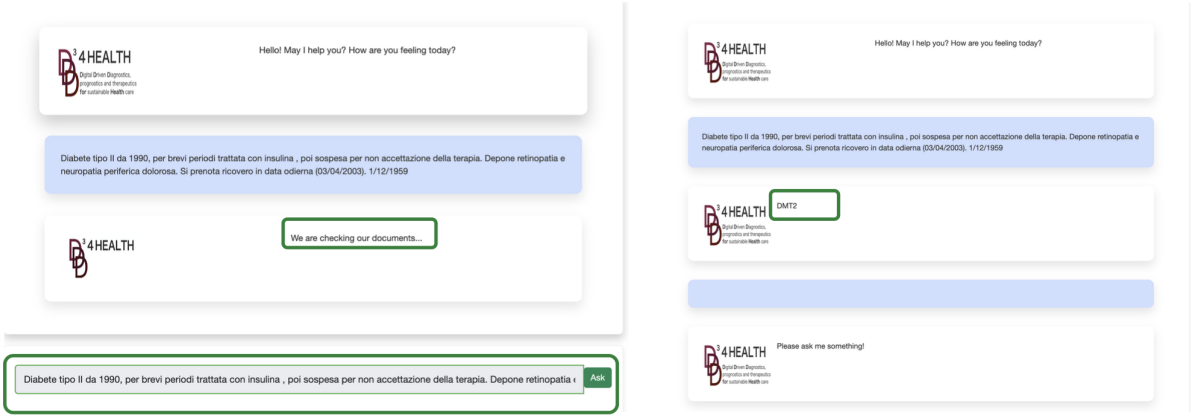

Fig. 5. An example response from DiabeTalk. in the image on the left, you can see the interface. Initially there is the initial message from the bot and the text input box (see second green box below). After the input has been entered and the 'Ask' button has been pressed, the input is checked (see first green box). In the image on the right you can see the answer given in the output (see green box on the right image).

[12] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani *et al.*, "Natural language processing of clinical notes on chronic diseases: systematic review," *JMIR medical informatics*, vol. 7, no. 2, p. e12239, 2019.

[13] R. G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, A. Roberts, R. J. Dobson, and R. Stewart, "Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project," *BMJ open*, vol. 7, no. 1, p. e012012, 2017.

[14] G. K. Savova, E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris, H. Hochheiser, C. Lin, G. Chavan *et al.*, "Deepphe: a natural language processing system for extracting cancer phenotypes from clinical records," *Cancer research*, vol. 77, no. 21, pp. e115–e118, 2017.

[15] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," in *Recent Developments in Machine Learning and Data Analytics: IC3 2018*. Springer, 2019, pp. 67–78.

[16] A. Guazzo, E. Longato, G. P. Fadini, M. L. Morieri, G. Sparacino, and B. Di Camillo, "Deep-learning-based natural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visits' text," *Scientific Reports*, vol. 13, no. 1, p. 19132, 2023.

[17] M. Pourbehzadi, G. Javidi, K. Johnson, and O. Roberts, "Detecting ineffective self-management in diabetic patients: A data mining perspective," in *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2023, pp. 1–5.

[18] R. Chaturvedi, M. Rashid, B. T. Layden, A. Boyd, A. Cinar, and B. Di Eugenio, "Sequential representation of sparse heterogeneous data for diabetes risk prediction," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 831–834.

[19] S. Zhang and J. Song, "A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model," *Scientific Reports*, vol. 14, no. 1, p. 17118, 2024.

[20] F. Naja, M. Taktouk, D. Matbouli, S. Khaleel, A. Maher, B. Uzun, M. Alameddine, and L. Nasreddine, "Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome," *European Journal of Clinical Nutrition*, pp. 1–10, 2024.

[21] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[22] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang, "Wizardlm: Empowering large language models to follow complex instructions," *arXiv preprint arXiv:2304.12244*, 2023.

[23] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf, "Prefix tag clouds," in *2013 17th international conference on information visualisation*. IEEE, 2013, pp. 45–50.

horized licensed use limited to: Ballari Institute of Technology & Management (formerly Bellary Eng College). Downloaded on December 19,2025 at 04:11:38 UTC from IEEE Xplore. Restrictions app