NeuroNest: Deep Learning Chatbot for Applied Data Science.

Aboli Wankhade, Deekshita Prakash Savanur, Dharmateja Kolluri, Gouri Benni, Rama Krishna Poluru, Uzair Riyaz Pachhapure

Department of Applied Data Science, San Jose State University

DATA 255 - Sec 11: Deep Learning Technologies

Prof. Mohammad Masum

September 26, 2023

Dataset Description:

This project involves acquiring a dataset by utilizing the web scraping capabilities of the Beautiful Soup tool on the San Jose State University's Applied Data Science department website. The dataset obtained from this process will be initially unstructured or semi-structured, mainly in the form of HTML content. To make this data suitable for training Language Models (LLMs), a preprocessing step will be performed. In essence, the project involves the transformation of raw web data into an organized dataset, which will serve as the foundation for training advanced language models.

Problem Definition:

Navigating the intricate landscape of the Applied Data Science department often involves addressing a multitude of distinct inquiries, from curriculum nuances, innovative research methodologies, faculty areas of expertise, to procedural intricacies. While the tried-and-true methods of direct email interactions or scheduled faculty discussions offer some solutions, they occasionally result in varied wait periods and differing interpretations. This scenario, reminiscent of specialized domains across the board, amplifies the necessity for a bespoke, department-focused communication mechanism. Traditional touchpoints, be it scheduled academic consultations or departmental correspondences, are often hampered by time-bound constraints, potential information disparities, and the daunting task of efficiently managing a vast query volume. In contrast, generic chatbot platforms, anchored in preset decision trees, might not capture the unique subtleties of academic discourse. Harnessing the prowess of deep learning, our vision is to sculpt a chatbot fine-tuned to the Applied Data Science department's needs. This digital assistant aspires to equip stakeholders with swift, uniform, and context-aware insights, while also being malleable enough to adapt to the department's fluid academic terrain.

Approach:

In this project, the objective is to develop an interactive chatbot for San Jose State University's Applied Data Science department. The approach involves leveraging Deep Learning techniques, namely LLM models, Transfer learning, and Reinforcement learning. The selection of these methods is based on their ability to effectively tackle the complications of interactive chatbot.

The project’s methodology starts with data collection that involves applying various web scraping techniques to obtain the data from the Applied Statistics department website. This step sets the groundwork for development of chatbot, ensuring that the collected data aligns with the requirements of the project.

After the data has been collected, the complex LLM models such as s Llama 2, GPT2, and Distilbert will be used to train the data. The selection of the most appropriate model will be based on performance metrics and computational efficiency, ensuring to make the most efficient use of available resources. After the models undergo training, the next step involves fine-tuning them. This phase relies on transfer learning, a technique that allows for the adaptation of these models to the precise demands of the department. This customization ensures that the chatbot becomes skilled at addressing queries related to the domain and understanding terminology effectively.

To enhance the chatbot's accuracy, an advanced recognition system is being introduced. This system, driven by similarity measures and Natural Language Processing (NLP), is crucial for classifying user queries. It enables the chatbot to categorize user questions into predefined intent with various department-related topics. These topics range from course information and faculty profiles to research areas and administrative queries. The objective is to ensure that the chatbot delivers accurate and contextually appropriate answers for a wide range of questions.

Data preprocessing in this project mainly involves cleaning and formatting web-scraped data. This includes parsing HTML to extract text, removing HTML tags and special characters, and tokenizing the text. Additional steps may involve removing stopwords, entity recognition, and lemmatization or stemming. Data validation checks for completeness and correctness. Preprocessing ensures that the data is ready for training a chatbot or language model, improving its ability to provide accurate responses.

Performance Evaluation:

When developing any chatbot it is important to evaluate its performance metrics effectively to get desired outcomes. It helps to measure the performance of the chatbot. For the chatbot there are several performance metrics that can be implemented. F1 score, BLEU score and Semantic Answer Similarity (SAS) will be used for performance evaluation.

F1 Score: It is commonly used evaluation metrics for chatbot responses. The F1 score will be calculated by comparing chatbot responses to the correct answers. It is very helpful when there is a discrepancy between answer kinds.

BLEU Score: The Bilingual Evaluation Understudy score evaluates the value of text produced by a chatbot by comparing them to human generated answers. It provides scores between 0 and 1, where higher value indicates better similarity.

Semantic Answer Similarity (SAS) Metric: It is a measure that is used in natural language processing and chatbot tasks. It compares and assesses the similarity between the generated answer of the chatbot or the natural language processing models and the expected answer. The SAS metric considers various metrics such as Semantic Equivalence, Synonymy, Paraphrasing and Semantic Distance.

Related Research Review:

The project by Arkhangelski et al. (2022) develops Hebron, a web-based chatbot that assists students seeking information about product availability in real-time, in order to solve the demand for an online inventory system at Covenant University Shopping Mall (CUSM). It includes designing a website, developing a chatbot using deep Natural Language Processing (NLP), creating a large item database, conducting extensive testing, including gathering user feedback through surveys to guarantee functionality, usability, and interface satisfaction. This survey-based evaluation yielded favorable results, with Hebron's conversational skills and practicality. The project by Hefny et al. (2021) features "Jooka," a bilingual university admission chatbot created to reduce high school students' stress during the application process and enhance user experience with an emphasis on user-specific demands, including bilingualism support, and human-like behavior cues. The findings of the evaluation showed that users were very satisfied, with mean SUS (System Usability Scale) and CUQ scores exceeding the required levels with mean SUS score was 88.5 and Mean CUQ score was 87.3, both over the 71.1 acceptance cutoff.

In this research paper Nguyen et al. (2021) talks about how chatbots that use artificial intelligence (AI) are in high demand. They have built an AI chat bot which helps students to get daily updates on curriculum, tuition fees and many more. They have built this chat bot by using Deep Learning models which were later integrated on the RASA framework. They manage to handle up to 50 varieties of student questions with an accuracy of 97.1% on test data. In this research paper Khin and Soe (2020) have built a chatbot for university (FAQ)that answers some basic questions about the university. The chatbot which they have built is in burmese language to help people get answers in their local language. They have used AIML (Artificial Intelligence Markup Language) for specifying chatbot content. Wang (2021) in their research investigates the application of generative models, specifically OpenAI's GPT-2, in developing therapeutic chatbots for caregivers. Given the pressing need for mental health resources globally, the study examines chatbots as a potential remedy. The paper contrasts the outcomes of the GPT-2 model in its original form and after being fine-tuned using therapy session transcripts between caregivers of dementia patients and their therapists. Findings indicated that while the fine-tuned model had a higher rate of non-English word outputs (40.6%) compared to the original model (5.8%), its response length was more aligned with human therapists. Windiatmoko et al. (2021) in their paper presents a novel chatbot system designed to provide assistance for various university-related queries. The primary focus of the paper is on evaluating the chatbot's performance in generating accurate responses to user inquiries. To achieve this, the paper employs deep learning models for tasks such as intent classification and entity recognition, implemented using Rasa NLU, along with dialogue policy management using Rasa Core. The results obtained from these models are notably strong, with precision, recall, and F1-scores approaching ideal values in many instances, underscoring the chatbot's capacity to deliver precise responses.

Adoma et al. (2020) explored the effectiveness of BERT, RoBERTa, DistilBERT, and XLNet models for text-based emotion recognition, highlighting the advantages of these models in addressing language understanding challenges.With the growth of social media users, recognizing emotions from text has become important, it evaluates the performance of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from text. RoBERTa outperforms the other models in recognizing emotions, achieving the highest accuracy. While DistilBERT is computationally efficient, it lags in accuracy. RoBERTa is recommended for emotion recognition in text, especially when optimization is a priority, supported by precision, recall, and F1-score results. In this research, Pramanik and Maliha (2022) developed two distinct supervised natural language models to enhance customer experience analysis through product reviews. Based on deep neural LSTM, the first model focused on word-level sentiment analysis using a dataset of 50,000 Amazon reviews. The second model leveraged the advanced DistilBERT, fine-tuning its multi-layered encoders for faster and more robust sentence-level sentiment analysis. Key contributions include a comparative evaluation favoring the Transformer-based model's superior performance and the introduction of a lightweight sentiment analyzer, DistilBERT. The research addresses this challenge by implementing two models. The first utilizes LSTM to predict sentiment by analyzing previous words in a sentence. In contrast, the second, based on DistilBERT, harnesses transformer-based bidirectional training and attention mechanisms for enhanced sentiment analysis. These models offer effective solutions for comprehensively mining opinions and aligning products with customer satisfaction goals. DistilBERT outperformed the LSTM model with and without Word2Vec word embeddings. Lyko et al. (2020) discuss the significance of a new ABR algorithm called "Llama" designed for low-latency live streaming scenarios. Llama uses two throughput measurements over different timescales to make quality decisions swiftly while maintaining stable video quality. It outperforms other ABR algorithms regarding P.1203 Mean Opinion Score (MOS) and reduces rebuffering by up to 33% with DASH and 68% with CMAF in low-latency situations. However, it may need adjustments for less latency-critical environments. Sultan and Abdullah (2022) discussed the use of Beautiful Soup, a web scraping tool, for data extraction from Google Scholar. It addresses the challenges of dealing with vast, unstructured internet data and highlights Beautiful Soup's role in efficiently retrieving and storing information from a researcher's Google Scholar page. The process involves specific steps like web data retrieval, element-based data selection, and result printing or CSV file storage. Python is employed for implementation due to its rich library resources.

Work Division:

All team members contributed to the documentation of the project. The duties of all team members are divided as follows.

Responsibilities	**Team Members**	**Start Date**	**End Date**
Data Collection: ● Identify and source relevant university webpages. ● Perform web-scraping to collect data.	Aboli and Deekshita	September 13th, 2023	September 16th, 2023
Pre-processing: ● Ensure data is readable. ● Standardize data if required. ● Split the data into training, validation, and test sets.	Gouri and Rama Krishna	September 22nd, 2023	September 30, 2023
Data Transformation: ● Convert data into a format suitable for modeling (e.g., one-hot encoding for categorical variables). ● Feature engineering to create new variables if necessary. ● Reduce dimensionality if required using techniques like PCA.	Dharmateja and Uzair	October 2nd, 2023	October 12th, 2023
Modeling using Deep Learning Techniques: ● Choose appropriate deep learning techniques based on the problem statement. ● Train initial models using the training dataset. ● Validate models using the validation set.	Aboli and Dharmateja	October 14th, 2023	October 30th, 2023
Model Evaluation and Fine-tuning: ● Evaluate model performance using appropriate metrics. ● Retrain models with optimized parameters.	Deekshita and Ramakrishna	November 1st, 2023	November 13th, 2023
Build ChatBot Application and Deployment : ● Build a ChatBot Application that is user-friendly. ● Deploy Application.	Uzair and Gouri	November 14th, 2023	November 30th, 2023

Bibliography

Khin, N. N., & Soe, K. M. (2020b). University Chatbot using Artificial Intelligence Markup Language. *University Chatbot Using Artificial Intelligence Markup Language*. https://doi.org/10.1109/icca49400.2020.9022814

Nguyen, T., Le, A. D., Hoang, H., & Nguyen, T. T. (2021b). NEU-chatbot: Chatbot for admission of National Economics University. *Computers & Education: Artificial Intelligence*, *2*, 100036. https://doi.org/10.1016/j.caeai.2021.100036

Oguntosin, V., & Olomo, A. (2021). Development of an E-Commerce chatbot for a university shopping mall. *Applied Computational Intelligence and Soft Computing*, *2021*, 1–14. https://doi.org/10.1155/2021/6630326

Hefny, W. E., Mansy, Y., Abdallah, M., & Abdennadher, S. (2021). Jooka: A bilingual chatbot for university admission. In *Advances in intelligent systems and computing* (pp. 671–681). https://doi.org/10.1007/978-3-030-72660-7_64

Adoma, A. F., Nunoo-Mensah, H., & Chen, W. (2020). Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. *Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition*. https://doi.org/10.1109/iccwamtip51612.2020.9317379

Pramanik, V., & Maliha, M. (2022). Analyzing Sentiment Towards a Product using DistilBERT and LSTM. *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. https://doi.org/10.1109/icccis56430.2022.10037634

T. Lyko, M. Broadbent, N. Race, M. Nilsson, P. Farrow and S. Appleby, "Llama - Low Latency Adaptive Media Algorithm," 2020 IEEE International Symposium on Multimedia (ISM), Naples, Italy, 2020, pp. 113-121, doi: 10.1109/ISM.2020.00027.

Wang, L. (2021, July 28). *An evaluation of Generative Pre-Training Model-based Therapy chatbot for caregivers*. arXiv.org. https://arxiv.org/abs/2107.13115

Windiatmoko, Y., Rahmadi, R., & Hidayatullah, A. F. (2021). Developing Facebook Chatbot based on deep learning using RASA framework for university enquiries. *IOP Conference Series*, *1077*(1), 012060. https://doi.org/10.1088/1757-899x/1077/1/012060

Sultan, N. A., & Abdullah, D. B. (2022). Scraping Google Scholar Data Using Cloud Computing Techniques. Scraping Google Scholar Data Using Cloud Computing Techniques. https://doi.org/10.1109/iccitm56309.2022.10032044