CHATBOT: A Comparative Analysis

Ankita Manjarekar
Department of Data Science
University of New Haven
amanj4@unh.newhaven.edu

Chetana Nannapaneni
Department of Data Science
University of New Haven
cnann1@unh.newhaven.edu

Abstract

This project endeavors to construct an AI-driven chatbot capable of conducting a comparative analysis of Natural Language Processing (NLP) models across diverse datasets, focusing primarily on Mental Health and Customer Support datasets. Leveraging three pre-trained models—GPT-2, BERT (Bidirectional Encoder Representations from Transformers), and RoBERTa (a refined variant of BERT)—we fine-tune them on the datasets to assess their performance. Through rigorous experimentation and evaluation, we measure the accuracy of these models to discern their efficacy in handling specific datasets. We compare their performance to determine which model is better suited for a particular dataset, which should help with the selection of NLP models for various applications. This project not only contributes to advancing the field of NLP but also facilitates informed decision-making in deploying AI-driven solutions tailored to specific domains.

1 Introduction

In recent years, the proliferation of Natural Language Processing (NLP) technologies has revolutionized how we interact with machines, enabling advanced capabilities such as language understanding, generation, and sentiment analysis. These advancements have found wide-ranging applications across various domains, including healthcare, customer service, and education. One of the key challenges in harnessing the full potential of NLP lies in selecting the most suitable model for a given task or dataset.

This project addresses this challenge by developing an AI-driven chatbot tailored for comparative analysis of NLP models on two distinct datasets: Mental Health and Customer Support. The choice of these datasets reflects their significance in real-world applications and underscores the diverse contexts in which NLP technologies are deployed.

The primary objective of this study is to evaluate and compare the performance of three prominent pre-trained NLP models: GPT-2 (Generative Pre-trained Transformer 2), BERT (Bidirectional Encoder Representations from Transformers), and RoBERTa (a refined variant of BERT). These models, known for their effectiveness in various NLP tasks, are fine-tuned on both Mental Health and Customer Support datasets to adapt them to the specific characteristics of each domain.

By conducting a rigorous comparative analysis, we aim to determine which model exhibits superior performance in terms of accuracy, robustness, and efficiency for tasks within each dataset. Such insights are invaluable for practitioners and researchers seeking to deploy NLP solutions in specific domains, as they enable informed decision-making regarding model selection and deployment strategies.

Furthermore, this project contributes to advancing the broader field of NLP by providing empirical evidence on the efficacy of different models across diverse datasets. Additionally, the development of an AI-driven chatbot capable of conducting nuanced comparative analyses underscores the potential of AI technologies to enhance human-machine interaction and address complex challenges in various domains.

In the following sections, we delve into the methodology employed for fine-tuning the NLP models, the datasets used for evaluation, the experimental setup, and the results obtained. Through this comprehensive exploration, we aim to provide valuable insights into the performance of NLP models in real-world scenarios and contribute to the ongoing discourse on advancing AI-driven solutions for language understanding and interaction.

2 Related Work

Comparison of BERT and RoBERTa for Sentence Embeddings

This study delves into the performance of BERT and RoBERTa models in generating sentence embeddings and their suitability for various tasks. The results show that using RoBERTa in place of BERT (SRoBERTa) can generate similar sentence embeddings and achieve comparable performance. The paper also discusses the advantages of using a multilingual model like XLM-R, which can achieve similar performance on a single language as a monolingual model. This information is valuable for researchers and developers who need to choose the most appropriate model for their sentence embedding tasks, as it provides insights into the relative strengths and trade-offs of BERT and RoBERTa.

Supporting Mental Health Self-Care Discovery through a Chatbot

This research explores the use of chatbots as a tool for discovering self-care solutions in mental health. The chatbot is designed to provide users with a wide range of crowdsourced self-care methods, such as meditation and mindfulness exercises, to help them manage their mental health. The study investigates the factors that contribute to building trust between the user and the chatbot, which is crucial for the effective delivery of mental health support.

Chatbots for E-commerce Customer Support4

This case study explores how AI-driven chatbots can transform e-commerce customer support. The article highlights the challenges faced by companies in effectively addressing customer queries, concerns, and feedback in a timely manner. By implementing a chatbot, companies can streamline communication, reduce reliance on human resources, and provide rapid and accurate responses to customer inquiries. The chatbot can be trained on a customer support dataset to manage repetitive questions and free up employees for more strategic tasks.

3 Methodology

In this study, we adopt a systematic methodology to develop an AI-driven chatbot for comparative analysis of Natural Language Processing (NLP) models on Mental Health and Customer Support datasets. We begin by collecting anonymized text data from relevant sources for each domain. Subsequently, we select three state-of-the-art pre-trained NLP models—GPT-2, BERT, and RoBERTa—and fine-tune them on the acquired datasets using transfer learning techniques. Following fine-tuning, we evaluate the performance of each model using domain-specific evaluation metrics tailored to the characteristics of the datasets. Finally, we conduct a comparative analysis to assess the accuracy, robustness, and suitability of the models for tasks within each domain. This methodology enables us to provide valuable insights into the performance of NLP models in real-world scenarios and guide decision-making in deploying AI-driven solutions tailored to specific applications.

3.1 Datasets

3.1.1 Mental Health Dataset

The Mental Health dataset comprises conversations between patients and healthcare assistants discussing various mental health issues. The dataset serves as a valuable resource for understanding the dynamics of interactions between individuals seeking support and healthcare providers in the context of mental health. Each entry in the dataset consists of conversational pairs of questions and answers, reflecting the dialogue flow between the patient and the healthcare provider. The dataset offers insights into common concerns, inquiries, and responses encountered in mental health support settings. Researchers and practitioners can leverage this dataset to develop and evaluate AI-driven solutions for mental health support, including chatbots, sentiment analysis models, and dialogue systems. The dataset is publicly available and can be accessed via the provided link. The dataset is comprised of a single field:

- text: This field contains conversational pairs representing questions posed by patients and corresponding answers provided by healthcare assistants. These exchanges encompass various mental health issues, offering a comprehensive view of the language used in such conversations.

By leveraging this dataset, we aim to train NLP models that can effectively manage patient inquiries related to mental health, potentially offering support and resources to individuals seeking help.

3.1.2 Customer Support Dataset

The Customer Support dataset includes discussions between clients and help desk agents, documenting a variety of questions and exchanges that take place in the customer support area. The dataset provides a rich collection of dialogues representing various categories of customer queries with a total of 26,872 rows. Essential attributes like flags for each customer query, customer requests, high-level semantic categories for intents, specific intents corresponding to user instructions, and sample responses expected from the virtual assistant are included in every entry of the dataset. This dataset is a great tool for researching customer service interactions, determining what customers need, and creating AI-powered processes that automate customer service procedures. This dataset can be used by researchers and practitioners to test and train customer support-specific chatbots, language understanding models, and dialogue management systems. Each row is structured with the following fields:

- Flags: This field employs tags to categorize different customer queries, allowing for targeted analysis of specific support needs.
- Instruction: This field captures the actual customer request phrased in a natural language format, simulating real-world customer interactions.
- category: This field provides a high-level understanding of the customer's intent by assigning a broader semantic category to the request.
- intent: This field delves deeper, pinpointing the specific intent behind the customer's instruction, enabling more precise model responses.
- response: This field presents an example answer crafted by a human support agent, offering a benchmark for training the NLP model to generate appropriate responses.

Our goal is to train natural language processing (NLP) models that can manage a variety of customer support queries and provide better customer service by utilizing this dataset.

3.2 Model Training

In this project, we undertake a meticulous process of training and fine-tuning three leading pre-trained NLP models—GPT-2, BERT, and RoBERTa—on the Mental Health and Customer Support datasets. Leveraging transfer learning techniques, we initialize the models with pre-trained weights and fine-tune them on domain-specific data to capture nuanced language patterns and semantics within each domain. Throughout the iterative training process, we optimize hyperparameters, monitor training metrics, and conduct rigorous evaluations using task-specific metrics. This iterative refinement ensures that our models are effectively adapted to the characteristics of the datasets, enhancing their performance, and enabling more accurate analyses and responses tailored to the Mental Health and Customer Support domains.

3.2.1 Generative Pre-Trained Transformer 2 (GPT-2)

In our project, GPT-2 serves as one of the key pre-trained NLP models for comparative analysis on the Mental Health and Customer Support datasets. Due to its generative capabilities and versatility in understanding and generating human-like text, GPT-2 holds promise for tasks such as dialogue generation, response generation, and language understanding within the context of mental health support and customer service interactions.

In our project, we utilize the powerful capabilities of GPT-2 for training and fine-tuning on both the Customer Support and Mental Health datasets. The process involves several key steps, including dataset preparation, model initialization, optimization, and iterative training.

Dataset Preparation
We begin by loading the dataset for each domain—Customer Support and Mental Health—and organizing it into a format suitable for training the GPT-2 model. For the Customer Support dataset, we created a custom dataset class that manages the tokenization of questions and responses using the GPT2Tokenizer provided by the Hugging Face Transformers library. Similarly, for the Mental Health dataset, we structure the data to feed into the GPT-2 model for fine-tuning.

Model Initialization and Optimization
We begin by initializing the GPT-2 model using the GPT2LMHeadModel class from the Hugging Face Transformers library. This class provides a convenient interface for loading pre-trained GPT-2 models and fine-tuning them on domain-specific datasets. Additionally, we set up the optimizer (AdamW) and learning rate scheduler to facilitate model optimization during training. These steps enable us to leverage the pre-trained weights of the GPT-2 model and efficiently optimize its parameters to improve performance on the Customer Support and Mental Health datasets.

Fine-Tuning on Domain-Specific Datasets
We fine-tune the GPT-2 model on both the Customer Support and Mental Health datasets separately. This fine-tuning process involves adjusting the model's parameters to better suit the characteristics of each dataset, such as the language patterns, semantics, and context specific to each domain. By fine-tuning on domain-specific data, we aim to enhance the model's ability to generate contextually relevant and coherent responses tailored to the respective domains.

Evaluation and Monitoring
Throughout the training and fine-tuning process, we monitor various training metrics such as loss, accuracy, and convergence behavior to assess the model's performance and progress. Additionally, we utilize validation datasets or cross-validation techniques to evaluate the model's generalization ability and ensure robust performance on unseen data.

3.2.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary pre-trained language representation model developed by Google. Unlike traditional language models, BERT

utilizes a bidirectional approach, allowing it to capture context from both left and right contexts in a sentence. This bidirectional understanding enables BERT to generate more accurate and contextually relevant representations of text. BERT has demonstrated remarkable performance across a wide range of NLP tasks, including sentence classification, named entity recognition, and question answering. Its versatility and effectiveness have made it one of the most widely adopted models in the NLP community. In our project, we leverage the power of BERT for comparative analysis on the Mental Health and Customer Support datasets, aiming to assess its performance alongside other pre-trained models such as GPT-2 and RoBERTa.

Dataset Preparation
Split each dataset into training and validation sets. For the Customer Support dataset, organize the data into question-response pairs. For the Mental Health dataset, categorize the text into predefined classes related to mental health issues.

Tokenization and Encoding
Use the BERT tokenizer to tokenize and encode the text data for input into the model. Set appropriate maximum sequence lengths and handle padding and truncation as needed.

Model Initialization
Initialize the BERT model for sequence classification using the BertForSequenceClassification class from the Hugging Face Transformers library. Also, we ensured that the model architecture matches the task requirements for both datasets.

Model Training and Optimization
Set up the optimizer (e.g., AdamW) with an appropriate learning rate and the learning rate scheduler (e.g., linear scheduler with warmup) to facilitate model optimization during training. Then, iterate over batches of data from the training set. Forward pass the input through the model, compute the loss using the model's outputs, perform backpropagation, and update the model parameters using the optimizer.

Evaluation
Evaluate the model's performance on the validation set periodically during training. Compute validation metrics such as loss and accuracy to assess the model's generalization ability and identify overfitting. After training, assess the model's performance on held-out test data.

3.2.3 RoBERTa (Robustly Optimized BERT Approach)

RoBERTa (Robustly optimized BERT approach) is an enhanced version of BERT (Bidirectional Encoder Representations from Transformers), developed by Facebook AI. It builds upon the success of BERT by introducing improvements in training methodology and model architecture. RoBERTa adopts a larger training dataset, removes the next sentence prediction (NSP) task, and employs dynamic masking during pre-training. These enhancements result in a more robust and effective language representation model that achieves state-of-the-art performance across various natural language understanding tasks. In our project, we utilize RoBERTa for fine-tuning on both

the Customer Support and Mental Health datasets, aiming to leverage its advanced capabilities for improved contextual understanding and classification accuracy.

Data Preparation
We Begin by splitting each dataset into training and validation sets. For the Customer Support dataset, the data is arranged into question-response pairs, while for the Mental Health dataset, the text is organized into conversational snippets or categorical labels.

Tokenization and Encoding
We utilized the RoBERTa tokenizer to tokenize and encode the text data, ensuring compatibility with the model's input requirements. And set the appropriate parameters for maximum sequence lengths and manage any necessary padding and truncation.

Model Training and Optimization
Initialize the RoBERTa model, tailored to the specific task requirements for each dataset. Iterate over batches of data from the training set, passing them through the RoBERTa model. Monitor training metrics, such as loss and performance, to track training progress.

By rigorously training and fine-tuning the pre-trained NLP models on domain-specific datasets, we aim to enhance their performance and adaptability to the Mental Health and Customer Support domains, facilitating more accurate and contextually relevant analyses and responses within each domain.

4 Model Evaluation

In our project, we use accuracy as a metric to evaluate the performance of our model. Accuracy measures the proportion of correctly predicted labels among all the instances in the dataset. Specifically, it calculates the ratio of the number of correctly classified instances to the total number of instances evaluated.

Tokenization and Bag-of-Words Encoding
Each sentence in the test data is tokenized into individual words, and a bag-of-words representation is created. This representation captures the presence or absence of specific words in the sentence.

Inference and Prediction
The bag-of-words representation of each sentence is fed into the model for inference. The model predicts the tag associated with the input sentence based on its learned parameters.

Accuracy Calculation
The predicted tag is compared against the expected tag for each sentence. If the predicted tag matches the expected tag, the prediction is considered correct, and the accuracy count is incremented. Finally, the accuracy is calculated as the ratio of correctly predicted instances to the total number of instances evaluated.

4.1 Results

Mental Health Dataset

Among the models evaluated on the Mental Health dataset, RoBERTa emerged as the most effective, achieving an accuracy of 57.68%. This signifies RoBERTa's superior capability in comprehending and responding to conversations related to mental health issues compared to GPT-2 and BERT. Its optimized architecture enabled better contextual understanding and generation of relevant responses tailored to mental health queries and discussions.

Customer Support Dataset

For the Customer Support dataset, BERT demonstrated the highest accuracy of 63.64%, outperforming both GPT-2 and RoBERTa. BERT's robust performance suggests its proficiency in understanding and addressing customer inquiries and support-related interactions effectively. The model's bidirectional encoding and attention mechanisms facilitated comprehensive comprehension of customer queries and generation of accurate responses.

Accuracy	GPT 2	RoBERTa	BERT
Mental Health Dataset	40.56	57.68	38.46
Customer Support Dataset	41.67	60	63.43

Interpretation

- The success of RoBERTa on mental health conversations and BERT on customer support inquiries underscores the importance of task relevance in model selection. Tailoring the model choice to the task requirements enhances performance and ensures more accurate responses.
- The findings provide valuable insights for deploying AI-driven chatbots in mental health support and customer service applications, guiding the selection of models to optimize performance and user experience.
- RoBERTa's success on mental health conversations and BERT's proficiency in customer support inquiries stem from their ability to understand and utilize domain-specific vocabulary and nuances.
- RoBERTa's success on mental health conversations and BERT's proficiency in customer support inquiries stem from their ability to understand and utilize domain-specific vocabulary and nuances.

5. Challenges and Future Work

Challenges

1. **Hardware Limitations:** Fine-tuning large models like BERT, RoBERTa demands significant computational resources, such as GPUs or TPUs, to expedite training processes. However, limited access to high-performance hardware can hinder the efficiency and scalability of the fine-tuning process, slowing down experimentation and model optimization.
2. **Hyperparameter Optimization:** Achieving optimal performance with BERT necessitates fine-tuning hyperparameters, such as learning rates, batch sizes, and optimization algorithms. Conducting thorough experimentation to identify the most effective hyperparameter configurations requires extensive computational resources and time, which may be limited in practical settings.

To address these challenges and enhance model accuracy in future work:

1. **Hardware Infrastructure:** Investing in or gaining access to advanced hardware infrastructure, such as GPUs or TPUs, can significantly accelerate the fine-tuning process for large language models like BERT.
2. **Rigorous Hyperparameter Tuning:** Implementing rigorous hyperparameter tuning techniques, involving systematic exploration of hyperparameter spaces and validation on appropriate evaluation metrics, can streamline the process of identifying optimal configurations.
3. **Transfer Learning Strategies:** Exploring transfer learning strategies that leverage pre-trained models and domain-specific data can enhance model accuracy with fewer computational resources

Conclusion

The effectiveness of the GPT-2, RoBERTa, and BERT models adjusted for the Mental Health and Customer Support datasets was successfully assessed in this project. Through our investigation, we observed varying model performances, with RoBERTa excelling on the Mental Health dataset and BERT demonstrating superiority on the Customer Support dataset.

The results emphasize how crucial it is to choose a model according to the particular application domain.

- **BERT:** Exhibited complete dominance, demonstrating exceptional performance in both datasets, with a special emphasis on Customer Support tasks.
- **RoBERTa:** Perhaps because of its capacity for masked language modeling, it demonstrated promise for use in mental health tasks.
- **GPT-2:** Although attaining a moderate level of accuracy, it was not as successful as BERT and RoBERTa in both domains.

These findings offer practitioners using NLP solutions insightful information. Depending on the specific focus and desired functionalities, RoBERTa or BERT could be viable options for chatbots designed to support mental health. Based on the results of this project, BERT seems to be the most promising option for customer support chatbots.

Subsequent investigations could delve more deeply into variables affecting model performance, like dataset size, hyperparameter adjustment, and looking into different pre-trained models made especially for customer service or mental health tasks.

References

1. Joonas Moilanen, Niels van Berkel, Aku Visuri, Ujwal Gadiraju, Willem van der Maden and Simo Hosio Research Article in NLM **Supporting mental health self-care discovery through a chatbot.** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10028281/
2. Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha, **AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing**. https://arxiv.org/pdf/2108.05542g
3. Kamal Nayanam, Vatsala Sharma, **Towards architecting research perspective future scope with chat GPT.** https://www.researchgate.net/publication/378704631_TOWARDS_ARCHITECTING_RESEARCH_PERSPECTIVE_FUTURE_SCOPE_WITH_CHAT_GPT
4. Mohsen Khosravi, **Factors influencing patient engagement in mental health chatbots: A thematic analysis of findings from a systematic review of reviews.** https://www.researchgate.net/publication/380002830_Factors_influencing_patient_engagement_in_mental_health_chatbots_A_thematic_analysis_of_findings_from_a_systematic_review_of_reviews
5. Alok Pandhare, **Using BERT to Build Chatbots for E-Commerce.** https://www.researchgate.net/publication/379810195_Using_BERT_to_Build_Chatbots_for_E-Commerce
6. Siddhant Meshram; Namit Naik; Megha VR; Tanmay More; Shubhangi Kharche, **Conversational AI: Chatbots.** https://ieeexplore.ieee.org/document/9498508
7. Lorentsa Gkinko, Amany Elbanna; AI Chatbots sociotechnical research: An overview and Future Directions. https://ceur-ws.org/Vol-3239/paper17.pdf
8. Bayan Abu Shawar, Eric Atwell; **Chatbots: Are they Really Useful?** https://www.researchgate.net/publication/220046725_Chatbots_Are_they_Really_Useful
9. Jingyun Wang Gwo-Haur Hwang, Ching-Yi Chang: **Directions of the 100 most cited chatbot-related human behavior research: A review of academic publications.** https://www.sciencedirect.com/science/article/pii/S2666920X21000175
10. Chien-Chang Lin, Anna Y. Q. Huang andStephen J. H. Yang. **A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022).** https://www.mdpi.com/2071-1050/15/5/4012
11. Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis; An Overview of Chatbot Technology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256567/
12. Farhan Aslam; **The Impact of Artificial Intelligence on Chatbot Technology: A Study on the Current Advancements and Leading Innovations.** https://ajpojournals.org/journals/index.php/EJT/article/view/1561
13. Vajinepalli Sai Harsha Vardhan, Parsi Anurag, Richa Sharma; RULE BASED CHATBOT. https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2022/22117/final/fin_irjmets1651726377.pdf

14. Anupam Mondal; Monalisa Dey; Dipankar Das; Sachit Nagpal; Kevin Garda; **Chatbot: An automated conversation system for the educational domain.** https://ieeexplore.ieee.org/document/8692927
15. Kadir Uludag; **The Use of AI-Supported Chatbot in Psychology.** https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4331367
16. Uday Sehgal, Shweta Bhardwaj; Building a Chatbot using Natural Language Processing. https://www.researchgate.net/publication/378081706_Building_a_Chatbot_using_Natural_Language_Processing
17. Dr. R. Regin, Dr. S. Suman Rajesh, Shynu T, Jerusha Angelene Christabel G, Steffi. R; **An Automated Conversation System Using Natural Language Processing (NLP) Chatbot in Python.** https://cajmns.centralasianstudies.org/index.php/CAJMNS/article/view/1027.
18. Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Vasundhara Rathod, Shreya Bisen; **Implementation of a Chatbot System using AI and NLP.** https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531782
19. V. Adarsh, B. Koushik, D. Mahesh; CHATBOT USING NATURAL LANGUAGE PROCESS (NLP). https://www.irjmets.com/uploadedfiles/paper/issue_2_february_2023/33529/final/fin_irjmets1676472749.pdf
20. Varun Srivastava, Suraj Kumar Prajapati, Shri Krishna Yadav, Dr. Himani Mittal; **Healthcare Chatbot System using Artificial Intelligence.** https://www.ijirt.org/Article?manuscript=151926