CHARLESTON SOUTHERN UNIVERSITY

Graduate School

This is to certify that the thesis prepared

By: **Donald J. Lauer IV**

Entitled: **Comparative Analysis of Semantic Understanding: ChatGPT vs. BERT**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this university and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

____________________ Chair
Dr. Valerie Sessions, Ph.D

____________________ External Examiner
Dr. Emory Hiott, DBA

Megan J Busch ____________________ External Examiner
Dr. Megan Busch, Ph.D.

Songhui Yue ____________________ Examiner
Dr. Songhui Yue, Ph.D.

Approved by ____________________
Dr. Valerie Sessions,
Director of Graduate Study in Computer Science

4/17/24 2024

____
Dr. William T. Ashby, Dean
College of Science and Mathematics

Comparative Analysis of Semantic Understanding: ChatGPT vs. BERT

A thesis

by

Donald J. Lauer IV

Submitted to the Graduate School of

Charleston Southern University in fulfillment

of the requirement for the degree of

MASTER OF SCIENCE

February 2024

Major: Computer Science

Thesis Committee Members

Advisor: Dr. Valerie Sessions

Dr. Emory Hiott

Dr. Songhui Yue

Dr. Megan Busch

List of Equations, Figures, and Tables

Equation:

Figures

Tables:

Abstract

This thesis presents a comprehensive comparative analysis of two leading artificial intelligence (AI) language models, ChatGPT and BERT (Bidirectional Encoder Representations from Transformers), in the context of natural language processing (NLP), with a specific focus on sentiment analysis. The study aims to evaluate and contrast the capabilities of these models in understanding and processing sentiment-laden language, a crucial component of effective human-AI interaction. Employing a range of sentiment scenarios, including positive, negative, neutral, and mixed sentiments, and specialized datasets from the University of California, Irvine (UCI) Machine Learning Repository, the thesis provides an in-depth assessment of the models based on accuracy and F1-score metrics.

The results demonstrate BERT's remarkable consistency and proficiency across various sentiment contexts, reflecting its deep understanding of context in language. BERT excels in accurately processing a wide range of sentiments, maintaining high performance even in texts with nuanced or conflicting emotions. Conversely, ChatGPT, while showing strong performance in generating coherent and contextually appropriate responses, particularly in positive sentiment scenarios, exhibits variability in handling more complex emotional content. The analysis reveals its strengths in language generation and limitations in interpreting and processing negative, neutral, or mixed sentiments.

The thesis contributes to the field of NLP by highlighting each model's distinct capabilities and potential applications. It underscores the importance of continuous development in AI models for more nuanced, context-aware, and emotionally intelligent systems. The findings provide valuable insights for future advancements in semantic understanding, guiding the evolution of AI interactions in natural language processing. The study not only delineates the

current capabilities of ChatGPT and BERT but also sets the stage for future research, emphasizing the necessity for more refined AI systems for effective human-AI interaction. This thesis is foundational for ongoing research in the rapidly evolving domain of AI and NLP.

Furthermore, the analysis is primarily focused on analyzing the English language and relies heavily on quantitative metrics such as accuracy and F1-scores which are measurement of predictive performance. Precision is the ratio of true positive predictions to the total number of positive predictions made (including both true positives and false positives), and recall is the ratio of true positive predictions to the actual number of positive instance within the data (including both true positives and false negative). Although informative, these metrics may not fully capture the qualitative aspects of language understanding and generation. This quantitative focus might overlook some subtleties of semantic interpretation and contextual nuances that qualitative analysis could reveal. In the future, this comparative study could include more diverse and updated versions of the AI models used in this study, investigate the model’s performance on broader linguistic tasks beyond sentiment analysis, and incorporate more diverse and complex datasets such as multilingual and cross-cultural contexts.

Chapter 1: Overview

In the realm of natural language processing (NLP), the task of semantic understanding plays a pivotal role, particularly in the context of sentiment analysis. This paper focuses on contrasting these two advanced language models' abilities to comprehend and interpret sentiments expressed in text. This comparison is crucial, as sentiment analysis is critical in various domains, from market research to social media monitoring.

ChatGPT by OpenAI:

- Model Architecture: ChatGPT, a language model based on GPT-3.5 architecture developed by OpenAI (OpenAI, 2023), employs a transformer-based generative model. It is pre-trained on a diverse range of internet text and fine-tuned for specific tasks.
- Semantic Understanding in Sentiment Analysis: ChatGPT's strength lies in generating contextually coherent and relevant responses. Its performance in sentiment analysis is influenced by its generative nature, enabling it to infer sentiments by constructing responses based on the emotional tone of the input.

BERT by Google:

- Model Architecture: BERT (Bidirectional Encoder Representations from Transformers) by Google (Devlin et al., 2018) differs from traditional models by its bidirectional training, which allows it to understand the context of a word based on all its surroundings rather than just one direction.
- Semantic Understanding in Sentiment Analysis: BERT excels in extracting contextual meanings of words, making it highly effective in identifying nuanced sentiments. Its bidirectional nature allows for a more profound understanding of the sentiment expressed in complex sentences.

Comparative Analysis:

- Approach to Semantic Analysis: The comparative analysis will explore how each model processes and interprets emotional nuances in language. While OpenAI's ChatGPT (OpenAI, 2023) may generate responses indicative of sentiment understanding, Google's BERT (Devlin et al., 2018) architecture allows it to analyze text semantics deeply.
- Performance in Sentiment Analysis Tasks: This section will examine how each model performs in accurately categorizing sentiments in varied datasets, considering aspects like accuracy, precision, and recall.
- Handling of Contextual and Subtle Sentiments: The analysis will delve into the models' capabilities in understanding contextual cues and subtle emotional expressions, which are essential in sentiment analysis.
- Practical Implications: The comparison will also consider how these models can be applied in real-world sentiment analysis scenarios, evaluating their effectiveness and limitations.

Chapter 1.1 Introduction

Natural language processing (NLP) has witnessed substantial advancement in recent years, primarily attributed to the development of sophisticated language models such as OpenAI's ChatGPT (OpenAI, 2023) and BERT, developed by Google (Devlin et al., 2018). These models have revolutionized researchers' approach to understanding and processing the human language, offering remarkable capabilities in various applications, from automated customer service to enhanced text analysis. This thesis analyzes two pivotal NLP models: ChatGPT, a GPT (Generative Pre-trained Transformer) architecture variant, and BERT

(Bidirectional Encoder Representations from Transformers). The focus lies in examining their semantic understanding capabilities.

Semantic understanding in NLP pertains to the ability of a system to comprehend and interpret the meaning and context of words and sentences in human language. This capability is central to numerous NLP applications, including machine translation, sentiment analysis, and question-answering systems. ChatGPT, developed by OpenAI (OpenAI, 2023), represents a line of models known for their generative capabilities and for producing coherent and contextually relevant text based on given prompts. BERT, developed by Google (Devlin et al., 2018), is renowned for effectively handling context in language, which is achieved through its innovative bidirectional training approach.

The comparative analysis in this thesis is structured to address several vital dimensions: architectural differences between ChatGPT and BERT, their respective approaches to semantic understanding, performance benchmarks in various NLP tasks, and practical implications of their semantics capabilities in real-world applications. This analysis aims not only to highlight the strengths and limitations of each model but also to contribute to the broader understanding of current NLP technologies. This thesis aspires to shed light on the current state of semantic understanding in Artificial Intelligence (AI) and its potential future directions by dissecting the mechanisms through which these models process and understand language.

The preceding chapters meticulously prepared the groundwork for an in-depth examination of the models used in this study. This exploration commences with extensive background analysis, delving into the developmental history and theoretical foundations of ChatGPT and BERT. This is followed by a literature review that evaluates various themes pertinent to the evolution of artificial intelligence and machine learning. Subsequently, the methodology section will delineate the systematic approach adopted for comparing these models, culminating in a thorough analysis and discussion. The thesis will ultimately draw to a close with a reflective synthesis of the findings, contemplating their broader implications for the advancement of NLP and AI.

Chapter 2: Literature Review

Natural Language Processing (NLP) is a pivotal subfield of artificial intelligence that focuses on the interaction between computers and human language. It involves programming computers to process and analyze large amounts of natural language data. The goal of NLP is to enable computers to understand, interpret, and respond to human language in a valuable and meaningful way.

- Evolution of NLP: NLP has evolved significantly over the past few decades. Initially, NLP relied heavily on rule-based methods in which linguists manually coded language rules into systems. The advent of machine learning and, subsequently, deep learning transformer NLP enables more sophisticated and nuanced language understanding. These advancements led to the development of complex models capable of various tasks, from speech recognition and language translation to sentiment analysis and chatbots.
- Key Challenges: NLP encompasses various challenges, primarily due to the complexity and diversity of human language. These include handling ambiguous and context-dependent meanings, understanding idioms and colloquialisms, and dealing with different dialects and languages.

Initially, NLP systems were primarily rule-based, relying on linguists to manually input language rules. This approach had significant limitations, especially in handling the complexity and nuance of human language. Nadkarni et al. (2011) illustrated this early phase, emphasizing the shift towards statistical methods in the 1980s. This shift was facilitated by the introduction of machine learning techniques and the availability of large text corpora, which allowed for probabilistic approaches to language processing, significantly advancing the field's capabilities in understanding and generating human language.

The advent of deep learning and, more specifically, transformer architectures has marked a new era in NLP, offering unparalleled advancements in processing and analyzing natural language data. Wolf et al. (2020) documented how transformers have rapidly become the dominant architecture in NLP, surpassing traditional models like convolutional and recurrent neural networks in performance across a broad spectrum of tasks. This includes natural language understanding, language generation, and even machine translation. Transformers excel due to their ability to efficiently process large datasets, capture long-range dependencies in text, and facilitate more nuanced understanding and generation of language through advanced pretraining techniques.

2.1 ChatGPT: Generative Pre-trained Transformer

ChatGPT is a variant of the GPT (Generative Pre-trained Transformer) series developed by OpenAI (OpenAI, 2023) and represents a significant leap in language model capability.

- Model Architecture: ChatGPT is based on transformer architecture, a significant breakthrough in NLP due to its ability to handle long-range dependencies in text. Unlike earlier models that processed text linearly, transformers use attention mechanisms to weigh the importance of different words in a sentence, regardless of their position.
- Training and Functionality: ChatGPT is pre-trained on a vast corpus of text data, enabling it to understand and generate human-like text. This pre-training is followed by fine-tuning, where the model is further trained on specific tasks, enhancing its ability to respond contextually in conversations.
- ChatGPT in Semantic Understanding: The strength of ChatGPT lies in its ability to generate coherent, contextually appropriate text. It can engage in dialogue, answer questions, and even create content often indistinguishable from that written by humans.

Regarding semantic understanding, ChatGPT can grasp the nuances of language, including sentiment, tone, and intent.

As detailed by Ray (2023), the landscape of artificial intelligence (AI) and natural language processing (NLP) has been transformed by the advent of sophisticated language models like ChatGPT. Ray (2023) highlighted the emergence of generative AI models, particularly noting that “Generative AI refers to a class of artificial intelligence models that can create new data based on patterns and structures learned from existing data.” ChatGPT is among these AI models that have been developed into a powerful tool capable of being applied across various domains. This narrative traces the trajectory of ChatGPT from its inception, driven by advancements in the Transformer architecture, to its current state as a sophisticated model capable of understanding and generating human-like language, underscoring its profound impact on the field of NLP and AI.

In a complementary analysis, Bahrini et al. (2023) outlined the progression and scalability of the GPT series, leading to the creation of ChatGPT. The authors stated, “The architecture known as GPT initially introduced by OpenAI in 2018 serves as the basis for ChatGPT... It had advanced to version 3.5 by the time ChatGPT went public in November 2022” (Bahrini et al., 2023). The text further discussed the balance between the innovative capabilities of ChatGPT and the ethical considerations it necessitates. It underscored the model’s revolutionary role in automating tasks and its potential to provoke ethical concerns, advocating for responsible use to harness its benefits while mitigating risks. This dialogue between the technological evolution of ChatGPT and its societal implications presents a holistic view of the model’s journey from a novel AI experiment to a cornerstone in the current and future landscape of NLP and AI.

2.2 BERT: Bidirectional Encoder Representations from Transformers

BERT (Bidirectional Encoder Representations from Transformers), developed by Google, is another groundbreaking model in the field of NLP (Devlin et al., 2018).

- Innovative Architecture: What sets BERT apart is its bidirectional training, which allows the model to understand the context of a word based on all of its surrounding text rather than just the text that precedes it. This approach enables a deeper understanding of sentence structure and meaning.
- Pre-training and Contextual Understanding: BERT is pre-trained on a large corpus, similar to ChatGPT, but it is specifically designed to perform well on tasks that require a deep understanding of context and relationships within text. This includes question-answering, sentiment analysis, and language inference.
- BERT in Semantic Analysis: BERT's ability to analyze semantics is rooted in its understanding of the context in which words appear. This makes it particularly adept at tasks requiring nuanced understanding, such as identifying the sentiment expressed in a text or understanding the relationship between sentences.

The introduction of BERT by Google represented a significant evolution in the NLP field, as Rogers et al. (2020) highlighted. BERT is a model premised on transformer architecture and has been instrumental in enhancing the understanding of long-range dependencies in text, marking a departure from traditional models that rely on unidirectional text processing. Rogers et al. (2020) elucidated BERT's unique bidirectional training regime, which empowers the model to contextually interpret the text by considering all surrounding words, thereby enriching its semantic analysis capabilities: "Pre-training uses two self-supervised tasks: masked language modeling (MLM prediction of randomly masked input tokens) and next sentence prediction

(NSP predicting if two input sentences are adjacent to each other).” This architectural innovation allows BERT to excel in various NLP tasks, ranging from question-answering and sentiment analysis to language inference, by leveraging its profound understanding of language context and relationships within the text.

Furthermore, Koroteev (2021) expounded on BERT’s groundbreaking approach to pre-training on a large corpus of text, setting a new standard in the industry. As per the review, BERT’s method of masked language modeling and next-sentence prediction during its pre-training phase is pivotal for its unparalleled performance across diverse NLP applications. “BERT tries to get around this limitation by using learning according to the so-called ‘masked language models’ that is, the target function of learning a given representation formalizes the task of predicting a randomly selected and masked word in a text taking into account only the surrounding context. Thus a deep bi-directional transformer is trained” (Koroteev, 2021). This model achieves state-of-the-art results and simplifies the adaptation process for specific tasks by adding minimal task-specific layers. The paper emphasized BERT’s efficacy in understanding the nuances of language, making it exceptionally adept at tasks that require a nuanced understanding of text semantics. Through its comprehensive pre-training and fine-tuning processes, BERT has significantly advanced the field of NLP, demonstrating a deeper understanding of sentence structure and meaning than was previously possible.

2.3 Comparative Analysis: ChatGPT vs. BERT

In the comparative analysis of ChatGPT and BERT, several key aspects will be examined:

- Approach to Semantic Understanding: While both models are advanced in understanding language, their approaches differ. ChatGPT, with its generative nature, excels in creating contextually relevant responses, whereas BERT’s bidirectional understanding provides a deeper analysis of sentence structure and meaning.

- Performance Metrics: The comparison will be based on various NLP tasks that require semantics understanding, focusing on metrics such as accuracy, precision, recall, F1-scores, and the ability to handle complex and nuanced language.
- Application and Use Cases: The practical applications of both models in real-world scenarios, including their strengths and limitations, will be explored. This includes examining how these models are used in industry and academia for tasks like sentiment analysis, chatbots, and information extraction.

Delving deeper into the comparative analysis between ChatGPT and BERT, Zhong et al. (2020) provided an illuminating examination of their abilities across a spectrum of NLP tasks. The analysis revealed that ChatGPT, leveraging its generative capabilities, excels in inference tasks, showcasing its superior reasoning abilities. Zhong et al. (2023) highlighted that “ChatGPT surpasses all BERT-style models on natural language inference tasks i.e. MNLI and RTE, indicating its superiority on inference/reasoning.”

However, the study also uncovered ChatGPT’s challenges, particularly with paraphrase and similarity tasks, where it lags significantly behind BERT. This gap underscores the nuanced capabilities of BERT in understanding and analyzing text semantics, as BERT’s bidirectional design and attention mechanism provide it with a robust framework for deep semantic analysis. The performance discrepancy in tasks such as MRPC and STS-B illustrates the limitations of ChatGPT’s generative approach when precise semantic distinctions are required.

The practical implications of these findings are profound. For tasks requiring high levels of inference and reasoning, ChatGPT’s abilities make it a powerful tool for generating human-like text and engaging in complex dialogues. In contrast, BERT’s strengths in handling paraphrase and similarity tasks make it indispensable for applications requiring nuanced text understanding, such as document summarization and semantic search.

Further analysis within the paper explored the potential for improving ChatGPT’s performance on its weaker tasks through advanced prompting strategies. These strategies, including standard few-shot prompting, manual few-shot CoT (Chain of Thought) prompting, and zero-shot CoT, demonstrated significant improvements in ChatGPT’s performance, narrowing the gap with BERT-style models on some tasks. The study found that “with the help of these prompting strategies, ChatGPT can achieve significant performance improvements and even outperforms the powerful RoBERTa-large on some tasks” (Zhong et al., 2023).

Zhong et al.’s (2023) comprehensive analysis of ChatGPT and BERT highlighted their respective strengths and weaknesses across various NLP tasks and suggested pathways for leveraging advanced prompting techniques to enhance ChatGPT’s understanding capabilities. The study’s findings offer valuable insights for both academic research and practical applications, suggesting that the choice between ChatGPT and BERT should be guided by the specific requirements of the task at hand, considering the potential benefits of advanced prompting strategies to optimize performance.

This background section sets the stage for a thorough comparative analysis of ChatGPT and BERT, highlighting their unique features and contributions to the field of NLP, particularly in semantic understanding. The subsequent sections will delve deeper into their comparative performance and the implications of their respective approaches to semantic analysis.

The exploration of semantic understanding in the realms of ChatGPT and BERT necessitates a deep dive into the interconnected spheres of artificial intelligence (AI), machine learning (ML), and natural language processing (NLP). This literature review begins by tracing AI's evolution from its conceptual inception to its modern-day manifestations. It examines the progression from theoretical frameworks to practical applications, highlighting how AI has become indispensable to technological advancements. The review then transitions to focus on ML, a subset of AI that has been instrumental in enabling computers to learn from and make decisions based on data. In this context, the review scrutinizes critical methodologies and algorithms that have been pivotal in the development of advanced ML models. This exploration provides a backdrop to understanding how AI and ML synergize to drive innovations in NLP, setting the stage for the emergence of sophisticated models like ChatGPT and BERT.

Artificial intelligence (AI) is defined as "a system's ability to interpret external data correctly to learn from such data and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Haenlein & Kaplan, 2019). The evolution of artificial intelligence (AI) from its conceptual inception to its modern applications presents a fascinating journey through technological advancements. AI, defined as a system's ability to interpret external data correctly, learn from such data, and use those learnings to achieve specific goals through flexible adaptation, has transformed from a field of scientific obscurity into a cornerstone of contemporary technology. This transformation is rooted in the early 1940s, with significant milestones such as Isaac Asimov's formulation of the Three Laws of Robotics and Alan Turing's development of The Bombe, highlighting the foundational role of theoretical explorations in AI's history.

The journey of AI through periods of optimism and skepticism, notably the AI Spring marked by the Dartmouth Conference and the subsequent AI Winters, illustrates the cyclical nature of AI research and development. Despite early challenges, the revival of interest in AI, driven by advancements in deep learning and computing power, signifies a new era of possibilities. The transition from expert systems to contemporary AI applications, such as IBM's Deep Blue and Google's AlphaGo, showcases the shift from rule-based to machine learning-based approaches, emphasizing AI's capacity for innovation.

In exploring the definition and history of AI, it becomes clear that AI's development is not just a tale of technological advancements but also a reflection of the changing perceptions of intelligence, computation, and the potential for machines to emulate human cognitive processes. "It is often neglected that in scientific discussions there are (at least) two types of definitions with different properties: a dictionary definition is descriptive... while a working definition is prescriptive" (Wang, 2019). This literature review sets the stage for a detailed examination of NLP advancements, mainly through the lens of transformative models like ChatGPT and BERT, which represent the culmination of decades of AI research and development, embodying the principles of learning, adaptation, and the ongoing quest for machines that can understand and generate human language with unprecedented sophistication.

In the second phase, the review narrows its focus to natural language processing, a field at the intersection of linguistics, computer science, and AI. This section delves into the evolution of NLP, examining how it has transitioned from rule-based systems to more advanced machine learning-based approaches. The analysis here includes an exploration of the transformer architecture, a breakthrough in NLP that has been fundamental to the development of models like ChatGPT and BERT. The review analyzes the literature surrounding these models' particularity, emphasizing their architecture, training methodologies, and how they have redefined the benchmarks for semantic understanding in machines. This comprehensive overview not only contextualizes the capabilities of ChatGPT and BERT within the broader landscape of NLP but also sets a foundation for a detailed comparative analysis of their approaches and effectiveness in semantic understanding tasks.

2.4 Evolution of ChatGPT

This literature review section, we will discusses the evolution of ChatGPT and the development and application of GPT models that have significantly advanced the capabilities of conversational Artificial Intelligence (AI) capabilities. Wang et al. (2021) delineated the advancements in generative language models. Wang et al. (2021) emphasized that there is a shift from traditional retravel-based chatbots to more advanced, generative-based models, particularly highlighting the capabilities of OpenAI's GPT models. This is further expanded upon when the authors stated that "Generative-based approaches such as the OpenAI GPT models could allow for more dynamic conversations in therapy chatbot contexts than previous approaches" (Wang et al., 2021). This transition signifies a significant development in conversational AI, marking a move towards more nuanced and flexible interactions that better cater to the unique and varied demands of therapy contexts. Wang et al. focused on evaluating the GPT-2 model through specific metrics like non-word outputs, response length, and sentiment components, further illustrating the depth of analysis and sophistication that now characterizes the evaluation of conversational AI systems. This detailed examination underscored the ongoing evolution of ChatGPT, reflecting a broader trend towards more adaptive, responsive, and context-aware AI systems in various applications.

Following the exploration of generative language models in the therapy context from Wang et al. (2021), Khan and Uddin (2022) further exemplified the remarkable evolution of ChatGPT in a different domain, illuminating the impact of Codex, a model based on GPT-3, which significantly enhances the process of code documentation—a necessary aspect of software engineering. Khan and Uddin (2022) articulated, “Codex is a GPT-3 based model pre-trained on both natural and programming languages,” outperforming existing automatic code documentation techniques, even with basic settings like one-shot learning. This advancement indicates a substantial leap in the capabilities of GPT models, moving beyond conversational tasks to more technical applications. The comparative analysis with BERT-based models in this paper demonstrated the versatility and state-of-the-art performance of GPT-3, further underscoring its evolutionary journey. The application of GPT-3 in automating complex and labor-intensive tasks such as code documentation highlights its adaptability across different domains and marks a significant milestone in the evolution of ChatGPT, demonstrating its growing impact and potential in diverse fields.

Continuing the narrative of ChatGPT’s evolution from therapy and software engineering, Setianto et al. (2022) presented another pivotal advancement in cybersecurity. Setianto et al. (2022) leveraged GPT-2, a predecessor of GPT-3, for parsing and interpreting data from honeypot logs, an important component in detecting and analyzing cyber threats, highlighting that the system’s capability to parse dynamic logs generated by a Cowrie SSH honeypot effectively. They stated: “This article presents a run-time system (GPT-2C) that leverages a large pre-trained language model (GPT-2) to parse dynamic logs” (Setianto et al., 2022). This innovative application of GPT-2 in cybersecurity underscores the versatility of the ChatGPT series. The model’s success in accurately parsing and analyzing complex log data, as evidenced by its 89% inference accuracy, signifies a notable expansion of the GPT models' application spectrum. This extension from general-purpose language processing to specialized cybersecurity tasks showcases the adaptability of GPT models across diverse domains and cements ChatGPT's status as a multifaceted tool capable of addressing a wide array of complex, real-world problems.

Expanding upon the diverse application of ChatGPT models in cybersecurity and software engineering, Brown et al. (2020) further propelled the evolutionary trajectory of ChatGPT, particularly emphasizing the groundbreaking capabilities of GPT-3. Brown et al. (2020) delved into GPT-3's advanced few-shot learning abilities of GPT-3, where it remarkably performs various complex Natural Language Processing (NLP) tasks without requiring task-specific fine-tuning. They noted that "scaling up language models greatly improves task-agnostic few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches" (Brown et al., 2020). This leap in performance illustrates a significant milestone in the evolution of ChatGPT models, highlighting GPT-3's proficiency in adapting to tasks with minimal input and its capacity to process and generate human-like text. The research presented by Brown et al. (2020) not only exemplifies GPT -3's superiority in handling a broad spectrum of NLP challenges but also signifies its potential to revolutionize how AI systems learn and interact, setting a new benchmark in the field and marking a pivotal point in the ongoing development of ChatGPT technologies.

2.5 Natural Language Processing Task with ChatGPT

This section will delve into the intricate functionalities of ChatGPT in the Natural Language Processing (NLP) field. One notable task in NLP is the ability to translate various languages, highlighted in Jiao et al. (2023). This paper assessed the nuances and breadth of ChatGPT's application in NLP, particularly highlighting GPT's comparative performance with established translation systems. The authors emphasized that “By evaluating on a number of benchmark test sets, we find that ChatGPT performs competitively with commercial translation products (e.g., Google Translate) on high-resource European languages but lags behind significantly on low-resource or distant languages” (Jiao et al., 2023). This analysis allows for further understanding in ChatGPT’s sophisticated yet varied efficacy across different linguistic scenarios. Moreover, the paper delved into the intricacies of translation prompts and multilingual translation, noting that “ChatGPT is a single model handling various NLP tasks and covering different languages, which can be considered a unified multilingual machine translation model” (Jiao et al., 2023). The paper reinforced the comprehension of ChatGPT’s role in NLP, underscoring its capabilities, limitations, and challenges in bridging the gaps in machine translation.

Furthermore, Wang, Liang, Meng, Sun, et al. (2023) expanded upon the exploration of ChatGPT’s functionalities in Natural Language Processing (NLP) by further examining its role as an evaluator in Natural Language Generation (NLG). This study marks a significant shift in assessing ChatGPT’s capabilities, moving from translation to evaluation. The authors underscored the model’s effectiveness in evaluating NLG tasks, stating, “ChatGPT achieves state-of-the-art or competitive correlation with human judgments on various NLG tasks” (Wang, Liang, Meng, Sun, et al., 2023). This insight showcased ChatGPT’s versatility and advanced understanding of NLP, where it generates language and assesses the quality of generated content across different tasks. The authors emphasized ChatGPT’s adaptability and proficiency in various NLG tasks, such as text summarization, story generation, and data-to-text generation, providing a nuanced perspective on its performance as an NLG evaluator. These evaluations demonstrate the breadth of ChatGPT’s capabilities in NLP, highlighting its role not just as a creator but also as a discerning evaluator of language, adding complexity to our understanding of its applications in NLP.

Furthering the analysis of ChatGPT's multifaceted capabilities in NLP, Wang, Liang, Meng, Zou, et al. (2023) provided an innovative application of ChatGPT in Cross-Lingual Summarization (CLS). This transition from NLG evaluation to direct NLP task execution illustrates a significant expansion of ChatGPT's functional spectrum. The authors revealed that "ChatGPT originally prefers to produce lengthy summaries with more detailed information. But with the help of an interactive prompt, ChatGPT can balance between informativeness and conciseness and significantly improve its CLS performance" (Wang, Liang, Meng, Zou, et al., 2023). This observation highlights not only ChatGPT's advanced linguistic synthesis capabilities, but also its adaptability to the nuanced demands of CLS—a task that combines the complexities of translation with the preciseness required in summarization. The research further presented "a preliminary evaluation of ChatGPT's zero-shot CLS performance" (Wang, Liang, Meng, Zou, et al., 2023), thereby opening new avenues in the application of large language models like ChatGPT in diverse NLP tasks. The progression from language evaluation to active language processing tasks such as CLS underscores ChatGPT's evolving role in NLP, moving beyond content generation to more complex, multifaceted linguistic tasks. This advancement cements ChatGPT's standing in the NLP domain and enriches our understanding of its potential applications, highlighting its prowess in generating and analyzing language, adding a new dimension to our comprehension of its capabilities in NLP.

Another extension of ChatGPT's application in Natural Language Processing is the systematic review of literature searches offered by S. Wang et al. (2023). This study represents a shift in ChatGPT's role within NLP, showcasing its utility in a highly specialized and nuanced task of formulating Boolean queries. S. Wang et al. (2023) observed that "ChatGPT is capable of generating queries that lead to high search precision although trading-off this for recall," highlighting its precision in understanding and structuring complex query requirements. ChatGPT's functionality illuminates its potential as a valuable tool in academic and research-oriented settings, beyond general language tasks: "In terms of automatic query formulation, our results indicate that the use of ChatGPT compares favourably with the current state-of-the-art automated Boolean query generation methods in terms of precision at the expenses of a lower recall" (S. Wang et al., 2023). These findings underscore ChatGPT's advanced capabilities in parsing, interpreting, and formulating intricate search queries, a critical component of NLP. It shows that ChatGPT's application can be effectively extended to tasks requiring a high degree of accuracy and specificity in information retrieval, thus broadening the scope of its usage in NLP, and reinforcing its role as not just a language generator but as a sophisticated tool for specialized language processing tasks.

Various studies, notably the groundbreaking work in unsupervised neural machine translation by Han et al. (2021) show how generative models like ChatGPT are redefining the boundaries of what is achievable in NLP. Han et al. (2021) described a novel approach, stating, "We show how to derive state-of-the-art unsupervised neural machine translation systems from generatively pre-trained language models." The methodology employed in this research, such as few-shot amplification and back translation, highlights ChatGPT's advanced capabilities in handling intricate tasks like translation without direct supervision. This exploration underlines the model's remarkable adaptability and efficiency in a field traditionally dominated by supervised learning paradigms. As ChatGPT and similar models continue to evolve, they promise to enhance current NLP applications and open up new possibilities for handling complex linguistic tasks, thereby transforming our approach to language understanding and processing. This literature review thus encapsulates the significant strides made by ChatGPT in NLP, spotlighting its emerging role as a versatile, powerful tool capable of tackling an array of challenging language-related tasks.

2.6 Analyzing Emotion with Artificial Intelligence

In the evolving landscape of Artificial Intelligence (AI), analyzing emotions in textual data has become increasingly urgent, particularly in multilingual and culturally diverse digital communications. Lee and Wang (2015) offered a groundbreaking perspective in this area. They asserted, “Despite the important implications of code-switching for emotion analysis, existing automatic emotion extraction methods fail to accommodate for the code-switching content” (Lee & Wang, 2015). This statement underscores the central challenge addressed in their work: developing a robust framework capable of interpreting emotions in bilingual or code-switched social media content, explicitly focusing on Chinese-English texts. The paper is revolutionary in its approach, filling a gap in AI emotion analysis that has predominantly been oriented toward monolingual text processing. Lee and Wang’s methodology, encompassing “a multiple-classifier-based automatic detection approach to detect emotion in the code-switching corpus” (Lee & Wang, 2015), adeptly evaluated the effectiveness of both Chinese and English texts in capturing emotional nuances. This comprehensive approach not only enhances the capacity of AI in bilingual text processing but also underscores the necessity for AI systems to adapt to the complexities of linguistic diversity and cultural context for a more nuanced and empathetic understanding of human emotions.

Building on the theme of emotional analysis in multilingual contexts, Sonne and Erickson (2018) shifted the focus to image-centric social media, exploring how emotions are conveyed through text and visual elements. This study complements the findings of Lee and Wang (2015) by examining emotional expression in a different yet equally complex digital medium. Sonne and Erickson (2018) highlighted the nuanced relationship between textual and visual content in conveying emotions, noting, “Using open-ended coding of 651 Instagram posts, this study finds that women farmers use mostly neutral emotional tonality in their images and text content.” This finding challenges the conventional narrative of a positivity bias in social media, adding depth to our understanding of how emotions are presented and perceived in online environments. The authors further elaborated on the complexities of interpreting emotions in social media, stating, “Developing a means to disambiguate between authorial emotional expression and the experience of an emotion(s) on the part of an audience could help researchers, social media developers, and users alike better understand the social construction of emotions in social media environments” (Sonne & Erickson, 2018). This insight underscores the need for AI systems to recognize and classify emotions and discern the subtle interplay between author intent and audience perception, a critical aspect of any comprehensive AI-driven emotional analysis. Through this multimodal approach, Sonne and Erickson’s work extends the discussion of emotion analysis in AI, emphasizing the importance of context, medium, and the dynamic nature of emotional expression in digital spaces.

Building upon the foundation laid by studies like Sonne and Erickson (2018), Rao et al. (2023) presents a novel exploration of AI’s capabilities in a more linguistic domain. This research delved into using ChatGPT to analyze and interpret human emotions and personality traits through text. The study’s innovative approach presented “a generic evaluation framework for LLMs to assess human personalities based on Myers–Briggs Type Indicator (MBTI) tests” (Rao et al., 2023), which signifies a significant leap in AI’s ability to understand and categorize complex human psychological constructs. By employing ChatGPT in this unconventional application, the research offers a fresh perspective on how AI can move beyond mere sentiment analysis to more intricate personality assessments. This research highlights the capability of AI, particularly ChatGPT, to grasp the nuanced subtleties embedded in human language, as indicated by the observation that “ChatGPT is currently recognized as one of the most capable chatbots. It is able to perform context-aware conversations” (Rao et al., 2023). The paper expands the scope of AI’s application in emotional analysis to include sophisticated text-based personality profiling. This broader exploration contributes significantly to our understanding of AI’s role in deciphering the complex spectrum of human emotions and personalities, showcasing AI’s vast potential and evolving nature in the realm of psychological and emotional analysis.

Furthering the exploration into analyzing emotion through Artificial Intelligence (AI), Wei et al. (2023) extended the understanding of AI's potential in emotional analysis. This research ventured into new territory, differentiating itself from traditional emotion-centric AI studies by delving into how ChatGPT, within a zero-shot learning paradigm, can proficiently extract and contextualize emotional content from unstructured texts. The paper introduced a novel concept, stating, “We transform the zero-shot IE task into a multi-turn question-answering problem with a two-stage framework (ChatIE)” (Wei et al., 2023), which highlights ChatGPT’s advanced skill set. This skill set is not limited to mere recognition but extends to a detailed dissection and interpretation of emotional subtleties in textual data. The methodology employed in the study emphasized ChatGPT’s enhanced semantic understanding abilities, as evidenced by its impressive performance and higher consistency scores in diverse assessment scenarios. This capability is crucial in scenarios that demand an in-depth understanding of context, especially in parsing emotional nuances. This study represents a significant evolution in the emotional analysis of AI applications. It shifts that narrative from straightforward sentiment detection to a more dynamic, multifaceted process of extracting emotional information. This process involves interactive, multi-layered dialogues, signifying the growing complexity and contextual sensitivity of AI tools like ChatGPT. It showcase the progressive sophistication of AI in interpreting human emotions, a critical development for applications requiring nuanced emotional understanding and empathy in digital communication. Consequently, this study broadens the horizons of AI's applications in emotional analysis and sets a new benchmark for future AI research in understanding the intricate tapestry of human emotions in digital interactions.

Finally, Vijay et al. (2018) contributed a unique perspective by focusing on the complexities of emotion prediction within code-mixed language environments. Vijay et al. (2018) tackled the intricate task of discerning emotions in Hindi-English code-mixed social media texts, a notably complex and under-researched area. The authors elucidated their methodology, stating, "we analyze the problem of emotion identification in code-mixed content and present a Hindi-English code-mixed corpus extracted from Twitter and annotated with the associated emotion" (Vijay et al., 2018). This pioneering approach not only foregrounds the significance of understanding contextual and cultural subtleties in AI's analysis of emotions but also sheds light on the critical need for such technology in multilingual settings. When crafting a specialized corpus and integrating advanced machine learning techniques for emotion detection, this study significantly broadens the horizons of AI's interpretative abilities in the emotional domain. The research accounts for the rich tapestry of multilingual communication, highlighting the complexity and diversity of human emotional expression in digitally mediated interactions. Doing so extends AI's traditional scope, which predominantly focuses on monolingual text analysis, to embrace the nuanced linguistic variations found in code-mixed languages. Importantly, this research accentuates the imperative need for AI systems to be versatile and sophisticated enough to navigate the multifaceted linguistic landscapes of human communication. This need is particularly pressing in a globalized world where digital exchanges frequently span multiple languages and cultural contexts. Thus, the paper contributes significantly to the field of emotion analysis using AI and sets a precedent for future research to enhance AI's effectiveness in understanding and responding to emotional content in diverse language settings.

2.7 Sentiment Analysis Using ChatGPT and BERT

The intersection of ChatGPT and BERT in sentiment analysis represents a significant frontier in natural language processing (NLP). Both models, grounded in deep learning and transformer architectures, bring unique strengths to understanding and interpreting human emotions in text. This literature review section delves into the synergies between ChatGPT's generative capabilities and BERT's contextual understanding, highlighting advancements, comparative analyses, and the application of these models in extracting nuanced sentiment insights. The aim is to elucidate the evolving landscape of sentiment analysis, offering a comprehensive overview of current methodologies, challenges, and the potential for future research.

R et al. (2023) not only showcased a high degree of accuracy in sentiment analysis using BERT but also exemplified the model's capacity to handle the complexities of natural language understanding and emotion detection in digital communications. By achieving an accuracy of 96.49%, the study sets a new benchmark for sentiment analysis tasks, particularly conversations generated by AI technologies like ChatGPT. "The study compares the performance of ChatGPT against other models, including GPT-3.5, on various NLP tasks, with a focus on zero-shot learning capabilities… It emphasizes the effectiveness and limitations of ChatGPT, highlighting its strong performance on tasks that require reasoning capabilities, while also acknowledging challenges in specific areas such as sequence tagging" (R et al., 2023). This research marks a significant advancement in NLP by demonstrating how deep learning models, especially those built on transformer architecture, can effectively interpret and analyze human emotions conveyed through text. The comparative analysis provided against other machine and deep learning models underscores BERT's superior capability in capturing nuanced emotional contexts, which is critical for applications ranging from customer feedback analysis to monitoring social media sentiment. Such findings enrich the academic discourse on sentiment analysis and offer valuable insights for developers and researchers looking to leverage advanced NLP techniques for real-world applications.

The integration of ChatGPT and BERT in sentiment analysis illuminates the sophisticated capacity of AI to discern complex human emotions from text, as highlighted in Taherdoost and Madanchian (2023). They emphasized that "Artificial intelligence (AI) has improved the performance of multiple areas, particularly sentiment analysis. Using AI, sentiment analysis is the process of recognizing emotions expressed in text" (Taherdoost & Madanchian, 2023) and further elaborated on the categorization methodologies, stating, "Lexicon-based techniques and artificial intelligence (AI) or machine learning-based approaches may be used to categorize sentiment analysis methodologies" (Taherdoost & Madanchian, 2023). This accentuates the transformative impact of AI on understanding and analyzing sentiments, marking a pivotal advancement in the field. Their analysis suggests that the future of sentiment analysis lies in enhancing the precision of emotional detection through advanced AI models, underscoring the importance of AI in gaining insights into consumer sentiments and the broader implications for competitive research. This dual approach broadens the scope of sentiment analysis across various digital platforms and fine-tunes the accuracy of emotional inference, making it indispensable for applications ranging from customer service to social media monitoring. Such advancements underscore the evolving landscape of NLP, where integrating sophisticated AI models like ChatGPT and BERT paves the way for more profound and accurate interpretations of sentiment, contributing significantly to both the academic field and practical applications in industry settings.

Qin et al. (2023) provided valuable insights into the capabilities of ChatGPT and its comparative performance with BERT in sentiment analysis. The study highlighted that "ChatGPT performs well on many tasks favoring reasoning capabilities, while it still faces challenges when solving specific tasks such as sequence tagging" (Qin et al., 2023). This indicates the nuanced ability of ChatGPT to understand and interpret complex language tasks, which is further emphasized by its "superior reasoning capability" in tasks requiring logical deduction. Moreover, the comparison that "ChatGPT outperforms GPT-3.5 for natural language inference tasks and question answering tasks that favor reasoning capabilities" (Qin et al., 2023) showcases its advanced NLP proficiency. Despite these strengths, the acknowledgment that ChatGPT and GPT-3.5 "face challenges on certain tasks such as sequence tagging" (Qin et al., 2023) underscores the ongoing development needs in AI models. Notably, the finding that "ChatGPT's sentiment analysis ability is better than that of GPT-3.5" (Qin et al., 2023) positions ChatGPT as a potent tool for sentiment analysis, suggesting that when used alongside BERT, it could significantly enhance the accuracy and depth of sentiment analysis outcomes. This integrated approach could offer a more sophisticated understanding of text-based emotions, benefiting various applications from customer feedback analysis to social media monitoring.

In conclusion, the evolution of ChatGPT highlights significant strides in AI, evolving from basic text responses to complex, contextual-aware interactions, showcasing the continuous improvement in machine learning models. The exploration of natural language processing tasks with ChatGPT reveals its proficiency in understanding, generating, and interpreting human language, marking a pivotal shift in how machines process linguistic information. Additionally, the section on analyzing emotion with AI underscores the nuanced capabilities of ChatGPT in emotional intelligence, a critical aspect of semantic understanding, contrasting it with the capabilities of BERT. Lastly, the review of sentiment analysis utilizing ChatGPT and BERT demonstrates a significant advancement in natural language processing (NLP), illustrating the robust capabilities of these models in understanding and interpreting human emotions through text. This analysis not only highlights the complementary strengths of ChatGPT and BERT in enhancing sentiment analysis but also enables future research focused on refining AI's emotional intelligence and contextual awareness, thereby fostering the development of more nuanced and effective AI-driven sentiment analysis tools. This review delineates the current state of AI and NLP and sets the stage for future research by emphasizing the need for more nuanced, context-aware, and emotionally intelligent AI systems. The comparative analysis between ChatGPT and BERT across these dimensions sheds light on the evolutionary trajectory of language models, paving the way for more advanced and human-like AI interactions.

Chapter 3: Methodology

Informed by the extensive literature review, which highlighted the evolution of ChatGPT, its capabilities in natural language processing (NLP) tasks, and the emerging domain of emotional analysis in artificial intelligence, this thesis adopts a methodical approach to scrutinize and compare the semantic understanding of ChatGPT (Generative Pre-Trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers). This methodology's intent is to evaluate these AIs by precisely recognizing the significance of contextual awareness and emotional intelligence in semantic processing. To systematically compare these models, the research utilizes a robust dataset sourced from a combination of a third-party AI DeepAI (Baragona, 2023) and datasets specifically curated for emotional assessment in AI from the University of California, Irvine Machine Learning Repository (Kotzias, 2015). These datasets encompass diverse linguistic scenarios, from standard language understanding tasks to complex emotional and context-specific interactions. The processing of this test data involves rigorous computational techniques, including preprocessing for normalization and contextual alignment, followed by applying each model to the same set of linguistic inputs. The evaluation criteria are designed to measure the accuracy and fluency of responses and the depth of context comprehension and emotional responsiveness. The analysis of results employs statistical methods and qualitative assessments to derive insights into each model's semantic understanding capabilities. This dual-analysis approach ensures a comprehensive understanding of how each model processes and interprets language, thereby providing a clear comparison of their performance and shedding light on the nuances of advanced AI language models in the realm of semantic understanding.

3.1 Problem Statement

This thesis is anchored around a set of pivotal research questions that seek to unravel the capabilities of these advanced AI models in sentiment analysis and natural language processing (NLP). The primary problem statement evaluates whether ChatGPT can effectively analyze emotions from contextual entries through sentiment analysis. This inquiry is vital as it delves into ChatGPT's ability to understand the literal meaning of the text and interpret the underlying emotional tone, a critical aspect of human-like language understanding.

In order to add depth to the investigation, this thesis will compare the performance of ChatGPT in sentiment analysis with that of BERT, a well-established rival AI in the field of NLP. The comparison aims to determine how ChatGPT, with its unique language generation and processing capabilities, fares against BERT, primarily known for its deep understanding of contextual languages. This comparative analysis aims to identify the strengths and weaknesses of each model in interpreting and processing sentiment-laden language.

Finally, the thesis strives to answer a broader question: between ChatGPT and BERT, which AI demonstrates superior proficiency in performing sentiment analysis and NLP tasks as a whole? This question is integral, as it encompasses a holistic view of each model's capabilities, extending beyond sentiment analysis to include various aspects of NLP, such as context understanding, language generation, and response accuracy. By addressing this question, the thesis aims to contribute valuable insights into the evolving landscape of AI language models and their applicability in real-world scenarios.

This thesis employs a mixed-methods approach to address the various problem statements, combining quantitative and qualitative analyses. The quantitative analysis systematically tests ChatGPT and BERT against a carefully curated dataset, including sentiment-labeled sentences from the UCI Machine Learning Repository and contextual entries generated by DeepAI. The AI models' performance in accurately identifying and interpreting sentiments in these texts is quantitatively measured and compared. Metrics such as accuracy, precision, and recall are used to objectively assess the models' sentiment analysis capabilities.

Concurrently, a qualitative analysis is conducted to explore the nuances of each model's response to the test data. This involves a detailed examination of how each model interprets and processes the emotional content and context of the sentences. Special attention is paid to instances where the models exhibit strengths in understanding subtle emotional nuances or where they encounter challenges.

By integrating these methodological approaches, the thesis aims to comprehensively evaluate ChatGPT and BERT's abilities in sentiment analysis and overall NLP performance. This research contributes to the academic understanding of these AI models and has practical implications for the future development and application of AI in fields requiring nuanced language understanding.

3.2 Collection of Test Data

The collection of test data was a meticulous process that involved sourcing data from two distinct and reputable sources. This approach maintained objectivity and reliability in the comparative analysis of these advanced AI models.

The first source of data was derived from an independent AI platform, deepai.org (Baragona, 2023), specifically from their text-generation tool. The importance of using a third-party AI tool like DeepAI's text generator lies in its ability to produce diverse text samples spanning various styles, contexts, and complexities. These generated texts serve as a unique dataset that challenges the semantic understanding capabilities of both ChatGPT and BERT.

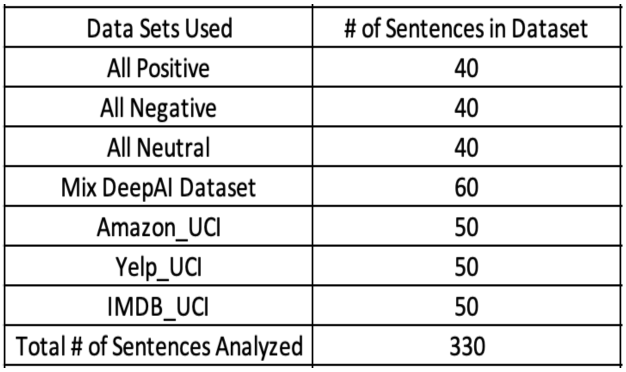
Using an external AI tool for data generation, the study avoids any potential bias that might arise from using internally generated datasets from the AI models being studied. This ensures that the evaluation of ChatGPT and BERT is conducted under fair and unbiased conditions, allowing for a genuine assessment of their capabilities in processing and understanding text generated by an independent AI system.

The second key source for test data was the Sentiment Labelled Sentences dataset from the UCI Machine Learning Repository. This dataset, created by Kotzias (2015), is a collection of sentences labeled with positive or negative sentiment, extracted from reviews of products, movies, and restaurants from websites like IMDB, Amazon, and Yelp. Each website contributed 500 positive and 500 negative sentences, randomly selected from larger datasets of reviews, with a precise aim of avoiding neutral sentences. The significance of incorporating this dataset lies in its real-world applicability and relevance to the everyday tasks these AI models are likely to encounter. Using this dataset, the thesis can effectively measure how well ChatGPT and BERT understand and process natural language with emotional nuances. The dataset's diverse sources – encompassing product, movie, and restaurant reviews – provide a broad spectrum of language use and sentiment, making it an ideal tool for testing the models' ability to interpret and respond to human emotions and opinions.

Together, these two sources provide a comprehensive and varied dataset for testing. The text generated by DeepAI offers a creative and unpredictable element, pushing the AI models to adapt to new and unseen text styles. In contrast, the Sentiment Labelled Sentences dataset from UCI brings a more structured and real-world aspect to the testing, focusing on the models' abilities to comprehend and respond to everyday language with emotional content. Combining these datasets enables a thorough and balanced evaluation of ChatGPT and BERT, shedding light on their strengths and weaknesses in different aspects of semantic understanding. This approach underscores the thesis's commitment to a rigorous and unbiased comparative analysis, leveraging independent research tools and data to achieve its objective.

For this thesis, a deliberate approach was taken to establish the sample size for the test data, balancing comprehensiveness with manageability. From the DeepAI (Baragona, 2023) text generator, a total of 120 statements with varying lengths and complexities were curated, comprising 40 positive, negative, and neutral statements using a prompt for the AI. An example of the prompt is:

DeepAI Prompt: "*Generate* 10 *short sentences with a positive tone*."



Data Sets Used	# of Sentences in Dataset
All Positive	40
All Negative	40
All Neutral	40
Mix DeepAI Dataset	60
Amazon_UCI	50
Yelp_UCI	50
IMDB_UCI	50
Total # of Sentences Analyzed	330

Table 1: Sentences Analyzed in Study

This selection was aimed at providing a diverse range of sentiments and contexts, challenging the language models with a spectrum of emotional and neutral language. Concurrently, the Sentiment Labelled Sentences dataset from the UCI Machine Learning Repository contributed 500 positive and negative sentences each from IMDb, Amazon, and Yelp. These datasets were specifically leveraged to provide real-world, sentiment-rich text samples. From each of these three sources, 25 positive and 25 negative statements were selected, amounting to 150 statements in total. This strategy of combining samples from different domains ensured a comprehensive and diverse dataset reflective of various styles and contexts of human language. The resulting dataset, comprising nearly 300 samples, was chosen to offer a robust and representative foundation for evaluating and comparing the semantic understanding capabilities of ChatGPT and BERT. This sample size was considered sufficiently large to provide meaningful insights while remaining manageable for detailed analysis.

3.3 Processing of Test Data

In order to process the test data found from DeepAI's text generator and the datasets from the UCI Machine Learning Repository, two different computational programs were used that formed the cornerstone for the subsequent quantitative and qualitative analyses. This initial step transforms the raw data into a format that is amenable to rigorous examination and interpretation. Utilizing specialized software and programming tools, the test data comprised various sentiments, and contextual entries were systematically processed. This involves normalization, tokenization, and other preprocessing techniques to ensure uniformity and comparability across the datasets used for testing ChatGPT and BERT. The processed data is then fed into each AI model under controlled conditions, allowing for the collection of comprehensive response data. This approach ensures that the ensuing analyses are grounded in a robust dataset, prepared meticulously to reflect natural language and sentiment nuances. The processed data sets the stage for a detailed exploration of each model's capabilities, enabling a thorough quantitative assessment of performance metrics such as accuracy, precision, and recall, as well as a qualitative evaluation of the models' proficiency in understanding and interpreting language and emotion. This methodical processing of test data is pivotal in ensuring that the analyses are accurate and deeply insightful, providing a clear picture of the comparative semantic understanding abilities of ChatGPT and BERT.

3.3.1 ChatGPT Computational Program

One of the most essential parts of this methodology involves an empirical assessment of ChatGPT's capabilities in sentiment analysis. To achieve this, a specific Python program was developed and integrated into our research framework. This program is designed to systematically evaluate the performance of the ChatGPT model in classifying sentiments as positive, negative, or neutral. The following detailed explanation of the program will elucidate its functionalities, the sequence of operations it performs, and the significance of each step in the context of our comparative analysis. This comprehensive understanding of the program is essential to appreciate how it contributes to our research findings and the conclusions drawn about the semantic understanding abilities of AI models. The ChatGPT analysis program is structured to assess the sentiment analysis capabilities of the model. Seen below is the assessment achieved through a series of well-defined steps, each of which plays a vital role in the overall evaluation process:

1. Library Importation and Setup:

   The code begins by importing essential libraries. The `'openai'` library is foundational for interfacing with the ChatGPT model. The `'sklearn.metrics'` library provides tools to calculate accuracy and F1 scores which are key metrics for evaluating model performance. The 'pandas' library is used for data handling and storage, facilitating the organization of test results.

2. OpenAI API Key Configuration:

   The script requires an OpenAI API key set at the beginning. This key is necessary to authenticate the user's access to OpenAI's services, including the ChatGPT model.

3. Data Preparation for Testing:

An important step is preparing the test data, consisting of a series of text samples and their corresponding true sentiment labels (positive, negative, or neutral). This dataset provides a basis for testing and evaluating the model's sentiment analysis capabilities.

4. Initializing Variables for Labels:

   The code initializes two lists: '`true_labels'` and '`predicted_labels`.' These lists are used to store the actual sentiment labels from the test data and the sentiment labels as predicted by ChatGPT, respectively.

5. Processing and Sentiment Analysis via ChatGPT:

   The script processes each text sample from the test data. For each sample, it utilizes ChatGPT to predict the sentiment label. This is done using the '`openai.Completion.create'` function, specifying the GPT model variant and providing a prompt for sentiment classification.

6. Label Extraction and Comparison:

   After receiving responses from ChatGPT, the script extracts the predicted sentiment labels and stores them. These predicted labels are then compared with the actual labels from the test data to assess the model’s accuracy.

7. Calculating Performance Metrics:

   The script computes performance metrics - accuracy and F1 score. The accuracy score provides a straightforward measure of the model’s overall correctness, while the F1 score offers a more nuanced view, considering both precision and recall, which is especially important in datasets with imbalanced classes.

8. Organizing and Storing Results:

Results are organized into a pandas DataFrame. This DataFrame includes the original text, true sentiment, and predicted labels, enabling a straightforward review and analysis of the model's performance.

9. Exporting Results for Documentation:

   The DataFrame is then exported to an Excel file. This step documents the results, allowing for further analysis, and providing a tangible output that can be referenced in the thesis.

10. Confirmation of Results Export:

    Finally, the script prints a message confirming the successful export of the results. This confirmation is a helpful indicator for the user to ensure that the process has been completed without errors.

The computational program for ChatGPT is used to evaluate the sentiment analysis capabilities of ChatGPT using the test data found from DeepAI (Baragona, 2023) and the UCI Machine Learning Repository (Kotzias, 2015). The program should systematically import any necessary libraries, set up API configurations, prepare and process test data, and accurately measure the performance of ChatGPT through objective metrics like accuracy and F1 score. The structured approach ensures a comprehensive and transparent assessment, from initializing data and label lists to exporting results into a well-organized Excel file. This process not only underscores the efficacy of ChatGPT in understanding and classifying sentiments but also provides a replicable and robust framework for comparing its performance with BERT. The insights gleaned from this analysis contribute significantly to the field of AI, offering a deeper understanding of the evolving capabilities of language models in semantic processing and sentiment analysis.

3.3.2 BERT Computational Program

Having thoroughly examined the Python code used to access the sentiment analysis capabilities of ChatGPT, we will now transition to the evaluation of BERT (Bidirectional Encoder Representations from Transformers) when performing a sentiment analysis. Similar to our approach with ChatGPT, a structured program was used to test BERT's ability to analyze and classify sentiment in a contextual setting. The two programs are similar in construction, so a fair and unbiased analysis between ChatGPT and BERT could be conducted. The figure below delves into the specifics of the code designed for the BERT analysis by highlighting the nuances and technicalities involved in leveraging this advanced language model for sentiment analysis.

The BERT analysis program is structured to assess BERT's sentiment analysis capabilities. The code involves several vital steps meticulously designed to test and assess BERT's performance on a given dataset. Seen below are the crucial steps taken to ensure the results are reliable and accurate:

1. Importing Libraries and Loading the Model:
    - Essential Python libraries and modules are imported, including `torch`, `transformers`, and `pandas`, which are fundamental for handling neural network operations, interfacing with the BERT model, and data manipulation, respectively.
    - The pre-trained BERT model (`bert-base-uncased`) and its tokenizer are loaded. This model is widely used for understanding general English text without case sensitivity. The tokenizer converts raw text into a format BERT can process.
2. Data Preparation:

- A sample dataset is created and then converted into a pandas DataFrame. This dataset comprises text samples and their corresponding sentiment labels.
- The dataset is then tokenized using BERT's tokenizer, which involves converting text into tokens and applying necessary padding and truncation.

3. Label Conversion and TensorDataset Creation:

   Sentiment labels are mapped to numerical values for processing by the model. A TensorDataset is created, which includes tokenized input IDs, attention masks, and labels. The attention mask informs the model which parts of the input are meaningful.

4. DataLoader Setup:

   A DataLoader is defined to handle data batching for the training and evaluation processes. This ensures efficient processing of data in manageable sizes.

5. Optimization and Training:

   - An optimizer (AdamW) and a loss function (CrossEntropyLoss) are set up for the training phase.
   - The model is fine-tuned over a specified number of epochs. In each epoch, the model learns by adjusting its parameters to reduce the loss. The total loss is calculated and averaged over all batches to monitor the training progress.

6. Model Evaluation:

   The model is switched to evaluation mode to test its performance on the dataset. The evaluation process involves feeding the data through the model and comparing the predicted labels against the actual labels.

7. Performance Metrics Calculation:

The accuracy and F1 scores are computed to quantitatively assess the model's performance. These metrics provide insight into how well the model can classify sentiments correctly.

8. Result Compilation and Export:
    - The results, including the texts, true labels, and predicted labels, are compiled into a pandas DataFrame for easy analysis and visualization.
    - The DataFrame is then exported to an Excel file for documentation and further reference.
9. Final Output:

    The code then prints the accuracy and F1-score and confirms the export of the results, which provides immediate insight into the model's performance and verifies that the results have been successfully saved for any further analysis.

The BERT computational program uses a rigorous and methodical approach to assessing the sentiment analysis capabilities of the BERT model. Through a series of well-defined and executed steps—from importing essential libraries and loading the model to processing and tokenizing the data, fine-tuning the model, and evaluating its performance on key metrics like accuracy and F1 score—the methodology ensures a thorough and empirical evaluation. This structured approach highlights the efficiency and effectiveness of BERT in sentiment analysis tasks and serves as a comparative point against ChatGPT. The results and insights from this analysis are invaluable, contributing significantly to the broader understanding of semantic processing capabilities in current AI language models. This detailed examination of BERT's performance underlines the importance of such comparative studies in advancing the field of natural language processing and artificial intelligence.

3.4 Analyzing the Processed Data

This section focuses on , the analysis of the process data obtained from the respective programs for both language models. This analysis is pivotal in deriving meaningful insight and understanding nuances of each model's performance in sentiment analysis tasks. To accomplish this, a suite of statistical metrics—accuracy, precision, recall, and F1-score—is employed. These metrics are chosen for their ability to comprehensively evaluate the models' performance, capturing both correctness and reliability aspects in their semantic understanding. Accuracy will offer a general view of the model's overall performance, while precision and recall will provide insight into the models' exactness and thoroughness, respectively. The F1-score, a harmonic mean of precision and recall, will be utilized to gauge the balance between these two metrics, offering a singular measure of the model's effectiveness in sentiment classification. This multi-metric analysis forms the backbone of the comparative study, enabling a robust and nuanced assessment of ChatGPT and BERT in the realm of natural language processing.

3.4.1 Accuracy

Accuracy is a fundamental metric in evaluating the performance of semantic analysis models, particularly in natural language processing (NLP) tasks such as sentiment analysis and text classification. The process of measuring accuracy involves several key steps, each contributing to a comprehensive evaluation of how effectively a model can classify and categorize textual data. This metric was chosen for its simplicity and interpretability since it is a good starting point when assessing the performance of NLP models. To correctly measure the accuracy, the following equation was used:

$$\text{Accuracy} = \frac{\text{Number of Correctly Classifed Instances}}{\text{Total Number of Instance}}$$

Equation 1: Accuracy

The theoretical aspect of calculating accuracy is essential for semantic analysis tasks, but it is instructive to speak about the practical application.

When calculating the accuracy through python the true and predicted labels need to be defined for a sentiment analysis task, which are essential for evaluating the model's performance. These labels are categorized into classes such as "positive," "negative," and "neutral." The program then initializes a pandas DataFrame to organize these labels for systematic analysis. For each class, it transforms the true and predicted labels into a binary format, assigning '1' if the label matches the current class and '0' otherwise. This binary conversion is crucial for accurately computing the class-wise accuracy. Utilizing the "accuracy_score" function from the "sklearn.metrics" library, the program calculates the accuracy for each class by comparing the binary true labels against the binary predicted labels.

Additionally, it computes the overall accuracy as the simple average of these class-wise accuracy scores. This overall score offers a holistic view of the model's performance across all sentiment categories. The accuracies are then appended to the DataFrame, providing a clear and organized presentation of the results. Finally, the program exports these results to an Excel spreadsheet, facilitating easy documentation and further analysis. This process exemplifies a practical approach to quantifying the accuracy of a semantic analysis model, providing a metric for assessing its performance in classifying sentiments. This program computes the accuracy metric for each sentiment class and provides an overall accuracy score, offering a tangible demonstration of how accuracy is operationalized in real-world data. Examining this code bridges the gap between theoretical understanding and practical application, facilitating insights into the nuances of accuracy calculation and its implications for evaluating the performance of natural language processing models.

3.4.2 F1 Score

In semantic analysis and natural language processing (NLP), the F1-score is a critical metric for evaluating the performance of models, classifiers, or systems, especially in tasks that involve binary classification or distinguishing between two distinct categories. The F1-score is a harmonic mean of precision and recall. Precision being the ratio of true positive predictions to the total number of positive predictions made (including both true positives and false positives). Recall is the ratio of true positive predictions to the actual number of positive instances within the data (including both true positives and false negative). The selection of the F1-score as an evaluation metric in NLP tasks is underpinned by its capability to equitably balance precision and recall, two pivotal components in assessing model performance. This balance is particularly crucial in NLP applications like sentiment analysis, entity recognition, and machine translation, where the implications of false positives and false negatives are significant. The F1-score's relevance is further amplified in scenarios involving imbalanced datasets, a common occurrence in NLP, where it provides a more accurate reflection of model performance on minority classes than metrics like accuracy. Its comprehensive consideration of all quadrants of the confusion matrix ensures a holistic view of the model's performance, unlike accuracy, which several factors can skew. The F1-score offers a balanced view of the model's effectiveness because it amalgamates precision and recall into a singular metric.

Precision:

- Precision is an essential metric to evaluate how accurately a model can identify texts (such as reviews, comments, or tweets) expressing a particular sentiment, typically positive, negative, or neutral. This metric becomes crucial in understanding the model's effectiveness in categorizing sentiments correctly without overgeneralizing or making too many errors. A high precision score suggests that when the model predicts a review to be positive, it is quite accurate in its prediction. This is particularly valuable in applications where it is critical to minimize the number of incorrect positive classifications, such as filtering out genuinely positive user feedback from a mix of reviews or comments.

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP}) + (\text{FP})}$$

Equation 2: Precision

- Consider a model classifying customer reviews into positive sentiments to break down how precision operates in sentiment analysis. In this scenario, precision is calculated by taking the total number of reviews that are correctly identified as positive (true positives (TP)) and dividing it by the total number of reviews the model classified as positive, which includes both correctly identified positive reviews and those that were incorrectly classified as positive also known as (false positives (FP)).

Recall:

- Recall, also known as the true positive rate or sensitivity, measures the model's ability to correctly identify all relevant instances of a specific sentiment within a dataset. Unlike precision, which focuses on the accuracy of positive predictions, recall assesses the model's capacity to capture all instances of a particular sentiment, ensuring no relevant information is missed. High recall is essential for ensuring that fewer positive or negative sentiments are missed when performing a sentiment analysis.

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP}) + (\text{FN})}$$

Equation 3: Recall

- Recall measures a model's effectiveness in identifying every instance of a particular class within a dataset, focusing specifically on positive instances. It is calculated by dividing the number of correct positive predictions (true positives (TP)) by the sum of all actual positive instances in the dataset, including the correctly identified positives and those positives the model failed to identify (false negatives (FN)). Essentially, recall provides insight into the model's thoroughness in capturing all relevant instances, highlighting its ability not to overlook or miss any positive cases.

Precision and recall are essential metrics in sentiment analysis for understanding how well a model performs in terms of its accuracy (precision) and completeness (recall) in identifying a specific sentiment. The distinction between precision and recall underlines the trade-offs between ensuring the accuracy of sentiment identification and the comprehensiveness of capturing relevant sentiments. Precision emphasizes the model's reliability in correctly identifying positive, negative, or neutral sentiments when it makes such a classification crucial for scenarios where the implications of incorrectly labeled sentiments are significant. On the other hand, recall highlights the importance of not missing any relevant sentiment expressions, which is essential for thorough sentiment monitoring and analysis. Balancing these metrics is critical, as the focus might shift based on specific project objectives, such as prioritizing the detection of all instances of negative customer feedback (emphasizing recall) or ensuring that only genuinely positive feedback is highlighted (prioritizing precision). The optimal balance depends on the application’s sensitivity to false positives versus false negatives, demonstrating the nuanced considerations in sentiment analysis projects.

Inside of the computational programs, a true label and predicted label system was used to analyze the data. The true label is the correct sentiment of the sentence all be positive, negative, or neutral. The predicted label is what ChatGPT and BERT thinks the sentiment should be based on the context of the sentence. Within the F1-Score program there were two functions imported from the ‘sklearn.metrics’ module in the scikit-learn library, a widely used Python library for machine learning named precision_score and recall_score. These functions are designed to evaluate the performance of models used for the analysis. Inside of the library it defines what precision and recall are based on the equations above and how to calculate them. The precision_score and recall_score reads the true label and predicted label and then performs the necessary calculations based on how precision and recall are defined in the Python library.

There is also a phenomenon where an analyzed dataset can have high precision alongside low recall, which stems from a model’s strategic prioritization. High precision indicates that when the model identifies an instance expressing a particular sentiment (positive, negative, or neutral), its identification is almost always accurate. This precision reflects the model’s effectiveness in accurately categorizing the sentiments it is confident about. However, the presence of low recall reveals a significant limitation: the model overlooks numerous instances that should also be classified under these sentiments, failing to capture a comprehensive snapshot of the dataset's emotional nuances. This particular phenomenon can be extremely complex especially when analyzing the sentiment of sentences with negative and neutral tones.

If the true label of a sentence is supposed to have a negative sentiment and the model correctly predicts it as negative, this scenario is categorized as a True Positive (TP). Despite the term “positive,” it simply means the prediction was correct (the positive outcome in terms of prediction accuracy). A False Negative (FN) occurs when the true label of a sentence is negative, but the model incorrectly predicts the sentiment as positive or neutral. It is considered “false” because the prediction was incorrect, and “negative” because the model failed to detect the presence of the condition being tested for (in this case, negative sentiment).

However, a False Positive (FP) would occur in the opposite scenario of a false negative. A false positive in sentiment analysis, when focusing on negative sentiments, happens when the model incorrectly predicts a text to have negative sentiment when its true label of the sentiment is actually positive or neutral. Imagine a review that says, “I could not be happier with the service!” If the model misinterprets this as negative, possibly due to keywords like “not,” without understanding the overall positive sentiment conveyed, the prediction is falsely identifying negative sentiment where there is none. This is a false positive because the model falsely flags the sentiment as being the one of interest (negative) when it is actually not.

This scenario unfolds due to the model’s conservative approach to classification. By prioritizing confidence over the detection range, the model opts for a safer route, which labels only those instances it is absolutely certain about. This cautiousness ensures the high precision of classified instances but also means that the model is hesitant to make judgments on instances where the sentiment is not as clear-cut. Consequently, while the model excels in minimizing errors in what it chooses to classify, it simultaneously misses out on a large portion of data that could provide a fuller understanding of the dataset’s overall sentiment landscape. It is important to look at the context of each sentence when performing a sentiment analysis to ensure that the data and the corresponding results can be interpreted correctly.

The F1-score is calculated using a harmonic mean of precision and recall, effectively combining these two metrics. The harmonic mean tends to favor the lower of the two numbers, which means that if the precision or recall is low, the F1-score will also be low. It is calculated with the following formula:

$$\mathrm{F1} - \mathrm{Score} = 2 \text{ x } \frac{\mathrm{Precision} \text{ x } \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Equation 4: F1-Score

This formula ensures that both precision and recall are given equal weight in the final score. The F1-score's unique property of being the harmonic mean gives more weight to lower values of either precision or recall. Therefore, the F1-score will be significantly affected if either metric is low. In semantic analysis tasks like sentiment analysis, the F1-score is a valuable measurement for accurately evaluating the model's ability to classify text into predetermined categories. It provides a more holistic view of the model's performance than using either precision or recall alone. By maximizing the F1-score, this thesis aims to achieve an optimal balance between making accurate positive predictions and correctly identifying all positive instances, which is essential for the effectiveness of NLP models in real-world applications.

This methodology section illustrates the integration of theoretical concepts and practical applications, which are then utilized in advanced machine learning models, ChatGPT and BERT, and subjected to rigorous testing using a well-structured dataset. The approach emphasizes the quantitative assessment of model performance through key metrics such as accuracy, precision, recall, and the F1-score, ensuring a comprehensive evaluation of each model's capabilities in sentiment analysis. The application of these metrics provides a nuanced understanding of the models' semantic processing strengths and limitations. This methodology not only facilitates a direct comparison between ChatGPT and BERT in terms of semantic understanding but also contributes valuable insights to the field of natural language processing. The results from this research endeavor aim to deepen the understanding of AI-driven language models and their evolving roles in interpreting and processing human language.

Chapter 4: Results and Analysis

The results section examines the empirical data obtained from the in-depth comparative analysis. This analysis encompasses a range of testing scenarios, including sentiment-specific contexts (positive, negative, and neutral) and mixed sentiment situations, as well as a specialized dataset focused on human-computer interaction (HCI). By employing robust statistical measures such as accuracy and F1-score, complemented by an evaluation of precision and recall, we aim to provide an extensive, multi-dimensional assessment of each model’s semantic understanding capabilities. The results presented herein are quantitatively rigorous and qualitatively insightful, offering a nuanced exploration of how ChatGPT and BERT perform under varying linguistic and contextual challenges. This section aims to uncover the intricate subtleties and operational distinctions between these advanced language models, thereby contributing to a deeper understanding of their respective strengths, limitations, and potential applications in the field of natural language processing.

During the testing phase, benchmark datasets were created encompassing all positive, negative, and neutral sentences for the comparative analysis of ChatGPT and BERT. The benchmark datasets were crafted from DeepAI and were used to discern if the models could distinguish and accurately classify sentiments in texts that may not adhere to typical human linguistic patterns. Having these benchmark datasets allows for a deeper understanding of each model’s capabilities and limitations in processing and interpreting complex, nuanced AI-generated language. This provides insights into their adaptability, robustness, and potential areas for enhancement in handling evolving linguistic trends. The benchmark datasets were then followed by an evaluation of the model’s ability to accurately interpret and classify a wide range of human emotions and nuances in language from websites such as Amazon, Yelp, and IMDB.

This approach addresses the complexity of real-world communication and the complexities of AI-generated text, where sentiments are not always clear-cut but often nuanced and context-dependent. Through rigorous testing of these models across all sentiment categories, this study can identify the strengths, weaknesses, and areas for improvement, leading to more refined and practical sentiment analysis tools. This benchmarking approach is instrumental in advancing NLP technologies towards greater accuracy and sophistication in sentiment analysis, ensuring they remain effective as the landscape of digital communication continues to evolve

4.1 Accuracy and F1-score Results

This section delves into an extensive analysis of the semantic understanding capabilities of ChatGPT and BERT. The performance of both models was evaluated across various scenarios, including sentiment-specific (all positive, all negative, all neutral) and mixed sentiment contexts, as well as specialized datasets from the UCI Machine Learning Repository. The primary metrics used for this comparative study are accuracy and F1-score, which include underlying precision and recall measures.

4.1.1 Accuracy Results

BERT's accuracy in different scenarios showcases its exceptional proficiency in semantic analysis across a spectrum of contexts. In controlled sentiment scenarios (positive, negative, and neutral), BERT achieved a perfect accuracy score of 100%. This indicates an extraordinary ability to correctly interpret and respond to clear, unambiguous sentiment cues. Such performance highlights the strength of BERT's bidirectional architecture, which allows it to contextualize words in a sentence more effectively than traditional, unidirectional models. This is particularly significant in scenarios where understanding the sentiment requires comprehending the context in which words are used.

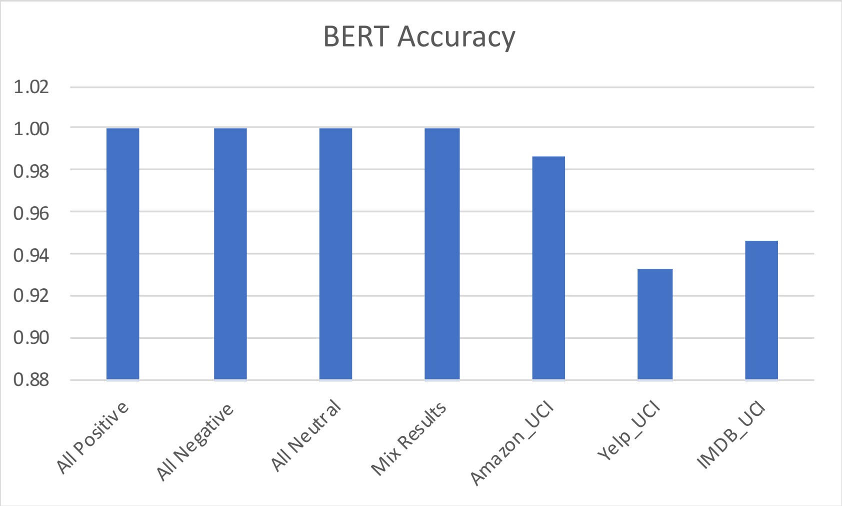
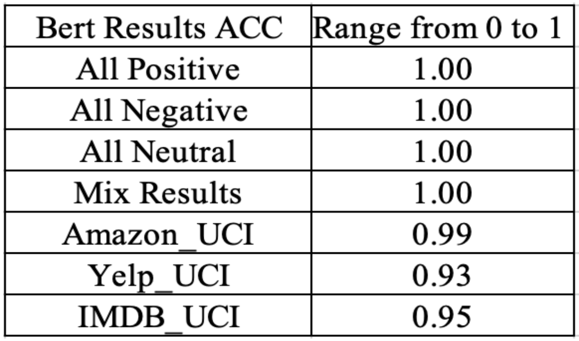


Figure 1: Bert Accuracy Bar

In mixed sentiment scenarios, maintaining this high accuracy is a notable achievement. It suggests that BERT is adept at handling complex inputs where multiple sentiments are interwoven, a common occurrence in real-world text. The ability to discern and accurately categorize such mixed sentiments is essential for applications like social media monitoring, where texts often contain varied emotions.



Bert Results ACC	Range from 0 to 1
All Positive	1.00
All Negative	1.00
All Neutral	1.00
Mix Results	1.00
Amazon_UCI	0.99
Yelp_UCI	0.93
IMDB_UCI	0.95

Table 2: BERT Accuracy Results

The Amazon_UCI dataset, more reflective of real-world complexities, showed a slight drop in accuracy (0.99%) for BERT. However, this score still represents a high level of competency in dealing with specialized, context-rich text. The minor decrease can be attributed to the inherent challenges in processing domain-specific language, which often includes nuanced expressions and technical terminology. BERT's strong performance in this dataset underlines its potential utility in nuanced fields such as customer service automation and interactive AI systems.

ChatGPT's performance, while showing strengths in certain areas, displayed more variability in accuracy across different scenarios. ChatGPT achieved its highest accuracy of 0.941667 in all positive sentiment scenarios, demonstrating a strong ability to generate and recognize positive sentiment. This indicates the model's training, which might have strongly emphasized producing coherent and contextually appropriate text, particularly effective in positive scenarios.

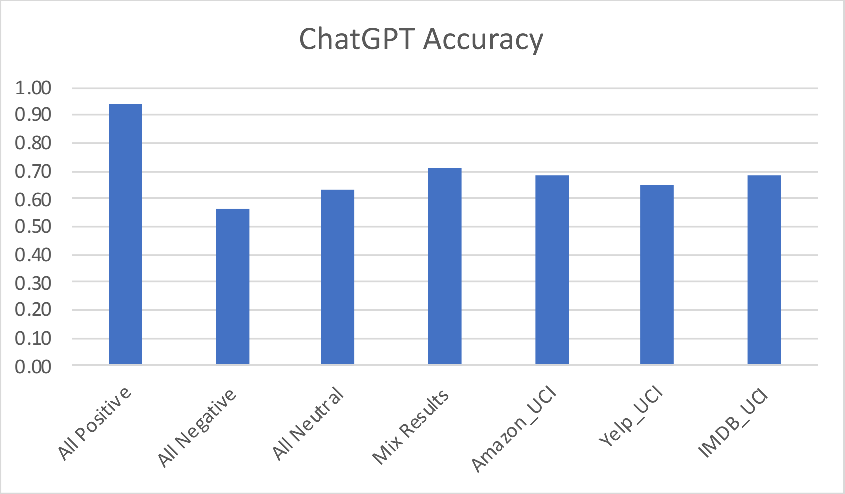
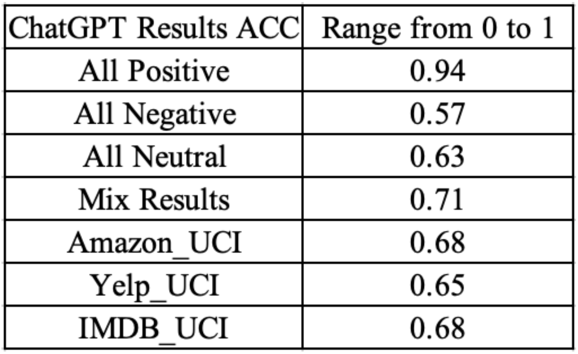


Figure 2: ChatGPT Accuracy Bar

However, ChatGPT's accuracy significantly dropped to 0.57 and 0.63 in all negative and all neutral scenarios, respectively. This variation highlights potential challenges in the model's ability to consistently interpret and generate less positive or more neutral text. Such a decline in performance suggests difficulties in handling texts lacking clear, affirmative emotional cues, which could be a limitation in applications requiring precise sentiment analysis across a broad emotional spectrum.



ChatGPT Results ACC	Range from 0 to 1
All Positive	0.94
All Negative	0.57
All Neutral	0.63
Mix Results	0.71
Amazon_UCI	0.68
Yelp_UCI	0.65
IMDB_UCI	0.68

Table 3: ChatGPT Accuracy Results

In mixed sentiment scenarios, ChatGPT achieved a moderate accuracy of 0.71. This indicates a reasonable capability to handle texts with varied emotions but also points to potential challenges in dealing with complex, sentiment-laden inputs. The model's architecture, while sophisticated, might be less adept than BERT's at navigating the intricacies of mixed sentiments.

Considering accuracy in the Amazon_UCI dataset, ChatGPT's was 0.68, lower than BERT's. This suggests that while ChatGPT can understand and respond to UCI-related text, it may struggle with the specialized language and nuanced expressions typical of such texts. This performance reflects the limitations of ChatGPT in contexts requiring a deep understanding of technical or domain-specific language.

The accuracy analysis between BERT and ChatGPT reveals distinct patterns in their semantic understanding capabilities. BERT's consistently high accuracy across various scenarios underscores its robustness and versatility in handling different types of text, especially those requiring a nuanced understanding of context and sentiment. On the other hand, ChatGPT, while showing strong capabilities in certain areas, exhibits variability in its performance, particularly in scenarios with mixed or negative sentiments and in processing specialized, domain-specific text. This comparative analysis provides valuable insights into the strengths and limitations of these models, guiding their potential applications and indicating directions for future improvement and development in the field of natural language processing.

4.1.2 F1-score Results

BERT's F1-score performance across various testing scenarios offers a comprehensive view of its semantic understanding capabilities. In scenarios with uniform sentiment (positive, negative, and neutral), BERT achieved a perfect F1-score of 1.00, indicating an exceptional balance between precision and recall. This suggests that BERT not only accurately identifies relevant sentiment cues but also maintains a high level of consistency in its responses, avoiding false positives and negatives. The model's bidirectional architecture likely contributes significantly here, allowing it to contextually analyze and interpret the sentiment in a holistic manner.

In mixed sentiment scenarios, maintaining this perfect F1-score is particularly impressive. It demonstrates BERT's ability to navigate complex, nuanced textual data where sentiments are intermingled. This capability is crucial for applications requiring nuanced sentiment analysis and contextual understanding, such as content moderation or customer feedback analysis. These results for the precision, recall, and F1-score can be seen in the chart below:

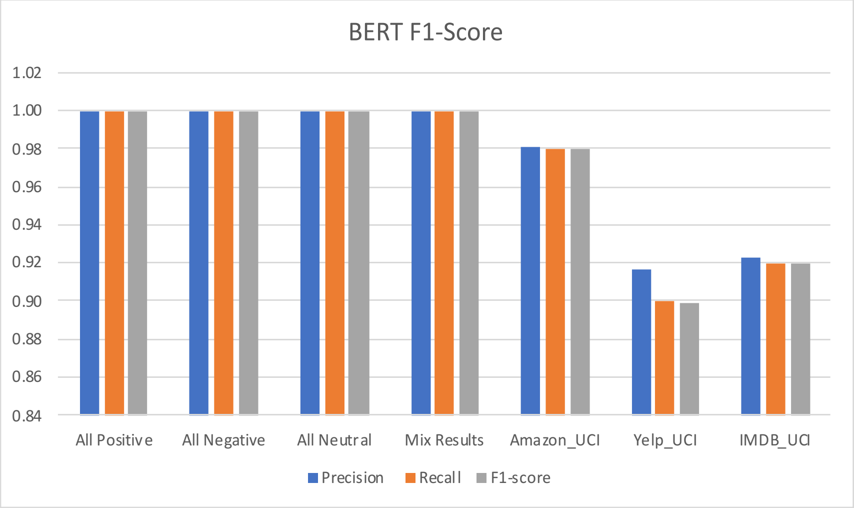
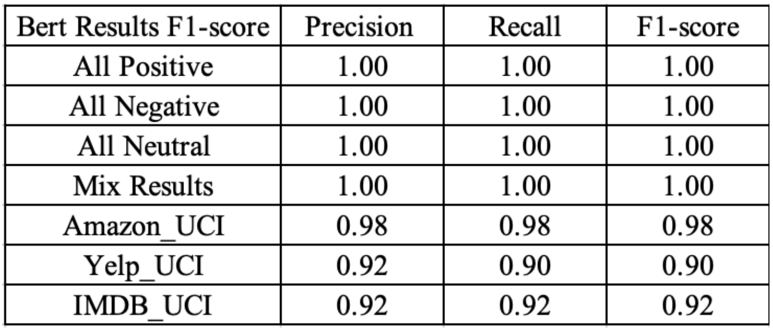


Figure 3: Bert F1-Score Bar

The performance in the Amazon_UCI dataset, with an F1-score of 0.98, slightly lower than in controlled sentiment scenarios, still reflects BERT's strong ability to handle specialized, context-rich text. This slight decrease might be attributed to UCI language's inherent complexity and variability, which often includes jargon, technical terms, or less standard language use. BERT's high precision (0.98) in this dataset indicates its effectiveness in accurately identifying relevant UCI terms and concepts, which is essential in domains like AI-driven customer support and user experience research.



Bert Results F1-score	Precision	Recall	F1-score
All Positive	1.00	1.00	1.00
All Negative	1.00	1.00	1.00
All Neutral	1.00	1.00	1.00
Mix Results	1.00	1.00	1.00
Amazon_UCI	0.98	0.98	0.98
Yelp_UCI	0.92	0.90	0.90
IMDB_UCI	0.92	0.92	0.92

Table 4: Bert F1-Score Results

ChatGPT's performance, as measured by F1-scores, shows a distinct pattern, highlighting its strengths and areas for improvement. ChatGPT achieved a high F1-score of 0.93 in all positive sentiment scenarios, signifying strong capabilities in generating and recognizing positive sentiment. This performance reflects the model's training emphasis on producing coherent, contextually appropriate text, particularly effective in positive scenarios.

However, in scenarios with all negative and neutral sentiments, the F1-scores dropped significantly to 0.46 and 0.41, respectively. This indicates a challenge for ChatGPT in handling texts with less explicit or more subdued emotional content. The lower F1-scores suggest difficulties in maintaining a balance between precision and recall, possibly generating either overly generic responses or missing the nuanced sentiment cues in such texts.

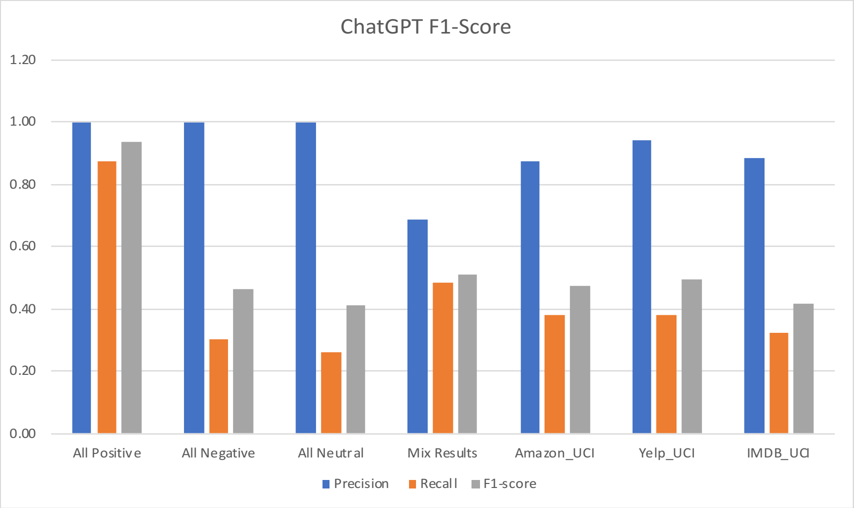


Figure 4: ChatGPT F1-Score Bar

In mixed sentiment scenarios, the F1-score of 0.50 for ChatGPT points to moderate effectiveness in dealing with complex emotional content. While reasonably capable, this suggests that ChatGPT might struggle with accurately interpreting and responding to texts with multiple sentiments, a common characteristic of natural, conversational language.

In the Amazon_UCI dataset, the F1-score of 0.47, with a precision of 0.88, indicates that while ChatGPT can understand and respond to UCI-related text, it may not capture such specialized language's full complexity and nuance. This is an important consideration for applications requiring deep domain knowledge and precise language understanding.

ChatGPT Results F1-score	Precision	Recall	F1-score
All Positive	1.00	0.88	0.93
All Negative	1.00	0.30	0.46
All Neutral	1.00	0.26	0.41
Mix Results	0.69	0.48	0.51
Amazon_UCI	0.88	0.38	0.47
Yelp_UCI	0.94	0.38	0.50
IMDB_UCI	0.88	0.32	0.42

Table 5: ChatGPT F1-Score Results

The performance discrepancy where ChatGPT exhibits high precision, but low recall can be attributed to its foundational architecture and training objectives. Precision in this context refers to the proportion of positive identifications that were actually correct, whereas recall measures the proportion of actual positives that were identified correctly. ChatGPT, trained on vast amounts of text data, is adept at generating coherent, contextually appropriate language responses. Its architecture, however, is not inherently designed for the granular detection and classification tasks that sentiment analysis often demands. This leads to scenarios where ChatGPT can accurately identify clear, explicit expressions of sentiment (thus high precision) but may overlook more nuanced or less explicitly stated sentiments (resulting in low recall).

ChatGPT's behavior is contrasted with models specifically fine-tuned for sentiment analysis, like BERT, which are optimized to understand and classify textual nuances more comprehensively. The F1-score analysis reveals significant differences in the performance of BERT and ChatGPT across various semantic contexts. BERT demonstrates a consistently high level of precision and recall, indicating robust capabilities in understanding and responding accurately to a wide range of text types. In contrast, ChatGPT, while effective in specific contexts, shows variability in its performance, particularly in more complex or less emotionally explicit scenarios. This analysis underscores the importance of considering both precision and recall in evaluating the effectiveness of language models, as it provides a more holistic view of their semantic understanding capabilities. The insights from this comparison are invaluable for guiding the application and further development of these models in various natural language processing tasks.

4.2 Answering the Problem Statement

The results indicate that ChatGPT has a strong capability in analyzing and generating positive sentiment, as reflected in its high accuracy and F1-score in all positive sentiment scenarios. However, its performance drops significantly in all negative or neutral scenarios. This suggests that while ChatGPT effectively identifies and responds to positive emotions, it faces challenges in accurately interpreting and processing negative or neutral emotions from contextual entries. The variability in its performance highlights a potential improvement in sentiment analysis capabilities, especially in texts lacking clear, affirmative emotional cues.

Compared to BERT, ChatGPT demonstrates a different pattern of strengths and weaknesses. BERT shows consistently high accuracy and F1-scores across various sentiment scenarios, including mixed sentiments and more complex datasets like the Amazon_UCI dataset. This suggests that BERT is superior in understanding and processing a wide range of sentiments, maintaining high performance even in texts with nuanced or conflicting emotions. In contrast, ChatGPT's performance, while commendable in certain areas, is less consistent, particularly in handling negative, neutral, or mixed sentiments.

The comparative analysis clearly outlines the strengths and weaknesses of each model. BERT's strength lies in its robust and consistent performance across different types of sentiment-laden language, indicating a deep understanding of context. ChatGPT, on the other hand, shows strength in generating coherent and contextually appropriate responses, particularly in positive sentiment scenarios. However, it exhibits weaknesses in contexts with less explicit or more complex emotional content.

In terms of sentiment analysis, BERT demonstrates superior proficiency over ChatGPT, as evidenced by its higher accuracy and F1-scores in a broader range of sentiment analysis tasks. BERT's ability to perform highly in complex and mixed sentiment scenarios indicates a more advanced capability in understanding and processing sentiment-laden language. However, it is essential to note that the overall proficiency in NLP tasks depends on the specific requirements of the task. While BERT excels in sentiment analysis and understanding context, ChatGPT's language generation capabilities may be more suited for tasks that require coherent and creative text generation.

In summary, this thesis provides a thorough comparative analysis of ChatGPT and BERT in sentiment analysis within natural language processing. The examination of their performances across various sentiment scenarios has revealed insightful distinctions. ChatGPT, adept in positive sentiment contexts and coherent text generation, demonstrates significant capabilities, yet it exhibits limitations in accurately processing negative, neutral, or mixed sentiments. In contrast, BERT emerges as a more robust model, consistently showing superior proficiency across a diverse range of sentiment analyses, including the ability to handle complex, nuanced language in mixed sentiment scenarios and specialized datasets like Amazon_UCI. This comparative study not only underscores each model's distinct strengths and weaknesses but also illuminates the broader landscape of NLP, highlighting the intricate challenges and potential areas for development in sentiment analysis. The findings from this analysis are pivotal, contributing valuable insights to the field of NLP and laying a foundation for future advancements in human-AI interaction and language understanding technologies.

Chapter 5: Conclusion

The conclusion synthesizes the findings from our in-depth analysis and provides insights into the capabilities of these advanced AI models in the domain of natural language processing (NLP), particularly in sentiment analysis.

BERT has demonstrated exceptional proficiency across various sentiment scenarios, achieving near-perfect accuracy and F1-scores in controlled and mixed sentiment contexts and in the specialized Amazon_UCI dataset. These results indicate BERT’s robust bidirectional architecture, which allows it to contextualize words and interpret complex, nuanced textual data effectively. BERT’s consistent performance highlights its deep understanding of context in language, making it highly effective in identifying nuanced sentiments and processing specialized, context-rich text, essential for applications in AI-driven customer support and user experience research.

In contrast, ChatGPT, while showing strong capabilities in generating coherent and contextually appropriate responses in positive sentiment scenarios, exhibits variability in handling negative, neutral, and mixed sentiments. The model’s training, focused on producing coherent text, is particularly effective in positive scenarios but less so in more complex emotional contexts. This suggests that while ChatGPT is adept at language generation, it may face challenges in consistently interpreting and processing less explicit or more complex emotional content. This is particularly evident in the Amazon_UCI dataset, where ChatGPT's performance was notably lower than BERT’s, highlighting its limitations in contexts requiring a deep understanding of technical or domain-specific language (Chapter 5).

The comparative analysis between BERT and ChatGPT underscores the importance of model architecture and training methodologies in shaping their semantic understanding capabilities. BERT’s bidirectional training and focus on contextual understanding give it an edge in accurately processing a wide range of sentiment-laden language. Meanwhile, ChatGPT’s generative nature, while effective in certain contexts, shows limitations in more complex sentiment analyses.

This thesis contributes valuable insights into the evolving landscape of AI language models and their applicability in real-world scenarios. It underscores the need for continuous development and optimization in AI models to achieve more nuanced, context-aware, and emotionally intelligent systems. The findings from this study pave the way for future advancements in semantic understanding, driving the evolution of more advanced, human-like AI interactions in the realm of natural language processing. The results delineate the current capabilities of ChatGPT and BERT and set the stage for future research, emphasizing the necessity for more refined, contextual-aware, and emotionally perceptive AI systems for effective human-AI interaction.

5.1 Comparison of Other Studies Focusing on Sentiment Analysis

This thesis distinctively explores the semantic understanding capabilities of ChatGPT and BERT, offering a comprehensive side-by-side evaluation that goes beyond traditional sentiment analysis. This research stands out by assessing both models through metrics like accuracy and F1-score across datasets enriched with a broad spectrum of sentiments. It reveals BERT’s adeptness at processing a wide array of sentiments due to its deep contextual understanding, whereas ChatGPT excels in generating coherent responses, particularly in scenarios involving positive sentiments.

In contrast, R et al. (2023) focused on a more specific application of BERT to understand public sentiment towards ChatGPT, as expressed through social media tweets. This approach is novel in its use of BERT for sentiment analysis and topic modeling of raw tweets, showcasing its effectiveness in capturing public opinions and identifying prevalent themes related to ChatGPT.

The distinction between the thesis and other research lies in the thesis's broader analytical scope and comparative nature. While R et al. (2023) demonstrated BERT's utility in social media sentiment analysis, this thesis provides a more nuanced examination of how ChatGPT and BERT perform across different sentiment analysis tasks. It compares their quantitative performance and discusses their qualitative strengths and limitations in understanding and generating language. This dual analysis emphasizes the complementary capabilities of both models, suggesting a potential for future AI advancements in creating more nuanced, context-aware, and emotionally intelligent systems. Through this comparison, the thesis contributes significantly to the field of NLP by highlighting the evolving capabilities of AI in interpreting and generating human-like responses, guiding future research toward enhancing semantic understanding in AI models.

5.2 Limitations and Future Works

The limitations of this thesis primarily stem from the inherent constraints associated with the comparative analysis of two sophisticated yet distinct AI language models, ChatGPT and BERT. Firstly, the evaluation is confined to the versions of ChatGPT and BERT available at the time of this study, which means that the findings may not fully encapsulate the ongoing developments and improvements made to these models. Additionally, while providing valuable insights, the focus on sentiment analysis limits the scope of the study and does not address the full spectrum of linguistic capabilities and applications of these models. While comprehensive, the datasets used, including the Amazon_UCI dataset, may not cover all nuances and variations of natural language, potentially affecting the generalizability of the results. Furthermore, the analysis relies heavily on quantitative metrics such as accuracy and F1-scores, which, though informative, may not fully capture the qualitative aspects of language understanding and generation. This quantitative focus might overlook some subtleties of semantic interpretation and contextual nuances that qualitative analysis could reveal. Finally, the study's conclusions are drawn based on AI's current state in NLP, a rapidly evolving field. The findings should be considered a snapshot within a dynamic and continuously advancing domain, necessitating ongoing research and evaluation to keep pace with technological advancements.

This thesis opens several avenues for future research in the field of natural language processing and AI-driven semantic analysis. Future work could expand on the comparative study by including more diverse and updated versions of language models like ChatGPT and BERT as these models continuously evolve. Investigating these models' performance on a broader range of linguistic tasks beyond sentiment analysis, such as intent recognition, sarcasm detection, and nuanced context understanding, could provide a more holistic view of their capabilities. Additionally, incorporating more diverse and complex datasets, including multilingual and cross-cultural content, could enhance understanding of these models' performance in varied linguistic contexts. Qualitative analyses, including user experience studies and interpretability assessments, could be conducted to complement the quantitative approach, providing deeper insights into how these models are perceived and interacted with by humans. Furthermore, it would be valuable to explore integrating these models into real-world applications, such as chatbots, virtual assistants, and content moderation tools, and assess their practical effectiveness and limitations. Finally, research into the ethical implications, biases, and fairness of these AI models in different applications could further enrich the discourse on responsible AI development and deployment in the realm of natural language processing.

Chapter 6: References

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023, April 14). *Chatgpt: Applications, opportunities, and threats*. arXiv.org. https://doi.org/10.48550/arXiv.2304.09103

Baragona, K. (2023). *Ai text generator - deepai*. DeepAI. https://deepai.org/chat/text-generator

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C.,...Amodei, D. (2020, May 28). *Language models are few-shot learners*. arXiv.org. https://arxiv.org/abs/2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Han, J. M., Babuschkin, I., Edwards, H., Neelakantan, A., Xu, T., Polu, S., Ray, A., Shyam, P., Ramesh, A., Radford, A., & Sutskever, I. (2021, October 11). *Unsupervised neural machine translation with generative language models only*. arXiv.org. https://arxiv.org/abs/2110.05448

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023, January 20). *Is chatgpt a good translator? yes with gpt-4 as the engine*. arXiv.org. https://arxiv.org/abs/2301.08745

Khan, J. Y., & Uddin, G. (2022, September 6). *Automatic code documentation generation using gpt-3*. arXiv.org. https://arxiv.org/abs/2209.02235

Koroteev, M. V. (2021, March 22). *Bert: A review of applications in natural language processing and understanding*. arXiv.org. https://doi.org/10.48550/arXiv.2103.11943

Kotzias, D. (2015). *Uci machine learning repository*. https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences

Lauer, D. J., IV. (2024, February 9). *Research_Code for masters thesis*. https://github.com/DonLauer19/Research_Code.git.

Lee, S., & Wang, Z. (2015, July 1). *Emotion in code-switching texts: Corpus construction and analysis*. ACL Anthology. https://aclanthology.org/W15-3116/

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

OpenAI. (2023). *Chatgpt: Optimized language model for dialogue*. https://openai.com/chatgpt

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023, February 8). *Is chatgpt a general-purpose natural language processing task solver?* arXiv.org. https://doi.org/10.48550/arXiv.2302.06476

R, S., Mujahid, M., Rustam, F., Shafique, R., Chunduri, V., Villar, M., Ballester, J., Diez, I., & Ashraf, I. (2023). Analyzing sentiments regarding chatgpt using novel bert: A machine learning approach. *Information*, *14*(9), 474. https://doi.org/10.3390/info14090474

Rao, H., Leung, C., & Miao, C. (2023, March 1). *Can chatgpt assess human personalities? a general evaluation framework*. arXiv.org. https://arxiv.org/abs/2303.01248

Ray, P. P. (2023). Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020, February 27). *A primer in bertology: What we know about how bert works*. arXiv.org. https://doi.org/10.48550/arXiv.2002.12327

Setianto, F., Tsani, E., Sadiq, F., Domalis, G., Tsakalidis, D., & Kostakos, P. (2022, January 19). *GPT-2C: a parser for honeypot logs using large pre-trained language models*. ACM Digital Library. https://doi.org/10.1145/3487351.3492723

Sonne, J., & Erickson, I. (2018, July 18). *The expression of emotions on instagram*. ACM Digital Library. https://doi.org/10.1145/3217804.3217949

Taherdoost, H., & Madanchian, M. (2023). Artificial intelligence and sentiment analysis: A review in competitive research. *Computers*, *12*(2), 37. https://doi.org/10.3390/computers12020037

Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June 1). *Corpus creation and emotion prediction for hindi-english code-mixed social media text*. ACL Anthology. https://aclanthology.org/N18-4018/

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023, March 17). *Is chatgpt a good nlg evaluator? a preliminary study*. arXiv.org. https://arxiv.org/abs/2303.04048

Wang, J., Liang, Y., Meng, F., Zou, B., Li, Z., Qu, J., & Zhou, J. (2023, February 28). *Zero-shot cross-lingual summarization via large language models*. arXiv.org. https://doi.org/10.48550/arXiv.2302.14229

Wang, L., Mujib, M. I., Williams, J., Demiris, G., & Huh-Yoo, J. (2021, July 28). *An evaluation of generative pre-training model-based therapy chatbot for caregivers*. arXiv.org. https://arxiv.org/abs/2107.13115

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Wang, S., Scells, H., Koopman, B., & Zuccon, G. (2023, February 3). *Can chatgpt write a good boolean query for systematic review literature search?* arXiv.org. https://arxiv.org/abs/2302.03495

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023, February 20). *Zero-shot information extraction via chatting with chatgpt*. arXiv.org. https://arxiv.org/abs/2302.10205

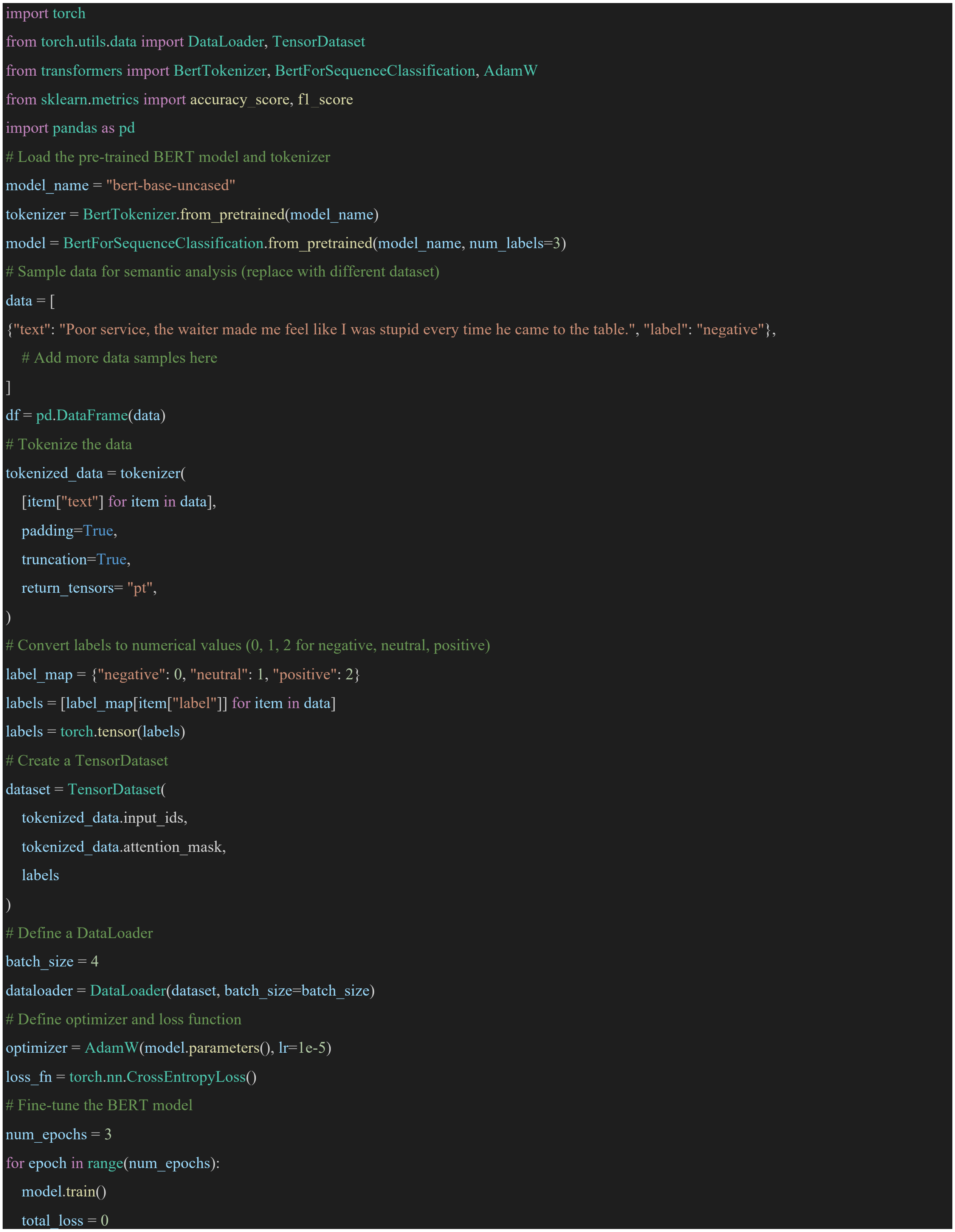
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. V., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S.,...Rush, A. (2020, October 1). *Transformers: State-of-the-art natural language processing*. ACL Anthology. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023, February 19). *Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert*. arXiv.org. https://doi.org/10.48550/arXiv.2302.10198

Appendix A: ChatGPT Analysis Program

```
import openai
from sklearn.metrics import accuracy_score, f1_score
import pandas as pd
# Set your OpenAI API key
api_key = "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"
openai.api_key = api_key
# Sample data for semantic analysis (replace with different data)
data = [
{"text": "Poor service, the waiter made me feel like I was stupid every time he came to the table.", "label": "negative"},
# Add more data samples here
]
# Initialize lists to store true labels and predicted labels
true_labels = []
predicted_labels = []
# Loop through the data and use ChatGPT to generate labels
#davinci-similarity
for item in data:
input_text = item["text"]
response = openai.Completion.create(
engine= "davinci-similarity",
prompt=f "Label the sentiment of the following text: '{input_text}' as positive, negative, or neutral.",
max_tokens=1,
)
# Extract the predicted label from the response
predicted_label = response.choices[0].text.strip() if response.choices else "unknown"
# Append true and predicted labels to the lists
true_labels.append(item["label"])
predicted_labels.append(predicted_label)
# Calculate accuracy and F1-score
accuracy = accuracy_score(true_labels, predicted_labels)
f1 = f1_score(true_labels, predicted_labels, average="weighted")
# Create a DataFrame to store the results
results = {
"Text": [item["text"] for item in data],
"True Label": true_labels,
"Predicted Label": predicted_labels,
}
results_df = pd.DataFrame(results)
# Specify the output file path
output_file = "GPT_YelpResults.xlsx"
# Save the results to an Excel spreadsheet
results_df.to_excel(output_file, index=False, engine="openpyxl")
print(f"Results exported to '{output_file}'")
```

Appendix B: Bert Analysis Program



```
import torch
from torch.utils.data import DataLoader, TensorDataset
from transformers import BertTokenizer, BertForSequenceClassification, AdamW
from sklearn.metrics import accuracy_score, f1_score
import pandas as pd
# Load the pre-trained BERT model and tokenizer
model_name = "bert-base-uncased"
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertForSequenceClassification.from_pretrained(model_name, num_labels=3)
# Sample data for semantic analysis (replace with different dataset)
data = [
{"text": "Poor service, the waiter made me feel like I was stupid every time he came to the table.", "label": "negative"},
    # Add more data samples here
]
df = pd.DataFrame(data)
# Tokenize the data
tokenized_data = tokenizer(
    [item["text"] for item in data],
    padding=True,
    truncation=True,
    return_tensors= "pt",
)
# Convert labels to numerical values (0, 1, 2 for negative, neutral, positive)
label_map = {"negative": 0, "neutral": 1, "positive": 2}
labels = [label_map[item["label"]] for item in data]
labels = torch.tensor(labels)
# Create a TensorDataset
dataset = TensorDataset(
    tokenized_data.input_ids,
    tokenized_data.attention_mask,
    labels
)
# Define a DataLoader
batch_size = 4
dataloader = DataLoader(dataset, batch_size=batch_size)
# Define optimizer and loss function
optimizer = AdamW(model.parameters(), lr=1e-5)
loss_fn = torch.nn.CrossEntropyLoss()
# Fine-tune the BERT model
num_epochs = 3
for epoch in range(num_epochs):
    model.train()
    total_loss = 0
    for batch in dataloader:
        input_ids, attention_mask, target_labels = batch
        optimizer.zero_grad()
        outputs = model(input_ids, attention_mask=attention_mask, labels=target_labels)
        loss = outputs.loss
        loss.backward()
        optimizer.step()
        total_loss += loss.item()
    avg_loss = total_loss / len(dataloader)
    print(f"Epoch {epoch + 1}/{num_epochs}, Loss: {avg_loss:.4f}")
# Evaluation
model.eval()
predicted_labels = []
true_labels = []
for batch in dataloader:
    input_ids, attention_mask, target_labels = batch
    with torch.no_grad():
        outputs = model(input_ids, attention_mask=attention_mask)
    logits = outputs.logits
    predicted_label = torch.argmax(logits, dim=1).tolist()
    predicted_labels.extend(predicted_label)
    true_labels.extend(target_labels.tolist())
# Calculate accuracy and F1-score
accuracy = accuracy_score(true_labels, predicted_labels)
f1 = f1_score(true_labels, predicted_labels, average="weighted")

print("BERT True Labels:", true_labels)
print("BERT Predictions:", predicted_labels)
print("BERT Accuracy:", accuracy)
print("BERT F1-Score:", f1)
# Create a DataFrame to store the results
results = {
    "Text": [item["text"] for item in data],
    "True Label": [label_map[item["label"]] for item in data],
    "Predicted Label": predicted_labels,
}
results_df = pd.DataFrame(results)
# Specify the output file path
output_file = "BERT_YelpResults.xlsx"
# Save the results to an Excel spreadsheet
results_df.to_excel(output_file, index=False, engine="openpyxl")
print(f"Results exported to '{output_file}'")
```

Appendix C: Accuracy Analysis Program

```
import pandas as pd
from sklearn.metrics import accuracy_score
# Define true labels and predicted labels for each class
true_labels = ["negative", "positive", "negative",]
predicted_labels = ["negative", "positive", "negative",]
# Define class names
classes = ["positive", "negative", "neutral"]
# Initialize a DataFrame to store results
results_df = pd.DataFrame({"True Label": true_labels, "Predicted Label": predicted_labels})
# Calculate accuracy for each class and store in a dictionary
accuracy_scores = {}
for current_class in classes:
    true_labels_class = [1 if label == current_class else 0 for label in true_labels]
    predicted_labels_class = [1 if label == current_class else 0 for label in predicted_labels]
    accuracy = accuracy_score(true_labels_class, predicted_labels_class)
    accuracy_scores[current_class] = accuracy
# Calculate overall accuracy (simple average)
overall_accuracy = sum(accuracy_scores.values()) / len(accuracy_scores)
# Add overall accuracy to the results DataFrame
results_df = results_df.append({"True Label": "Overall", "Predicted Label": "Overall"},
ignore_index=True)
results_df.loc[results_df["True Label"] == "Overall", "Accuracy"] = overall_accuracy
# Specify the output file path
output_file = "BERT_Ylp_Acc.xlsx"
# Save the results to an Excel spreadsheet
results_df.to_excel(output_file, index=False, engine="openpyxl")
print(f"Results exported to '{output_file}'")
```

Appendix D: F1-Score Analysis Program

```
import pandas as pd
from sklearn.metrics import precision_score, recall_score, f1_score
# Define true labels and predicted labels for each class
true_labels = ["negative", "positive", "negative",]
predicted_labels = ["negative", "positive", "negative",]
# Define class names
# I can take away to neutral class when it is time to analyze the results of the archive site
classes = ["positive", "negative", "neutral"]
# Initialize dictionaries to store precision, recall, and F1-score for each class
class_metrics = {}
# Calculate precision, recall, and F1-score for each class and store in dictionaries
for current_class in classes:
    precision = precision_score(true_labels, predicted_labels, labels=[current_class], average="weighted")
    recall = recall_score(true_labels, predicted_labels, labels=[current_class], average="weighted")
    f1 = f1_score(true_labels, predicted_labels, labels=[current_class], average="weighted")
    class_metrics[current_class] = {"Precision": precision, "Recall": recall, "F1-Score": f1}
# Calculate overall metrics (simple average)
overall_precision = sum(metric["Precision"] for metric in class_metrics.values()) / len(class_metrics)
overall_recall = sum(metric["Recall"] for metric in class_metrics.values()) / len(class_metrics)
overall_f1_score = sum(metric["F1-Score"] for metric in class_metrics.values()) / len(class_metrics)
# Add overall metrics to the class_metrics dictionary
class_metrics["Overall"] = {"Precision": overall_precision, "Recall": overall_recall, "F1-Score":
overall_f1_score}
# Convert class_metrics to a DataFrame for export
results_df = pd.DataFrame(class_metrics).T
# Specify the output file path
output_file = "BERT_Yelp_F1.xlsx"
# Save the results to an Excel spreadsheet
results_df.to_excel(output_file, engine="openpyxl")
print(f"Results exported to '{output_file}'")
```