CP3106 Independent Project

# Automation of Social Work BPSS Assessment Using a Hybrid BERT and GPT Model

By

Xianqing Zeng

Department of Computer Science

School of Computing

National University of Singapore

2024

CP3106 Independent Project

# Automation of Social Work BPSS Assessment Using a Hybrid BERT and GPT Model

By

Xianqing Zeng

Department of Computer Science

School of Computing

National University of Singapore

2024

Advisor: Assoc Prof Yi-Chieh Lee

Deliverables:

Report: 1 Volume
Program: 1 Diskette
Database: 1 Diskette

**Abstract**

In contemporary society, the rapid advancement of artificial intelligence technologies has positioned Natural Language Processing (NLP) as a pivotal bridge between human language and machine intelligence. The field of social work, in particular, faces a growing demand for comprehensive assessments of individuals' biological, psychological, social, and spiritual (BPSS) aspects. The quest for more efficient and precise assessment tools is urgent. This study explores the application of NLP technology, especially the Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) models, in automating BPSS assessments in social work. By analyzing the strengths of BERT in Natural Language Understanding (NLU) and the exceptional capabilities of GPT in Natural Language Generation (NLG), this study introduces a novel automated assessment method aimed at enhancing the efficiency and accuracy of individual assessments in social work. Experimental validation revealed that using the BERT model for text classification and organization, combined with the GPT model to generate coherent and detailed summaries, effectively accomplishes precise individual profiling and BPSS modeling. This hybrid approach enhances the efficiency and accuracy of social workers in conducting BPSS assessments, providing an innovative tool for rapid response to individual needs. The experimental results, based on authentic dialogue texts from social workers, demonstrate the potential application value of this hybrid method in the practical automation of assessments in social work.

Subject Descriptors:

| | |
|---|---|
| I.2.7 | Natural Language Processing |
| J.4 | Social and Behavioral Sciences |
| K.4.2 | Social Issues |

Keywords:

Artificial Intelligence, Natural Language Processing, Social Work, BPSS Assessment

Implementation Software and Hardware:

Microsoft Windows 11 Home Chinese Edition, 10.0.22631 Build 22631
AMD Ryzen 7 6800H with Radeon Graphics, 8 Cores, 16 Logical Processors

Python 3.10.11

# List of Figures

# List of Tables

# Table of Contents

# Chapter 1

# Introduction

In the context of the rapid development of modern information technology, Natural Language Processing (NLP) has emerged as one of the core research directions in the field of artificial intelligence. The broad application of NLP encompasses tasks such as machine translation(Montejo-Ráez & Jiménez-Zafra, 2022), language modeling(Montejo-Ráez & Jiménez-Zafra, 2022), text generation(Zhang, Song, Li, Zhou, & Song, 2023), sentiment analysis(Zhang et al., 2023), natural language understanding(Montejo-Ráez & Jiménez-Zafra, 2022), and question-answering systems(Sultana & Badugu, 2020). Its capabilities extend across multiple domains, demonstrating exceptional prowess. Specifically, in the field of social work, NLP has shown significant potential by assisting social workers in understanding and predicting individual needs and emotional states through text data analysis(Montejo-Ráez & Jiménez-Zafra, 2022). Additionally, generative artificial intelligence models and large language models are considered to have the potential to support safe and ethical decision-making in social work(Victor, Kubiak, Angell, & Perron, 2023), with systems like ChatGPT demonstrating above-average performance in assessments within this domain(Markovič, Daniel, 2024). Moreover, generative dialogue systems, such as chatbots, are increasingly being utilized to provide routine consultation and support, capable of simulating social worker interactions and offering preliminary mental health

support and crisis intervention(Montejo-Ráez & Jiménez-Zafra, 2022).

However, the application of this technology faces challenges, including the potential for information lag for social workers. The widespread use of chatbots may replace direct communication between social workers and their clients, impacting the timeliness with which social workers can obtain current information about their clients. Additionally, there is a paucity of research on individual assessments, which limits its effectiveness in social work applications. Particularly in the comprehensive assessment and intervention of individual, family, and community welfare, it is crucial for social workers to employ a comprehensive analysis using the biological, psychological, social, and spiritual (BPSS) assessment method. Traditional BPSS assessment methods are often time-consuming and subject to evaluator bias, reducing efficiency and potentially affecting the accuracy of results. Therefore, there is an urgent need to develop an efficient and accurate automated assessment tool. Such a tool can generate comprehensive analyses, significantly enhancing the accuracy and efficiency of chatbots in conversations, while also helping social workers to timely document and follow up on client situations, thereby optimizing social work practice.

The Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) represent two milestones in NLP technology in recent years, exhibiting outstanding performance in natural language understanding (NLU) and natural language generation (NLG) respectively. Addressing the aforementioned needs, this study introduces for the first time a hybrid assessment method based on BERT and GPT models, specifically optimized for the BPSS assessment scenario in social work. This method combines the efficient text classification capabilities of BERT with the text generation capabilities of GPT, aiming to significantly enhance assessment efficiency and accuracy through automation, thus better serving social work practice. Through this innovative approach, social workers and chatbots can conduct comprehensive and in-depth

assessments of individuals in a shorter time, providing strong support and scientific basis for practical social interventions.

## 1.1 Research Questions and Objectives

This study aims to explore the potential application of BERT and GPT models in conducting BPSS assessments in the field of social work and to achieve automated and precise individual assessments through the construction of a hybrid model. The research primarily focuses on the following questions:

1. How can the strengths of BERT and GPT models be effectively combined to construct a hybrid model suitable for BPSS assessments in social work?
2. What are the practical effects of this hybrid model in social work practice? Can it enhance the efficiency and accuracy of assessments?
3. What challenges and limitations are encountered in the practical application process, and how can they be overcome?

Based on these questions, the objectives of this study are:

1. To analyze and validate the application potential of BERT and GPT models in BPSS assessments within social work;
2. To design and implement an application framework for BPSS assessment using a hybrid of BERT and GPT models;
3. To evaluate the effectiveness of this hybrid model in BPSS assessments through experimental research, providing technical support and theoretical basis for social work practice.

The research methodology will combine literature review, theoretical analysis, and experimental validation. An in-depth analysis of relevant literature will be conducted, and based on the unique characteristics of the BERT and GPT models, an experimental design will be formulated. The model will be tested and validated using authentic dialogue texts from social work.

# Chapter 2

# Background

## 2.1 Overview of Natural Language Processing Technology

Natural Language Processing (NLP), as a computer technology that simulates human language understanding and generation, has become an important branch within the field of artificial intelligence. The development of NLP has not only promoted the widespread adoption of applications such as machine translation and sentiment analysis, but it has also demonstrated significant potential within the field of social work. As illustrated in Figure 2.1(Khurana, Koli, Khatter, & Singh, 2023), Natural Language Understanding (NLU) and Natural Language Generation (NLG) are the two main pillars of NLP, focusing respectively on the computer's ability to understand and generate human language.

In the field of social work, the application of NLU is primarily manifested in understanding communications with service recipients. For example, by analyzing consultation records or social media posts of service recipients, NLU can assist social workers in quickly identifying individual emotional states, needs, and problems (khurana2023natural). Additionally, NLU can aid in document classification and information extraction, thereby

enhancing the efficiency and quality of social work services. For instance, by automatically categorizing and summarizing a large volume of case records, social workers can devote more time and energy to direct service provision.

The application of Natural Language Generation (NLG) technology in the field of social work has demonstrated its potential to enhance communication efficiency and manage operations. NLG technology can be integrated into chatbots to provide immediate and natural responses, which is particularly useful for managing large volumes of cases and offering support outside of regular working hours. This not only increases client engagement but also allows social workers to focus their time and energy on more complex cases. Additionally, social workers often spend considerable time on case documentation and report generation. NLG tools can automate these processes by synthesizing narratives from structured data, significantly reducing the administrative burden on social workers and freeing up more time for direct client interaction. The application of these technologies not only improves work efficiency but also optimizes resource allocation, making social services more efficient and personalized.

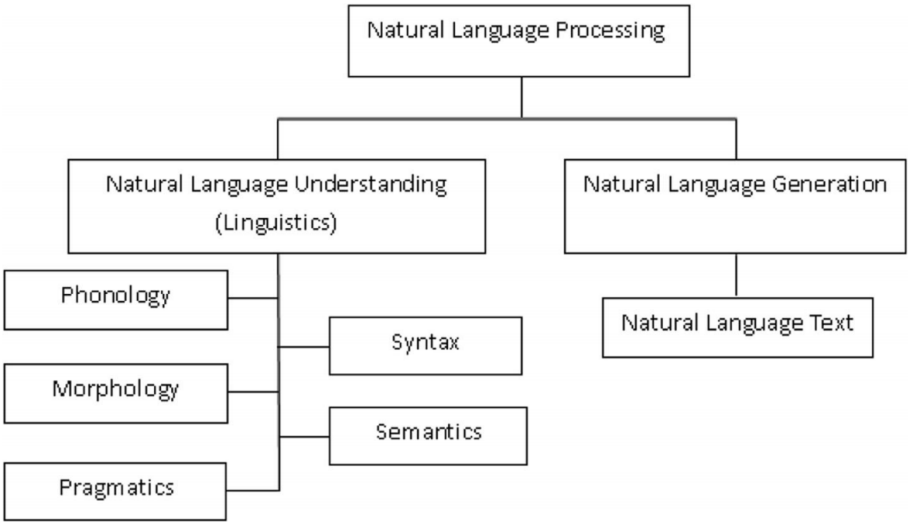



Figure 2.1: General Classification of NLP

## 2.2 BERT and ChatGPT Model

### 2.2.1 Overview of BERT Model

The Bidirectional Encoder Representations from Transformers (BERT) model, developed by Google's AI team, is a pre-trained language representation model(Devlin, Chang, Lee, & Toutanova, 2018). The core advantage of the BERT model lies in its bidirectional training mechanism, which allows it to consider the context before and after each word during the pre-training phase, thereby more accurately capturing the meaning of words in specific contexts. The BERT model is based on the Transformer architecture, which includes several layers of Transformer encoders. Each encoder consists of two parts: a multi-head self-attention mechanism and a feed-forward neural network. This structure enables BERT to capture complex relationships between words through large-scale parallel computations while processing text.

As illustrated in Fig 2.2(Devlin et al., 2018), during the pre-training process, BERT utilizes two strategies: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM enhances bidirectional training by randomly masking some words and training the model to predict these words, while NSP trains the model to predict whether two sentences appear coherently together, aiding in understanding the relationships between sentences. Once pre-trained, this model can be fine-tuned for various downstream natural language processing tasks, such as text classification, question-answering systems, and named entity recognition. BERT has demonstrated outstanding performance in these tasks. Through deep bidirectional understanding and extensive pre-training, BERT not only improves performance across various NLP tasks but also advances the field of natural language processing as a whole.

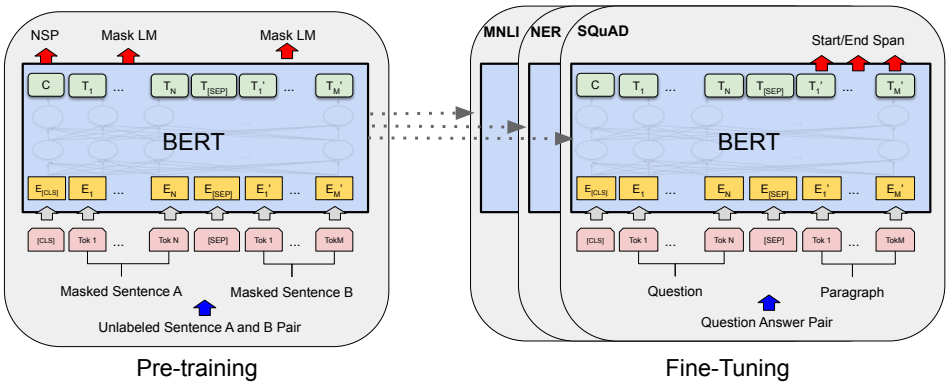



Figure 2.2: Overall pre-training and fine-tuning procedures for BERT

### 2.2.2

Overview of ChatGPT Model

ChatGPT, developed by OpenAI based on the GPT-3 model, is an advanced dialogue generation system that demonstrates significant improvements over previous language models in producing coherent, relevant, and diverse text. The training data for ChatGPT encompasses a wide range of topics and domains, enabling it to understand and generate texts across various styles and themes, thereby excelling in natural language generation (NLG) tasks.

A key feature of ChatGPT is its powerful language generation capabilities and flexible adaptability. In dialogue generation, ChatGPT can produce smooth and logical responses based on contextual prompts, making the conversation appear more natural and human-like. Additionally, ChatGPT maintains consistency and relevance across multiple rounds of conversation, which is crucial for building effective human-machine dialogue systems. In text generation, ChatGPT not only crafts stories, articles, and other creative texts but also assists in automatically generating technical documents, reports, and summaries.

Technically, ChatGPT utilizes a Transformer-based architecture that processes and generates text data through a self-attention mechanism, ideally suited for capturing complex relationships within the input data, thereby enhancing the quality and relevance

of the generated text(Liu, Han, Ma, Zhang, Yang, Tian, He, Li, He, Liu, & others, 2023). Furthermore, during the model training phase, ChatGPT integrates reinforcement learning techniques, allowing the model not only to learn language patterns during the pre-training phase but also to improve itself based on feedback, further enhancing the naturalness and adaptability of the dialogue. This strategy enables ChatGPT to perform better in multi-turn dialogues, more accurately meeting user needs(Liu et al., 2023).

Currently, ChatGPT has shown significant potential and practical value in various application fields. In chatting and interaction, it provides a highly human-like conversational experience, flexibly responding to user prompts to meet diverse dialogue needs. In the education sector, ChatGPT offers customized assistance by simulating different dialogue scenarios, greatly enriching teaching resources and methodologies. Additionally, ChatGPT is widely used in content creation, capable of automatically generating articles, creative texts, and literature summaries, particularly excelling in fields like news, marketing, and academic research. Combined with voice interaction technologies, such as OpenAI's Whisper system, ChatGPT can also perform voice input and recognition, further expanding its range of applications(Radford, Kim, Xu, Brockman, McLeavey, & Sutskever, 2023). By providing APIs and developer tools, ChatGPT enables developers to integrate its technology into various scenarios, such as customer service and online education platforms, demonstrating its wide adaptability and technical maturity.

### 2.2.3 Applications and Advantages of BERT and ChatGPT in NLP

In the field of Natural Language Processing (NLP), BERT and ChatGPT have demonstrated significant strengths in Natural Language Understanding (NLU) and Natural Language Generation (NLG) respectively. BERT leverages its deep bidirectional representation capability through the pre-training stages of the Masked Language Model

(MLM) and Next Sentence Prediction (NSP) to learn contextual information, which makes it highly effective in understanding tasks such as sentiment analysis, language acceptability judgments, and natural language inference(Devlin et al., 2018). This profound bidirectional understanding, combined with the flexibility of fine-tuning, enables BERT to adapt effectively to a variety of NLU tasks(Zhong, Ding, Liu, Du, & Tao, 2023).

On the other hand, ChatGPT has showcased its formidable capabilities in natural language generation. As a model based on GPT-3, ChatGPT can not only generate smooth, coherent, and highly relevant text but also excels in complex generation and dialogue tasks. Although GPT models are traditionally considered weak in understanding tasks, ChatGPT has actually outperformed all BERT models on some reasoning tasks, such as the natural language inference portion of the GLUE benchmark, demonstrating its advantages in reasoning and logical processing(Zhong et al., 2023). Furthermore, with the adoption of advanced prompting strategies, such as chain of thought (CoT), ChatGPT's understanding capabilities have significantly improved, enabling it to surpass even the powerful RoBERTa model in certain NLU tasks(Zhong et al., 2023; Wei, Wang, Schuurmans, Bosma, ichter, Xia, Chi, Le, & Zhou, 2022).

Overall, BERT's structure is particularly suited for NLU tasks that require fine-grained understanding, making it excel in single-sentence level text classification. This precise text classification ability provides a solid foundation for assessing an individual's biological, psychological, social, and spiritual states. Meanwhile, ChatGPT complements BERT's analysis through its outstanding ability in generating language and handling complex dialogues, helping to generate detailed and coherent individual assessment reports, thereby enhancing the overall efficiency and accuracy of assessments. In the design of this study, leveraging the complementary strengths of both models not only showcases the potential of large language models in handling complex linguistic tasks but also provides an innovative automated assessment tool for social work. With appropriate model

training and strategy adjustments, this hybrid approach offers broad prospects for application, helping to respond to and address individual needs more precisely.

## 2.3 BPSS Assessment Model

The Biological-Psychological-Social-Spiritual (BPSS) model is a comprehensive assessment framework extensively used in the fields of health and social sciences, particularly crucial in social work practice. This model integrates four dimensions: biological, psychological, social, and spiritual, acknowledging the interplay between physical, psychological, social, and spiritual aspects with patient care and welfare (Galbadage, Peterson, Wang, Wang, & Gunasekera, 2020). Thus, the model emphasizes the holistic consideration of an individual's multifaceted needs and resources during assessment and intervention processes to fully understand the state of the service recipients.

Within the context of social work, the BPSS model aids social workers in identifying the health issues, psychological states, social relationships, and spiritual beliefs of service recipients, thus enabling the formulation of more precise and personalized intervention plans. Traditional BPSS assessment methods often include the use of genograms to analyze family history and relationship patterns, as well as eco-maps to identify the interactions between an individual and their social environment (Cheung, Chin, Chua, Das, Fan, Mardiana, & Yong, 2023). These tools not only help social workers gather critical information but also facilitate a deeper understanding of the service recipients' living environments, providing more effective support and services. Through the application of the BPSS model, social work services can be more closely tailored to the actual needs of individuals, enhancing the quality and outcomes of the services provided.

As shown in Fig 2.3(Cheung et al., 2023), the BPSS model encompasses the following aspects:

1. **Biological**: This dimension focuses on an individual's basic physiological needs

(such as food, clothing, housing, and transportation), health issues (such as chronic diseases or disabilities), health status, and health-related behaviors (such as smoking and drinking). It also considers the individual's medical needs, including whether they have acute or chronic illnesses and their adherence to treatments.

2. **Psychological**: Covers mental health status, emotional control, thought processes, and cognitive functions. It also addresses emotional expression and regulation, mental and spiritual health conditions, such as self-identity, self-esteem, and coping mechanisms.

3. **Social**: Concentrates on an individual's social network, family relationships, employment status, social support systems, and cultural background. This includes examining family functionality, community interactions, economic status, and career aspects.

4. **Spiritual**: Explores an individual's beliefs, life purposes, moral values, and religious activities. This part also includes how individuals seek meaning, support, and coping strategies through spiritual activities.

In conducting BPSS assessments, social workers typically use various tools and methods. These include genograms to understand genetic and behavioral patterns by depicting relationships among family members; ecomaps to assess the impact of social relationships by illustrating the individual's interactions with their social environment; and detailed information about the individual gathered through interviews and surveys to comprehensively assess their biological, psychological, social, and spiritual health(Cheung et al., 2023).

The process of social work assessment generally involves three stages: preparation, implementation, and summary. In the preparation stage, social workers need to clarify the purpose, subject, and methods of the assessment and develop a detailed assessment plan

based on these. The implementation stage mainly focuses on collecting information and data, involving in-depth analysis of family relationships and social connections, as well as meticulous assessments of the individual's biological and psychological conditions. The final summary stage involves integrating all collected data and information, identifying the specific needs and available resources of the service recipients, and based on this, formulating appropriate intervention strategies.

Through the application of the BPSS model, social work services can not only gain a deeper understanding of the complex needs of service recipients but also enhance the personalization and effectiveness of services, ensuring that each intervention is as closely aligned with the individual's actual situation as possible. This multidimensional assessment approach makes social work practice more precise and effective, greatly enhancing social workers' capacity to handle diverse cases.

A biopsychosocial-spiritual model is a holistic approach that acknowledges the interaction between physical, psychological, social, and spiritual aspects to patient care and patient well-being(Galbadage et al., 2020).

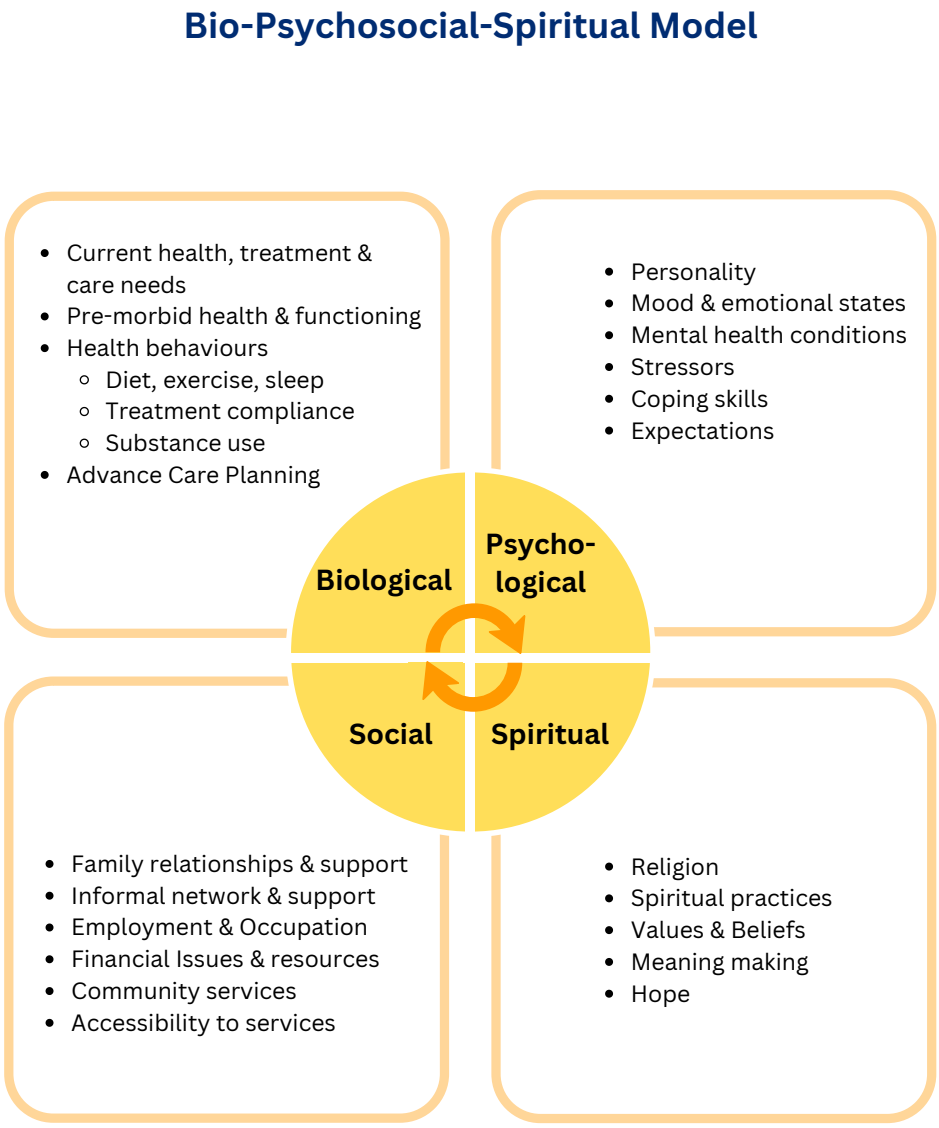



Figure 2.3: Four Dimensions of BPSS

# Chapter 3

# Research Methods and Design

## 3.1 BPSS Overview Generation Framework Combining BERT and ChatGPT

This study has developed an innovative text framework that leverages BERT's strong text classification capabilities and ChatGPT's advanced text generation technology. The aim of this framework is to automate the production of scenario overviews based on the Biological-Psychological-Social-Spiritual (BPSS) model to support professional practices in psychological counseling and social work. The mechanism for generating BPSS scenario overviews involves several key steps: data preprocessing, model fine-tuning, text rewriting, generation of classification results, and the construction of scenario overviews, as shown in the Figure 3.1.

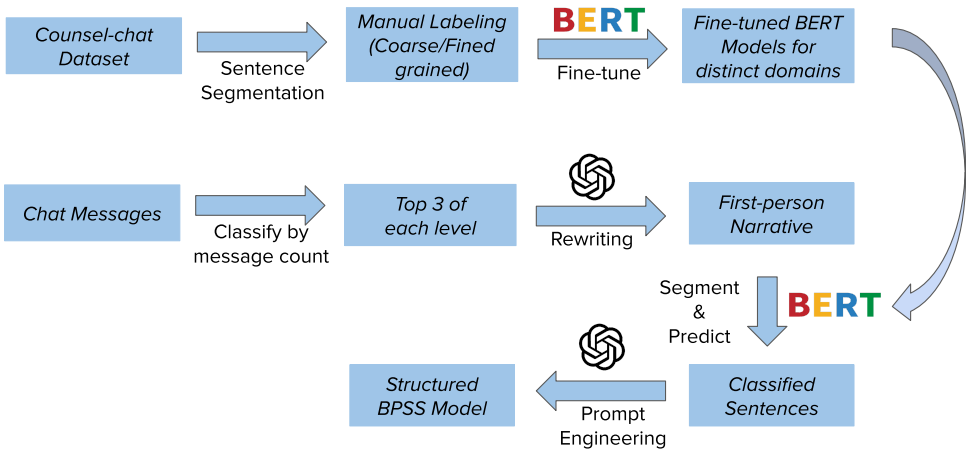



Figure 3.1: Automated BPSS assessment process based on BERT and GPT

## 3.2 Implementation Mechanism of the BPSS Scenario Generation Method

In the methodology of this study, initially, text preprocessing tasks are performed on the Counsel-chat dataset, which includes sentence breaking and dividing into coarse and fine-grained manual annotations to ensure the quality and consistency of the dataset, laying the foundation for model training. Further, we utilize the BERT model to capture and classify complex contextual information in the text, with each sentence finely categorized into the corresponding categories of the BPSS model. This model, through deep bidirectional learning, extracts text features to effectively identify content related to biological, psychological, social, and spiritual aspects of BPSS.

Upon fine-tuning BERT, the user dialogue text is rewritten in the first person by ChatGPT to enhance textual coherence and reduce redundant information. This step not only maintains consistency with the original dataset format but also ensures the integrity of the information and the coherence of the theme through the refining process. The rewritten text is then classified by the fine-tuned BERT model, and the output classification information is integrated to serve as input for ChatGPT to generate structured

and coherent overviews. In this process, ChatGPT is fine-tuned through domain-specific few-shot learning to produce scenario overviews that are consistent with the BPSS model while maintaining the original dialogue's emotional tone and sentiment. Ultimately, this method offers a new avenue for deep understanding of individual cases in the field of social work.

# Chapter 4

# Experimental Design and Evaluation

## 4.1 Selection and Validation of Classification Models

Before conducting the formal experiment, a preliminary experiment is necessary to identify the model that is most suitable for the BPSS classification task, which will provide technical support for the subsequent automated BPSS assessment method. The results of the preliminary experiment will directly influence the selection and optimization direction of subsequent experimental designs, particularly in choosing the best text classification technology and improving classification accuracy. The objective of the classification is to categorize single-sentence texts into one of the categories: Biological (B), Psychological and Spiritual (P), Social (S), or None of the above (N). The text data used is sourced from Hugging Face's "nbertagnolli/counsel-chat" dataset. The specific steps of the preliminary experiment are as follows and in Figure 4.1:

1. Data Acquisition and Preprocessing

   The data acquisition phase involves obtaining the "counsel chat" dataset from the Hugging Face platform, which includes counseling dialogue texts related to psychotherapy, supporting mental health research and applications. During the data preprocessing stage, the Python programming language is used for data cleaning,

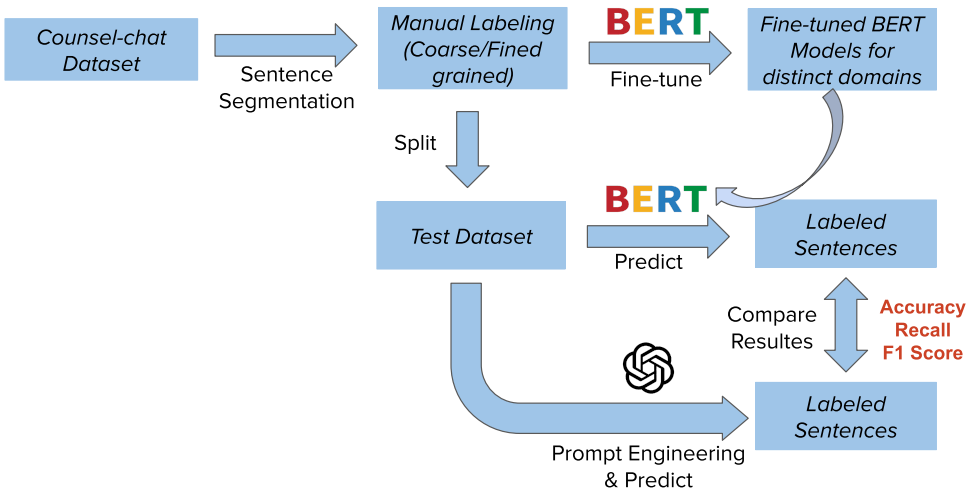



Figure 4.1: Preliminary BPSS classification workflow using BERT and ChatGPT

primarily involving the removal of invalid or duplicate texts from the dataset. Additionally, sentence segmentation is implemented to divide longer text paragraphs into individual sentences, facilitating more precise subsequent processing. The dataset is rigorously divided into training, testing, and validation sets, distributed in an 8:1:1 ratio to ensure that each category label is evenly distributed across the datasets, providing a balanced data foundation for model training and evaluation.

2. Manual Annotation

In the manual annotation phase, the annotation team meticulously categorized preprocessed sentences, covering over 400 sentences. Each sentence was assigned to biological (B), psychological and spiritual (P), social (S), or no relevant category (N), along with finer sub-category labels within these domains. To ensure consistency and accuracy in annotation, the process adhered to strict guidelines. After completing the annotations, inter-rater reliability tests were conducted to verify consistency among different annotators. This test helps to assess and ensure the quality of annotations, ensuring that the dataset's annotated results are highly reliable and statistically valid.

3. BERT Fine-Tuning

   For the BPSS classification task, a BERT base model ("google-bert/bert-base-uncased") obtained from Hugging Face was fine-tuned. This model, not case-sensitive, is suitable for English text processing tasks. The fine-tuning process included adjusting key model parameters such as learning rate, batch size, and training epochs to suit specific classification needs. Hyperparameter adjustments were based on performance on the validation set to achieve optimal model performance. Additionally, the random seed was fixed to ensure the reproducibility of the experiment. Strategies such as early stopping were employed to monitor the risk of overfitting during the model training process, ensuring the model's generalization ability.

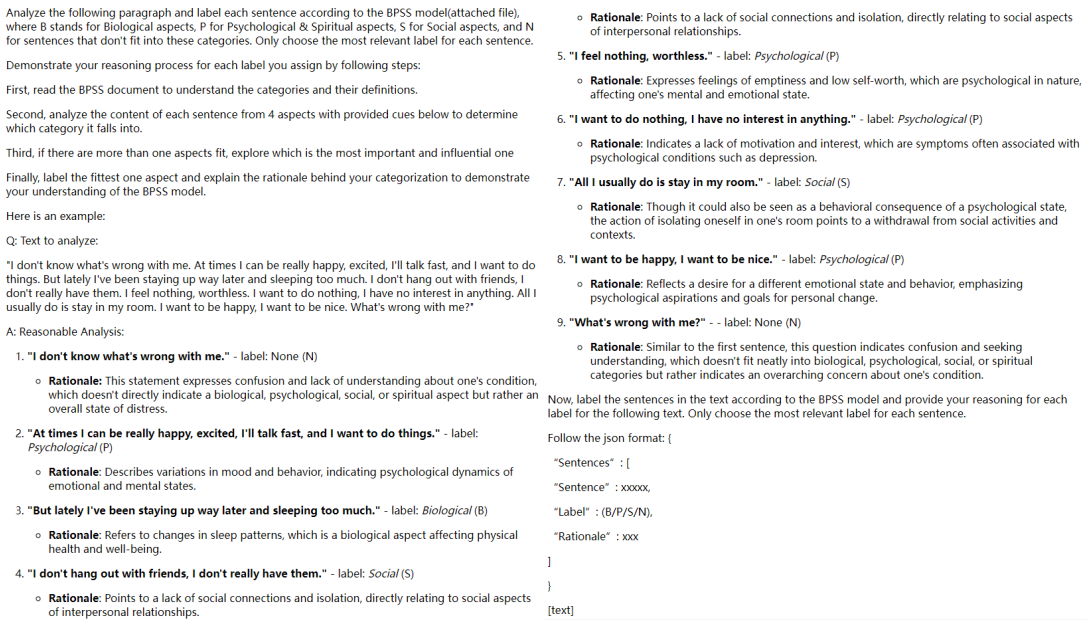
4. Prompt Engineering and GPT Annotation

   Interacting with the OpenAI API through Python scripts, prompt engineering techniques were employed for the GPT model, as shown in figure 4.2, incorporating chain-of-thought (CoT) and few-shot learning strategies to enhance the model's text understanding and analysis capabilities. This method, through detailed prompt design, guides the model to perform in-depth text analysis and understanding. The design of the prompt requires the model to first read and understand the category definitions of the BPSS model, then analyze the content of each sentence in the text, and finally select the most fitting category, elaborating on the logic and reasons for the selection. In practice, this strategy requires the GPT model to demonstrate its reasoning process when predicting categories, thereby improving the accuracy and interpretability of classification.

5. Performance Evaluation

   After the completion of model training, a comprehensive evaluation was conducted

on the performance of ChatGPT4-Turbo and the fine-tuned BERT model. The evaluation focused on key metrics such as accuracy, recall, and F1 score on an independent test set. These metrics measure the model's efficacy in the BPSS classification task from multiple perspectives, providing quantitative grounds for ultimately selecting the model best suited for automated BPSS assessments.



Analyze the following paragraph and label each sentence according to the BPSS model(attached file), where B stands for Biological aspects, P for Psychological & Spiritual aspects, S for Social aspects, and N for sentences that don't fit into these categories. Only choose the most relevant label for each sentence.

Demonstrate your reasoning process for each label you assign by following steps:

First, read the BPSS document to understand the categories and their definitions.

Second, analyze the content of each sentence from 4 aspects with provided cues below to determine which category it falls into.

Third, if there are more than one aspects fit, explore which is the most important and influential one

Finally, label the fittest one aspect and explain the rationale behind your categorization to demonstrate your understanding of the BPSS model.

Here is an example:

Q: Text to analyze:

"I don't know what's wrong with me. At times I can be really happy, excited, I'll talk fast, and I want to do things. But lately I've been staying up way later and sleeping too much. I don't hang out with friends, I don't really have them. I feel nothing, worthless. I want to do nothing, I have no interest in anything. All I usually do is stay in my room. I want to be happy, I want to be nice. What's wrong with me?"

A: Reasonable Analysis:

1. **"I don't know what's wrong with me."** - label: None (N)
   - **Rationale:** This statement expresses confusion and lack of understanding about one's condition, which doesn't directly indicate a biological, psychological, social, or spiritual aspect but rather an overall state of distress.
2. **"At times I can be really happy, excited, I'll talk fast, and I want to do things."** - label: *Psychological* (P)
   - **Rationale**: Describes variations in mood and behavior, indicating psychological dynamics of emotional and mental states.
3. **"But lately I've been staying up way later and sleeping too much."** - label: *Biological* (B)
   - **Rationale**: Refers to changes in sleep patterns, which is a biological aspect affecting physical health and well-being.
4. **"I don't hang out with friends, I don't really have them."** - label: *Social* (S)
   - **Rationale**: Points to a lack of social connections and isolation, directly relating to social aspects of interpersonal relationships.
5. **"I feel nothing, worthless."** - label: *Psychological* (P)
   - **Rationale**: Expresses feelings of emptiness and low self-worth, which are psychological in nature, affecting one's mental and emotional state.
6. **"I want to do nothing, I have no interest in anything."** - label: *Psychological* (P)
   - **Rationale**: Indicates a lack of motivation and interest, which are symptoms often associated with psychological conditions such as depression.
7. **"All I usually do is stay in my room."** - label: *Social* (S)
   - **Rationale**: Though it could also be seen as a behavioral consequence of a psychological state, the action of isolating oneself in one's room points to a withdrawal from social activities and contexts.
8. **"I want to be happy, I want to be nice."** - label: *Psychological* (P)
   - **Rationale**: Reflects a desire for a different emotional state and behavior, emphasizing psychological aspirations and goals for personal change.
9. **"What's wrong with me?"** - - label: None (N)
   - **Rationale**: Similar to the first sentence, this question indicates confusion and seeking understanding, which doesn't fit neatly into biological, psychological, social, or spiritual categories but rather indicates an overarching concern about one's condition.

Now, label the sentences in the text according to the BPSS model and provide your reasoning for each label for the following text. Only choose the most relevant label for each sentence.

Follow the json format: {

"Sentences" : [

"Sentence" : xxxxx,

"Label" : (B/P/S/N),

"Rationale" : xxx

]

}

[text]

Figure 4.2: Prompts for ChatGPT Rewriting

## 4.2 Method Implementation and Generation Process

As outlined in Chapter 3 above, the formal experimental design consists of the following steps:

**Dialogue Rewriting**:

Before the formal experiment, a dataset containing dialogues between social workers and a Chatbot is first obtained from the Hear4U organization. This dataset is categorized based on the number of messages, as shown in the table4.1, and dialogues from the top three categories with the highest message counts are selected. This selection criterion

is designed to ensure the breadth of the sample and maintain enough messages to preserve the integrity of the dialogues. The selected dialogue texts are then converted into first-person narrative form for subsequent experimental analysis. This process aims to retain the important details, emotions, and tones of the original dialogues while refining and compressing information based on question-and-answer interactions, omitting the responder's prompts, and focusing on preserving the original mode of expression. By simplifying sentence structures, avoiding complex clauses, and the use of present participles, the narratives are ensured to be direct and concise. Since the original training data from the "Counsel-chat Dataset" is primarily presented in the first-person narrative, rewriting the dialogue data in the first person helps the model to better learn and understand the data, providing a clear framework and context for subsequent text generation. Additionally, a characteristic of psychotherapy dialogue data is that it contains a large amount of question-and-answer and redundant information. The first-person rewrite allows for the distillation of key information from the dialogues, eliminating unnecessary parts, enhancing the density of information and thematic coherence of the text, and creating more accurate and efficient conditions for subsequent classification and scenario generation.

| messages | count |
|---|---|
| $x > 80$ | 3 |
| $80 \geq x > 50$ | 7 |
| $50 \geq x > 20$ | 20 |
| $20 \geq x > 10$ | 6 |
| $10 \geq x > 5$ | 10 |
| $5 \geq x > 0$ | 28 |

Table 4.1: Message count distribution in dialogues

**Data Preparation and Preprocessing**:

Using natural language processing tools such as NLTK, the collected data undergo cleaning and preprocessing. This includes removing duplicate texts and segmenting sentences to ensure the data format meets experimental requirements. This step guarantees the quality of the data, providing well-prepared inputs for subsequent models.

**BERT Fine-Tuning**:

The BERT model is fine-tuned for specific tasks to enhance its ability to recognize BPSS categories. During fine-tuning, learning rates, batch sizes, and training cycles are carefully selected, and the dataset is shuffled multiple times to prevent overfitting. Fine-tuning is divided into two stages: initially, the model is fine-tuned on broader categories to help it recognize and differentiate macro dimensions; subsequently, the model continues training on finer granularity to identify sub-domain labels under the main BPSS categories, significantly improving the model's accuracy in classifying specific BPSS subcategories.

**ChatGPT Training and Generation**:

Using the fine-tuned BERT model, new dialogue texts are classified into corresponding BPSS categories. These classification results are then used as input to train ChatGPT to generate scenario overviews that comply with the BPSS model. ChatGPT utilizes its excellent natural language generation capabilities, combined with domain-specific knowledge obtained from reading BPSS-related PDFs, to produce coherent and structurally complete overviews.

**Evaluation and Comparison**:

To verify the effectiveness of the proposed method, an exhaustive comparative analysis of the BERT classification results and the overviews generated by ChatGPT is conducted. Quantitative and qualitative evaluation metrics, including accuracy, consistency, and user satisfaction, are used to measure the performance of both methods.

At this stage, the overviews generated by the two methods are compared: one based on

texts that have been fine-tuned and classified by BERT and then generated by ChatGPT; the other using overviews generated directly from the original texts by GPT without rewriting and classification. This comparison aims to assess the impact of preprocessing and classification steps on the quality of the generated content.

By comparing their performance in terms of accuracy, consistency, and information density, the effect of preprocessing and precise classification on enhancing the quality of generated texts is evaluated. This analysis helps identify potential issues with generating overviews directly from original texts, such as information loss or thematic drift, and how texts optimized through preprocessing can more effectively guide the generation process.

Evaluation metrics include textual accuracy (reflecting the precision of original intentions and classification labels), consistency (coherence of internal logic and theme within the text), and semantic relevance. These metrics provide a comprehensive evaluation framework to quantify the specific contributions of processing steps to the final generated content.

# Chapter 5

# Results Analysis and Discussion

## 5.1 Consistency Assessment of Manually Annotated Dataset

In natural language processing and other related fields, manual annotation is a key step in acquiring high-quality training data. To ensure the reliability and statistical validity of the dataset, the annotation team meticulously categorized over 400 preprocessed sentences, covering major areas such as Biological (B), Psychological and Spiritual (P), Social (S), and Non-relevant (N) categories, along with their sub-labels. The entire annotation process was strictly conducted following established principles and referenced BPSS model-related materials to ensure high data quality(Cheung et al., 2023).

Inter-rater reliability (IRR) is an important method for assessing the consistency among different raters in scoring or classification tasks. IRR can be calculated in various ways, among which Cohen's Kappa coefficient is one of the most commonly used statistical tools. This coefficient measures the level of agreement between two or more raters beyond chance, providing a value from -1 to 1: 1 indicates perfect agreement, 0 indicates agreement that is no better than chance, and negative values indicate agreement below

chance level. This measure is more stringent than simple percentage agreement because it considers the possibility of chance agreement.

In this study, the Cohen's Kappa coefficient for the main domain labels was 0.3365, indicating a moderate level of agreement among annotators. The consistency assessment for sub-domain labels, considering only cases where the main domain labels matched and were not in the "N" category, yielded a Kappa coefficient of 0.2454, indicating lower consistency at this level. This suggests that while the consistency in main domain labeling is acceptable, the sub-domain labeling shows room for further standardization and training.

## 5.2 Coarse Classification Model Performance Comparison and Analysis

This experiment aims to compare the performance of the Google BERT model and the ChatGPT4-Turbo model in the BPSS classification task. In the experiment, single-sentence texts are classified into Biological (B), Psychological and Spiritual (P), Social (S), or None of the above categories (N). The dataset used includes data from the Counsel-chat dataset, which has been preprocessed and manually annotated to ensure accuracy and consistency in the classification task. The dataset was divided into training, validation, and test sets in an 8:1:1 ratio to assess the models' generalization ability on unseen data.

For this experiment, the base version of BERT, which is not case-sensitive, was used, with specific training parameters set including a learning rate of 2e-5, batch size of 4, training duration of 5 epochs, and weight decay of 0.01. Moreover, the best model saving strategy was adopted during the model training process, selecting the best model based on the F1 score. As a generative model, ChatGPT4-Turbo was adjusted for the classification task through specific prompt engineering.

| | BERT-Base-Uncased | ChatGPT4-Turbo |
|---|---|---|
| Test Accuracy | 0.8864 | 0.7045 |
| Test Recall | 0.8455 | 0.5948 |
| Test F1 Score | 0.8781 | 0.6010 |

Table 5.1: Performance comparison of BERT-Base-Uncased and ChatGPT4-Turbo

The experimental results in Table 5.1, by comparing the performance of the two models, show that the Google BERT model has a significant advantage in text classification tasks, particularly in terms of accuracy and model stability. The deep language understanding capabilities of the BERT model make it more suitable for handling social work texts with rich semantics and complex structures. In contrast, although the ChatGPT4-Turbo can provide reasonable classification results in some cases, its performance is more variable and may require further adjustments and optimizations to achieve optimal performance in such tasks.

Specifically, the BERT model achieved an accuracy of 0.886 on the test set, while the ChatGPT4-Turbo model had an accuracy of 0.705. Additionally, the recall rate for the BERT model was 0.846, compared to 0.595 for the ChatGPT4-Turbo model. The F1 scores also reflect the superiority of the BERT model, with a score of 0.878 compared to 0.601 for the ChatGPT4-Turbo model. These results indicate that the BERT model consistently demonstrated better classification performance throughout the experiment. When considering model selection, generalization ability is an important factor. The experimental results show that due to the superior performance of the BERT model on the test set, it has stronger generalization capabilities on unseen data. Therefore, based on the results of this experiment, it can be concluded that the BERT model is a more suitable choice for the BPSS classification task.

## 5.3 Fine-Tuning BERT Model for Detailed Classification Analysis

In the research methods of this project, fine-tuning the BERT model for both coarse and fine classification was employed to enhance the precision of automated BPSS assessments in social work. The main objective of fine classification is to perform more refined information extraction and analysis for the individual dimensions of biological, psychological and spiritual, and social aspects. This section will discuss the results of fine classification and the challenges encountered.

| | biology | psychology&spirit | social |
|---|---|---|---|
| Test Accuracy | 0.875 | 0.75 | 0.68 |
| Test Recall | 0.875 | 0.75 | 0.68 |
| Test F1 | 0.8214 | 0.6858 | 0.6594 |

Table 5.2: BERT model metrics by BPSS dimensions

The fine classification results are shown in the table. Within the Biology category, the fine-tuned BERT model displayed high accuracy and recall, both reaching 0.875, with an F1 score of 0.8214. These high-performance indicators reflect the model's strong capability in identifying information related to biological aspects, such as employment status, medical needs, and living environments, demonstrating the BERT model's advantage in handling specific and well-defined category information. In contrast, the accuracy and recall for the Psychology & Spirit category were both 0.75, with an F1 score of 0.6858. The performance in this category was lower, primarily due to the broader and more complex subcategories within the psychological and spiritual domain, which include emotional therapy, cognitive-behavioral therapy, and stages of psychological development. The ambiguous definitions of these areas increase the difficulty of the classification task. Lastly,

the Social category showed the most average performance, with both accuracy and recall at 0.68 and an F1 score of 0.6594. This category includes complex analyses of social structures and relationships such as culture, marital relations, and family dynamics, whose breadth and interactivity may lead to lower performance in this category.

Several factors may contribute to the results:

1. Complexity of Psychological and Social Categories: The complexity of psychological and social categories is one of the main reasons for the poor classification outcomes. The diversity and complexity of concepts and theories in psychology and sociology, which involve a wide range of interpersonal relationships and assessments of internal psychological states, significantly increase the difficulty of automated processing.

2. Limitations of the Dataset: The small size of the test dataset and potential imbalances limit the effectiveness of model fine-tuning. When data is insufficient or of low quality, the model struggles to learn enough representative features, impacting the final classification performance.

3. Impact of Annotation Quality: Errors in manual annotations could also affect the model's learning effectiveness. In psychological and social categories, due to unclear conceptual boundaries, subjective judgments among different annotators may vary, leading to inconsistent annotation quality in the training data, which in turn affects the accuracy and generalizability of model training.

Despite facing multiple challenges, the fine-tuned BERT model demonstrated good classification results in the biological category but still has room for improvement in the psychological and spiritual and social categories. Future research can enhance the model's performance in complex classification tasks by expanding and optimizing the training set, improving data annotation quality, and further optimizing the model structure and parameters. Additionally, given the special complexity of the psychological and social

categories, exploring new methods that incorporate more contextual information and model interpretability will be an important direction for further research.

## 5.4 BPSS Scenario Overview Quality Assessment

To assess the impact of preprocessing and precise classification on the quality of generated content, this study conducted a detailed comparative analysis of two different generation methods: one based on texts that have been fine-tuned and classified by BERT, with overviews generated by ChatGPT; the other using original texts that have not been rewritten or classified, with overviews generated by GPT. The evaluation analysis relies on Coh-Metrix-related indices, which provide a comprehensive framework for assessing natural language processing and text generation quality, including dimensions such as textual accuracy, consistency, and semantic relevance(McNamara, Graesser, McCarthy, & Cai, 2014).

1. Textual Accuracy reflects whether the generated text accurately conveys the original intent and meets the classification label requirements. Specific evaluation metrics include:

    - **PCNARz/PCNARp** (Text Easability PC Narrativity): Reflects the narrativity of the text, which can indirectly measure whether the text accurately conveys the original intent or meets specific classification labels.
    - **PCCNCz/PCCNCp** (Text Easability PC Word concreteness): Measures the concreteness of vocabulary, where specific words help more accurately reflect the intent of the original content.
    - **CRFNO1/CRFNOa** (Noun overlap, mean): Reflects the repetition of nouns within sentences or texts; high repetitiveness may indicate a strong emphasis, aligning closely with the original intent.

2. Consistency involves the internal logic and thematic coherence of the text, with evaluation metrics including:

    - **PCREFz/PCREFp** (Text Easability PC Referential cohesion): Referential cohesion, measuring how vocabulary in the text is related and builds themes.
    - **PCDCz/PCDCp** (Text Easability PC Deep cohesion): Deep cohesion, reflecting the coherence of internal logic and themes within the text.
    - **PCCONNz/PCCONNp** (Text Easability PC Connectivity): Connectivity, showing how parts of the text are logically connected.
    - **CRFCWO1/CRFCWOa** (Content word overlap, mean): Content word overlap, reflecting the consistency and coherence of themes between sentences or paragraphs.

3. Semantic Relevance focuses on how the meanings at the vocabulary and sentence levels correspond to the original text or contextual environment, with evaluation metrics including:

    - **LSASS1/LSASSp**(LSA overlap, mean): Latent Semantic Analysis overlap, displaying the semantic coherence between sentences or paragraphs.
    - **SMCAUSlsa/SMCAUSwn** (LSA/WordNet verb overlap): Semantic overlap of verbs, showing the semantic relationships between verbs through LSA or WordNet.
    - **SYNLE** (Left embeddedness, words before main verb, mean): Represents the depth of lexical embedding before the main verb, which may affect the complexity of semantic interpretation.
    - **DRPVAL**(Agentless passive voice density, incidence): Density of agentless passive voice, which may affect the clarity of semantics.

| | BERT | Non-BERT | Original |
|---|---|---|---|
| CRFCWO1 | 0.540816 | 0.443311 | 0.391156 |
| CRFNO1 | 0.619818 | 0.629139 | 0.145205 |
| DRPVAL | 0.394314 | 0.274228 | 0.055943 |
| LSASS1 | 0.853937 | 0.790655 | 0.095659 |
| PCCNCz | 0.638011 | 0.633481 | 0.213162 |
| PCCONNz | 0.262317 | 0.273080 | 0.767200 |
| PCDCz | 0.265725 | 0.398793 | 0.455249 |
| PCNARz | 0.079764 | 0.196386 | 0.942566 |
| PCREFz | 0.377516 | 0.409804 | 0.766013 |
| SMCAUSlsa | 0.440801 | 0.260474 | 0.198543 |
| SYNLE | 0.763296 | 0.427867 | 0.143766 |

Table 5.3: Coh-Metrix indices for BERT, Non-BERT, and Original texts

According to the analysis results of the normalized indices, as shown in the table, the impact of BERT processing on text quality can be deeply explored, particularly in terms of textual accuracy, consistency, and semantic relevance. Here is a specific comparative analysis of different versions (original text, text unprocessed by BERT, and text processed by BERT):

1. Textual Accuracy

   - **PCNARz (Narrativity)**: The original text significantly outperforms other versions in terms of narrativity, with a normalized score of 0.943, indicating that the original text retains a more complete narrative structure. In contrast, the text processed by BERT scored 0.080, suggesting that some narrativity may be sacrificed to adapt to specific content summarization and information compression.

- **PCCNCz (Word Concreteness)**: Text processed by BERT scored 0.638 in terms of word concreteness, slightly higher than the unprocessed text, indicating that BERT processing helps select more specific and direct vocabulary to convey information.
- **CRFNO1 (Noun Overlap)**: Text processed by BERT scored 0.620 in noun overlap, slightly lower than unprocessed text but much higher than the original text's 0.145, indicating that BERT processing effectively emphasizes key nouns.

2. Consistency

- **PCREFz (Referential Cohesion) and PCDCz (Deep Cohesion)**: The original text scores highest on these indices, showing high consistency in referential and deep cohesion. Text processed by BERT, although scoring lower, shows signs of improvement over unprocessed text.
- **PCCONNz (Connectivity)**: The original text scored the highest in connectivity (0.767), indicating that the original text may have a more complete logical structure and connections.
- **CRFCWO1 (Content Word Overlap)**: Text processed by BERT scored the highest in content word overlap (0.541), indicating good vocabulary consistency across different parts.

3. Semantic Relevance

- **LSASS1 (LSA Overlap)** and **SMCAUSlsa (LSA Verb Overlap)**: Text processed by BERT scored higher on these indices, especially in LSA overlap (0.854), indicating that BERT processing helps maintain semantic consistency between sentences and paragraphs.

- **SYNLE (Left Embeddedness)** and **DRPVAL (Agentless Passive Voice Density)**: Text processed by BERT also performed well on these indices of structural complexity, indicating that the processed text may be more refined and complex in linguistic complexity and sentence structure.

Combining these analysis results, it can be concluded that BERT fine-tuning and classification significantly enhance the overall quality of the text, showing clear advantages over the original unprocessed text, especially in terms of consistency and semantic relevance. Although there are sacrifices in maintaining original narrativity, overall, preprocessing and precise classification clearly optimize the quality of generated content, making it more suitable for specific application needs such as content summarization and information retrieval. These findings provide valuable insights for future text processing and generation strategies, demonstrating the critical role of precise preprocessing in automated text generation.

# Chapter 6

# Conclusion and Future Work

## 6.1 Research Summary and Main Findings

This dissertation has explored the integration of BERT and GPT models to automate the Biological, Psychological, Social, and Spiritual (BPSS) assessment in social work, resulting in the development of a novel hybrid model. This model effectively combines the analytical strengths of BERT in natural language understanding with the generative capabilities of GPT, facilitating enhanced efficiency and accuracy in social work assessments. The following are the main findings from this research:

1. Improved Assessment Efficiency and Accuracy: The hybrid model significantly outperforms traditional assessment methods by providing quicker and more accurate evaluations of clients' needs. This is primarily due to BERT's robust text classification and GPT's adeptness at generating coherent, contextually appropriate summaries of individuals' BPSS states.

2. Methodological Innovation: The study introduces an innovative automated approach that leverages the synergistic potentials of BERT and GPT. This methodology enables more nuanced and dynamic assessments, reflecting a significant advancement over static, manual methods traditionally used in social work.

3. Empirical Validation: The hybrid model's utility was empirically tested using real-world data from social work interactions, validating its effectiveness in practical settings. This not only demonstrates the hybrid model's operational viability but also underscores its potential to enhance the quality of BPSS assessments in actual social work environments.

## 6.2 Future Research Directions and Potential Applications

**Future Research Directions**:

1. Daily Social Work Practices: The hybrid model can be regularly employed in various social work settings to streamline BPSS assessments, significantly reducing time spent on paperwork and allowing social workers to dedicate more effort to direct client care.

2. Use in Crisis Situations: In emergency or crisis situations where swift decision-making is crucial, the hybrid model can provide rapid assessments, aiding in the quick formulation of intervention strategies.

3. Training and Education: The hybrid model can be used as an educational tool in social work training programs, helping students learn complex assessment processes through simulation of real-life scenarios.

4. Research in Social Sciences: It can also serve as a powerful research tool to analyze communication patterns and interactions within social work, thereby contributing to academic research and practical methodologies in the field.

Overall, the successful implementation of this hybrid model holds the promise not only to revolutionize BPSS assessments in social work but also to provide a framework

for future advancements in the integration of AI technologies in humanitarian disciplines.

# References


Cheung, S. L., Chin, E., Chua, E. C., Das, B. M., Fan, L. C., Mardiana, S., & Yong, J. (2023). *A bio-psychosocial-spiritual assessment guide for health and social work*. Singapore Association of Social Workers.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*, , 2018.

Galbadage, T., Peterson, B. M., Wang, D. C., Wang, J. S., & Gunasekera, R. S. (2020). Biopsychosocial and spiritual implications of patients with covid-19 dying in isolation. *Frontiers in Psychology*, *11*, , 2020, 588623.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), , 2023, 3713–3744.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, , , 2023, 100017.

Markovič, Daniel (2024). Current options and limits of digital technologies and artificial intelligence in social work. *SHS Web Conf.*, *184*, , 2024, 05003.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix*. Cambridge University Press.

Montejo-Ráez, A., & Jiménez-Zafra, S. M. (2022). Current approaches and applications in natural language processing. *Applied Sciences*, *12*(10), , 2022.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning* (pp. 28492–28518), PMLR, , 2023.

Sultana, T., & Badugu, S. (2020). A review on different question answering system approaches. In S. C. Satapathy, K. S. Raju, K. Shyamala, D. R. Krishna, & M. N. Favorskaya (Eds.), *Advances in Decision Sciences, Image Processing, Security and Computer Vision* (pp. 579–586), Cham, , 2020: Springer International Publishing.

Victor, B. G., Kubiak, S., Angell, B., & Perron, B. E. (2023). Time to move beyond the aswb licensing exams: Can generative artificial intelligence offer a way forward for social work? *Research on Social Work Practice*, *33*(5), , 2023, 511–517.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Vol. 35 (pp. 24824–24837), , 2022: Curran Associates, Inc.

Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, *56*(3), , oct, 2023.

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.