*applied sciences*

MDPI

*Article*

# DAFT: Domain-Augmented Fine-Tuning for Large Language Models in Emotion Recognition of Health Misinformation

**Youlin Zhao [1,*], Xingmi Zhu [1], Wanqing Tang [1], Linxing Zhou [2,*], Li Feng [3] and Mingwei Tang [4]**

[1] Business School, Hohai University, Nanjing 211100, China; 241320010038@hhu.edu.cn (X.Z.); 240220010004@hhu.edu.cn (W.T.)
[2] School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China
[3] Hohai University Library, Nanjing 211100, China; 20040603@hhu.edu.cn
[4] School of Engineering Audit, Nanjing Audit University, Nanjing 211815, China; tmw@nau.edu.cn
* Correspondence: sobzyl@hhu.edu.cn (Y.Z.); zlxshu@shu.edu.cn (L.Z.)

**Abstract**

This study proposes a domain-augmented fine-tuning strategy for improving emotion recognition in health misinformation using pre-trained large language models (LLMs). The proposed method aims to address key limitations of existing approaches, including insufficient precision, weak domain adaptability, and low recognition accuracy for complex emotional expressions in health-related misinformation. Specifically, the Domain-Augmented Fine-Tuning (DAFT) method extends a health emotion lexicon to annotate emotion-oriented corpora, designs task-specific prompt templates to enhance semantic understanding, and fine-tunes GPT-based LLMs through parameter-efficient prompt tuning. Empirical experiments conducted on a health misinformation dataset demonstrate that DAFT substantially improves model performance in terms of prediction error, emotional vector structural similarity, probability distribution consistency, and classification accuracy. The fine-tuned GPT-4o model achieves the best overall performance, attaining an emotion recognition accuracy of 84.77%, with its F1-score increasing by 20.78% relative to the baseline model. Nonetheless, the corpus constructed in this study is based on a six-dimensional emotion framework, which may not fully capture nuanced emotions in complex linguistic contexts. Moreover, the dataset is limited to textual information, and future research should incorporate multimodal data such as images and videos. Overall, the DAFT method effectively enhances the domain adaptability of LLMs and provides a lightweight yet efficient approach to emotion recognition in health misinformation scenarios.

**Keywords:** health misinformation; emotion recognition; large language models; domain adaptation; prompt tuning; fine-tuning; GPT-4o

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 28 October 2025
Revised: 23 November 2025
Accepted: 27 November 2025
Published: 29 November 2025

**Citation:** Zhao, Y.; Zhu, X.; Tang, W.; Zhou, L.; Feng, L.; Tang, M. DAFT: Domain-Augmented Fine-Tuning for Large Language Models in Emotion Recognition of Health Misinformation. *Appl. Sci.* **2025**, *15*, 12690. https://doi.org/10.3390/app152312690

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In recent years, the frequent occurrence of public health crises, coupled with the rapid expansion of social media, has markedly accelerated the dissemination of health misinformation [1,2]. Although social media platforms have become important channels for sharing health information, the lack of rigorous verification mechanisms enables large volumes of misleading or unverified content to spread rapidly, often leading to inappropriate health-related decisions and increased behavioral risks [3]. Health misinformation generally refers to health-related content that contradicts established scientific evidence and medical consensus [4,5]. It not only threatens individual and public health but also undermines health communication ecosystems by amplifying uncertainty and distrust.

Prior studies have shown that health misinformation frequently employs emotional manipulation strategies to attract attention and facilitate viral diffusion [6]. Emotionally charged content—such as messages evoking fear, anger, or sympathy—has been shown to be more likely to trigger user engagement and sharing behaviors, thereby amplifying misinformation cascades on social media [7]. Consequently, traditional approaches such as fact-checking and content-based detection alone are often insufficient to effectively curb its spread [8]. These limitations underscore the necessity of incorporating emotional analysis into misinformation detection frameworks, making emotion recognition in health misinformation an increasingly important area of research [6].

Existing methods for emotion recognition in health misinformation still face several challenges, including limited contextual understanding, insufficient domain adaptability, and limited capability for dynamic emotion analysis [6]. Traditional dictionary-based methods rely on predefined emotion lexicons but fail to capture complex semantic dependencies and often lack applicability in health-related contexts [9]. Machine learning-based approaches introduce automatic feature learning for emotion classification; however, they remain constrained by their reliance on large-scale annotated datasets and have difficulty recognizing fine-grained or implicit emotions [10]. While large LLMs excel in language understanding and generation, their performance in domain-specific tasks like health misinformation is suboptimal due to domain gaps in pre-training corpora and insufficient task adaptation. [11]. Consequently, fine-tuning strategies are required to enhance the domain adaptability of LLMs and to equip them with stronger capabilities for emotion recognition in health misinformation scenarios [12].

To adapt pretrained LLMs to domain-specific emotion recognition tasks, fine-tuning has become a widely adopted strategy. Full-parameter fine-tuning improves performance but is computationally expensive and prone to overfitting, particularly in low-resource domains [13]. To address these limitations, prompt-based tuning has emerged as an efficient alternative, using task-specific templates to guide model outputs and reduce dependence on large annotated datasets while preserving generalization [11]. Recent studies have attempted to systematize prompt design by introducing prompt pattern engineering to improve task adaptability [14]. However, current practices in prompt design still rely heavily on expert experience and lack unified optimization principles, which can lead to unstable model performance across domains [15]. Additionally, parameter-efficient tuning techniques such as adapter tuning freeze most model parameters and add lightweight trainable modules, reducing training complexity [16]. Despite these advances, achieving both high efficiency and robust task adaptability remains an open challenge in fine-tuning LLMs for domain-specific tasks [12].

Despite significant progress in emotion recognition and health misinformation detection, existing methods often struggle with domain-specific challenges, particularly in health misinformation. Traditional fine-tuning approaches are computationally expensive and may overfit when working with limited domain-specific data. Moreover, prompt-based strategies, while efficient, still face issues with domain adaptability and consistency. In response to these challenges, this study proposes a Domain-Augmented Fine-Tuning (DAFT) method for GPT-based LLMs to improve emotion recognition performance in health misinformation scenarios. DAFT enhances the domain adaptability of LLMs by integrating domain-specific emotional knowledge with parameter-efficient optimization strategies. The proposed framework consists of three key components: (1) Construction of a domain-specific emotional corpus, in which a health misinformation emotion corpus is developed using an extended domain-adapted emotion lexicon, enabling automatic emotion annotation and providing reliable supervision signals for model learning; (2) Task-oriented prompt template optimization, where customized prompt templates are designed for both

emotion classification and emotional intensity prediction tasks to improve semantic understanding in fine-grained emotion recognition; and (3) Parameter-efficient prompt-based fine-tuning, in which DAFT combines prompt tuning with selective parameter adjustment to reduce reliance on large annotated datasets and lower computational overhead while preserving model expressive capacity. Overall, this method integrates a health-specific emotion lexicon, task-optimized prompt templates, and parameter-efficient tuning to enhance the performance of large language models in emotion recognition tasks within health misinformation. DAFT not only reduces reliance on large annotated datasets but also improves the model's ability to adapt to complex and nuanced emotional expressions unique to health misinformation, offering a novel solution to existing limitations in the field.

## 2. Research Background and Related Work

### 2.1. Definition of Health Misinformation

Health misinformation is a widely discussed concept within the broader literature on misinformation. Fernandez and Alani define misinformation as false or misleading information that is created or disseminated without verification of its authenticity, often including fake news and rumors [17]. Related notions such as disinformation and rumor are frequently used in similar contexts but convey distinct meanings [18]. Disinformation refers specifically to intentionally fabricated content designed to mislead or manipulate public perception [19], while rumors consist of unverified information that spreads rapidly within social groups and may later be confirmed or disproven [20,21]. Recent research on misinformation increasingly focuses on the health domain, with topics such as vaccination, infectious diseases, cancer, and chronic illnesses becoming common subjects of investigation [22]. Within this line of work, health misinformation is generally viewed as a subset of misinformation that specifically concerns health-related content [23]. Unlike other forms of misinformation, health misinformation may arise from either unintentional information distortion or deliberate manipulation, without necessarily distinguishing between the motivations behind its dissemination [8]. In academic discourse, health misinformation is commonly defined as health-related information that contradicts scientific facts or medical consensus [4,5]. In line with these perspectives, this study defines health misinformation as false or misleading health information that deviates from evidence-based medical knowledge or established scientific consensus.

### 2.2. Health Misinformation Emotion Recognition Method

Emotional features have been widely recognized as salient indicators for the detection and analysis of health misinformation. Early studies relied primarily on manual annotation, where domain experts identified emotional cues such as exaggeration, fear appeals, or misleading sentiment based on professional judgment. However, manual methods are constrained by subjectivity and limited scalability. Subsequently, dictionary-based approaches emerged, using predefined sentiment lexicons to compute emotion polarity and classify health-related claims [24]. Although simple and interpretable, these methods struggle to capture contextual semantics and domain-specific emotional nuances. To improve detection performance, multimodal emotion recognition methods were later introduced, integrating emotional cues from text and images to enhance semantic understanding and classification robustness [25]. With advances in natural language processing (NLP), research attention has gradually shifted from surface-level lexical features to deeper semantic representations, enabling more accurate modeling of the emotional patterns present in misinformation [26]. In particular, deep learning techniques have

demonstrated strong capabilities in learning complex contextual dependencies from large-scale datasets [27]. For example, Bi-LSTM architectures have been used to jointly encode semantic and emotional information, thereby improving sentiment-aware misinformation classification [28]. Graph-based models such as GCNs and GNNs have also been employed to model relational structures in social media propagation networks, enabling more accurate recognition of emotional signals embedded in misinformation dissemination patterns [29].

Despite these advances, current emotion-based misinformation detection methods still exhibit significant limitations. General-purpose emotion lexicons often lack domain adaptability and frequently fail to accurately interpret medical terminology, leading to semantic bias and misclassification in health-related contexts [9]. Additionally, many pre-trained models rely heavily on superficial sentiment polarity features while underutilizing deeper semantic cues, which makes it difficult to detect misinformation that is masked by neutral or scientific language [6]. Furthermore, owing to limited annotated data, existing models often generalize poorly across languages, platforms, and communication scenarios [10]. These limitations highlight the need for new approaches that integrate specialized health knowledge with rich semantic modeling to improve domain adaptability in emotion recognition for health misinformation.

### *2.3. Prompt Fine-Tuning for LLMs*

Prompt-based fine-tuning has emerged as an efficient transfer-learning paradigm for adapting large language models (LLMs) to downstream tasks by guiding model behavior through task-specific prompt templates [13]. In prompt learning, prompts are often combined with few-shot examples to enhance task adaptability and reduce dependence on large annotated datasets [11]. Prompting strategies are typically classified into two categories: hard prompts and soft prompts. Hard prompts, also known as discrete prompts, consist of manually crafted textual instructions that steer model outputs but require extensive expert intervention and task-specific tuning [12]. In contrast, soft prompts—also referred to as continuous prompts—inject learnable embedding vectors into model inputs, enabling parameter-efficient tuning without modifying the full model architecture [30]. Due to their interpretability and ease of application, hard prompts have been widely adopted, particularly in generative systems such as ChatGPT which is based on the GPT-3.5 architecture [31].

Prompt design is a critical factor that influences model performance [15]. Recent research has demonstrated that LLMs are highly sensitive to the structure and wording of prompts, which has motivated the development of prompt engineering strategies aimed at enhancing prompt robustness and task alignment [14,32]. Semi-supervised prompt learning frameworks have also been proposed to improve generalization in few-shot classification tasks [33]. In parallel, retrieval-augmented prompting has been introduced as a means of incorporating external context and reduce factual errors and hallucinations in model outputs [34].

However, prompt-based fine-tuning still faces significant limitations. Prompt construction relies heavily on expert knowledge and lacks systematic optimization principles, which weakens domain transferability [14]. Hard prompts are highly sensitive to lexical variation and often yield unstable performance across tasks. Additionally, prompt tuning alone is insufficient to address hallucinations and provides limited domain specialization, particularly in complex fields such as healthcare and misinformation detection. These limitations highlight the need to integrate prompt-based tuning with domain-enhanced knowledge in order to further improve LLM adaptability and task reliability.

### 2.4. Review Summary

Research on health misinformation and emotion recognition has advanced substantially in recent years, with notable progress achieved in several related areas. (1) Scholars have clarified the conceptual distinctions among misinformation, disinformation, and rumor, and have provided more rigorous definitions of health misinformation within academic discourse [4,5,8,17–23]. (2) Emotion recognition techniques for health misinformation have evolved from manual annotation and lexicon-based methods to machine learning and deep learning approaches, thereby improving semantic understanding and feature extraction capabilities [24–29]. (3) Prompt engineering has emerged as a promising paradigm for adapting large language models to downstream tasks by leveraging hard and soft prompts to improve task alignment [11–15,30–34].

Despite these developments, several important research gaps remain. (1) Insufficient domain adaptability: General-purpose sentiment lexicons and pretrained models often fail to capture emotional expressions specific to medical discourse, resulting in limited recognition accuracy for health-related emotional cues [6,9,10]. (2) High data dependency and computational cost: Conventional deep learning methods rely heavily on large annotated datasets and typically require full-parameter fine-tuning, which is computationally expensive and prone to overfitting [13,27]. (3) Limited generalization of prompt tuning: Existing prompt design strategies depend strongly on manual heuristics and lack systematic adaptability across domains, which restricts their effectiveness in specialized fields such as health-related misinformation [14,15]. A summary of the representative studies reviewed above is presented in Table 1.

**Table 1.** Summary of representative studies in the literature review.

| Concept | Author, Year | Main Contribution | Limitations |
| --- | --- | --- | --- |
| Definition of Health Misinformation | Fernandez & Alani (2018) [17] | Clarifies core concepts and challenges of online misinformation | Does not focus on health domain or emotional mechanisms in misinformation |
| | Victoria L. Rubin (2019) [18] | Differentiates misinformation, disinformation and interventions | No empirical modeling of health-specific misinformation or emotion |
| | Chen et al. (2022) [4] | Systematic overview of health misinformation and detection approaches | Emotion is treated as a feature but not modeled as multi-dimensional vectors |
| Health Misinformation Emotion Recognition Method | Balshetwar et al. (2023) [24] | Shows emotional features can help fake news detection | Uses general emotion polarity; ignores domain-specific medical emotion expressions |
| | Kaur & Kautish (2019) [25] | Demonstrates benefit of combining text and image modalities | Not specific to health misinformation; no domain-adapted lexicon |
| | Hamed et al. (2023) [28] | Integrates sentiment of news and users' comments to detect fake news | Emotion representation still coarse-grained |
| | Xuewen et al. (2023) [29] | Uses graph structures and sentiment for rumor detection | Does not exploit LLMs or prompt-based fine-tuning for emotion vectors |

**Table 1.** *Cont.*

| Concept | Author, Year | Main Contribution | Limitations |
| --- | --- | --- | --- |
| Prompt Fine-tuning for LLMs | Liu et al. (2023) [11] | Systematically reviews prompting methods for PLMs | Focuses on general NLP tasks; no health-specific emotion recognition settings |
| | Parthasarathy et al. (2024) [13] | Summarizes full-parameter and parameter-efficient fine-tuning | Lacks concrete application to health misinformation emotion recognition |
| | Xu et al. (2023) [16] | Reduces training cost via dynamic capacity | Method is generic; not combined with emotion-specific lexicons or health misinformation |
| | Shin et al. (2025) [30] | Provides evidence on when prompt or fine-tuning works better | Application domain is software engineering, not health misinformation |

To address these challenges, this study proposes a Domain-Augmented Fine-Tuning (DAFT) method based on GPT models. Specifically, the proposed approach (1) enhances domain adaptability by constructing a domain-specific emotion corpus for health misinformation; (2) reduces computational cost through parameter-efficient fine-tuning; and (3) improves task relevance by integrating optimized prompt templates for emotion recognition sub-tasks. Overall, this framework provides a practical and effective pathway for improving emotion recognition in health misinformation.

## 3. DAFT Framework for Emotion Recognition in Health Misinformation

To address the limitations of existing methods for emotion recognition in health misinformation, this study proposes a Domain-Augmented Fine-Tuning (DAFT) framework designed to enhance the domain adaptability of large language models (LLMs) in health-related scenarios. As illustrated in Figure 1, the framework consists of two key components: (1) construction of a domain-specific emotion corpus for health misinformation via emotion lexicon expansion and automatic emotion annotation; and (2) domain-enhanced fine-tuning of LLMs using optimized prompt templates and parameter-efficient tuning strategies. By integrating domain knowledge from the health sector, the DAFT framework improves emotional semantic understanding and enhances recognition accuracy for implicit and fine-grained emotional expressions in misinformation.

### *3.1. Construction of Health Misinformation Emotion Corpus*

#### 3.1.1. Corpus Collection

As summarized in Table 2, text data related to health misinformation were collected from multiple publicly available datasets on the Kaggle platform (https://www.kaggle.com/datasets (accessed on 12 March 2025)). To ensure data reliability and relevance, a two-stage filtering and validation process was employed. In the first stage, six data analysts independently conducted preliminary data cleaning to remove duplicate entries, incomplete records, and irrelevant content. The results were then cross-validated, and any inconsistencies were resolved through group discussion to ensure overall consistency. In the second stage, a review panel consisting of four domain experts in health information and misinformation analysis systematically evaluated the filtered dataset based on established criteria for identifying health misinformation [35,36]. Only samples that met the requirements for domain validity and misinformation relevance were retained, resulting in the construction of a high-quality health misinformation corpus for subsequent emotion annotation and model fine-tuning.

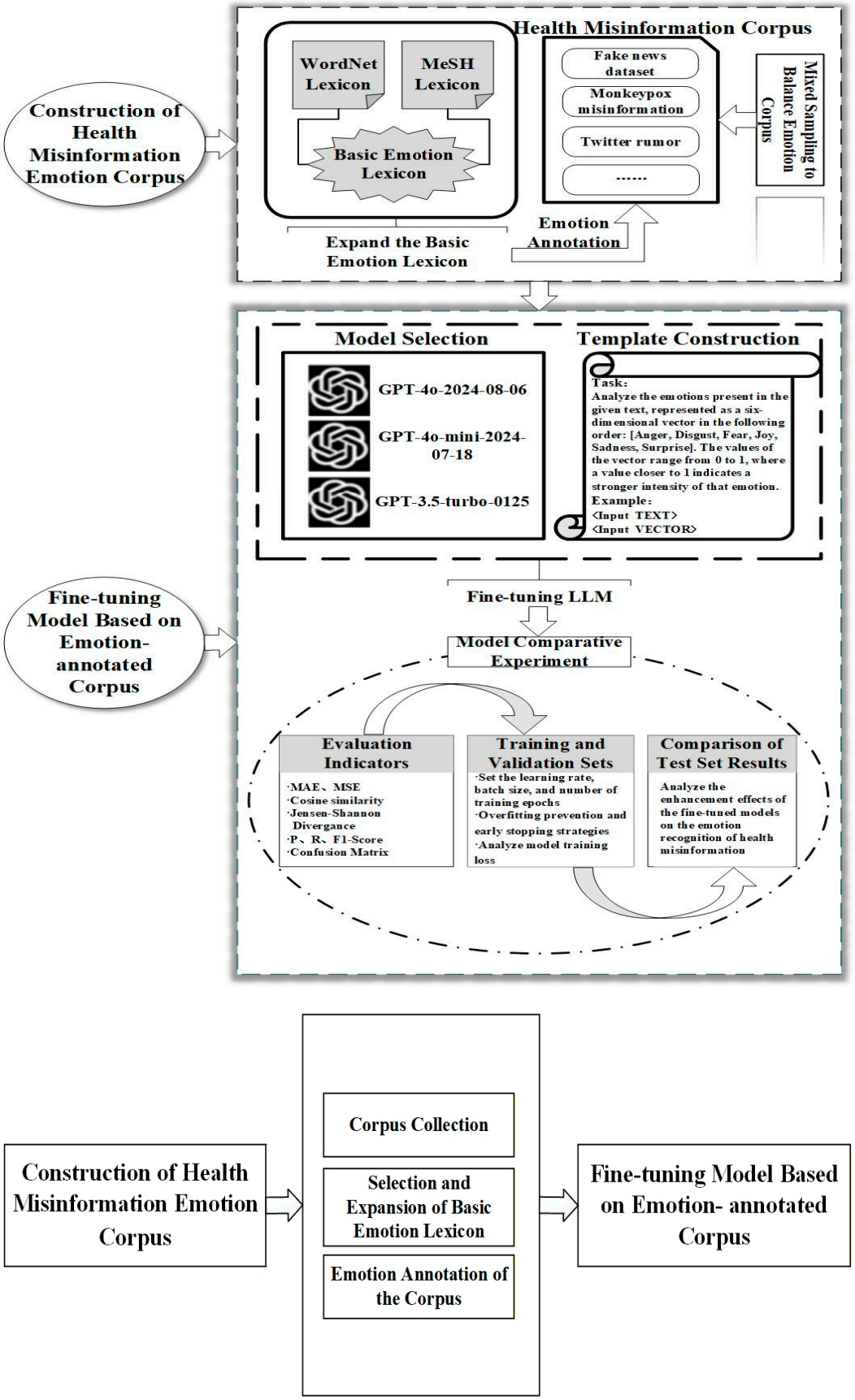



**Figure 1.** DAFT Framework for Emotion Recognition in Health Misinformation.

**Table 2.** Data Source.

| Misinformation Dataset | Type | Original Data Count and Filtered Data Count | Health Misinformation Assessment Criteria |
|---|---|---|---|
| Fake news dataset | News | 23502→2409 | ① Misleading headlines, pseudoscientific claims, and unverified reports; ② Incorrect information and claims regarding diseases; ③ Mixed disinformation, including political, medical, and conspiracy content; ④ Exaggerations regarding vaccine side effects and vaccine conspiracy theories; ⑤ False medical statements, such as incorrect treatment methods and fabricated drug names. |
| Twitter rumor | Social- Media | 3213→1120 | |
| Monkeypox misinformation | News | 6287→1163 | |
| DataSet Misinfo FAKE | News and Social-media | 55164→4463 | |
| Vaccine misinformation | News and Social-media | 1999→763 | |
| COVID fake news | News | 2140→835 | |
| CoAID | News and Social-media | 4251→1763 | |

#### 3.1.2. Selection and Expansion of Basic Emotion Lexicon

To address the limitations of traditional general-purpose emotion lexicons in health-related contexts, this study develops a domain-adapted emotion lexicon specifically for health misinformation analysis. The construction process integrates three lexical resources: the general-purpose emotion lexicon EmoLex [37], the semantic network WordNet [38], and the biomedical knowledge base MeSH (Medical Subject Headings) [39]. Lexicon expansion is carried out in two stages. First, WordNet is used to enrich the emotional vocabulary by incorporating semantically related words and synonyms for each basic emotion category, thereby improving lexical coverage. Second, domain-specific terminology from MeSH is incorporated to bridge the semantic gap between general emotional expressions and specialized health-related discourse, particularly in medical misinformation scenarios. WordNet contributes semantic completeness through lexical relations, whereas MeSH provides authoritative domain-specific conceptual mappings. Together, these resources enhance the contextual precision and domain applicability of the constructed emotion lexicon.

#### 3.1.3. Emotion Annotation of the Corpus

The emotion annotation process used in this study is illustrated in Figure 2. Consistent with prior research in affective computing, we adopt a categorical emotion model, which assumes that human emotions can be represented through a finite set of universal emotional categories [40]. Among various categorical models, Ekman's six basic emotions—anger, disgust, fear, joy, sadness, and surprise—are widely recognized for their cross-cultural validity and strong applicability in emotion classification tasks [41]. Previous studies have demonstrated that Ekman's framework provides robust semantic coverage for emotion analysis in textual data and is well suited for fine-grained emotion computation [42]. Therefore, this study adopts Ekman's six-dimensional emotional framework as the basis for text emotion representation.

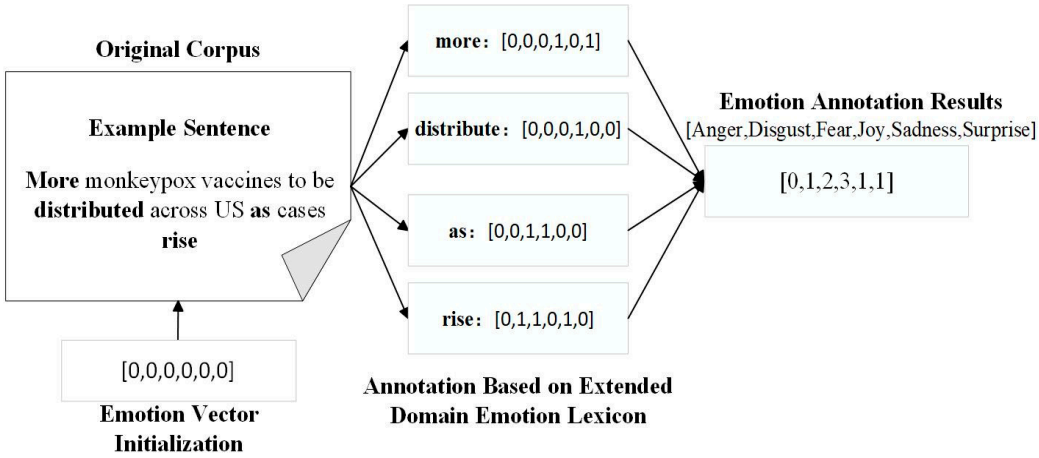



**Figure 2.** Emotion Annotation Process for the Corpus.

The objective of emotion annotation is to generate, for each text instance, an emotion vector that represents the intensity distribution across six emotion dimensions. Following established automatic emotion annotation methods [43], we first perform text preprocessing using the Stanford CoreNLP toolkit V4.5.10, including tokenization and lemmatization, to ensure lexical consistency. Each token is then matched against the expanded domain-specific emotion lexicon. When a match is found, the corresponding emotional weight is retrieved and accumulated into the emotion vector of the sentence. Finally, emotional intensity is aggregated across all matched tokens to produce a complete emotion vector for each text sample, which serves as the ground truth for model training and evaluation.

### *3.2. Fine-Tuning Model Based on Emotion-Annotated Corpus*

The first stage of the DAFT framework focuses on prompt construction, as illustrated in Figure 3. Prompts act as task instructions that guide large language models (LLMs) to generate outputs aligned with specific learning objectives. In the context of emotion recognition for health misinformation, prompt templates must incorporate domain-specific emotional cues derived from the emotion-annotated corpus to help the model accurately identify emotion categories. In this study, we adopt a task-clarification strategy for prompt design [44], enabling the model to explicitly recognize the target emotion categories. For categorical emotion prediction, prompt templates enumerate the six basic emotions (anger, disgust, fear, joy, sadness, and surprise) and provide representative domain-specific examples to improve semantic alignment during model inference. In addition, to support fine-grained emotional intensity prediction, continuous prompt templates are designed to output emotion intensity values within the range [0, 1], where 0 indicates the absence of an emotional feature and 1 denotes maximum emotional intensity. This dual template design enhances both classification precision and emotional nuance detection in health misinformation.

The second phase involves fine-tuning the LLM in a single-turn dialogue setting, where each training sample corresponds to one text segment that requires an emotion value assignment. (1) A task-specific training set is constructed and serialized in JSONL format to conform to the architectural characteristics of GPT-series models. Each record contains three fields: System (prompt defining task role and instruction scope), User (the input corpus text), and Assistant (the target output), which together specify the emotion vector assignment objective. (2) The model is then fine-tuned on this dataset to learn the mapping from input text to emotion vectors encoded in the annotations. The training process uses the OpenAI fine-tuning API, which adaptively adjusts the number of training epochs and the learning rate to improve convergence stability and task alignment.

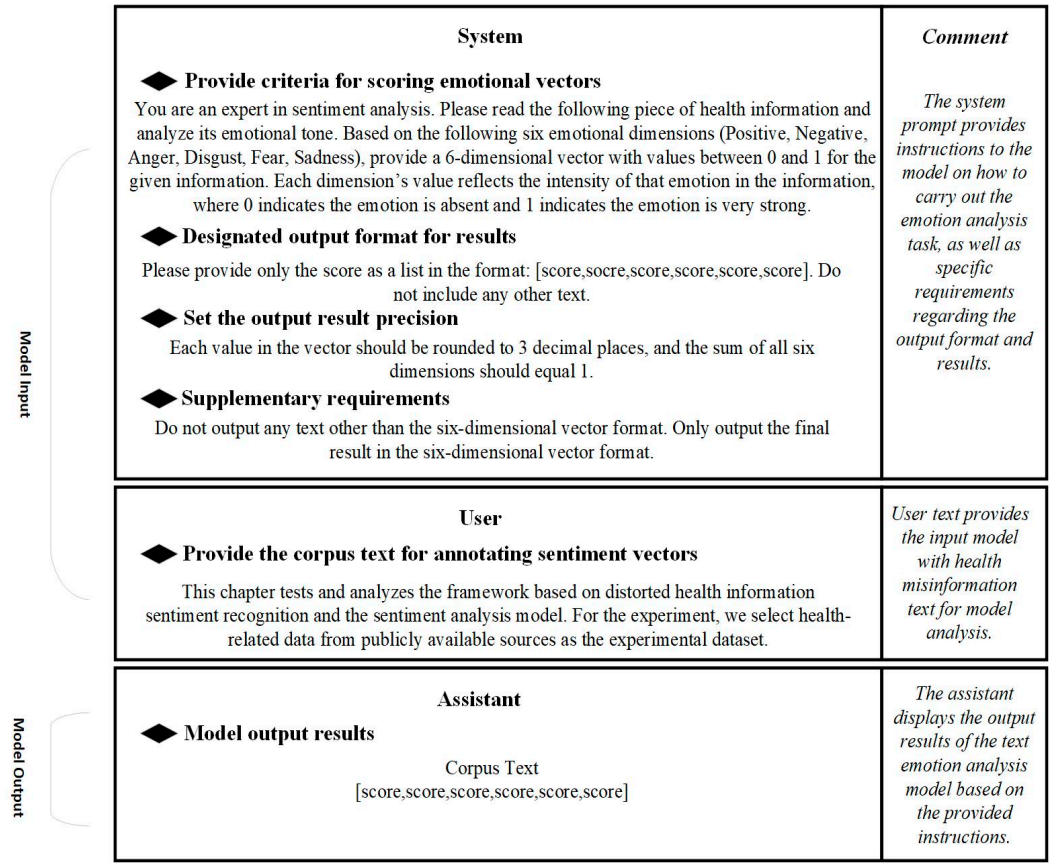



**Figure 3.** LLM Prompt Design.

### 3.3. Model Evaluation

The evaluation metrics selected for this study are presented in Table 3. The primary purpose of these metrics is to assess the accuracy, completeness, and overall performance of LLMs in determining the emotional tendencies of health misinformation. In multidimensional emotion analysis, particularly for tasks that predict emotion intensity, numerical accuracy constitutes a key criterion for evaluating model precision [45]. Health misinformation often contains complex contexts and subtle nuances, requiring the model to maintain output consistency in order to effectively identify the emotions associated with such information. A lack of consistency when faced with contextual changes may lead to erroneous judgments of emotional tendencies [46]. For emotion recognition in health misinformation, especially when the boundaries between emotion categories are relatively ambiguous, it is essential not only for the model to provide predictions of emotional tendencies but also to offer confidence estimates for those predictions. The model's ability to fit probabilities distributions reflects its capacity to handle complex emotional predictions and assists decision-makers in making more informed judgments [47]. Finally, classification accuracy is one of the foundational standards for evaluating classification models, and in emotion analysis tasks, the accurate identification and categorization of emotion categories are central to measuring the model's overall capability [48].

Therefore, this study designed an evaluation system based on four dimensions: numerical prediction accuracy, structural consistency, probabilistic fitting ability, and classification judgment accuracy, with five representative indicators selected.

**Table 3.** Model Evaluation Metrics.

| Indicator | Name | Meaning |
|---|---|---|
| Regression error metrics | MAE, MSE | Mean absolute errors and mean squared errors for each dimension |
| Structural similarity of vectors | Cosine similarity | Assessing the similarity in direction between predicted and actual distributions |
| Differences in probability distributions | Jensen-Shannon distance | Evaluating the overall distance between predicted and actual distributions |
| Classification support metrics | Precision, Recall, F1-Score | Analyzing the ability to identify whether emotion exists |
| Classification index | Confusion matrix | Analyzing misclassification cases to identify which emotion categories are prone to misclassification |

Regression error metrics: The model outputs a continuous emotional vector for each text across six emotional dimensions, categorizing it as a regression problem. Therefore, two error metrics are used to assess the model's capability in numerical prediction [49]. MAE (Mean Absolute Error) represents the average of the absolute differences between the predicted values and the actual values for each dimension, while MSE (Mean Squared Error) calculates the mean of the squared differences between the predicted and true values. Both metrics are utilized to gauge the extent of prediction errors by the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

In this formula, $y_i$ represents the actual value, $\hat{y}_i$ denotes the predicted value, and $n$ indicates the number of samples.

Structural similarity of vectors: In tasks involving multi-dimensional emotion vector outputs, it is essential to assess not only absolute errors but also the structural consistency between the predicted vector and the true emotion vector [50]. By measuring the cosine of the angle between the two vectors, it is possible to determine whether the model is capable of capturing the overall direction of emotion tendency. A value closer to 1 indicates greater consistency.

$$Cosine\ Similarity = \frac{A \cdot B}{\|A\|\|B\|} \tag{3}$$

In this formula, $A$ and $B$ are two vectors, $A \cdot B$ is their dot product, and $\|A\|$ and $\|B\|$ are their Euclidean norms.

Differences in probability distributions: The Jensen-Shannon distance is a symmetrical measure based on two probability distributions. To further assess whether the probability distribution of the model's output emotion intensity is close to the true emotion distribution, this distance metric is introduced [51]. It is based on the symmetrical form of the Kullback–Leibler Divergence (KL Divergence), which balances the information loss between the two distributions and is widely used for evaluating probabilistic models in natural language processing and machine learning.

$$Jensen - Shannon(P,\ Q) = \frac{1}{2}[D_{KL}(P\|M) + D_{KL}(Q\|M)] \tag{4}$$

In this formula, $D_{KL}(P\|M)$ represents the Kullback–Leibler Divergence between $P$ and $M$ (the average distribution of $P$ and $Q$), defined as follows:

$$D_{KL}(P\|M) = \sum_i P(i)\log\frac{P(i)}{M(i)} \quad (5)$$

Classification support metrics: Although the model outputs continuous emotion vector values, an important sub-goal of the task is to determine whether "a certain emotion exists," which can be transformed into a classification problem. Precision, Recall, and F1 score are commonly used evaluation metrics for classification tasks [52]. The F1 score is the harmonic mean of Precision and Recall, combining both metrics, and is particularly suitable for classifying imbalanced data, providing a more comprehensive assessment of model performance.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

In this formula, *Precision* represents precision, which measures the proportion of instances predicted as positive by the model that are actually positive; *Recall* represents recall, which measures the proportion of all actual positive instances that the model correctly predicts as positive.

Classification index: The confusion matrix is used to visualize the performance of a classifier across different classification categories [53]. By tallying the correctly and incorrectly classified emotions, it highlights instances where each true emotion is misclassified as another by the model. This is particularly useful in cases of class imbalance, as it aids in analyzing the results and causes of misclassification.

## 4. Experimental Results and Analysis

### *4.1. Experimental Data*

#### 4.1.1. Dataset Construction

To support model optimization and fine-tuning, a task-specific dataset was constructed consisting of 7277 training samples, 1250 validation samples, and 2267 test samples. The training set was used to enable the GPT model to learn emotional representations, while the validation set was used for hyperparameter tuning and to trigger an early stopping mechanism during training to prevent overfitting. The test set was reserved for independent evaluation on unseen data to ensure the objectivity and robustness of the model's emotion recognition performance.

#### 4.1.2. Dataset Balancing Processing

Table 4 presents the distribution of the primary emotional dimensions in the dataset. Health-related misinformation is predominantly associated with negative emotional expressions, making negative emotion the dominant category, while positive emotion occurs far less frequently [7]. In machine learning, data imbalance refers to the unequal distribution of class samples within a dataset. To alleviate this issue, oversampling and undersampling techniques are commonly used to address class imbalance and improve model performance [54]. In this study, undersampling was applied to reduce the proportion of negative-emotion samples, while oversampling was used to increase the number of positive-emotion samples, resulting in a mixed sampling strategy to balance the health misinformation dataset.

**Table 4.** The Frequency of the Main Emotions.

| Dimension | Occurrence Count (Before/After Balancing) | | Relative Proportion (Before/After Balancing) | |
|---|---|---|---|---|
| Anger | 1768 | 1768 | 14.35% | 15.96% |
| Disgust | 1460 | 1460 | 11.85% | 13.18% |
| **Fear** | **3998** | **2928** | **32.45%** | **26.43%** |
| Joy | 1563 | 2196 | 12.69% | 14.11% |
| Sadness | 2611 | 1963 | 21.19% | 17.72% |
| **Surprise** | **919** | **1397** | **7.46%** | **12.61%** |

As a result of this mixed sampling procedure, the distribution of negative and positive emotions is made more balanced. In particular, oversampling increased the representation of samples labeled with Joy and Surprise, which helps reduce model bias toward majority classes. This strategy prevents the model from over-relying on any single emotion category and enables it to learn more effectively across all emotion dimensions, thereby improving its ability to recognize a wider range of emotions.

### *4.2. Model Parameter Settings*

To standardize reporting, the model names before and after fine-tuning are unified as shown in Table 5. For model selection, we directly used the official APIs provided by OpenAI. Given that the corpus is in English and GPT-5 is currently not available for fine-tuning, we evaluated three OpenAI models that support fine-tuning: GPT-4o-2024-08-06, GPT-4o-mini-2024-07-18, and GPT-3.5-turbo-0125.

**Table 5.** Model Name.

| Original Model Name | Model Name Before Fine-Tuning | Fine-Tuned Model Name |
|---|---|---|
| GPT-4o-2024-08-06 | GPT-4o | 4o-FT |
| GPT-4o-mini-2024-07-18 | GPT-4o-mini | 4o-mini-FT |
| GPT-3.5-turbo-0125 | GPT-3.5-turbo | 3.5-turbo-FT |

Table 6 summarizes the hyperparameters used for fine-tuning. We primarily configured the learning rate, batch size, and number of epochs while keeping all other settings at their defaults. Considering the very large parameter scales of GPT models, an overly high learning rate can destabilize training; therefore, we initialized the learning rate at $2 \times 10^{-5}$ and applied a warm-up schedule with dynamic decay to $1 \times 10^{-5}$, thereby preserving pretraining capacity while adapting the model to the new task. The batch size was set to 32, which balances memory usage and throughput without inducing overfitting. The model was trained for 8 epochs with early stopping, using validation performance on 1250 samples as the stopping criterion: if the validation loss failed to improve for several consecutive epochs, training was terminated to prevent overfitting and reduce computational cost.

**Table 6.** Parameters Setting for Fine-tuning the Model.

| Parameter Name | Parameter Settings |
|---|---|
| model | in Table 4 |
| batch size | 32 |
| learning rate | $2 \times 10^{-5}$ |
| epoch | 8 |

### *4.3. Validation Set Performance Monitoring*

Figure 4 depicts the training loss curves for the three models. During fine-tuning, stepwise training losses were recorded and compared, with validation-set performance used to monitor and guide optimization. The results indicate that GPT-4o achieved the lowest training loss (0.02) and ultimately converged to a lower final loss (0.16) than the other models, demonstrating superior fit and convergence efficiency. The GPT-4o-mini curve is the most stable, exhibiting the lowest average training loss across steps and indicating steady optimization behavior. By contrast, GPT-3.5-turbo presents higher losses and larger fluctuations than the other two models, reflecting comparatively weaker training outcomes. With validation monitoring and early stopping in place, all three models avoided overfitting and achieved stable validation performance by the end of training.

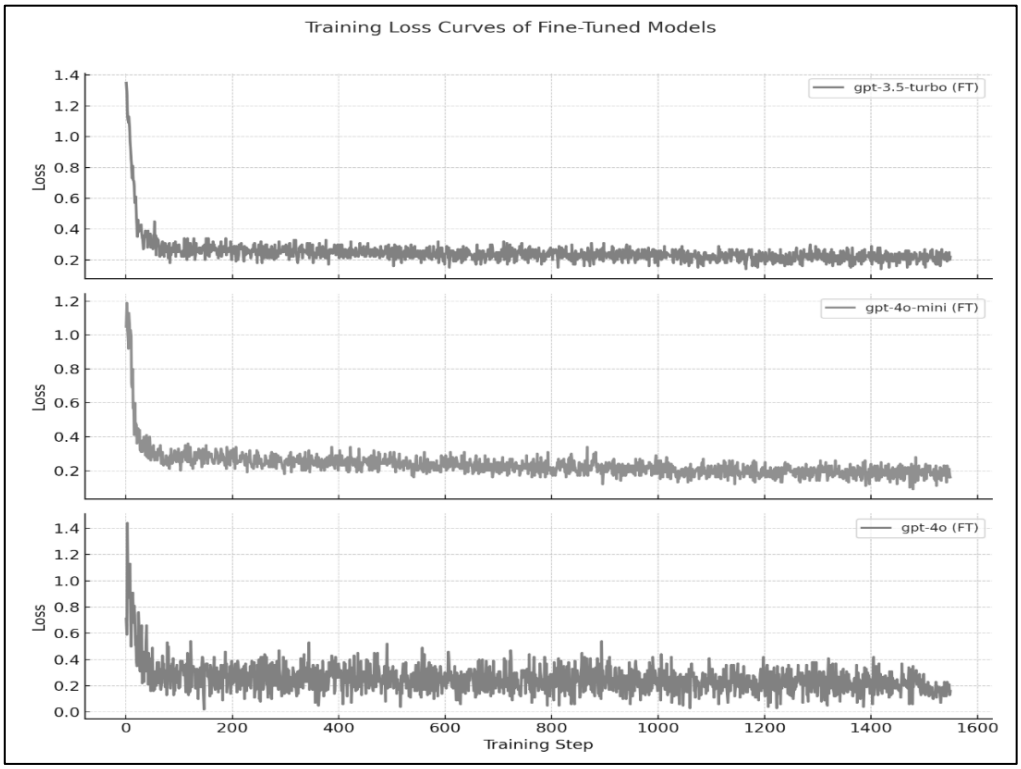



**Figure 4.** Model Training Loss Curve.

### *4.4. Model Comparison Experiment*

To evaluate the effectiveness of the proposed domain-enhanced fine-tuning method for emotion recognition in health misinformation, three baseline models—GPT-4o, GPT-4o-mini, and GPT-3.5-turbo—were fine-tuned and their performance was compared across 2267 test samples using multiple evaluation metrics.

Table 7 compares the MAE and MSE results of the three models—GPT-4o, GPT-4o-mini, and GPT-3.5-turbo—for emotion recognition in health misinformation. Bold values indicate the lowest MAE or MSE within each emotion dimension.

The GPT-4o model consistently achieves the lowest MAE and MSE values across the six emotional dimensions, particularly excelling in Disgust, Joy, and Surprise. These results indicate that, after fine-tuning, GPT-4o can identify diverse emotions with higher precision, reflecting stronger fitting capability and lower prediction error.

**Table 7.** Comparison of MAE and MSE among six models.

| Emotion Dimension | MAE | | | | | | MSE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **4o-FT** | **GPT-4o** | **4o-mini-FT** | **GPT-4o-mini** | **3.5-Turbo-FT** | **GPT-3.5-turbo** | **4o-FT** | **GPT-4o** | **4o-mini-FT** | **GPT-4o-mini** | **3.5-Turbo-FT** | **GPT-3.5-Turbo** |
| Anger | **0.104** | 0.132 | 0.105 | 0.157 | 0.121 | 0.141 | **0.024** | 0.029 | 0.025 | 0.039 | 0.030 | 0.035 |
| Disgust | **0.085** | 0.118 | 0.090 | 0.126 | 0.101 | 0.161 | **0.018** | 0.023 | 0.019 | 0.026 | 0.021 | 0.048 |
| Fear | **0.125** | 0.154 | 0.126 | 0.147 | 0.133 | 0.161 | **0.034** | 0.041 | **0.034** | 0.036 | 0.035 | 0.044 |
| Joy | **0.111** | 0.168 | 0.117 | 0.143 | 0.125 | 0.183 | **0.034** | 0.064 | 0.035 | 0.044 | 0.037 | 0.070 |
| Sadness | **0.123** | 0.151 | 0.125 | 0.167 | 0.131 | 0.172 | 0.035 | 0.037 | **0.034** | 0.043 | **0.034** | 0.044 |
| Surprise | **0.104** | 0.159 | **0.104** | 0.240 | 0.135 | 0.173 | 0.029 | 0.054 | **0.027** | 0.111 | 0.041 | 0.060 |

In contrast, although GPT-4o-mini performs slightly below GPT-4o overall, it demonstrates smaller errors and greater stability across most emotion categories. Its performance in the Disgust and Surprise dimensions is especially stable, suggesting strong adaptability. The GPT-3.5-turbo model shows only modest gains after fine-tuning—particularly limited improvements in Joy and Surprise—with marginal enhancements in Fear and Sadness. Overall, GPT-3.5-turbo remains comparatively weaker in fine-grained emotion recognition, highlighting existing constraints in capturing nuanced emotional variations.

Table 8 compares the cosine similarity of emotional vectors generated by the three models before and after fine-tuning. The 4o-FT model achieves the best overall performance, with cosine similarity increasing from 0.6532 to 0.7818 (+0.1286), indicating a stronger alignment between predicted and true emotion vectors. The 4o-mini-FT model also demonstrates substantial improvement, increasing from 0.5972 to 0.7731 (+0.1759), reflecting effective task adaptation. Although 3.5-turbo-FT shows some improvement (from 0.6022 to 0.7266), its performance remains lower than the other two models, suggesting limited structural consistency. Bold values indicate, for the three models, the highest cosine similarity before and after fine-tuning, as well as the largest improvement.

**Table 8.** Comparison of Model Cosine Similarity.

| Model Name | Before Fine-Tuning | After Fine-Tuning | Enhancement in Value |
|---|---|---|---|
| GPT-4o | **0.6532** | **0.7818** | +0.1286 |
| GPT-4o-mini | 0.5972 | 0.7731 | **+0.1759** |
| GPT-3.5-turbo | 0.6022 | 0.7266 | +0.1244 |

Overall, all models achieved varying degrees of improvement after fine-tuning, validating the effectiveness of the fine-tuning strategy. Fine-tuning not only enhanced the models' ability to capture emotional patterns but also strengthened the semantic alignment between predicted and true emotion vectors, resulting in emotional representations that are more directionally consistent with the ground truth.

Table 9 compares the Jensen–Shannon Distance (JSD) between the predicted emotional distributions of the three models and the true emotional distribution before and after fine-tuning. All models show a notable reduction in JSD following fine-tuning, indicating that the fine-tuning process effectively brought the predicted emotional distributions closer to the ground truth. Among them, the 4o-FT model achieves the lowest JSD value (0.3909), demonstrating the closest match to the true emotional distribution and thus the best overall performance. The GPT-4o-mini model records the largest reduction in JSD (0.1498), reflecting strong adaptability to health misinformation data and effective optimization of emotional prediction. In comparison, the GPT-3.5-turbo model shows a

smaller JSD reduction (0.0997), though still indicating measurable improvement. Bold values indicate, for the three models, the lowest JSD before and after fine-tuning, as well as the largest improvement.

**Table 9.** Comparison of Differences in Probability Distributions (JSD).

| Model Name | Before Fine-Tuning | After Fine-Tuning | Difference |
|---|---|---|---|
| GPT-4o | **0.4891** | **0.3909** | −0.0982 |
| GPT-4o-mini | 0.5513 | 0.4015 | **−0.1498** |
| GPT-3.5-turbo | 0.5398 | 0.4401 | −0.0997 |

Overall, fine-tuning improves the emotional distribution alignment of all three models, significantly reducing divergence between predicted and true emotional vectors. The reduction in JSD values demonstrates that the fine-tuned models not only better identify emotional categories but also more accurately represent emotional intensity across dimensions.

Table 10 and Figure 5 summarize the auxiliary classification metrics of the three fine-tuned models across six emotional dimensions. Based on the F1-scores, the 4o-FT model achieves the highest performance in Disgust, Joy, Sadness, and Surprise, demonstrating strong overall recognition capability. The 4o-mini-FT model performs best in the Anger and Fear dimensions and shows relatively stable recognition across categories. In contrast, the 3.5-turbo-FT model shows only comparable performance in the Fear category and lags notably behind the other two models in Surprise, indicating limited capability in recognizing nuanced emotional categories. Overall, the 4o-FT model provides the most stable performance across most emotional categories, followed by 4o-mini-FT, while 3.5-turbo-FT demonstrates the weakest results. These findings further confirm the effectiveness of fine-tuning in improving emotional classification accuracy and label coverage.

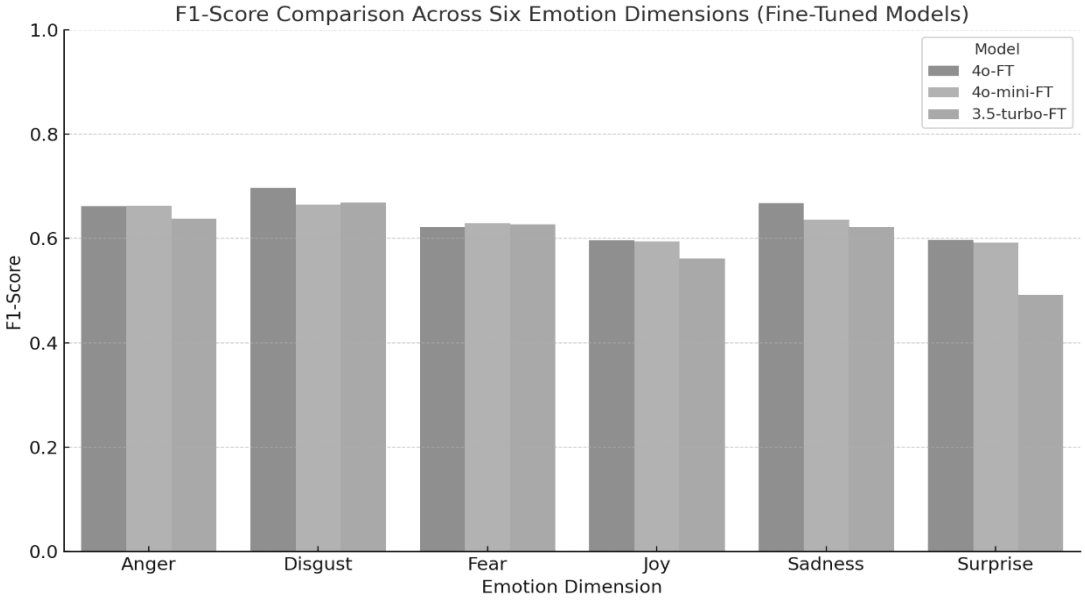



**Figure 5.** Comparison of F1-Scores of the Three Models After Fine-Tuning.

As shown in Table 11, this section systematically analyzes the impact of fine-tuning on model performance by comparing the differences in F1-scores before and after fine-tuning for different models, using statistical significance testing and confidence intervals. It can be observed that the $p$-values for the 4o model and the 4o-mini model are both less than 0.05, and the confidence intervals do not include 0, indicating that the differences before and after fine-tuning are statistically significant. The average differences for both models are

greater than 0.1, suggesting that fine-tuning improved model performance. In contrast, for the 3.5-turbo model, the differences observed are not statistically significant, as indicated by the *p*-value and 95% confidence interval. However, the average score improved by 0.0607 after fine-tuning, showing that the fine-tuning strategy did improve the model's performance, but the improvement was not statistically significant. This finding aligns with our observations in other metric tests, where the overall performance of fine-tuned 3.5-turbo was not as good as the other two models. This also suggests that for the relatively underperforming 3.5-turbo model, better fine-tuning methods should be employed to enhance the model's performance.

**Table 10.** Precision, Recall, and F1-Score of Six Models.

| Model Name | Emotion Dimension | Precision | Recall | F1-Score |
|---|---|---|---|---|
| 4o-FT | Anger | 0.852 | 0.541 | 0.662 |
| | Disgust | 0.852 | 0.590 | 0.697 |
| | Fear | 0.912 | 0.472 | 0.622 |
| | Joy | 0.816 | 0.469 | 0.596 |
| | Sadness | 0.902 | 0.531 | 0.668 |
| | Surprise | 0.752 | 0.495 | 0.597 |
| 4o-mini-FT | Anger | 0.835 | 0.550 | 0.663 |
| | Disgust | 0.835 | 0.553 | 0.665 |
| | Fear | 0.903 | 0.483 | 0.629 |
| | Joy | 0.782 | 0.479 | 0.594 |
| | Sadness | 0.878 | 0.499 | 0.636 |
| | Surprise | 0.752 | 0.488 | 0.592 |
| 3.5-turbo-FT | Anger | 0.808 | 0.527 | 0.638 |
| | Disgust | 0.789 | 0.580 | 0.669 |
| | Fear | 0.879 | 0.487 | 0.627 |
| | Joy | 0.730 | 0.457 | 0.562 |
| | Sadness | 0.855 | 0.489 | 0.622 |
| | Surprise | 0.612 | 0.412 | 0.492 |
| GPT-4o | Anger | 0.825 | 0.483 | 0.609 |
| | Disgust | 0.748 | 0.507 | 0.604 |
| | Fear | 0.885 | 0.466 | 0.610 |
| | Joy | 0.698 | 0.236 | 0.352 |
| | Sadness | 0.881 | 0.342 | 0.493 |
| | Surprise | 0.580 | 0.410 | 0.480 |
| GPT-4o-mini | Anger | 0.834 | 0.414 | 0.553 |
| | Disgust | 0.793 | 0.463 | 0.585 |
| | Fear | 0.892 | 0.430 | 0.580 |
| | Joy | 0.752 | 0.216 | 0.335 |
| | Sadness | 0.870 | 0.271 | 0.414 |
| | Surprise | 0.543 | 0.362 | 0.434 |
| GPT-3.5-turbo | Anger | 0.810 | 0.459 | 0.586 |
| | Disgust | 0.690 | 0.481 | 0.567 |
| | Fear | 0.889 | 0.471 | 0.615 |
| | Joy | 0.655 | 0.277 | 0.389 |
| | Sadness | 0.849 | 0.472 | 0.607 |
| | Surprise | 0.623 | 0.393 | 0.482 |

**Table 11.** Statistical significance testing and confidence intervals of F1-Score.

| Pair of Models | Paired *t*-Test | *p*-Value | F1-Score Mean Difference | 95% Confidence Interval |
|---|---|---|---|---|
| 4o-FT, 4o | 3.121 | 0.0262 | 0.1103 | (0.0195, 0.2012) |
| 4o-mini-FT, 4o-mini | 6.016 | 0.0018 | 0.1042 | (0.0597, 0.1487) |
| 3.5-turbo-FT, 3.5-turbo | 2.270 | 0.0724 | 0.0607 | (−0.0080, 0.1294) |

Table 12 presents the confusion matrices of emotion classification results for GPT-4o, GPT-4o-mini, and GPT-3.5-turbo before and after fine-tuning. The GPT-4o-FT model achieves the best overall performance in multi-dimensional emotion recognition, accurately identifying primary emotions such as Anger, Fear, and Sadness. The GPT-4o-mini-FT model also demonstrates good recognition capability in Fear and Sadness, although its performance declines for other emotion categories. In contrast, the GPT-3.5-turbo-FT model performs the weakest among the three, showing the lowest recognition accuracy for most primary emotions. These results indicate that GPT-3.5-turbo has limited capability in handling fine-grained emotional distinctions in multidimensional emotion recognition tasks, suggesting that further optimization is required to improve its primary emotion identification accuracy. We bold the maximum value in each column of the table to indicate, for each model, the highest number of emotions identified in each emotion dimension before and after fine-tuning.

**Table 12.** Confusion Matrix.

| Prediction / Real Numbers | Anger | | Disgust | | Fear | | Joy | | Sadness | | Surprise | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **4o-FT** | **GPT-4o** | **4o-FT** | **GPT-4o** | **4o-FT** | **GPT-4o** | **4o-FT** | **GPT-4o** | **4o-FT** | **GPT-4o** | **4o-FT** | **GPT-4o** |
| Anger | **219** | **118** | 32 | 40 | 156 | 177 | 35 | 72 | 65 | 32 | 20 | 88 |
| Disgust | 41 | 43 | **58** | **36** | 81 | 81 | 23 | 29 | 33 | 9 | 7 | 45 |
| Fear | 141 | 187 | 80 | 71 | **376** | **288** | 60 | 129 | 142 | 42 | 22 | 104 |
| Joy | 48 | 41 | 17 | 20 | 70 | 82 | **60** | **48** | 25 | 8 | 9 | 30 |
| Sadness | 35 | 108 | 19 | 28 | 147 | 109 | 24 | 55 | **143** | **33** | 13 | 48 |
| Surprise | 16 | 11 | 5 | 6 | 13 | 29 | 4 | 7 | 15 | 4 | **13** | **9** |
| | 4o-mini-FT | GPT-4o-mini | 4o-mini-FT | GPT-4o-mini | 4o-mini-FT | GPT-4o-mini | 4o-mini-FT | GPT-4o-mini | 4o-mini-FT | GPT-4o-mini | 4o-mini-FT | GPT-4o-mini |
| Anger | **51** | **53** | 13 | 3 | 58 | 30 | 15 | 6 | 13 | 4 | 1 | 33 |
| Disgust | 16 | 17 | **33** | **0** | 37 | 13 | 10 | 1 | 7 | 1 | 0 | 23 |
| Fear | 24 | 164 | 16 | 15 | **321** | **126** | 36 | 22 | 92 | 23 | 14 | 134 |
| Joy | 6 | 26 | 4 | 3 | 47 | 42 | **48** | **6** | 30 | 6 | 7 | 56 |
| Sadness | 10 | 119 | 12 | 1 | 115 | 80 | 19 | 14 | **165** | **22** | 17 | 101 |
| Surprise | 5 | 24 | 3 | 0 | 18 | 15 | 3 | 2 | 22 | 2 | **18** | **31** |
| | 3.5-Turbo-FT | GPT-3.5-turbo | 3.5-Turbo-FT | GPT-3.5-turbo | 3.5-Turbo-FT | GPT-3.5-turbo | 3.5-Turbo-FT | GPT-3.5-turbo | 3.5-Turbo-FT | GPT-3.5-turbo | 3.5-Turbo-FT | GPT-3.5-turbo |
| Anger | **49** | **27** | 6 | 31 | 37 | 24 | 13 | 23 | 31 | 4 | 12 | 18 |
| Disgust | 15 | 11 | **14** | **12** | 22 | 11 | 4 | 15 | 14 | 1 | 2 | 8 |
| Fear | 65 | 103 | 30 | 90 | **195** | **145** | 31 | 65 | 130 | 16 | 35 | 78 |
| Joy | 11 | 17 | 4 | 33 | 29 | 30 | **26** | **33** | 42 | 4 | 22 | 27 |
| Sadness | 32 | 61 | 11 | 64 | 54 | 92 | 22 | 56 | **143** | **17** | 52 | 68 |
| Surprise | 8 | 5 | 2 | 18 | 11 | 12 | 1 | 8 | 27 | 2 | **24** | **19** |

In summary, the experimental results demonstrate that the domain-enhanced fine-tuning strategy significantly improves the emotion representation and recognition capabilities of large language models in health misinformation tasks. Among the models tested, GPT-4o-FT achieves the highest overall performance, showing consistent advantages across prediction error metrics (MAE and MSE), structural similarity (Cosine), distributional consistency (JSD), and primary emotion classification (F1-score and confusion matrix analysis). The GPT-4o-mini-FT model exhibits moderate yet stable performance, while GPT-3.5-turbo-FT shows only limited improvement, indicating weaker adaptability to domain-specific emotional semantics. These findings confirm the effectiveness of the proposed DAFT method and provide a solid foundation for deeper analysis in the following section.

## 5. Discussion

This study evaluates the performance of three large language models—GPT-4o, GPT-4o-mini, and GPT-3.5-turbo—fine-tuned using the proposed DAFT method for the task of emotion recognition in health misinformation.

From the comparative analysis of the experimental results, three conclusions can be drawn:

(1) Effectiveness of fine-tuning: All models showed varying degrees of improvement after fine-tuning, with performance gains most evident in the recognition of negative emotions. The 4o-FT model achieved the lowest prediction error for negative emotions such as anger, fear, and sadness. This may be attributed to the dominance of negative emotions in health misinformation, which leads to a stronger representation of negative emotional features in the dataset. In addition, the GPT-4o model benefited from large-scale pretraining, which enhanced its capacity to recognize negative emotional cues.

(2) Model capability comparison: GPT-3.5-turbo performed worse than GPT-4o and GPT-4o-mini in emotion recognition, especially in identifying positive emotions such as joy and surprise, where its accuracy was significantly lower. This suggests that smaller pretrained models face challenges in capturing subtle emotional differences in multidimensional emotion recognition tasks, which negatively affects their overall performance.

(3) Impact of data balancing: The mixed sampling strategy effectively addressed emotional category imbalance in the dataset by increasing the number of positive emotion samples (e.g., joy and surprise) and reducing training bias. The confusion matrix results further demonstrate that, although negative emotions remain predominant, the distribution of true and predicted emotion labels became more consistent after fine-tuning, indicating that mixed sampling enhances the model's ability to learn emotional features across categories.

Based on the experimental results, three key findings can be summarized:

(1) The DAFT strategy significantly enhanced model performance in emotion recognition for health misinformation, improving prediction accuracy, precision, and emotional granularity. After fine-tuning, all three GPT models achieved reductions in MAE and MSE, consistent with findings that fine-tuning enhances task adaptation and semantic alignment in LLMs [12]. The F1-score of GPT-4o increased by 0.1156, and its accuracy improved from 76.95% to 84.77%, while GPT-4o-mini showed improved vector similarity (+0.1759) and reduced probability distribution divergence ($-0.1498$), demonstrating effective emotional representation learning.

(2) Model performance is positively correlated with pre-training scale, as larger-scale models can encode richer emotional semantics and contextual knowledge, leading to better task generalization [13]. This trend is evident in the confusion matrix analysis:

the GPT-4o model achieved higher recognition accuracy for primary emotions (219 Anger and 376 Fear correctly identified), with close alignment between predicted and true emotion counts. In contrast, the smaller GPT-3.5-turbo model struggled to capture complex emotional structures, correctly predicting only an average of 29 instances across Anger, Disgust, and Joy categories.

(3) Balanced data sampling improves performance consistency across emotional categories. The mixed sampling strategy effectively mitigated overrepresentation of negative emotions and enabled fairer learning across emotional dimensions, in line with previous work emphasizing the importance of addressing data imbalance in classification tasks [54]. Although variation in performance across emotion categories was observed, no extreme disparities were found. For example, the 4o-FT model reached a maximum accuracy of 0.912 and a minimum of 0.752 across emotional dimensions, with a median recall of 0.5 and only 0.101 variation between the highest and lowest F1-scores. These results suggest that balancing data distribution improves the overall stability of model recognition performance.

While the DAFT method significantly improves emotion recognition in health misinformation, there are several limitations and challenges that must be considered: (1) The dataset used for fine-tuning models in this study is restricted to textual health misinformation, which may limit the model's generalizability to more diverse forms of misinformation, including multimodal content such as images and videos; (2) Although this study focuses on publicly available datasets, there is still a risk of unintended data leakage or privacy violations, particularly when using social media-based datasets. Ensuring proper data handling and compliance with privacy standards is crucial in mitigating these risks; (3) Health misinformation is often not limited to text; images, videos, and audio are also widely used to spread misleading content. The DAFT method, which currently focuses on text-based misinformation, might face challenges when applied to multimodal misinformation detection. Despite these challenges, the DAFT framework represents a promising step forward in emotion recognition for health misinformation and provides a foundation for future research aimed at addressing these limitations.

In conclusion, the DAFT method proposed in this study effectively enhances the emotion recognition capabilities of large language models in the domain of health misinformation. Among the evaluated models, GPT-4o-FT achieved the best overall performance across all evaluation metrics, exhibiting high prediction accuracy, consistent emotional distribution alignment, and strong classification capability, making it the best-performing model in this experiment. These findings further validate the effectiveness of applying domain-adapted fine-tuning strategies to improve LLM performance in specialized tasks [12].

## 6. Conclusions

This study investigates emotion recognition in health misinformation and proposes a domain-enhanced fine-tuning method based on large language models (LLMs). The approach improves the emotional understanding capability of LLMs through task-oriented prompt design and requires only a small amount of training data to achieve stable performance. The experimental results confirm the effectiveness of the proposed method in emotion recognition for health misinformation. After applying DAFT, the models demonstrate improved accuracy in emotion label generation, reduced dependence on large annotated datasets, and enhanced classification efficiency. This study provides a feasible technical solution for emotion analysis in health misinformation and contributes to the development of lightweight fine-tuning strategies for LLMs.

The significance of this study is reflected in three main aspects:

(1) The proposed DAFT method effectively addresses challenges of data imbalance and domain adaptation in health misinformation emotion recognition, thereby improving the precision and robustness of LLMs in this task. (2) The method reduces data dependence and computational cost while maintaining model efficiency, providing a practical and generalizable fine-tuning strategy for LLM applications in low-resource scenarios. (3) The emotionally annotated health misinformation corpus constructed in this study provides a valuable resource for future research, with the potential to be extended to include additional emotion dimensions and more nuanced emotional categories, further enhancing multidimensional emotion modeling.

This study has constructed an emotion-annotated health misinformation corpus, which serves as a foundation for future work. In subsequent research, the scale of the corpus may be further expanded to cover additional emotional dimensions and more complex, subtle emotional categories, thereby improving the model's capacity for fine-grained emotion recognition. Furthermore, with the increasing diversity of social media content, the integration of multimodal information—including text, images, and videos—has become an important research direction in emotion analysis. Future studies may explore combining LLMs with multimodal fusion techniques to enhance emotion recognition in complex, real-world health misinformation scenarios.

**Author Contributions:** Conceptualization, Y.Z., L.Z., L.F.; Methodology, Y.Z. and X.Z.; Software, X.Z. and L.F.; Validation, W.T. and M.T.; Formal analysis, Y.Z., W.T., and L.F.; Investigation, L.F., X.Z. and L.Z.; Resources, Y.Z., L.Z., and M.T.; Data curation, Y.Z. and X.Z.; Writing—original draft preparation, Y.Z. and X.Z.; Writing—review and editing, L.F., W.T., and L.Z.; Visualization, W.T. and X.Z.; Supervision, L.Z. and M.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the National Social Science Foundation General Project (Grant No. 21BTQ055).

**Data Availability Statement:** Data will be made available from the corresponding author upon reasonable request, subject to privacy and licensing constraints.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Muhtar, S.M.; Amir, A.S. Utilizing social media for public health advocacy and awareness in digital health communication. *MSJ Major. Sci. J.* **2024**, *1*, 270–278. [CrossRef]
2. Islam, A.K.M.N.; Laato, S.; Talukder, S.; Sutinen, E. Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technol. Forecast. Soc. Change* **2020**, *159*, 120201. [CrossRef] [PubMed]
3. Zhou, X.; Jain, A.; Phoha, V.V.; Zafarani, R. Fake News Early Detection: An Interdisciplinary Study. *arXiv* **2019**, arXiv:1904.11679.
4. Chen, C.; Wang, H.; Shapiro, M.; Xiao, Y.; Wang, F.; Shu, K. Combating Health Misinformation in Social Media: Characterization, Detection, Intervention, and Open Issues. *arXiv* **2022**, arXiv:2211.05289. [CrossRef]
5. Robert, B.J.; Yan, Z.; Jacek, G. Predicting healthcare professionals' intention to correct health misinformation on social media. *Telemat. Inform.* **2022**, *73*, 101864.
6. Liu, Z.; Zhang, T.; Yang, K.; Thompson, P.; Yu, Z.; Ananiadou, S. Emotion detection for misinformation: A review. *Inf. Fusion* **2024**, *107*, 102300. [CrossRef]
7. Horner, C.G.; Galletta, D.; Crawford, J.; Shirsat, A. Emotions: The unexplored fuel of fake news on social media. In *Fake News on the Internet*; Routledge: London, UK, 2023; pp. 147–174.
8. Jawaher, A.; Suhuai, L.; Yuqing, L. A comprehensive survey on machine learning approaches for fake news detection. *Multimed. Tools Appl.* **2024**, *17*, 51009–51067.
9. Raymond, C.; Gregorious, S.B.; Sandeep, D. Combining Sentiment Lexicons and Content-Based Features for Depression Detection. *IEEE Intell. Syst.* **2021**, *36*, 99–105. [CrossRef]
10. Brauwers, G.; Frasincar, F. A Survey on Aspect-Based Sentiment Classification. *ACM Comput. Surv.* **2022**, *4*, 1–37. [CrossRef]

11. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
12. Peng, C.; Yang, X.; Smith, K.E.; Yu, Z.; Chen, A.; Bian, J.; Wu, Y. Model Tuning or Prompt Tuning a Study of Large Language Models for Clinical Concept and Relation Extraction. *J. Biomed. Inform.* **2024**, *153*, 104630. [CrossRef] [PubMed]
13. Parthasarathy, V.B.; Zafar, A.; Khan, A.; Shahid, A. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv* **2024**, arXiv:2408.13296. [CrossRef]
14. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv* **2023**, arXiv:2302.11382. [CrossRef]
15. Maharjan, J.; Garikipati, A.; Preet Singh, N.; Cyrus, L.; Sharma, M.; Ciobanu, M.; Barnes, G.; Thapa, R.; Mao, Q.; Das, R. OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci. Rep.* **2024**, *1*, 14156. [CrossRef] [PubMed]
16. Xu, H.; Elbayad, M.; Murray, K.; Maillard, J.; Goswami, V. Towards Being Parameter-Efficient: A Stratified Sparsely Activated Transformer with Dynamic Capacity. *arXiv* **2023**, arXiv:2305.02176. [CrossRef]
17. Fernandez, M.; Alani, H. Online misinformation: Challenges and future directions. In Proceedings of the Companion of the Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 595–602.
18. Victoria, L.R. Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *J. Doc.* **2019**, *5*, 1013–1034.
19. Karami, M.; Nazer, T.H.; Liu, H. Profiling Fake News Spreaders on Social Media through Psychological and Motivational Factors. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media, Online, 30 August–2 September 2021; pp. 225–230.
20. Bordia, P.; Difonzo, N. Rumor, Gossip and Urban Legends. *Diogenes* **2007**, *1*, 19–35. [CrossRef]
21. Wang, B.; Ma, J.; Lin, H.; Yang, Z.; Yang, R.; Tian, Y.; Chang, Y. Explainable fake news detection with large language model via defense among competing wisdom. In Proceedings of the ACM Web Conference 2024, Association for Computing Machinery, Singapore, 13–17 May 2024; pp. 2452–2463.
22. Yuxi, W.; Martin, M.; Aleksandra, T.; David, S. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Soc. Sci. Med.* **2019**, *240*, 112552. [CrossRef]
23. Sezer, K.; Adnan, K. A comprehensive analysis of COVID-19 misinformation, public health impacts, and communication strategies: Scoping review. *J. Med. Internet Res.* **2024**, *26*, e56931.
24. Balshetwar, S.V.; Abilash, R.S.; Dani, J. Fake news detection in social media based on sentiment analysis using classifier techniques. *Multimed. Tools Appl.* **2023**, *82*, 35781–35811. [CrossRef]
25. Kaur, R.; Kautish, S. Multimodal Sentiment Analysis: A Survey and Comparison. *Int. J. Serv. Sci. Manag. Eng. Technol. IJSSMET* **2019**, *10*, 38–58. [CrossRef]
26. Chen, H.; Zheng, P.; Wang, X.; Hu, S.; Zhu, B.; Hu, J.; Wu, X.; Lyu, S. Harnessing the Power of Text-image Contrastive Models for Automatic Detection of Online Misinformation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 1 June 2023; pp. 923–932.
27. Yousif, A.; Buckley, J. Impact of Sentiment Analysis in Fake Review Detection. *arXiv* **2022**, arXiv:2212.08995. [CrossRef]
28. Hamed, S.K.; Ab Aziz, M.J.; Yaakub, M.R. Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. *Sensors* **2023**, *4*, 1748. [CrossRef] [PubMed]
29. Xuewen, Z.; Yaxiong, P.; Xiao, G.; Gang, L. Sentiment analysis-based social network rumor detection model with bi-directional graph convolutional networks. In Proceedings of the International Conference on Computer Application and Information Security, Online, 21 March 2023.
30. Shin, J.; Tang, C.; Mohati, T.; Nayebi, M.; Wang, S.; Hemmati, H. Prompt Engineering or Fine-Tuning: An Empirical Assessment of LLMs for Code. In Proceedings of the 2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR), Ottawa, ON, Canada, 28–29 April 2025; pp. 490–502.
31. Ray, P.P. ChatGPT A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
32. Wang, Y.; Rao, S.; Lee, J.; Jobanputra, M.; Demberg, V. B-cos LM: Efficiently Transforming Pre-trained Language Models for Improved Explainability. *arXiv* **2025**, arXiv:2502.12992.
33. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L. A Brief Overview of ChatGPT:The History, Status Quo and Potential Future Development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [CrossRef]
34. Sharma, V.; Raman, V. A reliable knowledge processing framework for combustion science using foundation models. *Energy AI* **2024**, *16*, 100365. [CrossRef]
35. Briony, S.T.; David, L. Public Health and Online Misinformation: Challenges and Recommendations. *Annu. Rev. Public Health* **2020**, *41*, 433–451.

36. Victor, S.; Javier, A. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e17187.
37. Navarrete, A.S.; Martinez-Araneda, C.; Vidal-Castro, C.; Rubio-Manzano, C. A novel approach to the creation of a labelling lexicon for improving emotion analysis in text. *Electron. Libr.* **2021**, *39*, 118–136. [CrossRef]
38. Ordan, N.; Wintner, S. WordNet: A Lexical Database for the English Language. *Choice Rev. Online* **2007**, *45*, 45–1196. [CrossRef]
39. Tonin, F.S.; Gmünder, V.; Bonetti, A.F.; Mendes, A.M.; Fernandez-Llimos, F. Use of 'Pharmaceutical services' Medical Subject Headings (MeSH) in articles assessing pharmacists' interventions. *Explor. Res. Clin. Soc. Pharm.* **2022**, *7*, 100172. [CrossRef] [PubMed]
40. Karimi, Z.A.; Sayeh, M. A survey of aspect-based sentiment analysis classification with a focus on graph neural network methods. *Multimed. Tools Appl.* **2023**, *83*, 56619–56695.
41. Fugate, J.M.B.; O'Hare, A.J. A Review of The Handbook of Cognition and Emotion. *J. Soc. Psychol.* **2014**, *154*, 92–95. [CrossRef]
42. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
43. Lea, C.; Carlo, S.; Ester, B.; Patricio, M.B. Intensional Learning to Efficiently Build up Automatically Annotated Emotion Corpora. *IEEE Trans. Affect. Comput.* **2017**, *11*, 335–347. [CrossRef]
44. Lv, X. Few-Shot Text Classification with an Efficient Prompt Tuning Method in Meta-Learning Framework. *Int. J. Pattern Recognit. Artif. Intell.* **2024**, *38*, 2451006. [CrossRef]
45. Wankhade, M.; Kulkarni, C.; Rao, A.C.S. A survey on aspect base sentiment analysis methods and challenges. *Appl. Soft Comput.* **2024**, *167*, 112249. [CrossRef]
46. Krzysztof, K.; Agnieszka, C.; Karolina, B.; Łukasz, B. Vaccine misinformation on social media—Topic-based content and sentiment analysis of Polish vaccine-deniers' comments on Facebook. *Hum. Vaccines Immunother.* **2021**, *17*, 10–11.
47. Katarzyna, B.; Marcin, S. A dataset for Sentiment analysis of Entities in News headlines (SEN). *Procedia Comput. Sci.* **2021**, *192*, 3627–3636. [CrossRef]
48. Hemmatian, F.; Sohrabi, M.K. A survey on classification techniques for opinion mining and sentiment analysis. *Artif. Intell. Rev.* **2019**, *52*, 1495–1545. [CrossRef]
49. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
50. Kruthika, S.G.; Nagavi, T.C.; Mahesha, P.; Chethana, H.T.; Ravi, V.; Mazroa, A.A. Identification of Suspect Using Transformer and Cosine Similarity Model for Forensic Voice Comparison. *Secur. Priv.* **2025**, *8*, e70038. [CrossRef]
51. Nielsen, F. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy* **2019**, *21*, 485. [CrossRef]
52. Todinov, M.T. Probabilistic interpretation of algebraic inequalities related to reliability and risk. *Qual. Reliab. Eng. Int.* **2023**, *39*, 2330–2342. [CrossRef]
53. Deng, X.; Liu, Q.; Deng, Y.; Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* **2016**, *340*, 250–261. [CrossRef]
54. Chamidah, N.; Widiyanto, D.; Seta, H.B.; Aziz, A.A. The Impact of Oversampling and Undersampling on Aspect-Based Sentiment Analysis of Indramayu Tourism Using Logistic Regression. *Rev. D'Intelligence Artif.* **2024**, *38*, 795. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.