Article

Aspect-Based Sentiment Analysis of Patient Feedback Using Large Language Models

Omer S. Alkhnbashi [1,2,3,*], Rasheed Mohammad [4,*] and Mohammad Hammoudeh [2]

Citation: Alkhnbashi, O.S.; Mohammad, R.; Hammoudeh, M. Aspect-Based Sentiment Analysis of Patient Feedback Using Large Language Models. *Big Data Cogn. Comput.* **2024**, *8*, 167. https://doi.org/10.3390/bdcc8120167

Academic Editors: Salvador García López and Yoichi Hayashi

Received: 22 September 2024 Revised: 2 November 2024
Accepted: 15 November 2024
Published: 21 November 2024

[1] Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; mohammad.hammoudeh@kfupm.edu.sa
[2] Center for Applied and Translational Genomics (CATG), Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai Healthcare City, Dubai P.O. Box 505055, United Arab Emirates
[3] College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai Healthcare City, Dubai P.O. Box 505055, United Arab Emirates
[4] Department of Computer Sciences, College of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK
* Correspondence: omer.alkhnbashi@dubaihealth.ae (O.S.A.); rasheed.mohammad@bcu.ac.uk (R.M.)

Abstract: Online medical forums have emerged as vital platforms for patients to share their experiences and seek advice, providing a valuable, cost-effective source of feedback for medical service management. This feedback not only measures patient satisfaction and improves health service quality but also offers crucial insights into the effectiveness of medical treatments, pain management strategies, and alternative therapies. This study systematically identifies and categorizes key aspects of patient experiences, emphasizing both positive and negative sentiments expressed in their narratives. We collected a dataset of approximately 15,000 entries from various sections of the widely used medical forum, patient.info. Our innovative approach integrates content analysis with aspect-based sentiment analysis, deep learning techniques, and a large language model (LLM) to analyze these data. Our methodology is designed to uncover a wide range of aspect types reflected in patient feedback. The analysis revealed seven distinct aspect types prevalent in the feedback, demonstrating that deep learning models can effectively predict these aspect types and their corresponding sentiment values. Notably, the LLM with few-shot learning outperformed other models. Our findings enhance the understanding of patient experiences in online forums and underscore the utility of advanced analytical techniques in extracting meaningful insights from unstructured patient feedback, offering valuable implications for healthcare providers and medical service management.

Keywords: sentiment analysis; content analysis; patient feedback; medical forum; deep learning; large language model (LLM)

1. Introduction

1.1. Online Patients' Feedback

With increasing life expectancy, it is anticipated that there will be an associated increase in the prevalence of long-term medical conditions [1]. For healthcare to be effective there is a need for a partnership between both well-informed patients and the clinical practice teams [2,3]. In this context, self-management strategies recognize that individuals are experts in their own lives, having acquired the skills and knowledge to cope with their medical condition [3]. Such strategies enable people to further develop their skills, knowledge, and confidence, leading to improved health outcomes and patient experience. It has been found that they can reduce unplanned hospital admissions and improve adherence to treatments [4]. Nevertheless, despite the wealth of resources available, a GP survey reported that people living with long-term medical conditions still want more support in order to manage their health and well-being [5]. Actually, referring to online sources and communication with others (patients and medical professionals) has been encouraged in recent years [3,6] and has become valuable for healthcare providers to attract new clients [7]. However, Powell, et al. [8], Atherton, et al. [9] reported that medical staff in the UK are cautious about online feedback, yet, they read them, and most feedbacks were positively talking about medical staff, as Boylano, et al. [10] reported.

From the clinical perspective, understanding patients' level of satisfaction with their healthcare is very important [2]. Patient opinions can be collected by a range of means, including surveys and questionnaires [8,11]. However, there are limitations with this approach, as questions can be restricted to set criteria and categories, which can be too narrow when trying to gauge the entire patient experience [12]. Formal surveys can also be expensive to conduct, time-consuming

to undertake and analyze, and the findings are not always provided in a form that is easy for the clinical teams to interpret or take action on [8,13,14].

Apart from formal evaluation of patients' opinions, patients just want to be heard, ask questions and elicit advice [15]. One way of performing this is through patient stories. This is a powerful way of connecting with healthcare staff; however, in the clinical setting they are often seen as too informal to be considered as legitimate data [12]. There are also the challenges of managing and integrating large volumes of free-text feedback, preventing its widespread use in clinical settings [14]. On other hand; patient feedback is taken seriously and many medical education systems include training students to recognize the credibility of patients' feedback and other purposes [16,17].

In order to seek information, patients often turn to the internet for help. A large portion of internet users have used the internet for some form of information about health-related issues [18]. The internet is now accepted as a global medium for health information-seeking behavior [19]. Its passive content, traditionally seen in the early days, has now been complemented by active user-generated content in the form of social media and forums. These platforms provide different ways of spreading new ideas and practices, which can improve interpersonal and societal communication, interaction, and understanding [20]. This can empower patients by supplementing their existing knowledge received from healthcare professionals, as well as providing them with a means of psychosocial support [19,21].

Medical online forums and social media were found to be used by patients to seek information from people going through similar health issues [2,22] and by health establishments to measure patients' satisfaction with health services provided [15]. Moreover, patients also use the online community as a way of accessing more current information, as it can be disseminated faster online rather than during visits to their doctor [19,21,23]. Ziebland, et al. [24] elaborated this, explaining that patients visit these forums to look for medical information, receive support, create relationships, and use health services. On the other hand, while patients' feedback is essential to improve the quality of health services provided [25], medical staff prefer to receive informal feedback (any means except a formal survey) [26]. This encourages us to consider medical forums as among the main sources of feedback.

Natural language processing (NLP) has many possible applications in the medical domain [2]. It can perform serious tasks such as extracting medication items within clinical notes for the purpose of verification [27] or reporting adverse drug reaction [28]. The work of Mazanderani, et al. [29] focused on analyzing textual feedback provided by patients regarding the quality of health services provided. In a review of "The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication", Baker, Brookes and Evans [2] listed many possible NLP techniques to analyze patients' feedback; yet, the major concern was finding weaknesses in National Health services provided in the United Kingdom and how to utilize the result to satisfy patients by improving services. In other words, although the book provides a rich experience regarding the interaction between patients and the health services provided, there is room to dig more and find other perspectives that have not yet been covered. For instance, findings about the effectiveness of medicines and medical treatment received by patients are as important as knowing the patients' level of satisfaction with health services provided.

Eventually, extensive NLP research on the medical domain, such the work of Liu, et al. [30], Zeng, et al. [31], Zhang, et al. [32], Young-Min, et al. [33], Gu, et al. [34], Jin, et al. [35], Huy, et al. [36], Mondal, et al. [37], Bai, et al. [38], Bao, et al. [39], Chen, et al. [40], Mehta, et al. [41], Daniel, et al. [42], Mullick, et al. [43] have collectively pointed out the lack of labeled datasets for the medical domain. Moreover, most of the above-mentioned studies were mainly focused on detecting patient's intention for clinical purposes with very minimum focus on patient's opinions. Additionally, the input used was professional medical reports and papers. Therefore, this study first reports the aspects that patients discuss in the forums; this is performed via content analysis for large segments of posts collected and entity extraction tool such as Spacy (models for biomedical text processing). The results of the content analysis and entity extraction are then used to build a list of aspects and related keywords. Manual annotation has been conducted for classification columns, and an ABSA tool (deBERTa) has been used to calculate the

sentimental values related to each aspect for regression values. A multi-output regression model and classification model were built. In addition to employ Large Language Models.

1.2. Significance of the Study

This study investigates patient experiences through aspect-based sentiment analysis (ABSA) of patient feedback collected from medical forums. While previous work has addressed sentiment analysis in patient feedback, it has often been limited to assessing overall patient satisfaction with healthcare services [8]. In contrast, our study seeks to uncover specific insights related to patient sentiment across various health-related aspects, including medications, medical procedures, and experiences with healthcare staff, thereby offering a more detailed understanding of patient perspectives. This work used recommendations of Kotei and Thirunavukarasu [44] regarding the effectiveness of pre-trained models such as BERT and GPT if transfer learning is intelligently applied.

Our contribution lies in the development of an ABSA framework using deep learning techniques and large language models (LLMs) to classify and evaluate sentiments for specific aspects in patient narratives. Moreover, feedbacks' credibility and cohesion has been assessed to ensure the suitability of patients feedbacks for further advanced NLP analysis. The credibility analysis and cohesion of words are essential to avoid patients' bias and unrealistic feedback. By analyzing over 15,000 entries from patient.info, we demonstrate that advanced NLP methods can effectively extract sentiments for distinct health aspects. This study not only sheds light on patients' experiences but also presents actionable insights for healthcare providers aiming to improve quality of care based on patient-centered feedback.

2. Related Works

Medical online forums such as (patient.info) contain several sections, where each section concerns one type of medical condition. This facilitates direct access to the intended section. Although the dataset of this study was collected from several sections, a slight focus on one section was the choice in this study, because considering all aspects (e.g., medicines and medical procedures run to thousands if all sections are considered) would be timeconsuming and worth little, as the findings can be obtained if only a few sections are considered. This study chose to focus on bone-related medical conditions, due to the accessibility to a medical team specializing in these conditions.

Work by Smailhodzic, et al. [45] identified four common reasons for patients to use social media for health-related purposes. The first was emotional support, which can take the form of sharing and receiving personal stories from those who are going through similar experiences. The second was esteem support, where the patients are encouraged to implement changes in their healthcare behavior in order for them to recover from their conditions. This is closely linked to the third type of support, which is information support. Here, people who have been living with a medical condition for some time have a wealth of knowledge to share with those who have just received a diagnosis. Social media platforms also allow the user to search for historical information, which may not be available in face-to-face support groups [21]. The last category is network support, which reduces the feelings of loneliness and isolation.

Other reasons behind the use of social media include allowing patients to freely express their feelings without being afraid of any immediate negative reaction from those close to them [45,46]. Sometimes patients do not feel comfortable sharing the extent of their feelings in person because of social embarrassment [46]. Online patient forums are also a way of expressing negative feelings, regardless of whether anyone responds [45]. However, this can lead to a sense of anxiety and concern among the online community or frustration that once the negative feelings have subsided that person is no longer available to share how they overcame their problems [21]. Social media platforms have also been used as a way to be remembered should the prognosis of their medical condition be poor [47].

Medical conditions related to bones (including joints and muscles), particularly osteoarthritis, are long-term conditions where patients are known to experience a wide range of physical, mental, and social requirements that cannot always be fully satisfied by medical professionals. In the case of osteoarthritis, social support is particularly imperative in dealing with the psychosocial aspects of the medical condition. This includes physical, emotional, and informational support. Although

most patients with osteoarthritis derive their social support from family and peers, research shows they often have a preference for support from the latter because they perceive that those without the disease lack understanding of their condition [48–50]. Therefore, patients seek information from various communities about the types of healthcare services that are available to them and how to adjust their lifestyles and behaviors to live best with their medical condition [51]. In terms of long-term medical conditions like osteoarthritis, such online support may be even more helpful to these patients, as they can find it more difficult to attend face-to-face social support sessions because of mobility issues. Social media, therefore, has the potential to offer rich sources of information, which could support gaining an understanding of patients' views about their healthcare outside of the clinical setting. However, the volume of information contained within these sites can make manually reviewing them extremely time-consuming [14]. With advances in machine learning there is the potential to make it easier for healthcare providers to obtain feedback from patients and use it to improve their services [29].

There are various machine learning tools that can be used to help in exploring discussions on social media forums. These include sentiment analysis (SA), which is a subset of Natural Language Processing (NLP). It is a text categorization approach, which allows the mining of opinions. These opinions can be classified as being either positive, negative, or neutral. However, further experiments could analyze the effectiveness of aspect-based sentiment analysis (ABSA). In addition to classifying the sentiment of utterances, ABSA identifies which aspects of the utterances contain that sentiment, facilitating granular analysis of sentiments within an utterance, and key aspect classification to identify recurring aspects that have similar sentiment across multiple utterances [52]. These techniques have been used on social media posts from patients suffering from bone conditions (such as osteoarthritis) to gauge their opinions about their experiences with the disease [14]. The challenge is finding a specialized extraction tool for the entities found in the online patients' posts, because generic tools such as Flair, Spacy, and NLTK have a specific number of entity types that may not fit the medical information provided in the patients' post.

Overall, first; the extensive work performed by Powell, et al. [8] showed that there is 32% of patients look for other patients' feedback regarding medicines, treatment, and procedures; this percentage deserves (many studies mostly care about feedback related to staff and health services provided) to find solutions and tools to assist them in more quickly accessing this information. Second; although many instances were found employing NLP in the medical domain, they were used for some purely specific tasks in the medical field or to measure the level of satisfaction with health services, with less concern toward the actual feedback coming from patients. Such feedback contains valuable information regarding the effectiveness of medical treatment and medicine received by patients as well as judging the performance of medical establishments. Additionally, patients look for peers to share information and are more open to them than to their medical team. The online forum contains a store of unlimited information that could fill the gap that cannot be covered by regular surveys conducted by medical establishments to collect information from the patients. Accordingly, it is regarded as significant to dig inside medical forums and understand the type of information that could help medical establishments to provide better services.

Besides investigating the nature of information shared by patients, this study proposes an approach to tag every group of aspects with sentiment values, in order to directly spot the concerns (aspects) inside a patient's feedback and the level of sentiment (polarity). There is a need for a model that brings together the aspects and polarity at once instead of going through a series of separate actions such as entity extraction and sentiment analysis. In short, there is a need for a multidimensional tool that can reveal all possible information regarding the patient's feedback, together with the sentiment value. Although NIHR [14] delivered an extended review regarding the NHS effort to collect and store patients' feedback, the planning regarding how to utilize them effectively is not clear. Powell, et al. [8], Boylano, et al. [10] recommended investigating patients feedbacks about particular treatments or diagnostics. Additionally, the current sentiment analysis studies providing an overall polarity (positive, negative, neutral), which is not adequate [8]. This study aims to provide a proposal to provide a systematic approach to spot aspect types and their related sentimental values.

Boudjellal, et al. [53] reported that there are the following entities (topics)—gene, protein, chemical, disease, species, drug, body part, disorder, symptoms, and treatment methods—those

that the biomedical related studies recognized. Similar reports were established by Kalyan, et al. [54] regarding the entities in medical content. Moreover, the extensive survey by Nerella, et al. [55], Wang, et al. [56] barely mentioned a study that concerned patients' feedback on social media for the sake of reflecting their perceptions in medical services provided. The major topics (units) reported in prior studies were "disease, drugs, treatments, medical conditions, and symptoms" [55] and similar findings reported by Wang, et al. [56]. However, the inputs were from clinical narratives, and barely any units/topics in the patients' feedbacks were considered. Few exceptions were found, as reported in Nerella, et al. [55], that are related to patients' feedbacks such as monitoring adverse drug reactions, categorizing vaccine confidence, monitoring depression, and locating disease hotspots. It can be seen from that there are trends in considering patients feedbacks particularly in terms of adverse drug reactions, which is pure patients feedbacks input and for the sake of enhancing medicine provided to patients.

Boudjellal, et al. [53] were mainly concerned about identifying diseases and treatments in specialized medical inputs (medical journals). The ABioNER-modified Bert model [53] only identified two entities (disease and medicine), which is less effective in identifying all possible concerns in patients' feedback. Kalyan, Rajasekharan and Sangeetha [54] reported the following models: CT-BERT, BERTweetCOVID-19, BioRedditBERT, RuDRBERT, EnRuDR-BERT, and EnDR-BERT, which were concerned with social media and particularly health-related reviews in different languages such as English and Russian.

Regarding pre-trained models, Kotei and Thirunavukarasu [44] concluded that selfsupervised learning and knowledge transfer significantly enhance transformer models, allowing for efficient domain adaptation. Pre-trained models like ALBERT and RoBERTa demonstrate faster convergence and better performance in limited data contexts. Models
with smaller architectures, like ALBERT, are cost-effective while maintaining high accuracy.
However. the work of Kotei and Thirunavukarasu [44] technically reviewed pre-trained models, which is significantly important, yet, no specific domains have been highlighted where those pre-trained models can outperform others.

Regarding developing biomedical models that concerned health reviews in social media, Nguyen, et al. [57] proposed BERTWEET to perform the following NLP-related tasks:

Part-Of-Speech tagging, Named-Entity-Recognition, and text classification (i.e., sentiment analysis and irony detection). Yet, specific medical purpose was mentioned. Similarly, Müller, et al. [58] proposed CT-BERT model that classifies tweets related to COVID-19 into several sentimental categories. Yet, entities or topics inside tweets have been discussed or extracted. Additionally, Tutubalina, et al. [59] considered the patient's health conditions and adverse drug reactions from posts in the social media. Table 1 presents some available studies concerned processing aspects in posts/documents related to medical treatment/services.

Regarding aspect-base sentimental analysis for medical purposes, Zhao, et al. [52] developed a double-layer aspect recognition model (OMR-ARM) and an aspect-level sentiment analysis (OMR-ALSA) approach. This model uses a domain-specific ontology (OMR-Ontology) to identify and classify objects (e.g., doctors, hospitals) and associated aspects (e.g., medical ethics, cost). The OMR-ARM model is integrated with baseline models for fine-grained sentiment extraction. However, the aspects were considerably related to medical staff and less concern toward patients' topics (e.g., pain, effectiveness of treatment/medicine, etc.). On other hand, though the dataset (HCAHPS) used by Madan, et al. [60] contains patient feedback on services such as cleanliness, availability of doctors, and doctor–patient interactions, the aim was only predicting the sentimental value associated with patient's feedback with more focus on hospital's services.

Table 1. Studies concerned extracting topics/entities in medical materials online.

Author(s)	Aspects (Entities)	Training Data	Model	Findings	Weaknesses
[53]	Diseases and treatments	Arabic medical journals	BERT-based model (ABioNER)	ABioNER outperformed AraBERT and the multilingual BERT model, achieving a higher F1-score, especially in "Disease or Syndrome" and "Therapeutic or Preventive Procedure" categories, confirming the potential of small-scale, domain-specific pre-training	Limited to two entity types without evaluation on broader NER tasks due to data scarcity; lacks assessment of model performance on additional Arabic biomedical entity types or relation extraction tasks.
[58]	Nan	COVID-19 Category, Vaccine Sentiment (VC), Maternal Vaccine Stance (MVS), Stanford Sentiment Treebank 2 (SST-2) and Twitter Sentiment SemEval (SE)	BERT-based model CT-BERT	CT-BERT showed superior performance over the BERT-LARGE model with up to 30% improvement in classification tasks, especially in COVID-19 and health-related datasets. It yielded a mean F1 score improvement across datasets, with the most significant gains in COVID-19-specific contexts	Limited to classification tasks; not assessed for other NLP tasks like named entity recognition. Potential for further improvement through hyperparameter optimization. Additionally, only one COVID-19-specific dataset was available for fine-tuning and evaluation.
[57]	POS tagging, NER and textclassification	WNUT16 NER, WNUT17, oct27.traindev andoct27.test etc.	BERTweet-language model based on BERT	BERTweet outperformed baselines RoBERTa and XLM-R on tweet NLP tasks, achieving new state-of-the-art results, particularly in NER (+14% improvement) and text classification tasks for sentiment and irony detection (+5% and +4% improvement). "Soft" normalization outperformed "hard" lexical normalization for this domain	Although effective on the targeted tasks, BERTweet lacks evaluation on broader NLP tasks outside of Twitter and may not generalize beyond tweet-style text. No "large" version of BERTweet was included, limiting comparisons to larger models like RoBERTa-large

Table 1. *Cont.*

Author(s)	Aspects (Entities)	Training Data	Model	Findings	Weaknesses

[61]	medical terms: the Clinical finding, Substance, Body parts, Procedure, and pharmaceutical product	Corpus collected from social media health related posts	COMET-A, a corpus for medical entity linking in social media	Neural methods, combined with dictionary and string-matching baselines, were effective, though COMETA posed significant challenges, especially in zero-shot scenarios. The study observed a 28–46% gap from perfect performance, highlighting the complexity of social media language and the need for multi-view (text and graph) models for improved EL in health contexts.	Performance in zero-shot settings remains challenging due to the highly diverse and informal nature of social media language. The study noted limited performance improvements even with neural baselines, highlighting difficulties in capturing layman terminology in medical contexts.
[59]	patient's health conditions and adverse drug reactions	1.4 million health-related posts online	RuDR-BERT	RuDR-BERT outperformed multilingual and Russian-language BERT baselines, achieving higher F1 scores across both NER and multilabel sentence classification tasks, particularly in detecting ADRs and DI. The model revealed the difficulties of handling diverse regular person language used in patient reviews, particularly in ADRs.	Limited to Russian-language text, and the annotated portion includes only a subset of therapeutic categories, potentially limiting generalizability. Further limitations include the challenge of mapping informal language in user-generated content to formal medical terminology
[62]	Disease, drugs (chemical), drug (protein), species	NCBI, JNLPBA, BC2GM, LINNAEUS, Species-800, BC5CDR, BC4CHEMD	BioALBERT	BioALBERT outperformed other models, including BioBERT, on eight benchmark BioNER datasets, achieving significant F1 score improvements across all tested categories, especially in Drug/Chem and Disease categories. It also demonstrated faster training speeds and lower memory usage than BERT-based models.	BioALBERT's training is limited to biomedical corpora. Currently evaluated only on BioNER tasks.

Though there are many studies concerned with patients' feedback in social media, they are for the sake of medical system benefits such as drag reaction, monitoring diabetes, and monitoring depression. It is a noble objective, yet, there are enough information that include broad range of aspects (topics) that can help understanding patients, besides the aspects related to medical system enhancement. Many studies were particularly concerned with the British health system (NHS); they heavily reported patients' feedback that was collected via surveys, which are expensive, structured (i.e., limited the answers to specific group of questions), and cannot be mass collected. In contrast, medical forums and social media contain unlimited and free patients feedback, which are cost-effective to collect and rich in terms of aspects discussed.

This work found the study of Basaldella, et al. [61] can be considered as baseline for the proposal of this study. This is due to the fact that their corpus "COMETA" can recognize medical terms: the Clinical finding, Substance, Body parts, Procedure, and pharmaceutical product. This was in addition to the work of Tutubalina, et al. [59], which considered patient's health conditions and adverse drug reactions. Therefore, besides recognizing the medical terms, this study was concerned with more than just the feedback of patients regarding those terms for the sake of assisting medical teams' understanding of several things such as the effectiveness of medicine/procedures, the quality of service provided, etc.

Another approach, Large language models (LLMs), are AI-powered tools designed to handle and create text. They gained significant attention when OpenAI released ChatGPT to the public in

November 2022. These models are adept at answering questions, summarizing, paraphrasing, and translating text, achieving performance that closely mimics human abilities. The interactive nature of models like ChatGPT enhances their appeal across various sectors, including healthcare [63,64]. LLMs hold the promise of broadening access to medical knowledge and improving healthcare delivery [63,65]. Though the work in [63] has highlighted the role that can be played by LLMs in the medical domain, yet, it was light and lacked empirical results. The discussion in [63,65] broadly covered many aspects including Natural Language Processing but barely mentioned LLM practices in detecting the aspects and sentimental values attached. Similarly, a comprehensive review for ~50 models was performed by Wornow, et al. [66]; wherein they reveleaed works that extracted aspects and related sentiment values for the sake of medical service improvement, i.e., the work was medical staff-centered and barely discussed NLP related to patients' feedback. On other hand; the work in [64] proposed and tested a biomedical LLM (GatorTronGPT) that is based on ChatGPT 3.5. It was intended to cover several NLP tasks (drug–drug interaction, chemical-disease, drug-target interaction and Question-answering), but lacks of aspectsdetection and sentimental value related. However, it might be overcome with a few shots, assuming there is support for this.

3. Methodology

3.1. DataSet

Patient.info, established in 1996, has since grown to serve a broad audience, with forums covering a wide range of medical conditions and concerns [67]. This platform was specifically chosen as a data source because prior studies indicate that medical staff prefer feedback collected through specialized websites rather than through social media, as these sources are more focused and often have higher-quality responses [9]. The data used in this study were extracted from the publicly accessible forum at https://patient.info/ (accessed on 2 March 2024), where patients and family members discuss various medical situations and health issues. Importantly, user anonymity is maintained on this platform, with names hidden and profile information (such as gender and age) often concealed.

The dataset consists of over 15,000 posts; after removing advertisements and nonrelevant content, 12,103 posts were retained. A significant portion of the posts were replies to existing discussions rather than new, standalone posts, reflecting the interactive and supportive nature of the forum. Approximately 20% of the posts were manually reviewed to gain a better understanding of the context, identify appropriate data-cleaning procedures, and ensure that the posts accurately represented common patient concerns and narrative descriptions. This process also helped categorize common topics within patient feedback.

For the remaining 80% of the dataset, which was not manually annotated, we utilized the DeBERTa-v3-base-absa-v1.1 model, an advanced aspect-based sentiment analysis (ABSA) tool. This model is specifically tailored for extracting sentiments linked to predefined aspects. DeBERTa's unique transformer-based architecture allows it to handle detailed language, isolating each identified aspect and assigning a sentiment score—positive, neutral, or negative—within each sentence. This approach allowed for efficient annotation across a large number of posts, thus, ensuring scalability while maintaining accuracy.

The methodology prioritized sentiment accuracy but also presented potential biases. DeBERTa was pre-trained on a general-purpose dataset, which, while robust for many language tasks, may not fully capture the specific terminology or informal language frequently used in patient forums. This could result in biases, especially if the model misinterpreted nonstandard phrases or colloquial expressions as neutral or irrelevant. To mitigate this, the model's output was manually reviewed on a sample basis, ensuring that the sentiment distributions matched expected patterns within each aspect (e.g., more negative sentiments for 'medical procedures' due to complaints about pain).

To capture a broad spectrum of patient perspectives, additional posts were collected from sections dedicated to specific conditions such as cardiology, cancer, and diabetes. This inclusion strategy was designed to enhance the dataset's representativeness, aiming to cover diverse patient

experiences across multiple health domains. Using the Spacy library (specifically, the en_ner_bc5cdr_md model), we further extracted entities within these posts to capture pertinent details in patients' feedback about medicines, treatments, and procedures.

Despite the extensive nature of the forum, it is important to note that a relatively small percentage of patients actively provide feedback online; for instance, only 4% of patients surveyed by Powell, et al. [8] posted feedback on treatments, medicines, and procedures. Therefore, while our dataset from patient.info is substantial, it represents a subset of the patient population that actively engages online. This engagement profile may introduce certain biases, yet it provides valuable insights into patient experiences and concerns within the broader healthcare domain.

3.2. Methods

To execute a comprehensive sentiment analysis for patients' feedback, the review was initially segmented into individual sentences using the NLTK Tokenizer. This step was crucial for enhancing the model's performance [27]. Following this, the textblob library was employed to correct any misspellings in the sentence set. The subsequent phase involved the extraction of a predefined list of aspects, as detailed in Table 2. These aspects included 'medicine', 'human parts (organs) complaints', 'medical procedures', 'diagnosis', 'medical staff', 'food and supplements', and 'physical activities', among others. These were identified within each sentence, with a majority being manually generated, while entity extraction tool in [61] contributed to identifying a few. Upon detecting any of these aspects in feedback sentences, assessors were labeled as either positive, negative, or neutral (or not existant). The end of this process was a 7-feature matrix that effectively captured the polarity of the mentioned aspects within the forum posts. This matrix played a pivotal role in training the models of this study.

Table 2. Themes (aspects) found in patients' feedback (entities/topics reported in this table only related to osteoarthritis posts as a sample of entities under every category).

Topic	Aspects (These Words Were Used by Patients and Bearing Misspelling)-Sample from Feedbacks Under Bone-Related Disease/Disorder
Medicines	cortisone, acupuncture, MSM, Pentosan Polysulfate (PPS), Celebrex, Pazital and Lyrica, icyhot cream/icyhot patches, Biofreeze, Co-codamol, Naproxen and codeine, Bedranol, opiods, opioids, tramadol, tapered steroid, NSAID Etova, vicodin, prednisone, polymyalgia rheumatica, acetaminophen, glucosamine, prescription, zomorph, Flexiseq dihydrocodeine naproxen gabapentin amytriptaline
Complaints (parts of human body about which patients complain)	collar bone, palm, spinal circles, toes, leg, thumb, neck, low back, hip & pelvis, thumbs, right knee, right hip, right shoulder, SI joints, spine, ankle, fingers, jaw and foot, Joint, joint replacement relief, thumb, wrist, left thumb, cramps in my legs, tingling, finger, pinky, bone spur, hip bursitis
Medical procedures	ultra-sound, X ray, ACDF surgery, bilateral TKR, knee replacement, adipose stem cell & PRP, MRI and a spinal surgeon, chiropractor, Cac surgery, Pentosan Polysulfate, physical therapy, suture anchor, ankle fusion, kinsiology tape, Brodtens bracelet, X-rays and blood tests
Diagnoses	arthritis, osteoarthritis, polymyalgia rheumatica, osteoporosis, OE, Fibromyalgia, sciatica, Inflammation, Stiffness, herniated disc, MCAS, osteopenia, rhematic, osteo
Medical staff	orthopaedist, doctor, surgeon, physios, drs, DOC, rheumatologist, GP, physiotherapist
Activities	stretching, & walk when I can. Have a heating blanket, sports
Food and herbs	cider vinegar, fish oils, CBD oil, glucosamine sulphate, Capsaicin cream, turmeric

The main goal of these efforts was to develop a model adept at recognizing various aspects within the forum posts and accurately determining their associated polarity. The summary of methodology steps of this study is presented in Figure 1.

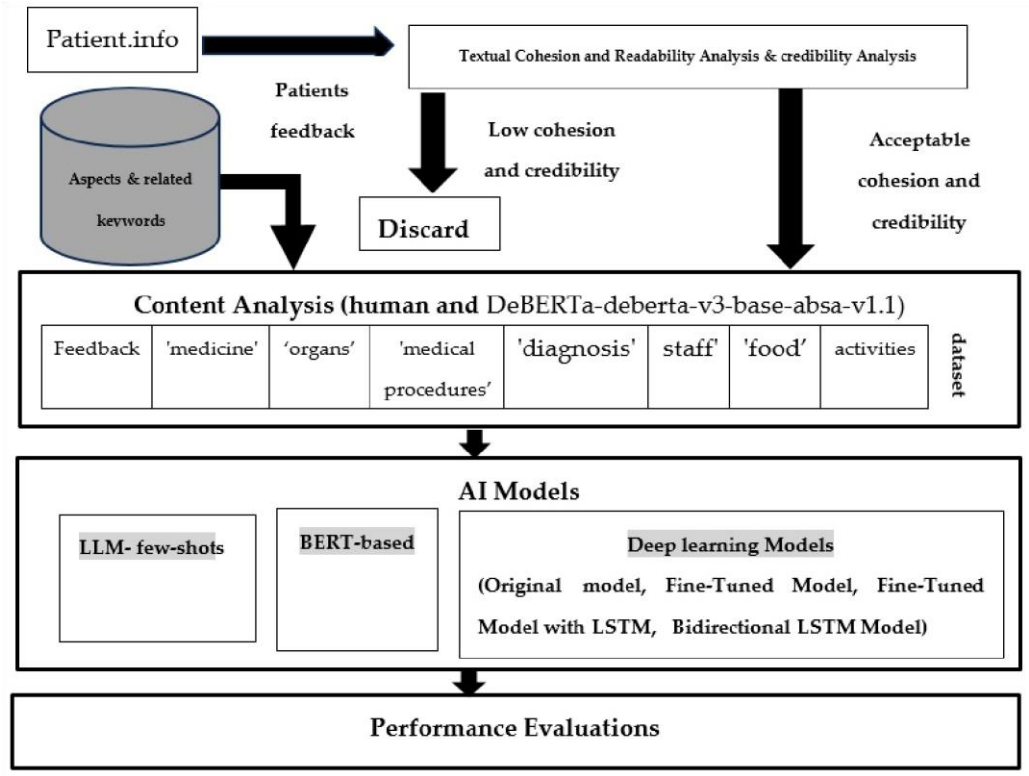

Figure 1. A conceptual architectural diagram for the proposed methodology.

3.2.1. Content Analysis

To effectively understand the aspects discussed by patients in medical forums, a thorough examination of the posts is crucial for creating an exhaustive list of potential aspects. Given that many entity recognition and aspect-based sentiment analysis tools are limited in their scope, often excluding medical entities, there is a pressing need for a comprehensive list of entity types relevant to medically related posts. Content analysis, a widely used methodology for understanding textual content context, is employed for this purpose, as evidenced in studies like those by Bez, et al., [15], Bond and Ahmed [68]. This study adopted a conventional approach to content analysis [69], suitable for its exploratory nature relying on pre-existing findings from prior studies that reported entities such as body parts, medicine, procedures etc. inside posts. Initially, approximately 10% of the posts from each assessor (two medical PhD students) were read repetitively to gain an overall understanding. This process involved categorization, coding, and establishing patterns [70], with a focus on identifying potential aspects present in patient posts on medical forums. Given the vast range of detailed characteristics found in these posts, the goal was to distill this information into a concise and focused list of main categories.

The identified aspects, as mentioned in the patients' feedback, fell into several categories [29], including: (a) medicine, (b) complaints about pain in organs, (c) activities, (d) medical procedures, (e) food, (f) diagnosis, and (g) medical staff, as detailed in Table 2. To ensure coding accuracy, intercoder reliability was assessed, with two more medical PhD students independently coding 10% of the cases. The resulting percentage agreement was commendably high, exceeding 97% for all codes. Then, comprehensive terms collection took place where the two PhD students went through patients feedbacks to collect terms related to each category.

In considering feedbacks related to disease/disorder: Further details of each category revealed that the 'medicine' category predominantly featured discussions about painkillers. The 'complaints about pain' category focused on pains in specific body parts. For instance; the feedbacks related to osteoarthritis mainly reported pain in hips, knees, and legs. 'Medical procedures' included reports of common treatments like bilateral Total Knee Replacement (TKR).

'Diagnosis' mainly involved conditions like arthritis and osteoarthritis, while 'medical staff' emerged as a category due to frequent mentions of healthcare professionals. The final two categories, 'food' and 'activities', encompassed dietary recommendations for pain management and refered to activities either as helpful in alleviating pain or hindered by it.

3.2.2. Data Pre-Processing

In contrast to platforms that provide only a short space to post, medical forums, particularly patient.info, provide enough space adequate to deliver the patient's story. Accordingly, a few issues related to the text were noticed, such as using abbreviations and misspelling of medicines and medical procedures, as can be seen in Table 2. Therefore, misspelling correction was conducted first, to avoid removing necessary words when data cleansing procedures take place. Textual cohesion and readability and credibility analyses were conducted. After this, data pre-processing was conducted and stopwords and punctuations were removed together with application of stemming techniques.

3.2.3. Textual Cohesion and Readability Analysis

The readability of patient feedback can provide insights into how well individuals articulate their health experiences and how accessible or complex their narratives are [71]. This analysis applies the Flesch Reading Ease score, a widely recognized metric for evaluating text readability [72]. A higher score indicates that the text is easier to read, while a lower score suggests more complexity, often requiring higher reading comprehension. This procedure is crucial for healthcare applications that may rely on patients' feedback in related NLP-based healthcare task such as drug adverse detection.

The Flesch Reading Ease score was chosen for its interpretability and widespread use in academic literature [71]. This score computes readability based on sentence length and word complexity (e.g., syllable count), producing a score typically ranging from 0 to 100.
The formula for Flesch Reading Ease is [72]:

$$206.835 - 1.015\left(\frac{\textit{total words}}{\textit{total sentences}}\right) - \left(84.6\frac{\textit{syllables}}{\textit{total words}}\right)$$

A higher score indicates easier readability (e.g., a score of 70–80 is easily readable by most 13- to 15-year-olds), while a lower score indicates more complex text. The average readability score was 65.05 (Table 3), indicating that, on average, the feedback was fairly readable (comprehensible to a wide audience). The median score was 76.11, suggesting that half of the posts scored higher than this, indicating better readability. The lowest score was −3939.59 (Table 3), which appears to be an outlier or erroneous data, likely due to extreme length or formatting issues. Finally, the highest score was 206.84 (Table 3), suggesting some posts were extremely simple in structure and language.

Table 3. Readability Score Statistics.

Measure	Value
mean	65.05
std	105.39
min	−3939.59
max	206.84

A mean readability score of 65.05 suggests that, on average, the text is accessible to individuals with a high-school level of education, which aligns with the intended audience for most healthcare communication. The interquartile range (IQR), represented by the 25th and 75th percentiles (66.54 to 83.46), indicates that most posts fall into a range of moderate readability, with some easier-to-read and some more complex narratives. The standard deviation of 105.39

suggests a high degree of variation in readability, meaning some posts were extremely complex while others were much simpler.

3.2.4. Credibility Analysis

Bias or credibility analysis in patient feedback aims to assess the presence of extreme or exaggerated language that could affect the reliability of the feedback. Identifying posts with biased phrases helps to determine whether the narrative reflects balanced opinions or if it leans toward overly positive or negative extremes, which might reduce its credibility [73,74].

The average bias count [73] is 0.27 (Table 4), indicating that, on average, there are very few biased phrases present in the posts. The median bias count is 0, which means that more than half of the posts contain no biased phrases. The standard deviation is 0.54 (Table 4), showing some variation in bias across the dataset. The maximum bias count is 4 (Table 4), indicating that the most biased post contains four biased phrases.

Table 4. Credibility related Statistics.

Measure	Value
mean	0.27
std	0.54
min	0.0
max	4.00

These statistics suggest that most posts contain neutral language, with only a small proportion of posts exhibiting extreme or biased phrases. A threshold of 1.0 (95th percentile) was established, meaning that any post with more than 1 biased phrase was considered potentially biased. A total of 621 posts had bias counts greater than 1. These posts were flagged as having high bias, and their credibility might be questioned. Examples of flagged posts include:

Post with a bias count of 2: “I am 28 and software engineer by profession. I think this is a miracle cure for anyone who tries it”.

Post with a bias count of 3: “I was just prescribed lisinopril today for my BP and it’s already showing amazing results, it’s the best decision ever”.

These posts contain language that might exaggerate the effectiveness of treatments or dismiss their downsides, raising concerns about their objectivity.

The histogram visualized the distribution of bias counts across all posts (Figure 2), showing that the majority of posts had 0 biased phrases, indicating neutral language. A small tail in the histogram represents the posts with higher bias counts, suggesting that a minority of posts might be overly positive or negative.

Accordingly, the feedbacks that contain high bias and less cohesion (based on the value of The Flesch Reading Ease score) were excluded, and there were less.

3.2.5. Aspect-Based Sentiment Analysis

Sentiment analysis of patient feedback is crucial for assessing the quality of service provided by medical staff, as highlighted by Dhajate and Joshi [1]. Aspect-based sentiment analysis, a specialized branch of natural language processing (NLP), is geared towards pinpointing sentiments directed at specific aspects or entities within a text [75]. This technique deconstructs text into various aspects or categories, assigning each a sentiment polarity—positive, negative, or neutral. For this purpose, we employed DeBERTa (debertav3-base-absa-v1.1), a sophisticated deep learning model renowned for its ability in NLP tasks, including aspect-based sentiment analysis [27,76].

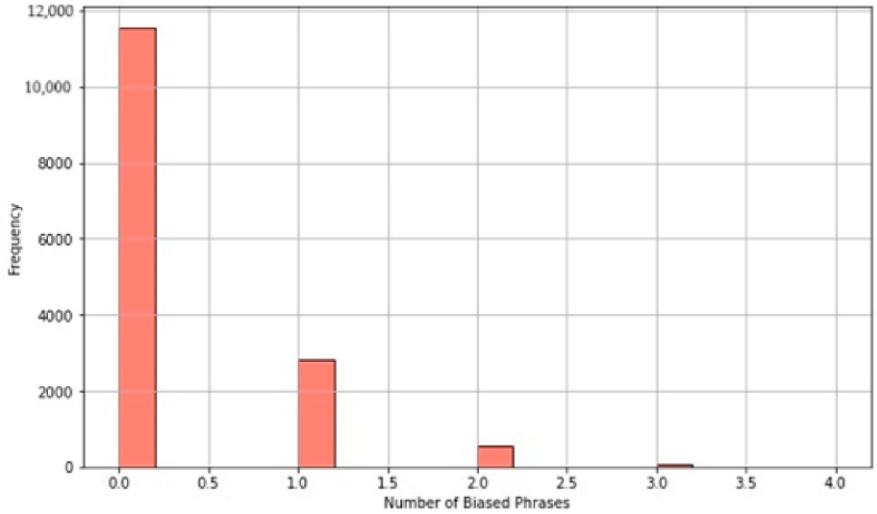

Figure 2. Distribution of bias counts in patient feedback.

Built on the BERT (Bidirectional Encoder Representations from Transformers) framework, DeBERTa surpasses its predecessor by integrating advanced features like disentangled attention and enhanced decoding capabilities. It has been trained and fine-tuned on extensive datasets, totalling over 210 k entries [76]. The model excels in discerning and interpreting sentiments associated with various aspects in a text. It employs a transformerbased architecture to contextualize the textual content, focusing specifically on relevant aspects through its 'disentangled' attention mechanism. This feature enables DeBERTa to isolate and analyze sentiments related to each aspect distinctly. The model's decoding mechanisms further refine these sentiment predictions, ensuring precise and nuanced sentiment polarity scores [27,77].

DeBERTa has set new benchmarks in NLP, outshining many traditional and transformerbased models in tasks like text classification and sentiment analysis [77]. Despite its strengths, challenges such as computational demands, interpretability, and fine-tuning complexity persist, especially given its training on non-medical datasets [76]. Continuous performance monitoring is essential, as its effectiveness might vary across domains [78,79]. The goal in using DeBERTa is to accurately tag aspects in patient feedback, laying the groundwork for training a deep learning model capable of comprehensive predictions (both aspects and sentiment values) in medical contexts. The gap this study addresses is the prediction of aspects and their sentiment polarities in patient posts using deep learning. We aim to provide a holistic sentiment analysis for medical-related units in patient feedback, bypassing the need for iterative word-by-word analysis using aspect-based sentiment analysis. This novel approach is elaborated in subsequent sections. Additionally, while Spacy's biomedical models identify a limited range of entities, our content analysis has yielded a more extensive list of seven categories, offering a broader spectrum of aspects relevant to patient feedback.

3.2.6. Architectures of Deep Learning Models

The selection of BERT, LSTM, and Bidirectional LSTM models in this study was guided by the unique strengths each architecture offers in handling text data, particularly in extracting sentiments within aspect-based sentiment analysis (ABSA). BERT was chosen for its transformer-based architecture, which allows it to capture contextual relationships between words in a sentence effectively. This capability is particularly valuable in sentiment analysis as it can better interpret patient language. BERT's ability to consider the entire context of a sentence enables more accurate sentiment labeling, which is crucial for this dataset's complex medical terminologies and patient experiences.

LSTM (Long Short-Term Memory) models were incorporated to handle the sequential nature of text data and to capture dependencies between words in longer sentences. LSTMs are widely recognized for their effectiveness in processing sequence-based data, making them suitable for tasks requiring attention to word order and sentence flow. Bidirectional LSTM extends the functionality of LSTM by allowing information to flow both forwards and backwards in a sentence. This bidirectionality is particularly useful in healthcarerelated texts, where context from surrounding words (both prior and subsequent) can significantly affect sentiment interpretation.

In this study, the development of a classification model was taken place with a ascertain the efficacy of various architectures as recommended by Kamıs¸ and Goularas [80]. The initial model employed a deep learning architecture designed for multi-output classification, commencing with an Embedding layer for textual input representation. This was followed by a Flatten layer to convert the sequences into a flat array, proceeding to two Dense layers with ReLU activation functions for extracting higher-level features. The output layer comprised seven nodes, each representing one of the classes identified in the content analysis subsection. Optimized using the Adam optimizer and Mean Squared Error (MSE) as the loss function, this model was trained for 20 epochs with a batch size of 32, without incorporating any regularization techniques.

Subsequently, modifications were introduced to formulate a Fine-Tuned Model. An additional Dense layer was added to augment the model's capacity for capturing complex data patterns, maintaining the same activation functions and loss function as the original. This model, too, was optimized with the Adam optimizer and trained similarly for 20 epochs with a batch size of 32, without any regularization techniques in the fine-tuning process.

Further refinements led to the integration of Long Short-Term Memory (LSTM) layers in the model to effectively handle the sequential nature of textual input. The enhanced architecture began with an Embedding layer and incorporated an LSTM layer for retaining sequence information. To counteract overfitting, Dropout layers were added. The processing sequence continued with a second LSTM layer and another Dropout layer, followed by a Dense layer with ReLU activation before the final output layer. This LSTM-based model was optimized using Adam and trained for 20 epochs at a batch size of 32, and utilized Dropout at a rate of 0.2 for regularization.

The final iteration involved designing a Bidirectional LSTM model to capture bidirectional contextual information from the textual input. Starting with an Embedding layer, the model included a Bidirectional LSTM layer, augmented by a Dropout layer for regularization. The second Bidirectional LSTM layer processed the outputs of its predecessor, followed by another Dropout layer. A subsequent Dense layer with ReLU activation led to the final output layer. Like the LSTM-based model, this model was trained for 20 epochs with a batch size of 32 and optimized using Adam, incorporating a Dropout rate of 0.2. Detailed specifications and configurations of these four models are described in Table 5.

Table 5. Deep learning architectures used in this study.

Architecture	Original Model	Fine-Tuned Model with LSTM	Fine-Tuned Model	Bidirectional LSTM Model
Architecture	Embedding	Embedding	Embedding	Embedding
			LSTM	Bidirectional
	Flatten	Flatten	Dropout	LSTM Dropout
	Dense	Dense	LSTM	Bidirectional
	Dense	Dense	Dropout	LSTM Dropout
	Output	Dense	Dense	Dense
		Output	Output	Output

Activation Functions	ReLU (except for the Output layer)			
Loss Function:	Sparse Cross Entropy Loss			
Optimizer/Epochs/Batch Size	Adam/20/32			
Regularization	None	None	Dropout (0.2)	Dropout (0.2)

During the training and testing phases, it was observed that the original model, while highly accurate in training (99%), demonstrated an overfitting issue, as evidenced by a reduced testing accuracy of 82%.

In response, the fine-tuned architecture, which incorporated regularization techniques, presented a more balanced performance, achieving 85% accuracy in training and 80% in testing. This adjustment effectively mitigated the overfitting problem while maintaining commendable accuracy. However, the LSTM architecture with pre-trained embedding was not exempt from overfitting, showing a disparity between training (90%) and testing (80%) accuracies. Further fine-tuning, which included integrating LSTM, CNN, GloVE embedding, and regularization, enhanced the model's performance. This comprehensive approach yielded an 85% accuracy rate, successfully overcoming the overfitting challenge.

In conclusion, the classification predictions using the proposed models achieved a satisfactory average accuracy of 85%. This result underlines the effectiveness of the models in handling classification tasks, particularly when modifications and fine-tuning are applied to address specific challenges such as overfitting.

3.2.7. Evaluation of ABSA Model Based on BERT

The model was fine-tuned on an ABSA (aspect-based sentiment analysis) task utilizing a BERT-based architecture. This fine-tuning process was aimed at adapting the pre-trained BERT model to specifically recognize and evaluate aspects and their corresponding sentiments within text data. The model underwent training for a single epoch, utilizing a custom configuration designed to address the multi-label nature of the ABSA task, where each aspect sentiment is categorized into one of three classes: −1 (negative), 0 (neutral), or 1 (positive). The model's performance was assessed using a comprehensive set of evaluation metrics: precision, recall, and F1-score, all of which yielded an identical score of approximately 0.733. This uniformity across precision, recall, and F1 indicates a balanced ability of the model to accurately identify relevant aspects and their sentiments while maintaining low false positive and false negative rates. Precision achieved was a value of 0.7327, indicating that, on average, 73.27% of the aspects identified by the model across all classes were correctly classified. Recall was 0.7327, suggesting that the model was able to correctly identify 73.27% of all relevant aspects present in the data. Finally, F1-Score—the harmonic mean of precision and recall—was 0.7327, reflecting the model's balanced performance in terms of both precision and recall.

The evaluation was conducted with an eval_loss of −4.2967. Although typically loss values are positive, this specific configuration and evaluation methodology resulted in a negative loss, which can be attributed to the particular characteristics of the loss function and model output interpretation in this context. The evaluation was performed over a duration of 1188.5878 s (~19.81 min), processing samples at a rate of 0.607 samples per second and achieving a computation throughput of 0.077 steps per second. This computational performance reflects the resource-intensive nature of processing and evaluating a complex model like BERT, particularly in tasks requiring nuanced understanding of text data, such as ABSA.

The fine-tuned BERT-based model demonstrates promising results in the ABSA task, with a solid balance between precision and recall as evidenced by the uniform F1-score. This performance suggests the model's effectiveness in identifying and classifying aspect sentiments from text data accurately.

3.2.8. Overfitting Handling

Overfitting, a common issue in deep learning models, was observed in some architectures during initial training, particularly in the BERT and LSTM-based models. To address this, several regularization techniques and validation methods were implemented to enhance model generalization on unseen data.

Regularization Techniques

Dropout was applied at rates of 0.2–0.3 in the LSTM and Bidirectional LSTM models. By randomly deactivating a portion of neurons during each training epoch, dropout helps prevent the model from becoming overly reliant on specific features, reducing the risk of overfitting. This approach was particularly effective in improving the robustness of the LSTM models, as evidenced by reduced training-testing performance gaps. Additionally, L2 regularization, or weight decay, was incorporated in the BERT model by adding a penalty term to the loss function that discourages large weight values. This constraint helps smooth the model by preventing extreme weight adjustments, thus, enhancing generalization. Finally, early stopping was implemented across all models, whereby training stopped if validation loss did not improve over five consecutive epochs. This method prevents the models from learning noise in the data as training progresses, further helping to mitigate overfitting.

Cross-Validation Approach

A 5-fold cross-validation was employed to validate each model's performance. This method involves dividing the dataset into five equal parts, using four parts for training and one for validation iteratively. Cross-validation provides a robust assessment of the model's ability to generalize across different data splits and mitigates the risk of overfitting to a particular subset of data. Additionally, cross-validation offers insight into model variance, helping us ensure stable performance across various data distributions.

Evaluation on a Held-Out Test Set

After implementing regularization and cross-validation, we evaluated the models on a held-out test set, comprising 25% of the dataset that was not used during training or validation. This final evaluation allowed us to assess how well each model was generalized to entirely unseen data. Notably, the LSTM model with dropout and BERT with L2 regularization demonstrated improved generalization with minimal accuracy loss between validation and test sets, indicating effective overfitting mitigation.

These regularization and validation steps collectively ensured that the models retained high accuracy while improving their generalization capabilities on new data, thus, making them suitable for deployment in real-world sentiment analysis tasks.

3.2.9. ChatGPT Few-Shots

In contrast to the proposal of Singhal, et al. [81] that built LLM for medical questionanswering, which is expensive to build, in terms of using LLMs for aspect-based sentimental analysis, Lu, et al. [82] saw that few-shot aspect-based sentiment analysis (ABSA) is challenging due to the lack of labeled data and the complexity of extracting fine-grained sentiment elements (e.g., aspect terms, opinion terms). Large language models (LLMs) struggle with ABSA in few-shot scenarios, often producing inconsistent or incomplete outputs [82]. However, the work of Lu, Liu, Cong, Yang, Gan, Fang and Wu [82] was trained on restaurant and laptop domains, which may limit its specificity to specific domain. Accordingly, this work aimed to mainly focus on patient's feedback and aspects included.

In this study, ChatGPT-3.5-turbo was employed using a few-shot learning approach rather than traditional fine-tuning. Few-shot learning involves providing the model with a small set of representative examples to guide its understanding of task-specific requirements, which is especially useful in cases where domain-specific training data are limited or where full fine-tuning is computationally intensive [65], which is the case of this study.

Accordingly, this study used few-shot to guide ChatGPT 3.5 turbo to detect medical aspects and their related sentiment values in the patients' feedback. Few-shot learning with ChatGPT-3.5-turbo offers a flexible and efficient way to make predictions with limited examples. When applied to the dataset, it allowed for accurate predictions with minimal training data. We supplied ChatGPT-3.5 with a few carefully crafted examples representative of the medical forum posts, covering common aspects (e.g., 'medicines', 'procedures', 'medical staff') and illustrating the sentiment categories (positive, neutral, and negative). This setup effectively guided the model on how to interpret patient language within the medical domain, enabling it to make informed predictions without extensive retraining.

Few-shot learning proved effective in this context because it allowed ChatGPT to generalize from the few provided examples to interpret similar expressions in other posts accurately [83]. For instance, in responses regarding 'medicine' and 'medical procedures', the model accurately captured common sentiment tones such as frustration or satisfaction with treatments. However, occasional misclassifications were observed in posts with more ambiguous language, suggesting that while few-shot learning achieved high accuracy overall, further domain-specific examples could enhance performance.

Advantages of few-shot over fine-tuning: The few-shot method provided the flexibility to use ChatGPT's capabilities without the need for large-scale medical domain training. This approach was particularly advantageous in terms of computational efficiency and model adaptability, as it required minimal setup and allowed for efficient adaptation to our specific sentiment analysis tasks. This few-shot approach demonstrated strong alignment with human-labeled data and provided an efficient alternative to fine-tuning, making it a practical choice for exploratory studies and applications requiring adaptability across varied patient feedback data.

The accuracy was high; ChatGPT 3.5 turbo on average achieved 90% in terms of accuracy, 0.90 for Precision, Recall, and F1-Score. Moreover, when ChatGPT 3.5 turbo was used to recognize aspects; particularly health related aspects such as medicines, body parts, medical procedures, medical staff, food and daily activities; it shows more detailed list of aspects compared to annotations provided by BERT-based models in [53,57–59,61]. The capability of LLMs such as ChatGPT is represented by facilitating the process of optimizing the output according to the required needs.

Though ChatGPT 3.5 turbo outperformed other models, many false classifications were noticeable, particularly within the medicine category. Many terms have been classified as medicines, yet, they are sort of food or any none-medical terms.

3.2.10. Comparison Between the Models

ABSA Model (BERT-Based): utilizing a pre-trained BERT model fine-tuned for the ABSA task. This approach benefits from BERT's deep understanding of language details and context but requires careful adjustment and can be computationally intensive. Moreover, it provides less accuracy compared to the other models. This might be due to specific features of the dataset used, which mainly contains patients online feedbacks.

Multilabel Multiclass Model (LSTM and CNN): Utilizes a hybrid architecture combining LSTM for capturing sequential data dependencies, CNN for local pattern recognition within sequences, and GloVe for utilizing pre-trained word embeddings. This model is customized for the task and shows adaptability through the integration of various techniques to enhance performance and reduce overfitting.

ChatGPT with few-shot achieved high accuracy (90%), which was outperforming the other models. This is might be due to the powerful capabilities of LLMs.

While models aim to address complex classification tasks, their architectural differences lead to distinct strengths and challenges. The BERT-based ABSA model shows promising precision, recall, and F1-scores, important for tasks requiring fine-grained sentiment analysis. In contrast, a custom LSTM and CNN model demonstrates high accuracy and effective generalization from training to testing, showcasing the benefits of a hybrid, fine-tuned approach to mitigate overfitting.

Finally, LLMs with few-shot provides the less complex configuration as it provides efficient approach to customize the output with less burden to the other aspects (except the payment to subscribe to the service). Moreover, the few-shot was given a competitive accuracy compared to other types of models.

In terms of performance, a comparative analysis of training time and parameter counts revealed the following: (a) BERT: High parameter count (~110 M) and longest training time, reflecting its heavy computational needs; (b) LSTM: Moderate parameter count (often ~10 M) and significantly reduced training time, making it more efficient for scenarios with limited computational resources; and (c) Bidirectional LSTM: Slightly higher parameter count and training time than LSTM due to its bidirectional architecture, yet notably faster and less demanding than BERT.

This choice of models reflects a trade-off between accuracy and efficiency. BERT's interpretive depth was beneficial for detailed sentiment analysis, while the LSTM-based models offered computationally efficient alternatives. Yet, GPT outperforms in terms of accuracy regardless the computational cost.

3.3. Human Validation

To assess how well the models' sentiment predictions aligned with human interpretations, we conducted external validation through a human evaluation process. Approximately 10% of the dataset was independently reviewed by two expert annotators with backgrounds in healthcare and natural language processing. These annotators manually assigned sentiment labels (positive, neutral, or negative) to the same set of aspects identified by the models. Each annotator independently reviewed a subset of the modelannotated data, focusing on seven primary aspects (e.g., 'medicines', 'medical procedures', 'diagnoses'). The annotators then assigned sentiment labels to these aspects based on the context within each forum post. Inter-annotator agreement was calculated using Cohen's kappa [69,84] to ensure consistency between human evaluations, yielding an average kappa score of 0.87, indicating strong agreement.

After the human annotations were complete, we compared these with the sentiment labels generated by the models (BERT, LSTM, Bidirectional LSTM, ChatGPT 3.5). The comparison revealed a high alignment between human labels and model predictions, particularly for the ChatGPT3.5 model, with an average accuracy of 93% in sentiment prediction against human-annotated labels. Discrepancies between model predictions and human annotations were further analyzed. Most mismatches occurred with complex or ambiguous sentiment expressions, such as mixed sentiments in discussions of 'medical procedures' where both positive and negative aspects were mentioned within the same post. These cases highlighted the models' occasional difficulty in handling complex, mixed sentiments, which are often better interpreted by humans due to their contextual judgment.

This human validation process not only provided an additional layer of verification for the model outputs but also offered insights into areas where the models could be further refined to improve alignment with human sentiment interpretation, especially in handling mixed sentiment cases. This alignment analysis highlights the models' robustness and highlights areas for future enhancement to improve interpretative accuracy in sentiment analysis of patient feedback.

4. Results

4.1. Patient Profiles Findings

During the content analysis of patient feedback in medical forums, a variety of patient profiles emerged, revealing interesting patterns in how individuals share and discuss their medical experiences. Predominantly, patients reported on their own medical conditions, with the next most common reports coming from partners detailing their significant others' health issues. Another notable group consisted of those discussing their parents' medical situations, while a smaller fraction mentioned the health conditions of grandparents.

The profiles essentially fell into two broad categories. The first category included patients who directly reported their own or their family members' medical cases. Analysis of approximately 10% of the posts from each author in this category brought to light several key observations: firstly, newly diagnosed patients often expressed confusion and sought advice about the medical procedures and medications prescribed by their doctors. Their posts reflected a strong desire to learn from others who had undergone similar health experiences. Secondly, a significant number of these patients expressed dissatisfaction with the effectiveness of their medications, especially painkillers. Thirdly, there was noticeable hesitancy to undergo surgery, with many seeking advice and alternatives to such invasive procedures. Lastly, these patients frequently shared their personal experiences with pain and the outcomes of their medical treatments, which, in many cases, involved reports of unsuccessful treatments, though a few did report successful outcomes.

The second profile consisted of individuals, primarily former or current patients, who responded to posts from the first group. Their contributions to the forum were primarily advisory in nature, often cautioning against the overuse of painkillers or premature resort to surgical interventions, advocating for physical therapy first. They also shared their own medical experiences, including the results of the treatments they had undergone. Additionally, these respondents frequently recommended alternative approaches, such as specific treatments, dietary changes, or particular activities, as potential solutions to the issues raised by the first group.

This diversity in patient profiles and the nature of their interactions on medical forums provides valuable insights into the complexities and nuances of patient experiences and their approaches to managing health conditions.

4.2. Wording Findings

The majority of words were related to advice: "any advice", "any help", "suggestion", and "idea". Next were medical terms: medications, medical procedures, or diagnosis. Finally, they used words showing appreciation and gratefulness for advice or help. In posts that related to replies, the words were more related to giving advice: "it is better to. . ., I suggest . . ., do surgery . . ., do test . . ., having this food, doing this activity"; asking for updates: "please let me know what happened with you, any updates . . .". Finally, there were words wishing speedy recovery and good luck. The work of Dhajate and Joshi [1] revealed that patients are willing to share their medical experience to help other patients, which may explain their frequent posts. However, this study found that those who initiate by revealing their experiences are seeking help/advice, while those with similar experiences were responding. This difference is due to the different resources used by this study (patient.info), which is global, while Dhajate and Joshi [1] used only Indian online patients feedbacks regarding online consultation. This circumstance could explain the differences between this work and that of Dhajate and Joshi [1]. Moreover. the work of Bond and Ahmed [68] showed that patients share their experiences as well external information (gathered from general reading or heard from a health professional). This study chimes with the findings of Bond and Ahmed [68] in terms of confirming that patients often share their experience. However, this study also found a significant number of posts revealing that the experiences reported were related to a relative (husband, wife, mother, father).

The wide variability in the readability scores (as indicated by the high standard deviation) suggests that some patient narratives may require higher reading comprehension levels, while others are highly accessible. This could reflect differences in patients' health literacy, education levels, or the complexity of their medical conditions. The outlier (negative readability score) may warrant further investigation to determine if there were issues with data formatting or text structure that affected the readability calculation, which was performed before assessing any proposed models. Finally, improving the readability of patient feedback is crucial for healthcare providers and researchers, as clearer narratives could facilitate better communication and understanding of patients' experiences, indicating more investment in technologies that narratively analyze human input/writing.

The credibility analysis shows that most patient feedback (more than 75%) does not contain biased language, as reflected by the median and 75th percentile values of 0. However, a small subset of feedback (around 4.15%) includes multiple biased phrases, which may indicate exaggerated claims or emotional responses. Posts with higher bias counts could potentially be less reliable, as extreme language (either overly positive or negative) might not provide a balanced view of the patient's experience. These results suggest that while most feedback is credible, it is important to consider the language used in

posts to determine whether certain narratives are driven by strong emotions or unrealistic expectations, which could affect the credibility of the insights derived from these posts.

4.3. Aspect-Based Sentiment Analysis Findings

The average values (three values within the range 0~1, the highest among negative, neutral, or the highest positive is the one tagging the aspect) produced by the ABSA model regarding each group of aspects has been collected. Apart from a few, most of the medicines were tagged with negative sentimental values. In terms of mentioning human parts where patients complain about pain in or show improvement in those parts it is expected to see negative sentimental values because people come to a medical forum to complain and seek advice.

The medical treatments, including procedures, were tagged most of the time with negative sentiment values, apart from a few procedures. The diagnoses found most frequently in the posts were related to the subsection of the medical forum, i.e., posts in sub-section "arthritis" are considerably related to this sub-section. This means patients intentionally surf the forum and selected the most relevant sub-section to their posts. Regarding medical staff, doctors were mentioned in negative posts more in comparison to other staff such as physiotherapists—although these were mentioned in few but positive posts. Finally, food was the only aspect group, particularly herbs, which was positively tagged in the medical forum.

4.4. Deep Learning Models Findings

The values generated by DeBERTa while the dataset was annotated was used to train and test the regression models. The regression models were trained using 75% of the dataset with settings described in Table 6. We calculated and compared the Mean Squared Error (MSE) for each output variable and the overall MSE for the four models: the original model, the fine-tuned model with a different architecture, the fine-tuned model with LSTM layers, and the Bidirectional LSTM model with Dropout. The test dataset (25%) was used in the evaluation. In several outputs, the original model scored the lowest MSE values (outputs 2, 3 and 7 and the overall MSE), which makes it the best among the four models evaluated. However, the second and fourth models performed better with outputs 4 and 5 and outputs
1 and 6, respectively. Details are presented in Table 6. Technically, the regression-based models are very less common within sentimental analysis. Accordingly, further discussion
will be on classification.

Table 6. Performance comparisons among the four deep learning models.

Model	Output1 (Medicine): (MSE)	Output2 (Pain Locations): (MSE)	Output3 (Medical Procedures): (MSE)	Output4 (Diagnoses): (MSE)	Output5 (Medical Staff): (MSE)	Output6 (Food): (MSE)	Output7 (Activity): (MSE)	Overall MSE
Original model	0.1704	0.4423	0.2604	0.3082	0.2800	0.0174	0.1749	0.2362
Fine-tuned model	0.1777	0.4451	0.2830	0.3159	0.2752	0.0165	0.1817	0.2421
Fine-tuned model using LSTM layers	0.1737	0.5096	0.3716	0.3824	0.4106	0.0156	0.2374	0.3001
Bidirectional LSTM model with Dropout layers	0.1605	0.5564	0.3580	0.3813	0.3672	0.0142	0.1879	0.2894

Classification models went through the same former procedures, and the best accuracy that could be obtained without overfitting/underfitting was ~85%, which was achieved by LSTM with word embedding (GloVE).

With classification models, Receiver Operating Characteristic (ROC) Analysis was conducted with the best model (LSTM+ GloVE embedding) (Figure 3). The performance of this study's classification model was evaluated using ROC curves and AUC values for each class, representing different categories relevant to the patients feedbacks. The ROC curve is a graphical representation of the trade-off between the true positive rate and false positive rate at various threshold settings,

and the AUC provides a single measure of overall accuracy that is independent of a specific threshold. Higher AUC values indicate better model performance. The results for each class are as follows: (a) Medicine: the ROC curve for the 'Medicine' class demonstrated excellent model performance with an AUC of 0.92. This indicates a high true positive rate and a low false positive rate, suggesting that the model is highly effective in identifying relevant instances in this category. (b) Organs: the AUC was 0.75. While this indicates acceptable performance, it suggests room for improvement, especially in reducing the false positive rate. (c) Medical Procedure: the 'Medical Procedure' class showed a good level of accuracy, with an AUC of 0.84. This reflects the model's robustness in correctly classifying procedural aspects within the dataset. (d) Diagnosed: The model exhibited strong performance in the 'Diagnosed' category, with an AUC of 0.90. This high value underscores the model's effectiveness in accurately identifying diagnosed conditions. (e) Staff: similarly, the 'Staff' class achieved an AUC of 0.84, indicating a good balance between true positive and false positive rates in classifying data related to healthcare staff. (f) Food: exceptionally, the 'Food' class achieved a perfect AUC of 1.0, indicating that the model was able to classify all relevant instances correctly without any false positives or false negatives. This is an ideal scenario, though rare in practical applications. (g) Activity: the 'Activity' class also demonstrated near-perfect model performance, with an AUC of 0.99. This suggests that the model is highly capable of discerning activities with great accuracy.

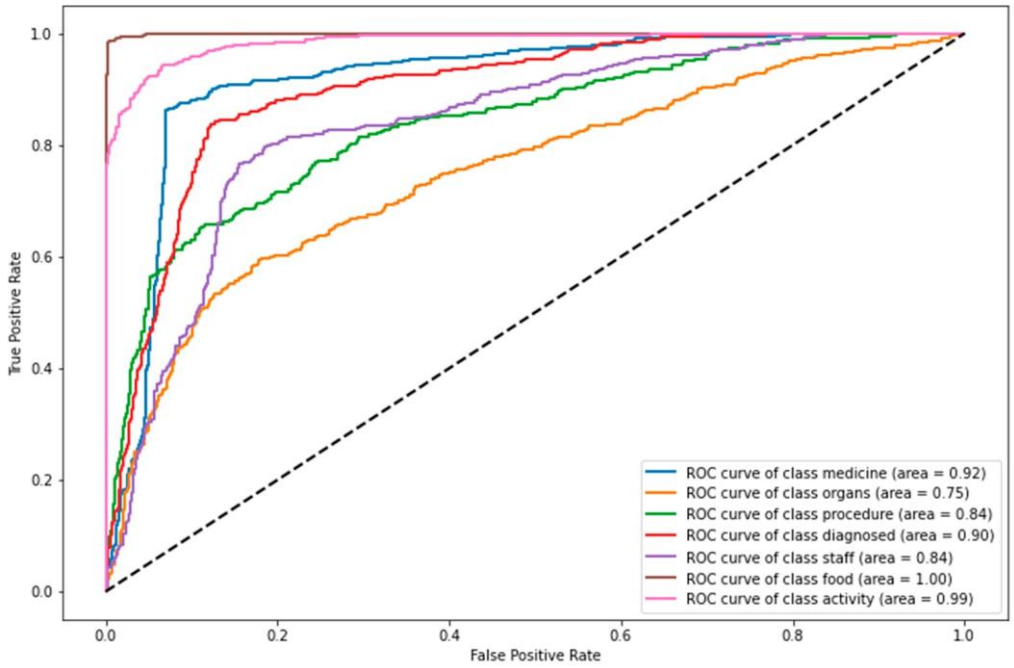

Figure 3. Receiver Operating Characteristic (ROC).

Overall, these results indicate that the developed models are highly effective in distinguishing between various patient-feedback-related categories, with particularly strong performance in classes like 'Food' and 'Activity'. While the performance in 'Organs' is comparatively lower, it still reflects a reasonable level of accuracy. These findings are promising for the application of this model in digital healthcare scenarios, particularly in automated classification tasks. Moreover; classification models are more related to the context of this study in contrast to regression models due to the fact that the concern is knowing the polarity (positive/negative/neutral) related to the aspect more than knowing the exact weight (±0~1), which eventually has no further usage.

5. Discussion

In this study, patients' shared experiences on medical forums are examined for the valuable information they contain, which can be crucial for health professionals and medical establishments. The coherency and credibility analyses reveal high trustworthy in patients' words.

While many previous studies have focused predominantly on patients' posts seeking support or commenting on health service quality, often using data from social media, this research diverges by concentrating on a specialized platform, specifically a medical forum. This approach utilizes content analysis to gain a comprehensive understanding of the types of information presented in patients' feedback. The study taps into a global information source, patient.info, where patients from various countries share their experiences and seek medical advice from peers.

Content analysis provided a deeper insight into patients' concerns compared to automated tools like Latent Dirichlet Allocation (LDA). This method accurately identified the number of aspect types within the feedback, along with related aspects, terms, and keywords. Seven aspect types were discerned (as shown in Table 2), facilitating the effective use of aspect-based sentiment analysis (ABSA), since the aspects to be assessed were already established. This foundation also supported the development of deep learning models to predict sentiment values associated with these aspects. The study draws inspiration from "The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication" and NIHR [14], but places a greater emphasis on patients' medical experiences rather than just service-related issues. One aspect type, "medical staff", aligns with findings in the aforementioned book.

Contrasting with Dhajate and Joshi [1], who used content analysis to collect nouns from patient feedback, creating a broad list of aspects later categorized into several groups, this study covers a wider domain, focusing on a more extensive range of patient interactions online. ABSA, known for its accuracy over classic sentiment analysis, requires training with domain-specific datasets (e.g., medical data) for optimal results [75]. This study sought to automate the process by generating a comprehensive list of aspects recognizable by a multi-label multiclass model, thus, laying the groundwork for building a specialized ABSA tool for patient sentiment.

During model testing, the CNN regression model demonstrated that CNNs can be as efficient as RNNs in processing textual data, particularly for sentiment analysis tasks. The CNN-based model slightly outperformed LSTM-based models. However, fine-tuning procedures did not significantly enhance model accuracy. Unlike previous studies, this research proposes and trains a deep learning model to automate the identification of aspects and their polarity in patient feedback. The model output provides quick insights into patients' intentions or complaints. CNN classification models experienced overfitting issues, but LSTM and LSTM with CNN achieved good accuracy, particularly when finetuned with techniques like regularization. These models show potential for increased accuracy with the addition of more data.

ChatGPT with few-shots has demonstrated great flexibility in terms of inputs and outputs. It is a powerful tool as it provides facilities for adjusting the output format and components required by users. When ChatGPT 3.5 is provided with few-shots and the required format, it generates output with high accuracy and minimal supervision. Moreover, in contrast to other AI models, the complexity of adjusting the output of ChatGPT is much easier while maintaining competitive accuracy. The experience of this study in employing ChatGPT with few-shots to extract aspects in patients' posts and related polarity (sentimental value) revealed competitive results and ease of use compared to in-house developed models. This is due to the unlimited resources available for large language models (LLMs) such as ChatGPT. This is in line with findings of Kotei and Thirunavukarasu [44] that highlighted the capabilities of GPT with NLP tasks.

When comparing ChatGPT with the BERT model, several key differences and performance metrics can be highlighted: Flexibility and Adaptability-ChatGPT offers greater flexibility in terms of input types and can generate more human-like responses. Its ability to handle few-shot learning makes it highly adaptable to various tasks without extensive retraining. Meanwhile, BERT, while powerful for specific NLP tasks, requires fine-tuning for each particular application. This fine-tuning process can be resource-intensive and time-consuming. Regarding the ease of use, ChatGPT is easier to use for generating text in different formats and adjusting components as needed. Users can guide the output by providing examples (few-shots), which reduces the need for complex adjustments. In contrast, BERT involves a more complicated setup for fine-tuning and requires a more detailed understanding of the model architecture and task-specific modifications. In

summary, while both ChatGPT and BERT are powerful NLP models, ChatGPT's ease of use and flexibility with few-shot learning make it particularly advantageous for tasks requiring adaptable and user-guided outputs. BERT, with its fine-tuning approach, remains a strong contender for high-accuracy tasks but may require more resources and expertise to achieve optimal performance.

5.1. Implications

Numerous studies have acknowledged the wealth of health-related information contained in social media and medical forums, highlighting the beneficial implications these data can have for patients and healthcare services alike [45,85]. The works of [8–10,14,29] have highlighted the value of patient feedback in pinpointing weaknesses in health services. However, online forums offer more than just critiques of health services; they serve as platforms where patients share experiences, including unconventional treatment options that might not be discussed in clinical settings. Effectively utilizing this information can prove highly beneficial.

By identifying and categorizing specific aspects of healthcare experiences (such as medications, procedures, and interactions with medical staff) alongside their associated sentiments, healthcare providers can gain a clearer understanding of common patient concerns. For example, if feedback consistently indicates negative sentiments regarding a particular medication or procedure, providers can investigate these concerns further and consider adjustments in treatment recommendations, patient education, or pain management strategies. Moreover, real-time monitoring of patient sentiment trends could allow healthcare providers to proactively address emerging issues. For example, if there is a noticeable increase in negative feedback related to a new treatment protocol, the organization can conduct an internal review or provide additional support resources to help patients manage any unintended effects.

Additionally, insights into patient interactions with medical staff, both positive and negative, can be used to inform targeted training programs. Understanding specific areas
where patients feel dissatisfied or report communication gaps enables healthcare organizations to train staff in empathetic communication, improve bedside manner, and provide clearer explanations about medical procedures, thus, fostering stronger patient–provider relationships. By integrating these sentiment analysis insights into existing healthcare management systems, providers can make data-informed decisions that directly address patient feedback, ultimately enhancing service quality, patient satisfaction, and clinical outcomes. These real-world applications highlight the potential of sentiment analysis in creating more responsive and patient-centered healthcare practices.

Contrary to the work in [2,8], who focused on patient satisfaction with medical staff and health services, this study investigated the health issues and medical experiences detailed in patient feedback. While Powell, et al. [8], Atherton, et al. [9] reported skepticism among medical staff about the value of online patient feedback due to its complexity, this study, along with Boylano, et al. [10], suggests otherwise.

Employing a deep learning model, this study presents detailed information about the topics and aspects found in patient posts, along with their associated polarity. This approach, also seen in research like that of Bez, Georgescu and Farazi [15], relies on manual content analysis to ensure comprehensive coverage of aspects. This method is favored over automated topic modeling approaches like LDA, which require pre-determining the number of topics. Content analysis proves more efficient in identifying a definitive list of aspects and related terms.

Moreover, manual analysis of patient feedback on medical forums provides an indepth understanding of the context and subtle themes that automated tools might miss. This gap was identified as many named entity recognition tools are not equipped to detect aspects specific to the medical domain, such as medications and diagnoses. Despite Spacy biomedical tools recognizing many medical entities, they are more geared towards entity recognition, aiding in automating note-taking rather than extracting patient sentiments.

This study aligns with previous research [2,25] on the significance of forums for patient knowledge exchange, emotional and social support, and self-management [45]. However, the

findings here suggest that patients seek tangible solutions and success stories more than emotional or social support. This is consistent with Hossain, et al. [22], who found limited evidence of patients seeking online emotional or social support, contrary to the findings of Smailhodzic, et al. [45], Armstrong and Powell [48], Hadert and Rodham [49], Shigaki, et al. [50]. The specialized nature of the medical forum examined in this study might explain this discrepancy, as it is likely a platform where tangible solutions are sought more than emotional support.

This study's findings, which include references to medical staff, align with prior research by Bez, Georgescu and Farazi [15], Mazanderani, et al. [29]. While Loo, et al. [86] focused on patient feedback regarding pharmacy services, this study extends to encompass all potential aspects of patient feedback, not just those pertaining to the quality of health services. The outcomes of this research can be seen as an effort to fulfill the recommendations set forth by Powell, et al. [8], Boylano, et al. [10] in developing techniques to extract patient opinions about medicines, treatments, and medical procedures.

Employing LLMs in inspecting patients online feedbacks in this study was among the earliest effort in employing LLMs for deep analysis of patients' feedback online. Moreover, it was found in this study that the few-shot is a significant approach in terms of ease of use and accuracy. The highlighted aspects found in this study can be utilized in medical application that concern patients' feedback regarding medical services provided, staff performance, and effectiveness of medical procedures and treatments. In contrast to the work of Lu, et al. [82] that emphasized augmenting feedback's text before passing them to LLMs for the ABSA task, the patients feedbacks analyzed in this study were sufficiently providing the opinions of patients. Therefore, the LLMs with few-shot used in this study produce high accuracy in terms of aspects and their associated sentimental value.

5.2. Future Directions

While this study offers valuable insights, there are several areas for future research that could address current limitations and expand the applicability of this work across other healthcare domains. The proposed directions for future research include:

Model Improvement and Domain Adaptation: Future research could focus on finetuning large language models specifically for the healthcare domain. While few-shot learning proved effective, further fine-tuning on a larger, more specialized dataset would likely enhance the model's accuracy and robustness in understanding patient language. Additionally, domain-specific pre-training of models like BERT or DeBERTa on medical corpora could improve their performance on sentiment classification tasks involving complex medical terminology.

Incorporation of Multimodal Data: Expanding the model to analyze multimodal data, such as combining text-based patient feedback with audio or video content (e.g., patient interviews or consultations), could offer deeper insights into patient sentiment. This approach would capture more nuanced emotional and behavioral indicators, providing a more comprehensive understanding of patient experiences.

Bias and Credibility Adjustments: To improve the handling of biased or exaggerated feedback, future research could focus on developing algorithms that automatically detect and adjust for emotional intensity in patient posts. Building on the current bias analysis, incorporating a weighting mechanism could enable the model to treat highly biased posts with adjusted sentiment weights, thereby improving the reliability of sentiment scores across diverse feedback.

Expansion to Additional Healthcare Domains: The methodology developed in this study could be applied to feedback from other healthcare sectors, such as mental health, chronic disease management, or pediatric care. Each of these areas has unique patient concerns and terminologies that would benefit from targeted sentiment analysis, contributing insights that healthcare providers can use to improve specialized services.

Real-Time Sentiment Monitoring System: Based on this study's findings, a real-time monitoring system could be developed to detect changes in patient sentiment trends over time.

Such a system would enable healthcare providers to proactively respond to emerging issues in patient satisfaction, potentially improving outcomes and operational efficiency.

Integration with Predictive Analytics: Future work could explore integrating sentiment analysis with predictive analytics to assess the potential impact of patient sentiment on health outcomes, treatment adherence, or patient retention. This would provide a more actionable framework, allowing healthcare providers to anticipate patient needs and refine care strategies accordingly.

By pursuing these future research directions, the field could further enhance the accuracy, adaptability, and practical applications of sentiment analysis models, ultimately driving improvements in patient-centered healthcare.

6. Conclusions

Previous studies have primarily focused on analyzing patient feedback to assess health service quality and explore the nature of patient support for one another. This study, however, extends beyond these traditional boundaries, affirming that patient feedback encompasses much more. While sentiment analysis has been conventionally applied to determine the polarity of feedback, particularly concerning health services, our research endeavors to present a broader and more inclusive list of aspects commonly found in patient feedback. This approach, although extensively tested on feedback specific to osteoarthritis, cardiology, and diabetes, identified seven potential categories that could be applicable in broader contexts if more data were gathered from various medical forums. The diversity of aspects uncovered indicates that medical forums are a richer source of patient stories about medical conditions than what is typically found in social media, as most previous studies have reported. These forums tend to capture a more comprehensive range of experiences, extending beyond just emotional or social support aspects.

A significant empirical contribution of this study is demonstrating the viability of using deep learning techniques and LLMs to identify both the aspects and their corresponding polarities in patient feedback. This advancement aids in better understanding the effectiveness of treatments and medications as experienced by patients. The ongoing work in this study aims to further automate the process of aspect identification and polarity analysis through a deep learning model, with an ultimate goal to develop a more specialized aspect-based sentiment analysis (ABSA) tool. This work, therefore, not only contributes to the field of sentiment analysis but also offers valuable insights for healthcare professionals and medical service providers, enhancing our understanding of patient experiences in medical forums.

Author Contributions: Conceptualization, R.M. and O.S.A.; methodology, R.M.; software, O.S.A.; validation, O.S.A., M.H. and R.M.; formal analysis, R.M.; investigation, M.H.; resources, O.S.A.; data curation, R.M.; writing—original draft preparation, R.M. and O.S.A.; writing—review and editing, M.H. and R.M.; visualization, R.M.; supervision, O.S.A.; project administration, M.H.; funding acquisition, O.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) grant number JRCAI-RG-08 And The APC was funded by SDAIA-KFUPM Joint Research Center for Artificial Intelligence.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: These data were derived from the following resources available in the public domain: https://patient.info/forums (accessed on 21 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dhajate, N.; Joshi, R. Classification of reviews of e-healthcare services to improve patient satisfaction: Insights from an emerging economy. *J. Bus. Res.* **2023**, *164*, 114015. [CrossRef] [PubMed]
2. Baker, P.; Brookes, G.; Evans, C. The language of patient feedback: A corpus linguistic study of online health communication. *Engl. Specif. Purp.* **2019**, *56*, 1–2. [CrossRef]

3. Farnood, A.; Johnston, B.; Mair, F.S. A mixed methods systematic review of the effects of patient online self-diagnosing in the 'smart-phone society' on the healthcare professional-patient relationship and medical authority. *BMC Med. Inf. Decis. Mak.* **2020**, *20*, 1–14. [CrossRef]
4. NHS. *Involving People in Their Own Health and Care: Statutory Guidance for Clinical Commissioning Groups and NHS England*; NHS: London, UK, 2017.
5. NHS. GP Patient Survey 2015–2016. Available online: https://www.england.nhs.uk/statistics/2016/07/07/gp-patient-survey2015-16/ (accessed on 4 May 2023).
6. Cuteanu, A.; Seguin, M.; Ziebland, S.; Pope, C.; Leydon, G.; Barnes, R.; Murray, E.; Atherton, H.; Stevenson, F. Qualitative study: Patients' enduring concerns about discussing internet use in general practice consultations. *BMJ Open* **2021**, *11*, e047508. [CrossRef]
7. Hedges, L.; Couey, C. How Patients Use Online Reviews. Available online: https://www.softwareadvice.com/resources/howpatients-use-online-reviews/ (accessed on 12 March 2024).
8. Powell, J.; Athertono, H.; Williams, V.; Mazanderani, F.; Dudhwala, F.; Woolgaro, S.; Boylano, A.-M.; Fleming, J.; Kirkpatrick, S.; Martin, A.; et al. Using online patient feedback to improve NHS services: The INQUIRE multimethod study. *Health Serv. Deliv. Res.* **2019**, *7*, 1–150. [CrossRef]
9. Atherton, H.; Fleming, J.; Williams, V.; Powell, J. Online patient feedback: A cross-sectional survey of the attitudes and experiences of United Kingdom health care professionals. *J. Health Serv. Res. Policy* **2019**, *24*, 235–244. [CrossRef]
10. Boylano, A.-M.; Turk, A.; van Velthoven, M.H.; Powell, J. Online patient feedback as a measure of quality in primary care: A multimethod study using correlation and qualitative analysis. *BMJ Open* **2020**, *10*, e031820. [CrossRef]
11. Dudhwala, F.; Boylano, A.-M.; Williams, V.; Powell, J. VIEWPOINT: What counts as online patient feedback, and for whom? *Digit. Health* **2017**, *3*, 2055207617728186. [CrossRef]
12. Entwistle, V.; Firnigl, D.; Ryan, M.; Francis, J.; Kinghorn, P. Which experiences of health care delivery matter to service users and why? A critical interpretive synthesis and conceptual map. *J. Health Serv. Res. Policy* **2012**, *17*, 70–80. [CrossRef]
13. Vahdat, S.; Hamzehgardeshi, L.; Hessam, S.; Hamzehgardeshi, Z. Patient Involvement in Health Care Decision Making: A Review. *Iran. Red Crescent Med. J.* **2014**, *16*, e12454. [CrossRef]
14. NIHR. Improving Care by Using Patient Feedback. Available online: https://content.nihr.ac.uk/nihrdc/themedreview-04327PE/Patient-Feedback-WEB.pdf (accessed on 2 April 2024).
15. Bez, S.M.; Georgescu, I.; Farazi, M.S. TripAdvisor of healthcare:Opportunities for value creation through patient feedback platforms. *Technovation* **2023**, *121*, 102625. [CrossRef]
16. Björklund, K.; Stenfors, T.; Nilsson, G.; Leanderson, C. Learning from patients' written feedback: Medical students' experiences. *Int. J. Med. Educ.* **2022**, *31*, 19–27. [CrossRef] [PubMed]
17. Eijkelboom, M.C.L.; de Kleijn, R.A.M.; van Diemen, W.J.M.; Maljaars, C.D.N.; van der Schaaf, P.; Frenkel, P. Patients as Feedback Providers: Exploring Medical Students' Credibility Judgments. *Perspect. Med. Educ.* **2023**, *12*, 129–140. [CrossRef] [PubMed]
18. Fox, S. *The Social Life of Health Information*; Pew Research Center: Washington, DC, USA, 2011.
19. Jia, X.; Pang, Y.; Liu, L. Sally Online Health Information Seeking Behavior: A Systematic Review. *Healthcare* **2021**, *9*, 1740. [CrossRef]
20. Gupta, A.; Tyagi, M.; Sharma, D. Use of social media marketing in healthcare. *J. Health Manag.* **2013**, *15*, 293–302. [CrossRef]
21. Coulson, N.S. How do online patient support communities affect the experience of inflammatory bowel disease? An online survey. *JRSM* **2013**, *4*, 2042533313478004. [CrossRef]
22. Hossain, S.N.; Jaglal, S.B.; Shepherd, J.; Perrier, L.; Tomasone, J.R.; Sweet, S.N.; Luong, D.; Allin, S.; Nelson, M.L.; Guilcher, S.J.; et al. Web-Based Peer Support Interventions for Adults Living with Chronic Conditions: Scoping Review. *JMIR Rehabil. Assist. Technol.* **2021**, *8*, e14321. [CrossRef]
23. Bartlett, Y.K.; Coulson, N.S. An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication. *Patient Educ. Couns.* **2011**, *83*, 113–119. [CrossRef]
24. Ziebland, S.; Powell, J.; Briggs, P.; Jenkinson, C.; Wyke, S.; Sillence, E.; Harris, P.; Perera, R.; Mazanderani, F.; Martin, A.; et al. *Examining the Role of Patients' Experiences as a Resource for Choice and Decision-Making in Health Care: A Creative, Interdisciplinary Mixed-Method Study in Digital Health*; NIHR Journals Library: Southampton, UK, 2016.
25. Harrington, M.; Heywood, B.; Rura, S.; Wicks, P. PRM95—An open research exchange for online patient feedback in pro development. *Value Health* **2014**, *17*, A197. [CrossRef]
26. Vilendrer, S.; Levoy, E.; Miller-Kuhlmann, R.; Amano, A.; Brown-Johnson, C.; De Borba, L.; Luu, J.H.; Sakamuri, S.; Gold, C.A. Physician Perceptions of Performance Feedback and Impact on Personal Well-Being: A Qualitative Exploration of Patient Satisfaction Feedback in Neurology. *Jt. Comm. J. Qual. Patient Saf.* **2023**, *49*, 138–148. [CrossRef]
27. Schäfer, H.; Idrissi-Yaghir, A.; Bewersdorff, J.; Frihat, S.; Friedrich, C.M.; Zesch, T. Medication event extraction in clinical notes: Contribution of the WisPerMed team to the n2c2 2022 challenge. *J. Biomed. Inform.* **2023**, *143*, 104400. [CrossRef] [PubMed]
28. McMaster, C.; Chan, J.; Liew, D.F.; Su, E.; Frauman, A.G.; Chapman, W.W.; Pires, D.E. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. *J. Biomed. Inform.* **2023**, *137*, 104265. [CrossRef] [PubMed]

29. Mazanderani, F.; Kirkpatrick, S.F.; Ziebland, S.; Locock, L.; Powell, J. Caring for care: Online feedback in the context of public healthcare services. *Soc. Sci. Med.* **2021**, *285*, 114280. [CrossRef] [PubMed]
30. Liu, W.; Tang, J.; Qin, J.; Xu, L.; Li, Z.; Liang, X. MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System. *arXiv* **2020**, arXiv:2010.07497.
31. Zeng, G.; Yang, W.; Ju, Y.; Wang, S.; Zhang, R.; Zhou, M.; Zeng, J.; Dong, X.; Zhang, R.; Fang, H.; et al. Meddialog: Large-scale medical dialogue datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 9241–9250.
32. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. Cblue: A Chinese biomedical language understanding evaluation benchmark. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 7888–7915.
33. Kim, Y.M.; Lee, T.H.; Na, S.O. Constructing novel datasets for intent detection and Ner in a Korean healthcare advice system: Guidelines and empirical results. *Appl. Intell.* **2022**, *53*, 941–961. [CrossRef]
34. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2020**, *3*, 1–23. [CrossRef]
35. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Hong Kong, China, 7 November 2019; pp. 2567–2577.
36. Huy, T.D.; Tu, N.A.; Vu, T.H.; Minh, N.P.; Phan, N.; Bui, T.H.; Truong, S.Q. ViMQ: A Vietnamese Medical Question Dataset for Healthcare Dialogue System Development. In *Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2023.
37. Mondal, I.; Ahuja, K.; Jain, M.; O'Neill, J.; Bali, K.; Choudhury, M. Global Readiness of Language Technology for Healthcare: What Would It Take to Combat the Next Pandemic? In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 4320–4335.
38. Bai, G.; He, S.; Liu, K.; Zhao, J. Incremental intent detection for medical domain with contrast replay networks. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022.
39. Bao, Q.; Ni, L.; Liu, J. Hhh: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In Proceedings of the Australasian Computer Science Week Multiconference, Melbourne, VIC, Australia, 4–6 February 2020; pp. 1–10.
40. Chen, Q.; Zhuo, Z.; Wang, W. BERT for Joint Intent Classification and Slot Filling. *arXiv* **2019**, arXiv:1902.10909.
41. Mehta, D.; Santy, S.; Mothilal, R.K.; Srivastava, B.M.L.; Sharma, A.; Shukla, A.; Prasad, V.; Sharma, A.; Bali, K. Learnings from technological interventions in a low resource language: A case-study on Gondi. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 2832–2838.
42. Daniel, J.E.; Brink, W.; Eloff, R.; Copley, C. Towards automating healthcare question answering in a noisy multilingual lowresource setting. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 948–953.
43. Mullick, A.; Mondal, I.; Ray, S.; Raghav, R.; Chaitanya, G.S.; Goyal, P. Intent Identification and Entity Extraction for Healthcare Queries in Indic Languages. *arXiv* **2023**, arXiv:2302.09685.
44. Kotei, E.; Thirunavukarasu, R. A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. *Information* **2023**, *14*, 187. [CrossRef]
45. Smailhodzic, E.; Hooijsma, W.; Boonstra, A.; Langley, D.J. David Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Serv. Res.* **2016**, *16*, 1–14. [CrossRef]
46. Broom, A. The eMale: Prostate cancer, masculinity and online support as a challenge to medical expertise. *J. Sociol.* **2005**, *41*, 87–104. [CrossRef]
47. Chiu, Y.C.; Hsieh, Y.L. Communication online with fellow cancer patients: Writing to be remembered, gain strength, and find survivors. *J. Health Psychol.* **2013**, *18*, 1572–1581. [CrossRef] [PubMed]
48. Armstrong, N.; Powell, J. Patient perspectives on health advice posted on Internet discussion boards: A qualitative study. *Health Expect.* **2009**, *12*, 313–320. [CrossRef] [PubMed]
49. Hadert, A.; Rodham, K. The invisible reality of arthritis: A qualitative analysis of an online message board. *Invis. Real. Arthritis: A Qual. Anal. Online Message Board* **2008**, *6*, 181–196. [CrossRef] [PubMed]
50. Shigaki, C.L.; Smarr, K.L.; Gong, Y.; Donovan-Hanson, K.; Siva, C.; Johnson, R.A.; Ge, B.; Musser, D.R. Social interactions in an online self-management program for rheumatoid arthritis. *Chronic Illn.* **2008**, *4*, 239–246. [CrossRef]
51. Lempp, H.; Scott, D.; Kingsley, G. The personal impact of rheumatoid arthritis on patients' identity: A qualitative study. *Chronic Illn.* **2006**, *2*, 109–120. [CrossRef] [PubMed]
52. Zhao, Y.; Zhang, L.; Zeng, C.; Lu, W.; Chen, Y.; Fan, T. Construction of an aspect-level sentiment analysis model for online medical reviews. *Inf. Process. Manag.* **2023**, *60*, 103513. [CrossRef]
53. Boudjellal, N.; Zhang, H.; Khan, A.; Ahmad, A.; Naseem, R.; Shang, J.; Dai, L. ABioNER: A BERT-based model for Arabic biomedical named-entity recognition. *Complexity* **2021**, *2021*, 6633213. [CrossRef]

54. Kalyan, K.S.; Rajasekharan, A.; Sangeetha, S. AMMU: A survey of transformer-based biomedical pretrained language models. *J. Biomed. Inform.* **2022**, *126*, 103982. [CrossRef]
55. Nerella, S.; Bandyopadhyay, S.; Zhang, J.; Contreras, M.; Siegel, S.; Bumin, A.; Silva, B.; Sena, J.; Shickel, B.; Bihorac, A.; et al. Transformers in Healthcare: A Survey. *arXiv* **2023**, arXiv:2307.00067.
56. Wang, B.; Xie, Q.; Pei, J.; Chen, Z.; Tiwari, P.; Li, Z.; Fu, J. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ACM Comput. Surv.* **2023**, *56*, 1–52. [CrossRef]
57. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. Bertweet: A pre-trained language model for English tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 9–14.
58. Müller, M.; Salathé, M.; Per, E.; Kummervold, P.E. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Front. Artif. Intell.* **2023**, *6*, 1023281. [CrossRef] [PubMed]
59. Tutubalina, E.; Alimova, I.; Miftahutdinov, Z.; Sakhovskiy, A.; Malykh, V.; Nikolenko, S. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* **2021**, *37*, 243–249. [CrossRef] [PubMed]
60. Madan, M.; Madan, R.; Thakur, P. Analysing the patient sentiments in healthcare domain using Machine learning. *Procedia Comput. Sci.* **2024**, *238*, 683–690. [CrossRef]
61. Basaldella, M.; Liu, F.; Shareghi, E.; Collier, N. Cometa: A corpus for medical entity linking in the social media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3122–3137.
62. Naseem, U.; Khushi, M.; Reddy, V.; Rajendran, S.; Razzak, I.; Kim, J. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021.
63. Clusmann, J.; Kolbinger, F.R.; Muti, H.S.; Carrero, Z.I.; Eckardt, J.N.; Laleh, N.G.; Löffler, C.M.L.; Schwarzkopf, S.C.; Unger, M.; Veldhuizen, G.P.; et al. The future landscape of large language models in medicine. *Commun. Med.* **2023**, *3*, 141. [CrossRef]
64. Peng, C.; Yang, X.; Chen, A.; Smith, K.E.; PourNejatian, N.; Costa, A.B.; Martin, C.; Flores, M.G.; Zhang, Y.; Magoc, T.; et al. A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **2023**, *6*, 210. [CrossRef]
65. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
66. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.; Steinberg, E.; Fleming, S.; Pfeffer, M.A.; Fries, J.; Shah, N.H. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit. Med.* **2023**, *6*, 135. [CrossRef]
67. Patient.Info. About-Us. Available online: https://patient.info/about-us (accessed on 15 December 2023).
68. Bond, C.S.; Ahmed, O.H. Can I help you? Information sharing in online discussion forums by people living with a long-term condition. *J. Innov. Health Inf.* **2016**, *23*, 620–626. [CrossRef]
69. Hsieh, H.F.; Shannon, S.E. Three Approaches to Qualitative Content Analysis. *Qual. Health Res.* **2005**, *15*, 1277–1289. [CrossRef]
70. Glenn. Analyzing Open-Ended Questions. Available online: http://intelligentmeasurement.net/2007/12/18/analyzing-openended-questions/ (accessed on 20 May 2024).
71. Akgül, Y. Evaluating the performance of websites from a public value, usability, and readability perspectives: A review of Turkish national government websites. *Univers. Access Inf. Soc.* **2024**, *23*, 975–990. [CrossRef]
72. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *23*, 221–233. [CrossRef] [PubMed]
73. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending against neural fake news. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
74. Fogg, B.J. *Persuasive Technology: Using Computers to Change What We Think and Do*; Morgan Kaufmann: Burlington, MA, USA, 2003.
75. Jin, W.; Zhao, L.; Liu, C.; Yu, H. Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis. *Inf. Process. Manag.* **2023**, *60*, 103260. [CrossRef]
76. Yang, H.; Zeng, B.; Xu, M.; Wang, T. Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning. *arXiv* **2021**, arXiv:2110.08604.
77. Cohen, S.; Presil, D.; Katz, O.; Arbili, O.; Messica, S.; Rokach, L. Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time. *Inf. Fusion* **2023**, *99*, 101887. [CrossRef]
78. Rolczyn´ski, R. Do You Trust in Aspect-Based Sentiment Analysis? Testing and Explaining Model Behaviors. In *Rafał Rolczyn´ski*; SCALAC SP. Z O. O.: Warsaw, Poland, 2021; Volume 2023.
79. Wang, H.; Li, J.; Hua, W.; Hovy, E.; Sun, Y. Pre-Trained Language Models and Their Applications. *Engineering* **2022**, *25*, 51–65. [CrossRef]
80. Kamıs¸, S.; Goularas, D. Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data. In Proceedings of the 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 26–28 August 2019.
81. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef]
82. Lu, H.Y.; Liu, T.C.; Cong, R.; Yang, J.; Gan, Q.; Fang, W.; Wu, X.J. QAIE: LLM-based Quantity Augmentation and Information Enhancement for few-shot Aspect-Based Sentiment Analysis. *Inf. Process. Manag.* **2025**, *62*, 103917. [CrossRef]

83. Zhang, H.; Zhang, Y.; Zhan, L.M.; Chen, J.; Shi, G.; Lam, A.; Wu, X.M. Effectiveness of Pre-Training for Few-Shot Intent Classification. *arXiv* **2021**, arXiv:2109.05782.
84. Petricˇ, G.; Cugmas, M.; Petricˇ, R.; Atanasova, S. The quality of informational social support in online health communities: A content analysis of cancer-related discussions. *Digit. Health* **2023**, *9*, 20552076231155681. [CrossRef] [PubMed]
85. Srimarut, T.; Techasatian, K. Use of social media in health care by patients and health care professionals: Motives & barriers in Thailand. *Utopía Y Prax. Latinoam.* **2019**, *24*, 215–223.
86. Loo, J.; Greaves, G.; Lewis, P.J. Exploring patients' pharmacy stories: An analysis of online feedback. *Int. J. Clin. Pharm.* **2021**, *43*, 1584–1593. [CrossRef] [PubMed]