Journal of Universal Computer Science, vol. 30, no. 1 (2024), 3-24
submitted: 26/3/2023, accepted: 12/9/2023, appeared: 28/1/2024 CC BY-ND 4.0

What is the Consumer Attitude toward Healthcare Services? A Transfer Learning Approach for Detecting Emotions from Consumer Feedback

Bashar Alshouha
(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain
https://orcid.org/0000-0001-6475-4248, bashar.alshouha@alu.uclm.es)

Jesus Serrano-Guerrero
(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain
https://orcid.org/0000-0002-6177-8188, jesus.serrano@uclm.es)

David Elizondo
(De Montfort University, Institute of Artificial Intelligence, School of Computer Science and Informatics, Leicester, United Kingdom
https://orcid.org/0000-0002-7398-5870, elizondo@dmu.ac.uk)

Francisco P. Romero
(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain
https://orcid.org/0000-0002-6993-2434, franciscop.romero@uclm.es)

Jose A. Olivas
(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain
https://orcid.org/0000-0003-4172-4729, joseangel.olivas@uclm.es)

Abstract The capability of offering patient-centered healthcare services involves knowing the consumer needs. Many of these needs can be conveyed through opinions about services that can be found on social networks. The consumers/patients can express their complains, satisfaction, frustration, etc. in terms of feelings and emotions toward those services; for that reason, it is pivotal to accurately detect them. There are many recent techniques to detect sentiments or emotions, but one of the most promising is transfer learning. This allows adapting a model originally trained for a task to a different one by fine-tuning. Following this idea, the primary objective of this research is to study whether several pre-trained language models can be adapted to a task such as patient emotion detection in an efficient manner. For this purpose, seven clinical and biomedical pre-trained models and four domain-general models have been adapted to detect multiple emotions. These models have been tuned using a dataset consisting of real patient opinions which convey several emotions per opinion. The experiments carried out state the domain-specific pre-trained models outperform the domain-general ones. Particularly, Clinical-Longformer obtained the best scores, 98.18% and 95.82% in terms of accuracy and F1-score, respectively. Analyzing the patient feedback available on social networks may provide valuable knowledge about consumer sentiments and emotions, especially for healthcare managers. This information can be very interesting for purposes such as assessing the quality of healthcare services or designing patient-centered services.

Keywords: Transfer learning, Emotion recognition, Consumer feedback, Pre-trained model
Categories: H.3.0, H.3.3, J.7, J.4
DOI: 10.3897/jucs.104093

1 Introduction and background

Public opinions expressed on blogging sites and social networking platforms can be a valuable source of information related to the feelings and emotions of the masses toward subjects in the fields of medicine, governance, or e-commerce, among others. Particularly, in the field of medicine, knowing the patient emotions or feelings can be a crucial point for determining what a user thinks about the services delivered by a particular healthcare organization. Accordingly, healthcare organization managers could comprehend the patients' expectations by analyzing their feelings or independent evaluation organizations could rank hospitals or medical services.

The use of online health communities can improve patient-peers exchange beneficial support but also, the provided text could be used to analyze the patient satisfaction, and not necessarily using questionnaires as in [Haldar et al. 2020], but using automatic sentiment analysis techniques. Furthermore, from these online health communities, it could be also possible to study the evolution of the emotions, how the patient mood can evolve over time. This fact could be useful, for instance, for implementing patient-centered services depending on those emotions. For that reason, it is paramount to have automatic mechanisms for emotion detection that can take advantage of the information available on social networks.

Natural language processing (NLP) techniques have been broadly applied to clinical text, for example, for analyzing the content of electronic health records (EHRs) which can usually contain clinical information and free-text notes about treatments, progress, family history, etc. These techniques have been mainly used to automate tasks or exploit the available information [Khattak et al. 2019]. Furthermore, these records can contain information about the patient sentiments, which can be useful, for instance, to detect suicidal behaviors [Bittar et al. 2021] as well as the probability of dying after being discharged [Smith et al. 2018]. In this case, it is necessary to resort to a closely related field called sentiment analysis, whose main techniques are based on machine learning and deep learning, lexicons, or hybrid approaches [Birjali et al. 2021].

Regarding machine learning techniques, studies such as [Alemi 2012, Greaves et al. 2013] applied diverse techniques such as naïve Bayes, decision trees, and support vector machines (SVM) to classify free text in clinical notes as a positive or negative opinion. In other studies [Sanglerdsinlapachai et al. 2021, Niu et al. 2005, Sarker and Paris 2011, Carrillo et al. 2018], Unified Medical Language System (UMLS) was used to generate features for machine learning-based sentiment analysis. Previous studies have emphasized that the incorporation of UMLS semantic types with content-based features improved the analysis of sentiments from clinical narratives texts [Sanglerdsinlapachai et al. 2021, Yuan et al. 2022]. Jimenez-Zafra et al. presented a study about the use of lexicon based-approaches for clinical texts in Spanish to detect negative or positive reviews about drugs and physicians [Jiménez-Zafra et al. 2019]. On the other hand, deep learning mechanisms arose to deal with some machine learning's drawbacks. For instance, in the area of medicine, different deep learning architectures have been proposed to deal with sentiment detection on drug reviews [Basiri et al. 2020, Zhang et al. 2019].

Yadav et al. proposed a deep convolutional neural network (CNN) model and compared it with traditional ML algorithms for detecting a variety of medical sentiments based on the users' medical conditions and medications [Yadav et al. 2018].

Moreover, new trends related to deep learning are arising such as transfer learning, which is a new paradigm whose fundamental goal is to reuse the knowledge learned for one task to tackle other tasks [Nourani and Reshadat 2020]. Recently, transformer models are being increasingly utilized for NLP tasks such as text classification, clinical concept extraction or named entity recognition (NER) [Qiu et al. 2020]. Over the last few years, extensive approaches have been developed to generate pre-training contextual representations, such as bidirectional encoder representations from transformers (BERT) [Devlin et al. 2018], robustly optimized BERT pretraining approach (RoBERTa) [Zhuang et al. 2021], ELMo [Cassani et al. 2017], or ULMFiT [Howard and Ruder 2018]. Most of the publicly available pre-trained language models (PLMs) are trained on general domain corpora such as Wikipedia; nevertheless, medical texts greatly differ from general text because of their domain-specific vocabulary. Consequently, the performance of general PLMs is limited in many tasks [Moradi et al. 2020]. To address this issue, several PLMs trained on medical corpora have been proposed such as Clinical-Longformer [Yikuan et al. 2023], CODER [Yuan et al. 2022], BioBERT [lee et al. 2020], SciBERT [Beltagy et al. 2020] or Bio_ClinicalBERT [Alsentzer et al. 2019], among others.

The PLMs in NLP have primarily evolved into three main types: transformer decoders-only, transformer encoders-only, and transformer encoder-decoders, each serving as a powerful framework for training PLMs that have been outperformed in various NLP tasks [Wang et al. 2022]. An example of a transformer decoder-only model is GPT (Generative Pre-trained Transformer), which used a unidirectional transformer decoder to generate text token by token [Radford et al. 2018]. BERT is a prominent model in the transformer encoders-only category, which utilized bidirectional transformer encoders for masked language modeling and bidirectional context understanding [Devlin et al. 2018]. On the other hand, transformer encoder-decoder frameworks, represented by models such as T5 (Text-to-Text Transfer Transformer) [Raffel et al. 2020] and ERNIE 3.0 [Sun et al. 2021], have been aimed at pre-training sequence-to-sequence generation models, with T5 framing all NLP tasks as text-to-text problems and ERNIE 3.0 outstanding in capturing bidirectional context understanding and generating contextually rich word embeddings.

In addition to the architectural differences, there is a distinction between short-sequence and long-sequence models. Short-sequence models such as BERT [Devlin et al. 2018], BioBERT [lee et al. 2020], or ClinicalBERT [Alsentzer et al. 2019] have been designed to efficiently process relatively small input sequences with standard transformer architectures. On the other hand, long-sequence models such as Longformer [Beltagy et al. 2020] and BigBird [Zaheer et al. 2020] have been tailored to handle much longer input sequences, such as lengthy documents or narratives, by incorporating novel attention mechanisms to efficiently process and understand long texts.

These PLMs have been used in many tasks, for instance, to classify lifestyle factors such as physical activity and excessive diet from clinical texts [Shen et al. 2021], to identify medication mentions in clinical notes [Schäfer et al. 2023], to extract oncologic outcomes [Araki et al. 2023], to detect clinical NER, classify biomedical text and predict disease diagnosis [Ni et al. 2021], to categorize sentiments (positive or negative) [Punith et al. 2021] or to classify texts into clinical specialties in different languages [Pomares-Quimbaya et al. 2021], but little work can be found on emotion detection.

The assessment of hospital services quality has been primarily done by manual questionnaires or interviews; nonetheless, the availability of online opinions from millions of healthcare users/patients is an interesting information source to assess the performance of a hospital. The detection of the emotions towards the different services of a hospital can be a valuable indicator about the quality of the offered services. Some studies can be found in the field of sentiments (positive or negative values) or emotions considering bag-of-words (BOW), term frequency–inverse document frequency (TFIDF), and Word2vec [Khaleghparast et al. 2023], deep learning architectures [Serrano-Guerrero et al. 2022] or fuzzy approaches [Serrano-Guerrero et al. 2023, Serrano-Guerrero et al. 2022], but not studying mainly the quality of the hospitals from the point of view of the emotions and particularly, using transfer learning techniques.

Therefore, to the best of our knowledge, there is little research on sentiment detection and let alone on automatic emotion (rage, joy, sadness, etc.) detection in clinical texts. Particularly, there is little research using both pre-trained domain-general language models [Acheampong et al. 2021] and domain-clinical or biomedical pre-trained language models [Wang et al. 2023, Kalyan et al. 2022, Saffar et al. 2023], in spite of the fact that the patient opinions on healthcare online communities are full of emotions, more than in other classical platforms such as Amazon or Tripadvisor. For that reason, the objective of this work to study the effect of transfer learning as a mechanism to detect multiple emotions from patient opinions and understand the quality of the healthcare services from them. To do so, it is necessary to evaluate the quality of the different pretrained language models, domain-general and domain-specific, for detecting emotions and determine the quality of the health care services depending on these emotions. As a result, the main contributions of this study are:

- To study the effectiveness of transfer learning as a tool for detecting emotions on patient opinions.
- To determine what type of PLMs are more effective, domain-general and domain-specific, to predict multiple emotions in clinical texts.
- To assess the quality of different PLMs specifically trained on clinical text.
- To provide a set of experiments dealing with opinions collected from real patients to support all conclusions.
- To present a case study to assess the services of a hospital through the patient emotions using the best pre-trained language model.

The rest of the study is organized as follows: Section 2 describes the methods and materials used to perform this study; Section 3 presents the experiments and results; Section 4 describes a study case and finally, Section 5 points out the reached conclusions.

2 Materials and methods

2.1 Dataset

The dataset was collected from a website called Careopinion[1] which enables patients to share their opinions about their experiences in a hospital. These free-text reviews are

[1] https://www.careopinion.org.uk Retrieved on july 23th, 2023

especially interesting because can describe many emotions at once over the different stages of a hospital stay (diagnosis, surgery, treatment, rehabilitation, etc.). 53,475 opinions were collected for training and testing the proposed methods to predict the patient emotions. The opinions were labeled using 8 emotions according to the terms used by the patient at the section "how did you feel" of every opinion (see Table 1), which is especially designed to capture his/her feelings when commenting his/her experiences. Those terms were searched on SenticNet [Cambria et al. 2020], and their associated emotions (joy, calmness, sadness, fear, eagerness, pleasantness, anger, or disgust) were used to label every opinion. Any opinion not including the section "how did you feel", it was removed.

Emotion	**Stories**	%
Anger	7,371	0.1378
Fear	15,007	0.2806
Joy	22,016	0.4117
Sadness	20,686	0.3868
Disgust	10,746	0.2009
Pleasantness	5,802	0.1084
Eagerness	15,691	0.2934
Calmness	17,025	0.3183

Table 1: Emotion distribution in the dataset

One of the most important challenges to be faced by all assessed models are the different lengths of the opinions. Overall, the length is long, 131.96 words per opinion on average; nonetheless, there are around $5,965$ opinions whose length is under 40 words, around 10% of the entire dataset.

2.2 Models

To categorize multiple emotions, seven available PLMs (Clinical-Longformer, CODER, BlueBERT, SciBERT, BioMed-RoBERTa, Bio_ClinicalBERT, BioBERT) which constitute a comprehensive view of the state-of-the-art in biomedical and clinical NLP, have been used and compared along with another four domain-general PLMs (BERT, RoBERTa, ELMo, ULMFiT) that have also achieved remarkable contributions in NLP tasks.

- **BlueBERT**. It is a domain-specific PLM based on the BERT-base model [Devlin et al. 2018] which was developed for the biomedical language understanding evaluation (BLUE) benchmark [Peng et al. 2019]. Its aim is to create a standardized benchmark that can be used to compare various models. It consists of a set of five tasks (document multi-label classification, sentence similarity, relation extraction, named entity recognition, and inference) derived from ten datasets that vary in genre, size, and difficulty, covering both biomedical and clinical texts. For our experiments, the "BlueBERT-Base, uncased, PubMed[2]" version has been used.

[2] https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-12_H-768_A-12 Retrieved on july 23th, 2023

- **BioMed-RoBERTa** [Gururangan et al. 2020]. It is a recent model based on the RoBERTa-base architecture [Zhuang et al. 2021]. It is initialized from RoBERTa-base and trained using a corpus of 2.7 million scientific papers from Semantic Scholar, with an additional pre-training of 12.5K steps and a batch size of 2,048 [Ammar et al. 2018]. For our experiments, the "biomed_roberta_base"[3] version has been selected.

- **ClinicalBERT** [Alsentzer et al. 2019]. It is a domain-specific PLM trained on clinical texts from around 2 million clinical notes of the publicly available MIMIC III database. For the experiments, the "Bio_ClinicalBERT"[4] checkpoint, which is initialized from BioBERT and trained on all MIMIC notes for 150K steps with a batch size of 32, has been used.

- **BioBERT** [lee et al. 2020]. It is a domain-specific PLM initialized from the BERT-base model [Devlin et al. 2018], with additional pre-training on Wikipedia, BooksCorpus [Zhu et al. 2015] and, both PubMed abstract and PMC full-text articles in the biomedical domain. In our experiments, the implementation used was "biobert-base-cased-v1.1"[5], trained over 200K steps on PubMed and 270K steps on PMC, followed by an additional 1M-step training on PubMed, and utilizing the same hyperparameter tunning for BERT-base.

- **SciBERT** [Beltagy et al. 2020]. It is a masked language model trained on a corpus of 1.14 M full-text papers from Semantic Scholar [Ammar et al. 2018], of which 82% belong to the biomedical domain and 18% are computer science papers to improve the performance on downstream scientific NLP tasks. There are several versions of SciBERT, the selected one for the experiment was "SciBERT-scivocab-cased"[6], which trained BERT from scratch on the scientific publication corpus along with its separate vocabulary called Scivocab [Beltagy et al. 2020].

- **Clinical-Longformer** [Yikuan et al. 2023]. It is a specialized language representation model designed for clinical text analysis. It is built upon the Longformer architecture, which enables it to handle long-term dependencies in medical documents effectively. It is pre-trained on a large corpus of clinical text. This model demonstrates remarkable performance in various clinical NLP tasks such as named entity recognition, medical code prediction, or clinical text classification. Its unique design and pre-training on domain-specific data make it a valuable tool for extracting meaningful information from complex medical texts. For the experimenta section, the "yikuan8/Clinical-Longformer"[7] version has been utilized for the experiments.

- **Cross-lingual knowledge-infused medical term embedding (CODER)** [Yuan et al. 2022]. The CODER model, particularly the UMLSBert_ENG variant, is a powerful language representation model designed for medical coding and clinical text understanding. It utilizes contrastive learning on the medical knowledge graph UMLS to train the model. This training approach calculates similarities using both terms and relations from the knowledge graph, effectively providing medical knowledge into the model embeddings. By incorporating such relations, the

[3] https://huggingface.co/allenai/biomed_roberta_base Retrieved on july 23th, 2023

[4] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT Retrieved on july 23th, 2023

[5] https://huggingface.co/dmis-lab/biobert-base-cased-v1.1 Retrieved on july 23th, 2023

[6] https://huggingface.co/allenai/scibert_scivocab_cased Retrieved on july 23th, 2023

[7] https://huggingface.co/yikuan8/Clinical-Longformer, Retrieved on july 23th, 2023

CODER model aims at providing improved machine-learning features for tasks such as medical term normalization, semantic similarity, or relation classification. With its advanced capabilities, the CODER model demonstrates superior performance compared to other biomedical and contextual embeddings in various benchmarks. Its ability to effectively capture cross-lingual medical term representations and leverage domain-specific knowledge makes it a valuable asset for medical text analysis, clinical decision support systems, and healthcare applications. The "GanjinZero/UMLSBert_ENG"[8] version has been utilized for the experiments.

- **BERT** [Devlin et al. 2018]. It is a language representation model introduced in [Devlin et al. 2018]. It is pre-trained on a vast, unannotated dataset. It excels in understanding language and performs exceptionally well in various NLP tasks. Using masked language modeling (MLM) and next sentence prediction (NSP), it achieves significant improvements in transfer learning and achieves remarkable results, particularly in text classification. These pre-training tasks enable it to grasp complex linguistic relationships and generalize effectively to downstream NLP tasks. It is available in two versions: BERT-Base Cased and Uncased. The "bert-base-cased"[9] version has been utilized for the experiments.

- **RoBERTa** [Zhuang et al. 2021]. It is a language representation model that builds on BERT and addresses its limitations. It is pre-trained on a larger corpus of data, uses dynamic masking, and is trained for a longer duration. This allows RoBERTa to capture a deeper understanding of the language and improve performance on NLP tasks such as text classification, named entity recognition, and sentiment analysis. For our experiments, the "roberta-base"[10] version has been used.

- **Embedding from Language Models (ELMo)** [Cassani et al. 2017]. It is a deep bi-directional language model trained on a large text corpus. It is a contextualized embedding at the word and character level. Instead of assigning a fixed embedding to each word, ELMo was designed to consider the entire context of the text. To create the embeddings, ELMo uses a bi-directional recurrent neural network (RNN) trained on a specific task. Since it uses a bidirectional architecture, the embedding relies on both the next and previous words in a sentence. For that reason, it surpasses preceding approaches when it comes to addressing the issues of polysemous phrases.

- **Universal Language Model Fine-Tuning (ULMFiT)** [Howard and Ruder 2018]. It was one of the first efficient approach for language model fine-tuning and can be used for many tasks in NLP. Its architecture is primarily based on an ASGD Weight-Dropped LSTM (AWD-LSTM) [Mikolov et al. 2013], without tuning the hyper-parameters except the dropout parameters. The model includes three main stages; first, it is pre-trained on Wikitext-based text, then it is fine-tuned on a target task, and finally, the classifier on the target task is also fine-tuned.

2.3 Evaluation metrics

As the goal of this study is to categorize multiple emotions from text, the following multilabel classification measures are used [Sorower 2010]:

[8] https://huggingface.co/GanjinZero/UMLSBert_ENG Retrieved on july 23th, 2023
[9] https://huggingface.co/bert-base-cased Retrieved on july 23th, 2023
[10] https://huggingface.co/roberta-base Retrieved on july 23th, 2023

- **Exact match ratio (EMR)** represents the ratio of instances that have all their labels correctly classified:

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \tag{1}$$

  where n is the number of instances, I is the indicator function, y_i is the number of true labels for the $i-th$ instance, and $\hat{y}_i$ the number of correctly predicted labels for the $i-th$ instance.

- **Zero_one_loss** calculates the ratio of instances whose actual value is not equal to the predicted value:

$$Zero\_one\_loss = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2}$$

- **Hamming loss (HL)** represents the ratio of emotions incorrectly predicted over the total number of emotions:

$$HL = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I(y_j^i \neq \hat{y}_j^i) \tag{3}$$

  where L is the number of emotions.

- **Accuracy (ACC)** calculates the ratio of correctly classified labels over the total number of labels:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i \cap \hat{y}_i}{y_i \cup \hat{y}_i} \tag{4}$$

- **Precision (P)** calculates the ratio of correctly identified labels over the total number of expected labels, averaged over all instances:

$$Precision(P) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i \cap \hat{y}_i}{y_i} \tag{5}$$

- **Recall (R)** It is the ratio of correctly identified labels to the total number of predicted labels, averaged over all instances:

$$Recall(R) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i \cap \hat{y}_i}{\hat{y}_i} \tag{6}$$

- **F1-score (F1)** is the harmonic mean of precision and recall:

$$F1 - score = \frac{1}{n}\sum_{i=1}^{n} \frac{2y_i \cap \hat{y}_i}{y_i + \hat{y}_i} \tag{7}$$

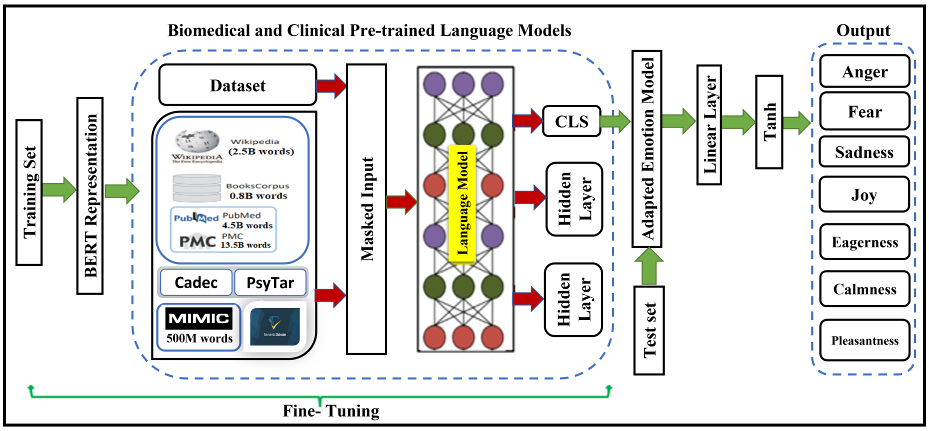

Figure 1: Fine-tuning method for all biomedical and clinical PLMs

– **Weighted average** considers label imbalance and can result in a value of F-weighted that is not between P-weighted and R-weighted. P-weighted, R-weighted, and F1-weighted are calculated as follows:

$$\mathbf{P}_{weighted} = \sum_{i=1}^{L} (P(i) \times weight(i)) \tag{8}$$

$$\mathbf{R}_{weighted} = \sum_{i=1}^{L} (R(i) \times weight(i)) \tag{9}$$

$$\mathbf{F1}_{weighted} = \sum_{i=1}^{L} (F1(i) \times weight(i)) \tag{10}$$

3 Experiments and results

3.1 Experimental setup

The models explained in subsection 2.2 were fine-tuned using the Careopinion dataset to detect multiple emotions from the patient opinions. Fig. 1 shows the necessary steps to detect the different emotions fine-tuning the biomedical and clinical PLMs. The training set (80% of the dataset) has been used to train them, and according to the instances of the training set (20% of the dataset), the weights of the biomedical and clinical PLMs have been optimized to classify the multiple emotions from the reviews. Finally, the test set has been utilized to evaluate the performance of the customized models.

Transfer learning adapts the knowledge obtained from a basic task to a target task. Specifically, in this research, the biomedical and clinical PLMs (Clinical-Longformer, CODER, Bio_ClinicalBERT, BioBERT, SciBERT, BlueBERT, BioMed-RoBERTa) have been adapted to perform multilabel emotion classification in clinical text. The fine-tunning and adaptation process consists of the following three stages. First, it is necessary to select the appropriate PLM (Clinical-Longformer, CODER, Bio_ClinicalBERT, BioBERT, SciBERT, BlueBERT, BioMed-RoBERTa) to perform transfer learning. The PLM enriches their data representations from the training set capturing the semantic and syntactic properties of the words. Second, it is necessary to identify a suitable layer of the model for adapting the knowledge to perform the emotion detection process. The last layer of all PLMs has been selected as the adaptation layer due to its simplicity and effectiveness [Gu et.al 2021]. Finally, it is necessary to identify the transfer strategy to implement between fine-tuning and feature extraction. In this case, fine-tuning has been selected because it is more suitable for classification tasks [Vrbancic and Podgorelec 2020].

The prior knowledge in the clinical and biomedical PLMs was fine-tuned for the task of emotion classification. The special classification token [CLS] is a special symbol added in front of every sequence in the BERT representation, and it is used for classification tasks [Devlin et al. 2018]. Accordingly, for emotion classification, the [CLS] token in the last hidden state of the model has been utilized for fine-tuning. The [CLS] token output of BERT's last layer is passed through a simple linear layer and the *tanh* activation function is applied to get the probability that an entry belongs to a specific emotion. In this manner, the linear layer weights from biomedical and clinical PLMs have been reconfigured according to the emotion classification task.

The proposed approaches for all of the biomedical and clinical PLMs were implemented using Pytorch by means of the SimpleTransformer[11] library and the transformer architecture for the fine-tuning processes [Vaswani et al. 2017]. The model hyperparameters were tuned using 'optimizer: Adam', 'learning-rate: 1e−5', 'batch-size:8', and 'num-train-epoch:3'.

For ULMFiT, the entire model was developed using Pytorch and fine-tuned with the Fastai[12] library. The model hyperparameters were tuned using this configuration: 'backpropagation through the time:70', 'batch size for the training PLM: 128', 'batch size for classifier:32, 'learning-rate: 1e-3', and 'num-train-epochs:5'.

For ELMo, the Tensorflow hub[13] was used. ELMo embeddings are PLMs available on Tensorflow Hub. The model hyperparameters were tuned using 'num-train-epochs:3', and 'batch size: 16'.

All hyperparameters were selected for this experimental section after trying several configurations, being these ones the most effective. Google Collab was the platform used to implement and perform all of the experiments.

3.2 Results and discussion

After carrying out the experiments, the results state the proposed biomedical and clinical PLMs (Clinical-Longformer, CODER, BioBERT, Bio_ClinicalBERT, SciBERT, BioMed-RoBERTa and BlueBERT) adapted for multiple emotion detection on patient feedback achieved substantially better results than general PLMs (BERT, RoBERTa, ELMo, ULMFiT) in terms of accuracy and F1-weighted average, as it is shown in Table 2 and 3. Furthermore, Clincal-Longformer obtained the best performance compared with the rest of the biomedical models in terms of accuracy and F1-weighted, 98.18% and 96.56%, respectively.

[11] https://simpletransformers.ai/ Retrieved on july 23th, 2023

[12] https://docs.fast.ai/ Retrieved on july 23th, 2023

[13] https://www.tensorflow.org/hub/overview Retrieved on july 23th, 2023

Models / Metrics	Bio_Clinical BERT	BioMed-RoBERTa	SciBERT	BlueBERT	BioBERT	Clinical-Longformer	CODER
EMR	85.22	79.89	82.87	81.56	86.43	**91.94**	87.65
0/1 loss	14.78	20.11	23.12	23.43	13.56	**8.06**	12.35
Ham. loss	0.0473	0.0678	0.0537	0.0543	0.0321	**0.0181**	0.0280
Accuracy	95.27	93.22	94.63	94.57	96.79	**98.18**	97.19
Precision	92.16	89.99	91.95	91.32	95.07	**96.16**	93.76
Recall	89.79	85.99	88.52	89.49	93.61	**95.50**	93.28
F1-score	90.10	86.79	89.20	89.42	93.82	**95.82**	93.51
P-weighted	93.58	90.52	92.69	91.53	95.32	**96.75**	95.05
R-weighted	88.41	83.41	86.62	87.81	92.85	**96.38**	94.38
F1-weighted	90.87	86.61	89.48	89.61	93.86	**96.56**	94.69

Table 2: Performance results for biomedical and clinical PLMs

Models/ metrics	ELMo	ULMFIT	Bert	Roberta
EMR	66.12	55.16	77.19	73.42
0/1 loss	33.87	44.83	22.81	26.58
Hamm. loss	13.92	18.27	0.0766	0.0921
Accuracy	86.08	81.73	92.34	90.79
Precision	74.12	65.52	82.03	82.52
Recall	73.93	65.59	86.13	80.10
F1-score	73.25	64.54	82.93	80.21
P-weighted	74.17	65.49	83.17	83.44
R-weighted	73.69	65.75	88.44	82.81
F1-weighted	73.92	65.61	84.83	82.25

Table 3: Performance results for general PLMs

Clinical-Longformer significantly outperformed BERT, RoBERTa, ELMo and ULMFiT, achieving significant improvements of 5.84%, 7.39%, 12.1%, and 16.45% for accuracy, respectively. One of the primary reasons why biomedical PLMs provided more promising results is that these models have gained domain-specific knowledge through pre-training on large volumes of biomedical text.

Even though every opinion contains several emotions, the clinical and biomedical PLMs perform quite well as the exact match ratio corroborates. This measure calculates the percentage of labels perfectly classified per opinion, that is, it is possibly one of the strictest metrics, and most of the results are over 80%, except for domain-general models, whose results are very low, especially, ULMFiT. This fact is also corroborated by the Hamming loss measure, which indicates the ratio of incorrect labels, and on which, the biomedical and clinical PLM obtained results under 0.1%.

Analyzing the rank-based metrics, Clinical-Longformer and CODER obtained very accurate results, 8.06 and 12.35, respectively, according to the zero_one_loss metric, which measures whether all emotions of a sample have been completely detected or not.

Regarding the classification capabilities of the PLMs with respect to each individual emotion (see Table 4 and 5), the results in terms of the precision show that Clinical-Longformer achieved the highest precision score for “anger”, “fear”, “sadness”, “calm-

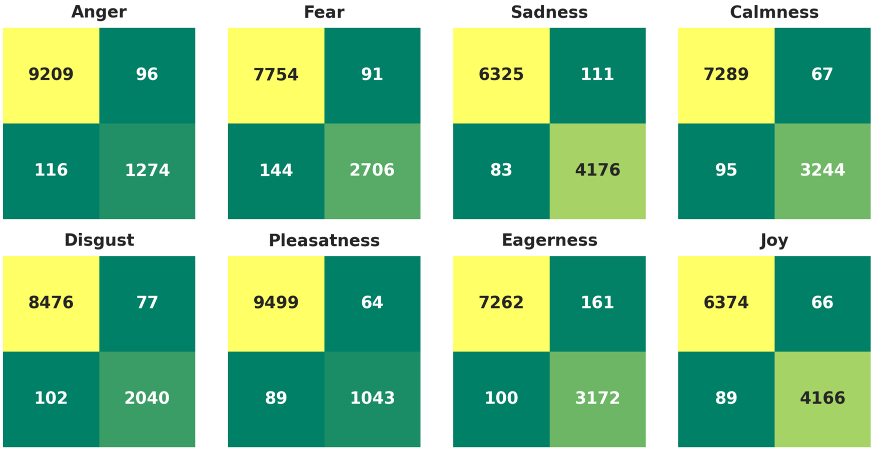

Figure 2: Confusion matrices for the best model (Clinical-Longformer)

ness", "disgust" and "joy", whereas BioBERT obtained the highest precision score for "pleasantness", and "eagerness". Regarding the best recall score, CODER just achieved the highest performance score for "eagerness", whereas Clinical-Longformer for the rest of the emotions. According to the F1-score, Clinical-Longformer achieved the highest performance score for all emotions, which makes it the most appropriate model for detecting emotions.

Observing the previous results, on the one hand, the label with the largest number of instances in the clinical dataset was "joy", comprising 41.17% of the instances in the dataset, which obtained the highest performance results in terms of precision and F1-score for all PLMs. On the other hand, "pleasantness" and "anger" had the smallest number of instances, comprising 10.84% and 13.78% of the instances in the dataset, respectively, and obtained the lowest performance results in terms of recall and F1-score. Hence, the more available instances there are, the better the adaption process is for the emotion detection task.

Since the best results were obtained by the biomedical and clinical PLMs, it is necessary to perform a deep analysis to understand their performance and associated weaknesses. To do so, the errors made when classifying have been also analyzed and the number of misclassified emotions has been summarized in Table 6. According to the three best classification models (Clinical-Longformer, CODER, BioBERT), the percentage of the instances mislabeled was 1.81%, 2.8%, and 3.34%, respectively. These percentages remark that the number of correctly predicted labels is quite large.

Corroborating the results in Table 4 and 5, the lowest performance (accuracy) was obtained by "fear" and "eagerness", which obtained the largest number of mislabeled instances. On the contrary, "joy" and "pleasantness" achieved the best accuracy, having the lowest number of mislabeled instances. Therefore, these ones seem easier to be classified by most of the PLMs in comparison with the other emotions. This fact is particularly clear for the best model, Clinical-Longfromer, as it can be seen in Fig. 2 which depicts the corresponding confusion matrices for each emotion.

The performance of PLMs is significantly influenced by several factors such as the architecture of the model, the domain specificity of the corpora used, and the size of the

Labels / metrics	**Anger**	**Fear**	**Sadness**	**Calmness**	**Disgust**	**Pleasantness**	**Eagerness**	**Joy**
SciBERT								
Accuracy	94.26	93.60	94.21	95.12	95.18	95.58	93.52	95.47
Precision	82.30	91.41	93.52	94.25	93.86	86.16	94.48	94.78
Recall	71.96	85.27	91.63	90.10	81.65	70.95	83.14	94.03
F1-score	76.78	88.23	92.56	92.12	87.33	77.81	88.44	94.40
BioMed-RoBERTa								
Accuracy	92.47	91.51	93.15	94.21	94.40	94.63	91.49	93.79
Precision	75.29	83.13	90.83	94.38	93.92	84.11	93.64	94.98
Recall	63.80	87.63	91.87	86.87	77.51	62.63	76.86	89.43
F1-score	69.07	85.32	91.34	90.47	84.93	71.80	84.32	92.12
Bio_ClinicalBERT								
Accuracy	96.64	93.91	94.88	95.55	95.59	95.98	94.43	95.09
Precision	87.71	89.91	93.24	96.15	95.36	84.55	95.44	96.53
Recall	86.65	88.26	93.79	89.53	82.34	77.37	85.42	91.18
F1-score	87.18	89.08	93.51	92.73	88.37	80.80	90.16	93.78
BlueBERT								
Accuracy	94.11	93.94	93.93	95.25	95.05	95.03	93.52	95.63
Precision	76.97	90.13	94.20	93.55	92.51	80.22	91.30	95.79
Recall	78.99	88.10	90.11	91.31	82.34	72.32	86.55	93.34
F1-score	77.96	89.10	92.11	92.41	87.13	76.07	88.86	94.55
BioBERT								
Accuracy	96.70	96.10	96.55	97.22	96.89	97.26	96.40	97.16
Precision	91.06	95.15	94.01	97.74	90.25	**94.54**	**96.79**	97.86
Recall	83.17	90.76	97.45	93.38	94.98	79.43	90.94	95.09
F1-score	86.94	92.90	95.70	95.51	92.56	86.35	93.77	96.46
Clinical-Longformer								
Accuracy	**98.01**	**97.80**	**98.18**	**98.48**	**98.32**	**98.56**	**97.55**	**98.55**
Precision	**92.99**	**96.74**	**97.41**	**97.97**	**96.36**	94.21	95.16	**98.44**
Recall	**91.65**	**94.94**	**98.05**	**97.15**	**95.23**	**92.13**	96.94	**97.90**
F1-score	**92.31**	**95.83**	**97.72**	**97.56**	**95.79**	**93.16**	**96.04**	**98.17**
CODER								
Accuracy	96.87	96.49	97.20	97.67	97.55	97.67	96.35	97.75
Precision	88.32	94.85	96.12	96.76	94.99	89.97	91.37	97.71
Recall	87.42	91.90	96.87	95.72	92.84	87.83	**96.95**	96.65
F1-score	87.87	93.39	96.49	96.24	93.90	88.89	94.07	97.17

Table 4: Results for each individual emotion for Biomedical and clinical PLMs

Labels / Metrics	**Anger**	**Fear**	**Sadness**	**Calmness**	**Disgust**	**Pleasantness**	**Eagerness**	**Joy**
ELMo								
Accuracy	91.44	84.43	81.84	83.96	88.30	92.25	84.02	82.31
Precision	67.21	72.29	76.25	75.76	70.05	66.03	73.47	79.10
Recall	67.54	71.73	77.68	73.72	72.12	62.02	72.66	77.80
F1-score	67.37	72.01	76.96	74.73	71.07	63.96	73.06	78.45
ULMFiT								
Accuracy	88.72	79.62	76.11	78.97	84.59	90.17	79.13	76.48
Precision	55.47	64.22	69.22	67.13	59.48	54.62	64.15	71.13
Recall	57.35	62.97	70.26	66.16	63.22	52.68	64.78	71.00
F1-score	56.39	63.59	69.74	66.64	61.29	53.63	64.46	71.16
BERT								
Accuracy	91.23	90.93	91.39	91.78	93.45	95.29	89.13	95.44
Precision	88.32	73.29	81.59	77.74	74.11	93.76	69.69	97.77
Recall	60.89	90.18	96.39	94.77	91.16	70.51	93.85	91.32
F1-score	72.08	80.86	88.37	85.41	81.76	80.49	79.99	94.44
RoBERTa								
Accuracy	89.64	88.92	90.21	90.59	91.89	93.61	87.43	94.00
Precision	88.46	74.79	81.71	77.77	74.25	93.85	71.49	97.82
Recall	56.05	81.33	93.01	90.50	82.99	62.81	85.81	88.29
F1-score	68.62	77.92	86.99	83.65	78.38	75.26	78.00	92.81

Table 5: Results for each individual emotion for general PLMs

Model/ Label	**Clinicalbert**	**biomed_ roberta**	**Scibert**	**bluebert**	**Biobert**	**Clinical-Long former**	**CODER**	**BERT**	**RoBERTa**	**ELMo**	**ULMFIT**
Anger	555	805	613	629	352	212	334	937	1,108	915	1,473
Fear	651	907	684	649	376	235	375	970	1,185	1,665	2,886
Sadness	547	732	619	648	368	194	299	920	1,047	1,942	2,541
Calmness	475	619	521	507	297	162	249	879	1,006	1715	2,897
Disgust	471	598	515	529	332	179	261	700	867	1,251	2,002
Pleasantness	429	574	472	531	293	153	249	503	683	828	1,169
Eagerness	595	910	693	692	385	261	390	1162	1,344	1709	2,934
Joy	525	664	484	467	303	155	240	487	641	1,891	2,293
#Total errors	4,248 (4.96%)	5,809 (6.78%)	4,601 (5.37%)	4,652 (5.43%)	2,859 (3.34%)	**1,551 (1.81%)**	2,397 (2.8%)	6,558 (7.66%)	7,881 (9.21%)	1,1916 (13.92%)	1,8195 (21.26%)

Table 6: Number of misclassified opinions per emotion

corpora. Table 7 provides a summary of the text corpora utilized for pre-training the PLMs, while Table 8 describes the specific combinations used in our implementation.

To understand why certain models outperform others, several factors can be considered. For instance, Clinical-Longformer's architecture was specifically designed to handle long sequences effectively, allowing it to capture the contextual information and dependencies present in the stories of the used dataset, thus leading to superior performance. The CODER model, on the other hand, achieved the second-best performance. As opposed to the masked language model task, CODER utilized medical standard terms obtained from Cadec and PsyTar datasets to learn how to normalize terms and generalize

Corpus	# Words	Domain
English wikipedia	2.5B	General
Wikitext-103	103M	General
Books corpus	0.8B	General
News Crawl data from WMT 2011	800M	General
PubMed abstracts	4.5B	Biomedical
PMC full-text articles	13.5B	Biomedical
Clinical notes (MIMIC-III)	>500M	Clinical
Cadec	6,754 medical terms	Biomedical
PsyTar	6,556 medical terms	Biomedical
Articles from Semantic Scholar	7.55B	Biomedical
Articles from Semantic Scholar	8.10B	Computer science (CS)
Careopinion dataset (see subsection 2.1)	∼3.6M	Opinions about medical experiences

Table 7: Description of the corpora used for pre-training the models

Model version	Corpora combination	# Words
BioBERT	English Wikipedia+ Books+ PubMed+ PMC+ Careopinion dataset	21.3B
Clinical-Longformer	MIMIC-III +Careopinion dataset	>503M
CODER	Cadec+ PsyTar+ Careopinion dataset	13,310 medical terms
BlueBERT	PubMed+ MIMIC-III+ Careopinion dataset	∼5B
Bio_ClinicalBERT	MIMIC-III + Careopinion dataset	>503M
SciBERT	Semantic Scholar- 18% (CS) and 82% Biomedical domain + Careopinion dataset	3.1B
BioMed-RoBERTa	Semantic Scholar-biomedical domain + Careopinion dataset	7.55B
BERT	Book +English Wikipedia+ Careopinion dataset	3.3B
RoBERTa	Book+ English Wikipedia+ CC-News+ OpenWebText+ Stories+ Careopinion dataset	>10B (160GB of text)
ULMFiT	Wikitext-103+ Careopinion dataset	106M
ELMo	News Crawl data from WMT 2011+ Careopinion dataset	803.6M

Table 8: Text corpora used for each pre-trained model implemented in the experiments

the terms that appear in social media. This approach not only helps the model learn how to normalize terms but also improved its ability to generalize to terms present in various contexts.

Additionally, it is important to note that all models, namely Clinical-Longformer, CODER, BioBERT, and Bio_ClinicalBERT, have been trained on the MIMIC-III dataset. The MIMIC-III dataset is a valuable resource in the medical domain, containing a vast collection of clinical notes, laboratory results, and other healthcare-related information. By leveraging this domain-specific dataset for pre-training, these models gain a deep understanding of the unique language patterns, terminology, and contextual nuances specific to the medical field. This further contributes to their improved performance compared to the more general-purpose language models in the medical domain.

Regarding the other domain-specific PLMs (BlueBERT, BioMed-RoBERTa, SciBERT), all have a similar number of biomedical and clinical words that were used for pre-training, and achieved similar results in terms of accuracy and F1-score. Overall, the results confirm that the models pre-trained on biomedical and clinical domains were significantly superior to the ones pre-trained on a general domain. Since the opinion dataset consists of many terms in the medical domain, this fact might explain why biomedical and clinical PLMs outperform BERT, RoBERTa, ELMo and ULMFiT, which could not benefit from their vocabularies.

Looking in detail at the vocabulary coverage of the domain-specific PLMs (see Table 9), BioBERT and Bio_ClinicalBERT provide the greatest coverage, including 15,132 words, that is, 14.07% of the Careopinion dataset. This fact may be another of the primary reasons why these models achieve the best performance. In this sense, BlueBERT and SciBERT are the third and fourth ones with highest coverage with 11,483 (10.68%) and 8,926 (8.02%) words, respectively. Hence, the use of a biomedical PLM is proved to be a useful technique when dealing with multi-label emotion detection in clinical and biomedical texts.

Model	Vocabulary	Included	%	Not included	%
BlueBERT	30,511	11,483	10.68	96,004	89.32
BioBERT	28,987	15,132	14.07	92,355	85.93
SciBERT	31,046	8,629	8.02	98,858	91.98
Bio_ClinicalBERT	28,987	15,132	14.07	92,355	85.93
CODER	28,895	16,468	15.32	91,019	84.67
Clinical-Longformer **BioMed-RoBERTa**	50,253	4,248	3.95	103,239	96.05

Table 9: Number of words of the biomedical and clinical PLMs included (or not) in the Careopinion dataset

It is also worth noting regarding BioMed-RoBERTa that despite having the largest number of words and ability to adapt to different domains, it just includes 4,248 (3.95%) words, has the weakest performance among the rest of models. This weakness can be explained because RoBERTa's vocabulary terms were obtained from different domains such as news, reviews, or biomedical sources, written in different languages. Nonetheless, Clinical-Longformer, having the same vocabulary, achieved the best performance among the evaluated models. This fact is considered one of its main limitations by its own authors because its vocabulary is derived from RoBERTa model, which uses 5,000 sub-word units primarily designed for non-clinical corpora [Yikuan et al. 2023]. Nevertheless, the training process is so powerful that can let it achieve the best performance. For that reason, the vocabulary coverage seems to be less important than being able to discover other characteristics such as lexical or syntactical structures.

One important advantage of biomedical and clinical PLMs is their efficiency when fine-tuning the data. The Clinical-Longformer, CODER, and rest of the clinical PLMs, for example, required approximately 18 hours, 8 hours, and one hour of training, respectively, utilizing a graphics processing unit (GPU) on Google Colab. This demonstrates the relatively quick training process enabled by these PLMs, whereas ELMo or UMLFiT needed to be trained for over 36 hours.

4 Study case: Assessing the services quality of a hospital

Once the quality of the different models has been tested, it is possible to find application for the emotions detected like assessing the quality of the hospitals according to their patients. To do so, from the dataset a hospital from Wishaw (Scotland) has been selected with the aim of analyzing the emotions regarding several aspects of the hospital. $6,308$ opinions in total were crawled.

A deep manual analysis of the patient reviews as well as the literature [Behdioğlu et al. 2019, Kuo et al. 2011, Raziei et al. 2018, Zarei 2015, Meesala and Paul 2018] was carried out to conclude that the best dimensions used to assess the quality of a hospital are:

- **Responsiveness:** It relates the hospital commitment to deliver its services with promptness and willingness.
- **Tangibles:** This dimension relates the physical appearance of buildings, equipment, uniforms, etc. and their maintenance.
- **Reliability:** This dimension focuses on the consistency and accuracy of the hospital services as well as the ability of the staff to deliver services as required.
- **Professionalism:** The expertise level and promptness of the hospital staff is vital when assessing a hospital.
- **Assurance:** How safe and confident the patients feel when they are in the hospital, is a very valuable dimension.
- **Empathy:** The staff must be careful and comprehensive when dealing with the patients.

These dimensions have been automatically detected using the topic modeling algorithm proposed in [Serrano-Guerrero et al. 2023, Serrano-Guerrero et al. 2022] and the corresponding texts to each dimension have been analyzed using the best algorithm found in the previous section: Clinical-Longformer.

Most of the expressions used by the patients convey emotions related to sadness, joy and fear. Analyzing the obtained results, these show that the dimension “professionalism” is very neutral for most of the patients, little emotional level has been conveyed regarding it. All emotions are under 0.2, expect joy 0.38. This fact contrasts with other dimensions such as “tangible” or “care” which have positive feelings, joy is 0.7 and 0.76, respectively. Nonetheless, other negative emotions such as anger, disgust, fear or sadness obtained relatively high values 0.26, 0.24, 0.73, and 0.67 for “tangible” and, 0.24, 0.2, 0.61, 0.51 for “care”, respectively. These values can alert the hospital managers about the fact that there are positive feelings in some opinions, nevertheless, some patients are not having the best experiences and it is necessary to analyze those values to understand why. “Responsiveness” can be considered highly rated because not only the score for joy was 0.62, but also, the pleasantness obtained the highest score (0.224) among all dimensions. On the other hand, “reliability” presents almost the lowest value for joy (0.5) and clear negative signs such as anger (0.22), disgust (0.20), fear (0.53) and sadness (0.45). This can be another indicator that patients are not content with some of the received services. Finally, “assurance” presents positive values for joy (0.62) and below 0.4 for fear and sadness and less than 0.2 for the rest of the emotions, therefore, it could be considered as another dimension having a good performance but improvable.

Overall, the patient opinions convey neutral and positive values about the different aspects; nonetheless, relatively high values for emotions such as anger or disgust can mean a significant piece of data to warn the hospital managers about faulty or improvable services.

5 Conclusions and future work

This study provides insights for hospital staff and managers to understand that patients' emotions are a valuable factor to consider when assessing the quality of the services provided by any hospital. It also contributes to the literature on how to collect the patients' experiences from social media and process it to capture the emotions using transfer learning.

An emotion detection comparison using transfer learning has been studied. Different PLMs, domain-specific and domain-general, have been assessed, resulting in a better performance of the domain-specific ones, and particularly, Clinical-Longformer achieved the best results in terms of F1-score and accuracy. Furthermore, a case study has been presented to analyze the different dimensions about the hospital services from the point of the view of the patient emotions.

The findings of this study regard emotion detection can lead other researchers to continue studying on fields such as the design of patient-centered healthcare services, or mechanisms to accurately capture opinions and measure the associated feelings/emotions along with their intensities.

Acknowledgments

This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant SAFER: PID2019-104735RB-C42 (AEI/FEDER, UE).

References

[Acheampong et al. 2021] Acheampong, F. A., Nunoo-Mensah, H., and Chen, W.: "Transformer models for text-based emotion detection: a review of BERT-based approaches"; In Artificial Intelligence Review, 54, 8 (2021), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

[Alemi 2012] Alemi, F., Torii, M., Clementz, L., Aron, D. C.: "Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments"; Quality Management in Healthcare, 21, 1 (2012), 9-19. https://doi.org/10.1097/QMH.0b013e3182417fc4.

[Alsentzer et al. 2019] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., McDermott, M.: "Publicly available clinical BERT embeddings"; In Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics (2019), 72–78.

[Ammar et al. 2018] Ammar W., Groeneveld D., Bhagavatula C., Beltagy I., Crawford M., Downey D.: "Construction of the literature graph in semantic scholar"; In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018), 3, 84–91.

[Araki et al. 2023] Araki, K., Matsumoto, N., Togo, K., Yonemoto, N., Ohki, E., Xu, L., Miyazaki, T.: "Developing artificial intelligence models for extracting oncologic outcomes from japanese electronic health records"; Advances in Therapy, 40, 3 (2023), 934-950.

[Basiri et al. 2020] Basiri, M. E., Abdar, M., Cifci, M. A., Nemati, S., Acharya, U. R.: "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques"; Knowledge-Based Systems, 198, (2020), 105949. https://doi.org/10.1016/j.knosys.2020.105949.

[Behdioğlu et al. 2019] Behdioğlu, S., Acar, E., and Burhan, H. A.: "Evaluating service quality by fuzzy SERVQUAL: a case study in a physiotherapy and rehabilitation hospital"; Total Quality Management and Business Excellence, 30, 3-4 (2019), 301–319.

[Beltagy et al. 2020] Beltagy, I., Lo, K., Cohan, A.: "SciBERT: A pretrained language model for scientific text"; In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (2020), 3615–3620.

[Beltagy et al. 2020] Beltagy, I., Peters, M. E., and Cohan, A.: "Longformer: The long-document transformer"; ArXiv Preprint ArXiv:2004.05150, (2020).

[Birjali et al. 2021] Birjali, M., Kasri, M., Beni-Hssane, A.: "A comprehensive survey on sentiment analysis: Approaches, challenges and trends"; Knowledge-Based Systems, 226, (2021), 107134. https://doi.org/10.1016/J.KNOSYS.2021.107134.

[Bittar et al. 2021] Bittar, A., Velupillai, S., Roberts, A., Dutta, R.: "Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: corpus-based analysis"; JMIR medical informatics, 9, 4 (2021), e22397. https://doi.org/10.2196/22397.

[Cambria et al. 2020] Cambria, E., Li, Y., Xing, F. Z., Poria, S., Kwok, K.:"SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis"; In Proceedings of the 29th ACM international conference on information and knowledge management (October 2020), 105-114.

[Carrillo et al. 2018] Carrillo-de-Albornoz, J., Rodriguez Vidal, J., Plaza, L.: "Feature engineering for sentiment analysis in e-health forums"; PloS one, 13, 11 (2018), e0207996. https://doi.org/10.1371/journal.pone.0207996.

[Cassani et al. 2017] Cassani L, Tomadonim, B., Ponce, A., Agüero, MV., Moreira, MR.: "Elmo-Deep Contextualized Word Representations", Food and Bioprocess Technology, 10, 8 (2017), 1454–1465. https://doi.org/10.48550/arXiv.1802.05365.

[Devlin et al. 2018] Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: "Bert: Pre-training of deep bidirectional transformers for language understanding"; Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA (June 2-7, 2019), 1, 4171–4186.

[Greaves et al. 2013] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L.: "Use of sentiment analysis for capturing patient experience from free-text comments posted online"; Journal of medical Internet research, 15, 11 (2013), e2721. https://doi:10.2196/jmir.2721.

[Gu et.al 2021] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Poon, H.: "Domain-specific language model pretraining for biomedical natural language processing"; ACM Transactions on Computing for Healthcare (HEALTH), 3, 1 (2021), 1-23.

[Gururangan et al. 2020] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N. A.: "Don't stop pretraining: Adapt language models to domains and tasks"; Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, Association for Computational Linguistics (July 5-10, 2020), 8342–8360.

[Haldar et al. 2020] Haldar, S., Mishra, S. R., Kim, Y., Hartzler, A., Pollack, A. H., Pratt, W.: "Use and impact of an online community for hospital patients"; Journal of the American Medical Informatics Association, 27, 4 (2020), 549-557. https://doi.org/10.1093/jamia/ocz212.

[Howard and Ruder 2018] Howard, J., Ruder, S.: "Universal language model fine-tuning for text classification"; In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, (2018), 328–339.

[Jiménez-Zafra et al. 2019] Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Molina-González, M. D., Ureña-López, L. A.: "How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain"; Artificial intelligence in medicine, 93, (2019), 50-57. https://doi.org/10.1016/j.artmed.2018.03.007.

[Kalyan et al. 2022] Kalyan, K. S., Rajasekharan, A., and Sangeetha, S.: "AMMU: A survey of transformer-based biomedical pretrained language models"; Journal of Biomedical Informatics, 126, (2022), 103982. https://doi.org/10.1016/j.jbi.2021.103982

[Khaleghparast et al. 2023] Khaleghparast, S., Maleki, M., Hajianfar, G., Soumari, E., Oveisi, M., Golandouz, H. M., Noohi, F., Golpira, R., Mazloomzadeh, S., and Arabian, M.: " Development of a patients' satisfaction analysis system using machine learning and lexicon-based methods"; BMC Health Services Research, 23, 1 (2023), 1–12.

[Khattak et al. 2019] Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., Rudzicz, F.: "A survey of word embeddings for clinical text"; Journal of Biomedical Informatics, 100, (2019), 100057. https://doi.org/10.1016/j.yjbinx.2019.100057.

[Kuo et al. 2011] Kuo, R.J., Wu, Y.H., Hsu, T.S., and Chen, L.K.: "Improving outpatient services for elderly patients in Taiwan: a qualitative study"; Archives of Gerontology and Geriatrics, 53, 2 (2011), e209–e217.

[lee et al. 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J.: "BioBERT: a pre-trained biomedical language representation model for biomedical text mining"; Bioinformatics, 36, 4 (2020), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682.

[Li et al. 2023] Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., and Luo, Y.: "A comparative study of pretrained language models for long clinical text"; Journal of the American Medical Informatics Association, 30, 2 (2023), 340–347.

[Meesala and Paul 2018] Meesala, A., and Paul, J.: "Service quality, consumer satisfaction and loyalty in hospitals: Thinking for the future"; Journal of Retailing and Consumer Services, 40, (2018), 261–269.

[Mikolov et al. 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: "Distributed representations of words and phrases and their compositionality"; Advances in neural information processing systems, 26, (2013).

[Moradi et al. 2020] Moradi, M., Dorffner, G., Samwald, M.: "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization"; Computer methods and programs in biomedicine, 184, (2020), 105117. //doi.org/10.1016/j.cmpb.2019.105117.

[Ni et al. 2021] Ni, P., Li, G., Hung, P. C., Chang, V.: "StaResGRU-CNN with CMedLMs: A stacked residual GRU-CNN with pre-trained biomedical language models for predictive intelligence"; Applied Soft Computing, 113, (2021), 107975.

[Niu et al. 2005] Niu, Y., Zhu, X., Li, J., Hirst, G.: "Analysis of polarity information in medical text"; In: AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA (October 22-26, 2005), 2005, 570.

[Nourani and Reshadat 2020] Nourani, E., Reshadat, V.: "Association extraction from biomedical literature based on representation and transfer learning"; Journal of theoretical biology, 488, (2020), 110112. https://doi.org/10.1016/j.jtbi.2019.110112.

[Peng et al. 2019] Peng, Y., Yan, S., Lu, Z.: "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets"; Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, Association for Computational Linguistics (August 1, 2019), 58–65.

[Pomares-Quimbaya et al. 2021] Pomares-Quimbaya, A., López-Úbeda, P., Schulz, S.: "Transfer learning for classifying spanish and english text by clinical specialties"; In Public Health and Informatics, (2012), 377-381. https://doi.org/10.3233/SHTI210184.

[Punith et al. 2021] Punith, N. S., Raketla, K.: "Sentiment analysis of drug reviews using transfer learning"; Third International Conference on Inventive Research in Computing Applications (ICIRCA) (September 2021), 1794-1799.

[Qiu et al. 2020] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: "Pre-trained models for natural language processing: A survey"; Science China Technological Sciences, 63, 10 (2020), 1872-1897. https://doi.org/10.1007/s11431-020-1647-3.

[Radford et al. 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.: "Improving language understanding by generative pre-training"; San Francisco: OpenAI, (2018).

[Raffel et al. 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J.: "Exploring the limits of transfer learning with a unified text-to-text transformer"; The Journal of Machine Learning Research, 21, 1 (2020), 5485–5551.

[Raziei et al. 2018] Raziei, Z., Torabi, S. A., Tabrizian, S., and Zahiri, B.: "A hybrid GDM-SERVQUAL-QFD approach for service quality assessment in hospitals"; Engineering Management Journal, 30, 3 (2018), 179–190.

[Saffar et al. 2023] Saffar, A. H., Mann, T. K., and Ofoghi, B.: "Textual emotion detection in health: Advances and applications"; Journal of Biomedical Informatics, 137, (2023), 104258. https://doi.org/10.1016/j.jbi.2022.104258.

[Sanglerdsinlapachai et al. 2021] Sanglerdsinlapachai, N., Plangprasopchok, A., Ho, T. B., Nantajeewarawat, E.: "Improving sentiment analysis on clinical narratives by exploiting UMLS semantic types"; Artificial intelligence in medicine, 113, (2021), 102033.https://doi.org/10.1016/j.artmed.2021.102033.

[Sarker and Paris 2011] A., Molla, D., Paris, C.: "Outcome polarity identification of medical papers"; In Proceedings of the Australasian language technology association workshop, Canberra, Australia (December 2011), 105-114.

[Schäfer et al. 2023] 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"; Journal of Biomedical Informatics, (2023), 104400.

[Serrano-Guerrero et al. 2022] Serrano-Guerrero, J., Bani-Doumi, M., Romero, F. P., and Olivas, J. A.: "A fuzzy aspect-based approach for recommending hospitals"; International Journal of Intelligent Systems, 37, 4 (2022), 2885–2910. https://doi.org/10.1002/int.22634

[Serrano-Guerrero et al. 2022] Serrano-Guerrero, J., Bani-Doumi, M., Romero, F. P., and Olivas, J. A.: "Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions"; Artificial Intelligence in Medicine, 128, (2022), 102298. https: //doi.org/10.1016/j.artmed.2022.102298

[Serrano-Guerrero et al. 2023] Serrano-Guerrero, J., Bani-Doumi, M., Romero, F. P., and Olivas, J. A.: "Selecting the Best Health Care Systems: An Approach Based on Opinion Mining and Simplified Neutrosophic Sets"; International Journal on Artificial Intelligence Tools, 32, 2 (2023), 2340007.

[Shen et al. 2021] Shen, Z., Yi, Y., Bompelli, A., Yu, F., Wang, Y., Zhang, R.: "Extracting Lifestyle Factors for Alzheimer's Disease from Clinical Notes Using Deep Learning with Weak Supervision"; ArXiv, (2021). https://doi.org/10.48550/arXiv.2101.09244.

[Smith et al. 2018] Waudby-Smith, I. E., Tran, N., Dubin, J. A., Lee, J.: "Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients"; PloS one, 13, 6 (2018), e0198687. https://doi.org/10.1371/journal.pone.0198687.

[Sorower 2010] Sorower, M. S.: "A literature survey on algorithms for multi-label learning"; Oregon State University, Corvallis, 18, 1 (2010), 25.

[Sun et al. 2021] Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., and Lu, Y.: "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation"; https://doi.org/10.48550/arXiv.2107.02137, (2021), 02137.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.: "Attention is all you need"; Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA (December 4-9, 2017), 30, 5998-6008.

[Vrbancic and Podgorelec 2020] Vrbančič, G., Podgorelec, V.: "Transfer learning with adaptive fine-tuning"; IEEE Access, 8, (2020), 196197-196211.

[Wang et al. 2022] Wang, H., Li, J., Wu, H., Hovy, E., and Sun, Y.: "Pre-trained language models and their applications"; Engineering, (2022).

[Wang et al. 2023] Wang, B., Xie, Q., Pei, J., Tiwari, P., and Li, Z.: " Pre-trained language models in biomedical domain: A systematic survey"; ACM Computing Surveys (In press), (2023). https://doi.org/10.1145/3611651

[Yadav et al. 2018] Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P.: "Medical sentiment analysis using social media: towards building a patient assisted system"; In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (May 2018), 2790–2797.

[Yikuan et al. 2023] Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., Luo, Y.: "A comparative study of pretrained language models for long clinical text"; Journal of the American Medical Informatics Association, 30, 2 (2023), 340-347.

[Yuan et al. 2022] Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., Yu, S.: "CODER: Knowledge-infused cross-lingual medical term embedding for term normalization"; Journal of biomedical informatics, 126, (2022), 103983.

[Zaheer et al. 2020] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., and Yang, L.: "Big bird: Transformers for longer sequences"; Advances in Neural Information Processing Systems, 33, (2020), 17283-17297.

[Zarei 2015] Zarei, E.: "Service quality of hospital outpatient departments: patients' perspective"; International Journal of Health Care Quality Assurance, 28, 8 (2015), 778–790.

[Zhang et al. 2019] Zhang, M., Geng, G.: "Adverse drug event detection using a weakly supervised convolutional neural network and recurrent neural network model"; Information, 10, 9 (2019), 276. https://doi.org/10.3390/info10090276.

[Zhu et al. 2015] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books"; In Proceedings of the IEEE international conference on computer vision (2015), 19-27.

[Zhuang et al. 2021] Zhuang, L., Wayne, L., Ya, S., Jun, Z.: " A robustly optimized BERT pre-training approach with post-training"; In Proceedings of the 20th Chinese national conference on computational linguistics (August 2021), 1218-1227.