Enriching Deep Learning with Frame Semantics for Empathy Classification in Medical Narrative Essays

Priyanka Dey
Computer Science Department
University of Illinois, Urbana-Champaign
pdey3@illinois.edu

Roxana Girju
Department of Linguistics,
Computer Science Department,
Beckman Institute,
University of Illinois, Urbana-Champaign
girju@illinois.edu

Abstract

Empathy is a vital component of health care and plays a key role in the training of future doctors. Paying attention to medical students' self-reflective stories of their interactions with patients can encourage empathy and the formation of professional identities that embody desirable values such as integrity and respect. We present a computational approach and linguistic analysis of empathic language in a large corpus of 440 essays written by pre-med students as narrated simulated patient – doctor interactions. We analyze the discourse of three kinds of empathy: cognitive, affective, and prosocial as highlighted by expert annotators. We also present various experiments with state-of-the-art recurrent neural networks and transformer models for classifying these forms of empathy. To further improve over these results, we develop a novel system architecture that makes use of frame semantics to enrich our state-of-the-art models. We show that this novel framework leads to significant improvement on the empathy classification task for this dataset.

1 Introduction

Empathy is a complex phenomenon concerning how we seek to understand and experience, to some extent, the experiences of others (Ratcliffe, 2017) – i.e., having a sense of the other's story and the context in which it takes place. One way to get to an appreciation of one's complex situation (i.e., the embodied actions and the contexts within which they act) is through narratives of lived experience (McIntyre, 1981; Gallagher, 2012). Self-reflective (i.e., first person) narratives, for instance, offer a wide range of resources for empathy, as they bring together one's inner and outer worlds, thus giving meaning to experience (Mattingly, 2000). In this respect, narratives seem necessary for empathy, as our first-person experience is grounded in the contextualized content of the narrative. They also provide a form or structure that allows us to frame an understanding of others, together with a learned set of skills and practical knowledge that shapes our understanding of what we and others are experiencing.

Reflective writing is a dynamic process that allows for an active engagement with knowledge and experience, being widely used in clinical practice (Jasper et al., 2013; Burkhardt et al., 2019; Artioli et al., 2021). Putting into words the focused inspection of their thoughts, feelings, and events enables one to reprocess the experience, build new insights, and new ways to conceive reality (Artioli et al., 2021). Thus, narrative exercises like self-reflective stories can help medical students recognise and derive meaning from key experiences, which in turn can support critical thinking, self-consciousness, and the development of personal skills, communication and empathy skills, self-knowledge, professional identify development, and instill behavior change (Craft, 2005; Borgstrom et al., 2016; Mintz-Binder et al., 2019; Allan and Driscoll, 2014; Peterson et al., 2018; Liu et al., 2016; Bekker et al., 2013). Such writing can lead to an increase in experience-taking skills (Kaufman and Libby, 2012) and can decrease stereotyping, prejudice, and racial bias in healthcare (Williams and Wyatt, 2015).

In this research, we take a narrative approach to empathy and explore the experiences of premed students at a large university by analysing their self-reflective writing portfolios (a large corpus of first-person essays written by premed students in narrated simulated patient-doctor interactions). Specifically, we introduce an exploratory study of empathy in clinical encounters paying attention to the discourse of three types of empathy: *cognitive* (the drive and ability to identify and understand another's emotional or mental states), *affective* (the capacity to experience an appropriate emotion in response to another's emotional or mental state), and *prosocial behavior* (a response to having identified

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 207 - 217

the perspective of another with the intention of acting upon the other's mental and/or emotional state), following established practices in psychology (Cuff et al., 2016; Eisenberg et al., 2006; Rameson et al., 2012). We introduce a set of informative baseline experiments using state-of-the-art recurrent neural networks and transformer models for classifying the various forms of empathy. As initial experiments show relatively low scores, we explore a novel FrameNet-based system architecture where we use sentence frames to extract additional semantic features. We apply this framework to state-of-the-art and representative neural network models and show significant improvement in the empathy classification task for this dataset. Although previous research suggests that narrative-based interventions tend to be effective education-based methods, it is less clear what are some of the mechanisms through which narratives achieve such an effect, which is another contribution of this research.

2 Related Work

In spite of its increasing theoretical and practical interest, empathy research in computational linguistics has been relatively sparse and lacks cohesion. Even more so, investigations of empathy as it relates to clinical practice have received even less attention mainly due to data and privacy concerns.

Most of the research on empathy detection has focused on conversations or interactions, as dialogue systems (Zhong et al., 2020; Chen et al., 2022a; Samad et al., 2022), or in online platforms (e.g. (Pérez-Rosas et al., 2017; Khanpour et al., 2017; Otterbacher et al., 2017; Sharma et al., 2020; Lahnala et al., 2022; Sharma et al., 2021; Hosseini and Caragea, 2021), a few on news stories and other narratives (Buechel et al., 2018; Wambsganss et al., 2021b; Sedoc et al., 2020; Mundra et al., 2021; Guda et al., 2021), and even less on empathy in clinical settings (Zhou et al., 2021; Shi et al., 2021). Buechel et al. (2018) used crowd-sourced workers to self-report their empathy and distress levels and to write empathic reactions to news stories. Wambsganss et al. (2021b) built a text corpus of student peer reviews collected from a German business innovation class annotated for cognitive and affective empathy levels. Furthermore, using Batson's Empathic Concern-Personal Distress Scale (Batson et al., 1987), Buechel et al. (2018) have focused only on negative empathy instances (i.e., pain and sadness "by witnessing another person's suffering"). This year, the WASSA shared task focused on predicting empathy, emotion, and personality in reaction to news stories (Barriere et al., 2022; Vasava et al., 2022). The dataset is an extension of Buechel et al. (2018)'s dataset – i.e., it includes news articles that express harm to an entity (e.g. individual, group of people, nature). Each article comes with reaction essays in which authors expressed their empathy and distress toward these news articles. Each essay is annotated for empathy and distress, and with authors' personality traits and demographic information (age, gender, ethnicity, income, and education level). Here, we could not compare our models with the WASSA results – our dataset does not capture the meta-data in WASSA. Moreover, our empathy instances are not always negative (Fan et al., 2011): a dataset reflecting empathetic language should ideally allow for expressions of empathy that encompass a variety of positive and negative emotions. We could not compare against its best performing system due to limited reproducibility (Chen et al., 2022b).

In multimodal research, R. M. Frankel (2000) and Cordella and Musgrave (2009) identify sequential patterns of empathy frequently expressed in video-recorded exchanges by medical graduates interacting with a cancer patient. Sharma et al. (2020) analyzed the discourse of conversations in online peer-to-peer support platforms. They successfully trained novice writers to improve low-empathy responses by giving the writers feedback with examples of sentences that are typical of recognition and interpretation of others' feelings or experiences. In a subsequent set of experiments (Sharma et al., 2021), they suggested that empathic written discourse should be coherent, specific to the conversation at hand, and lexically diverse.

To our knowledge, no self-reflective narrative text corpora have been developed for computational linguistics investigations of clinical student training. Adding to the scarcity of empathy-dedicated resources, there is also a lack of understanding of which linguistic features might contribute to the various types of empathy, like cognitive, affective, and prosocial behavior.

3 Self-reflective Narrative Essays in Medical Training

In this research, we focus on self-reflective narratives written by premed students given a simulated scenario. Simulation is strongly set on our first-

person experiences, relying on resources that are available to the simulator. In a simulation process, the writer puts themselves in the other's situation and asks what "I would do if I were in that situation." Perspective taking (i.e., cognitive empathy) is crucial for fostering affective abilities, enabling writers to imagine and learn about the emotions of others and to share them, too. As empathy is other-directed (De Vignemont and Jacob, 2012; Gallagher, 2012), this means that we, as narrators, are open to the experience and the life of the other, in their context, as we can understand it.

This study's intervention was designed as a written assignment in which premed students were asked to consider a hypothetical scenario where they took the role of a physician breaking the news of an unfavorable diagnosis of high blood cholesterol to a middle-aged patient[1]. They were instructed to recount (in first person voice) the hypothetical doctor-patient interaction where they explained the diagnosis and prescribed medical treatment using layman terms and language they believed would comfort as well as persuade the hypothetical patient to adhere to their prescription.

With the students' consent, we collected a corpus of 774 essays over a period of one academic year (Shi et al., 2021). Following a thorough annotation process, annotators (undergraduate and graduate students in psychology and social work)[2] labeled a subset of 440 randomly selected essays (henceforth, "the corpus"). Using a rich color code schema, each sentence in every essay was labeled as either cognitive empathy (green; e.g., "She looked tired"), affective empathy (yellow; e.g.: "I felt the pain"), or prosocial behavior (cyan; e.g.: "I reassured her this was the best way") (everything else was "no empathy") (Cuff et al., 2016; Eisenberg et al., 2006; Rameson et al., 2012). The six paid undergraduate students were trained on the task and instructed to annotate the data. Two meta-annotators, paid graduate students with prior experience with the task, reviewed the work of the annotators and updated the annotation guidelines at regular intervals, in an iterative loop process after each batch of essays[3]. The meta-annotators reached a Cohen's kappa of 0.82, a good level of agreement. Disagreed cases were discussed and mitigated. At the end, all the essays were re-annotated per the most up-to-date guidelines. The resulting annotated data shows an uneven label distribution in the annotated corpus (11,763 total): 667 (cognitive), 1,659 (affective), and 723 (prosocial) sentences (and 8,714 non-empathy sentences).

[1]The patient was referred to as Betty or John.

[2]The students were hired based on previous experience with similar projects in social work and psychology.

[3]10 essays per week

4 Empathy Classification Task

In this research, our goal is to explore machine learning models of empathy classification in narrative essays to better our understanding of the mechanisms through which empathy can be expressed. Since we are interested in the linguistic expressions of empathy, we zoom in to the sentence level. Given such a corpus of essay sentences, we first build a binary classifier which can be useful in applications requiring a general linguistic understanding of the presence of empathy. In some cases such as medical communication training of pre-med students, a more fine-grained understanding of different kinds of empathy is useful. Thus, we also build a classifier that can identify each type of empathy: cognitive, affective, and prosocial.

For both types of classifiers, we first experiment with several state-of-the-art statistical and machine learning models. As our research is focused on the subcategorization of empathy, we seek to improve our multi-label classifier. Thus, we introduce a new and better performing system architecture by employing FrameNet (Baker et al., 1998), the research and development project which builds on the theory of frame semantics. Using a state-of-the-art FrameNet sentence parser (Swayamdipta et al., 2017), we extract semantic frames from each sentence in our corpus and use this resource to enhance our original (baseline) models with these additional knowledge. As we will show in Subsection 4.4, incorporating FrameNet semantics into state-of-the-art deep learning models leads to an increase in empathy classification results.

4.1 Baseline Models

We started with the following representative baseline models: Naive Bayes (NB), support vector machines (SVM), and logistic regression (logR). We are also interested in observing the performance of deep learning methods and, among them, we experiment with long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and bidirectional long-short term memory (bi-LSTM) (Graves and Schmidhuber, 2005) models; additionally, we use the transformer neural network models BERT

(Devlin et al., 2018) and RoBERTa (Liu et al., 2019). We used unigrams as our features. We also initialized the embedding layers in our neural models (LSTM and bi-LSTM) with GloVe embeddings since the expression of empathy involves larger units than words, and embeddings are known to better capture contextual information. For the transformer models, we use the default BERT embeddings. Since our dataset is imbalanced, we report the precision, recall, and F1-score (harmonic mean of the precision and recall).

We identify sentences with empathy by using the annotator's highlights – e.g., a sentence containing cyan and green highlights is considered a prosocial and cognitive empathy sentence. For our binary empathy classification, we use colored sentences as empathy sentences. We consider sentences with no highlights as no empathy sentences.

For the NB, logistic regression, and SVM models, we generate binary classifiers for each type of empathy. For all the neural network models, we generate multi-label classifiers. For each type of empathy highlighted sentences, we reserve 80/20 training/test ratio, with 5-fold cross validation. For the logistic regression models, we use a L2 regularization and for the SVM models, a linear kernel function. We decided to apply an attention layer for the LSTM and bi-LSTM models to learn patterns that may improve the classification. For our final output layer, we use the sigmoid activation function, as we are dealing with a multi-label classification task. For the BERT and RoBERTa models, we apply a dropout layer with probability 0.4 which helps to regularize the model; we use a linear output layer and apply a sigmoid on the outputs.

For our binary empathy classification task, we find that the imbalanced dataset greatly affects the performance of most models; the best performing model: BERT achieves an F1-score of 0.56 for empathy sentences and 0.79 for no empathy sentences. To combat this imbalance, we randomly downsampled the no empathy sentence dataset (to get an equal number of empathy and no-empathy sentences). This resulted in an improved BERT model (0.72 F1 for empathy and 0.79 F1 for no empathy sentences). For our second empathy classification task, we again downsample the total number of no empathy sentences, resulting in a final dataset of 1,659 affective empathy sentences, 723 prosocial sentences, 667 cognitive sentences, and 1,659 no empathy sentences. Table 1 shows the precision, recall, and F1-measure scores for these baseline experiments. As only 5.81% of our sentences contain multiple types of empathy, we only present collapsed results for each category. We leave the study of these sentences for future research.

The Naive Bayes, SVM, and logistic regression models all overfit the training data and, in general, do not handle the imbalanced dataset well. The neural network models provide more promising results, with affective empathy even reaching 0.81 F1 scores. Prosocial empathy seems to be the most difficult to identify, with the highest F1 of 0.73 as obtained by the BERT model. Overall, the transformer models, BERT and RoBERTa, achieve the best performance across all three types of empathy.

4.2 Incorporating FrameNet to Improve Empathy Classification

In our attempt to improve the classification of our empathic narrative sentences, we decided to explore feature generation to further enhance these models. Since empathy is a highly complex semantic-pragmatic phenomenon, one intuition is that semantic knowledge should help the classifiers. One linguistic theory called frame semantics deconstructs a sentence into predicate-argument structures that describe meaning not at the level of individual words, but is instead based on the concept of a scenario, scene, or event called a frame. Frames are defined by the group of words that evoke the scene (frame-evoking elements or FEEs), as well as by their expected semantic arguments (frame elements). A JUDGMENT frame, for instance, has FEEs like *praise.v, criticize.v*, and *disapprove.v*, and frame elements such as Cognizer, Evalueee, Expressor, Reason. The Berkeley FrameNet project (Baker et al., 1998; Ruppenhofer et al., 2016) is the most well-known lexical resource of frame semantics, with definitions for over 1200 frames.[4]

To generate new features, we leverage frame semantics to identify all the frames that occur in a sentence. Each sentence in our essay corpus is parsed with the Frame-Semantic Parser (Swayamdipta et al., 2017), which is based on a softmax margin segmental recurrent neural network model. Specifically, we use the FrameNet 1.7 pretrained models to predict frames for each of our sentences. For instance, for "He played an important role in preventing her from becoming depressed", the frame

[4]We used the release 1.7 which has 1,222 frame annotations (http://framenet.icsi.berkeley.edu).

Classifier	Cognitive			Affective			Prosocial			None		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
NB	0.03	0.18	0.05	0.05	0.38	0.09	0.14	0.05	0.07	1.0	0.72	0.84
SVM	0.30	0.19	0.23	0.46	0.50	0.48	0.44	0.37	0.40	0.80	0.71	0.75
LogR	0.44	0.38	0.40	0.74	0.58	0.65	0.20	0.25	0.22	0.77	0.71	0.74
LSTM	0.62	0.72	0.67	0.63	0.61	0.62	0.51	0.59	0.55	0.71	0.76	0.73
biLSTM	0.64	0.71	0.67	0.79	0.62	0.69	0.59	0.62	0.60	0.78	0.74	0.76
BERT	0.74	0.78	0.76	0.92	0.73	0.81	0.72	0.75	0.73	0.75	0.84	0.79
RoBERTa	0.74	0.83	0.78	0.77	0.78	0.77	0.69	0.68	0.68	0.77	0.80	0.78
FN-LSTM	0.73	0.73	0.73	0. 83	0.68	0.75	0.66	0.78	0.72	**0.79**	0.77	**0.78**
FN-biLSTM	0.71	0.88	0.79	0.85	0.78	0.81	0.72	0.86	0.78	0.73	0.75	0.74
FN-BERT	**0.78**	**0.89**	**0.83**	**0.88**	**0.79**	**0.83**	**0.82**	**0.88**	**0.85**	0.71	**0.80**	0.75
FN-RoBERTa	0.73	0.88	0.80	0.85	**0.79**	0.82	**0.82**	0.86	0.84	0.71	**0.80**	0.75

Table 1: Precision, recall and F1 scores of all baseline and FrameNet-incorporated classifiers on the test dataset: 133 cognitive, 332 affective, 145 prosocial, and 332 no-empathy sentences. Bolded numbers indicate best performance.

semantic parser identifies four frames: PERFORMERS_AND_ROLES (i.e., he played a role), IMPORTANCE (i.e., important role), THWARTING (i.e., preventing her), EMOTIONS_BY_POSSIBILITY (i.e., becoming depressed).

Given the parser's extraction of 669 unique frames from our entire sentence dataset, we explore the most common frames present in sentences containing each type of empathy (Table 2). Many of the frames exhibited in cognitive empathy sentences focus on *speaking*, *supporting*, and *seeing*, while affective empathy sentences contain frames related to *responses*, *stimulating emotions*, and *perceiving emotions/states*. Many of the prosocial empathy sentences include frames that discuss a form of action e.g. *trying to [perform an action], reassure, seek to achieve*, etc.

To use the frame identification as a feature in our models, we generate a frequency vector to encode the occurrences of a frame in a sentence. For example, if we had a total of 3 frames *Fa*, *Fb*, *Fc*, and sentence x contained one mention of frame *Fa*, 2 mentions of *Fb*, and no *Fc*, our encoding vector would be: [1, 2, 0], representing their frequencies. Thus, we generate a vector of size 669 for each of our sentences in the whole essay dataset.

In our quest for improved empathy classification, we focus on our neural network (LSTM, bi-LSTM, BERT, and RoBERTa) models as these proved to perform best in our baseline experiments. For our LSTM and bi-LSTM models, we use GloVe embeddings to encode the processed sentence, and then add the FrameNet encoding vector to the end of the embedding vector. We then apply the LSTM or bi-LSTM layer followed by the attention layer and transform outputs using the sigmoid activation to get class probabilities (Figure 1 shows the system architecture for this framework).

For the BERT and RoBERTa models, we first input the processed sentence and extract the textual embeddings, and append the FrameNet encoding vector to the embedding vector. We then apply a feedforward neural network – i.e., a multi layer perceptron (MLP) with a sigmoid activation function – to get predictions (Figure 2 shows the system architecture for this framework).

4.3 Constructing a Frame Lattice

An initial exploration of the FN parser shows that our training dataset contained a total of 616 unique frames, roughly 50% of them appearing only in at most 5 sentences. To optimize learning in the neural network models, we identify a lattice of frames from our training corpus that most improves the classification performance. To do this, we iterate through each combination of subsets of size K of the identified frames in our training dataset. We then compute weighted average accuracy scores for empathy classification using the training dataset and identify the set of frames most influential in each of the four models considered. An initial set of exploratory experiments has shown that lattices of sizes between 5 and 20 yield the highest improvement. Frame lattices of size 2, 3, and 4 did not show any significant improvement (i.e., no increase in score above 0.01). Latices larger than 20 become very noisy and, thus, negatively impact performance. Thus, we decided to further explore this

Cognitive	Affective	Prosocial
JUDGMENT (159)	RESPONSE (831)	GESTURE (458)
MAKE_COGNITIVE_CONNECTION (143)	COMMUNICATION_RESPONSE (368)	DESIRABILITY (290)
SPEAK_ON_TOPIC (134)	PERCEPTION (293)	REASSURING (274)
SUPPORTING (108)	STIMULATE_EMOTION (209)	FACIAL_EXPRESSION (231)
SEE_THROUGH (96)	SOCIABILITY (148)	AWARENESS (173)

Table 2: Most common frame classes for each empathy class

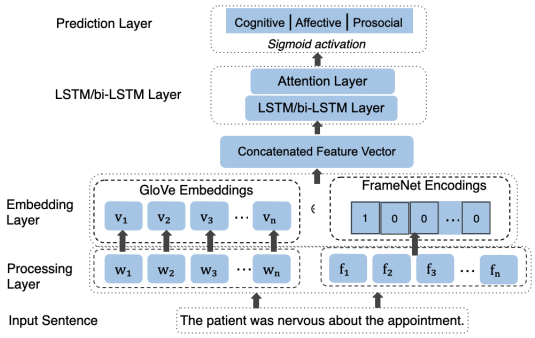

Figure 1: Architecture for LSTM & bi-LSTM models

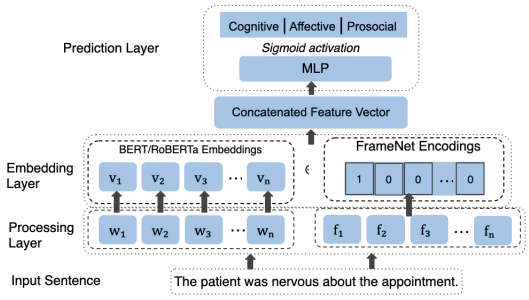

Figure 2: Architecture for BERT & RoBERTa models

5-20 range. Specifically, we iterate through all possible combinations of 5 frames that appeared in the training corpus. We then increment the frame size by 1 in each iteration, and recompute performance. Results on test data are shown in Fig. 3.

Since we wanted to use a metric that would measure the performance for all three empathy types together, we did not use the individual F1 scores for our categories. The closest measurement was the macro-F1 score, but this is still an unweighted average (since we have already had good performance for affective empathy, using this metric, the results would not increase by much). Thus, the weighted average made more sense to identify the best lattice.

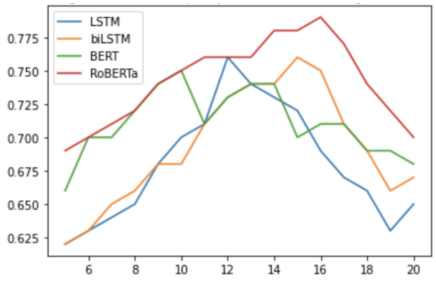

Figure 3: Weighted average scores for varying lattice sizes: 5 to 20

4.4 FrameNet Experiments' Results

To improve classification, we thus incorporate each neural network's best performing lattice and build a frame encoding vector for each sentence in our dataset. We then follow the system architectures in Figures 1 and 2 and compute the performance for each model (See Table 1).

The experimental results show that the inclusion of the FrameNet lattices improves performance considerably. The best models are FrameNet-BERT and FrameNet-RoBERTa, for which all the metric scores significantly improve with this additional feature. We also notice that the classification performance for prosocial empathy significantly improved over the baseline models (0.85 vs. previous score: 0.73). The enhanced BERT model yields the highest F1 scores for all three empathy types, with all empathy categories scoring above 0.8; the no-empathy category however does drop in performance (0.75 vs. previous score: 0.79).

These experiments indicate that our system learns best from a lattice of different sizes for each learning model. Table 3 shows the specific frames per model. Many of the learning architectures choose the same frames in their lattices, e.g. INTENTIONALLY_ACT, GESTURE, SOCIABILITY, PERCEPTION, SENSATION. Interestingly, the transformer models select some additional frames directly linked to certain types of empathy: cognitive (MAKE_COGNITIVE_CONNECTION, MENTION), affective (PERCEPTION, RESPONSE), and prosocial (SEEKING_TO_ACHIEVE). These frames are possibly somewhat tied to our specific dataset and narrative genre, issue we leave for future research.

FrameNet-BERT vs. BERT

We also examined a bit closer the results to get

LSTM (latice size = 13)	**bi-LSTM** (latice size = 14)
CAUSE_EMOTIONS, INTENTIONALLY_ACT, GESTURE, JUDGMENT, DESIRABILITY, PERCEPTION, COMMUNICATION_RESPONSE, SEEKING_TO_ACHIEVE, SENSATION, SOCIABILITY, TELLING, WORRY)	EMOTIONS, EMOTIONS_BY_POSSIBILITY, EVOKING, GESTURE, JUDGMENT, MENTION, OPINION, INTENTIONALLY_ACT, PERCEPTION, RESPONSE, RESPOND_TO_PROPOSAL, COMMUNICATION_RESPONSE_SCENARIO, SENSATION, SOCIABILITY, STIMULATE_EMOTION, SEEKING_TO_ACHIEVE
BERT (latice size = 14)	**RoBERTa** (latice size = 12)
CAUSE_EMOTIONS , EMOTIONS_BY_POSSIBILITY, EVOKING, GESTURE, INTENTIONALLY_ACT, MAKE_COGNITIVE_CONNECTION, MENTION, OPINION, PERCEPTION, RESPONSE, SENSATION, SOCIABILITY, SUPPORTING, WORRY	CAUSE_EMOTIONS, EMOTIONS_BY_POSSIBILITY, GESTURE, FACIAL_EXPRESSION, INTENTIONALLY_ACT, JUDGMENT, MAKE_COGNITIVE_CONNECTION, MENTION, PERCEPTION, RESPONSE, SEEKING_TO_ACHIEVE, SPEAK_ON_TOPIC

Table 3: Best frame latices for each learning model

more insights into the contribution of the FrameNet external semantic knowledge to the task of empathy classification. Specifically, we wanted to see what kinds of examples FrameNet-BERT classifies correctly over the baseline transformer BERT.

Overall, there was a total of 197 instances (affective: 76; cognitive: 59; prosocial: 62) that FrameNet-BERT classified correctly and BERT incorrectly. A look at these sentences shows a balanced combination of frames like MEDICAL_CONDITIONS, DIFFICULTY, QUESTIONING, BIOLOGICAL_CLASSIFICATION, EXPLAINING_THE_FACTS, CURE, as well as AWARENESS, EMOTION_DIRECTED, COMING_TO_BELIEVE, EXPERIENCER_FOCUS, EXPERIENCER_OBJ, FEAR. These empirical results support new evidence in medical education (Warmington, 2019; Warmington et al., 2022) – meaning, they highlight how important it is for future doctors to focus and reflect not only on how to diagnose and provide proper treatment to the patient, but also to develop an awareness of how patients experience their illness and focus on how patients need their experience of illness acknowledged.

In addition to these frames, a specific subset deserves particular attention and discussion, subset which works best in combination with those mentioned above. Table 4 lists the most frequent frames of non-verbal communication that tend to occur in true positive test instances as identified by FrameNet-BERT. These results indicate that, even in self-reflective narratives, both verbal and non-verbal aspects of interaction play an important role. What we wear and the way we physically interact with others communicate a great deal about who we are (Iedema and Caldas-Coulthard, 2008). Such narratives include information about non-verbal communication and impressions of other aspects of the context. For instance, the importance of the senses of sight and sound in building up a rich description of both the setting and events is well recognised (i.e., laughter, cry, the tone or volume of voices). These empirical results indicate that cognitive and sensory self-awareness are critical to the clinical encounter process. Doctors paying close attention to their patients' as well as to their own sensations, perceptions and emotional responses picture a process that emphasizes the importance of self-awareness and awareness of others, both indispensable in effective empathic communication.

5 Discussion and Conclusions

Medical education should and can incorporate guided self-reflective practices that show how important it is for the students to develop an awareness of the emotional and relational aspects of the clinical encounter with their patients (Warmington, 2019). The way people identify themselves and perform in particular roles and in relations to others brings together a specific set of values, attitudes, and competencies that can be supported through ongoing self-reflection. Thus, students learn not only how to diagnose and treat patients' medical conditions, but also how to witness the patient's illness experience. In practice, they often switch between these positions: witnessing what it is like for the patient, as well as understanding what they need medically.

Often, clinical encounters can be highly charged emotionally especially for patients in case of serious illness. Unfortunately, medicine lags behind other health professions (like nursing, social work, psychology) which learn from reflective practice and respect it from the beginning. Yet, acknowledging the patient's situation, who they are and their experience can make a huge impact on the quality

Frame	Examples	Count
BODY_PARTS	I noticed Betty fidgeting and clasping her hands, and so I tried to reassure her we would work together and develop a recovery plan.	59
SENSATION	After he left the meeting room, I began feeling very helpless.	71
BODY_MOVEMENT	He seemed to almost roll his eyes at that moment which I don't blame him for.	41
CHANGE_POSTURE	He quietly sat down with his hands folded without responding to my remark.	19
FACIAL_EXPRESSION	I noticed after I told her the news, her mouth forming into a frown and she seemed very depressed.	38
GESTURE	I proceeded with the diagnosis to explain the severity of elevated levels but stopped as she waved her hand.	83
BREATHING	Betty and her family both sighed a breath of relief.	40
SOUND_LEVEL	After I told him the bad news, my patient became silent.	16

Table 4: Examples of empathy sentences with non-verbal communication frames

of that relationship and the trust that is built up for the patient. Narrative-based interventions and activities can facilitate self-reflection and enrich medical students' professional identity formation.

Computational approaches to empathy can be very valuable, but it is clear that such AI initiatives must be multidisciplinary, using and developing a variety of core sets of requirements and expertise and engaging many participants, e.g. AI designers, developers, frontline clinical teams, ethicists, humanists, patients, caregivers (Matheny et al., 2019).

The research experiments and findings summarized in this paper are part of a larger interdisciplinary and highly collaborative project where we analyze both self-reflective narratives of simulated interactions, as well as multimodal patient-doctor encounters in real clinical settings (Girju, 2021; Girju and Girju, 2022). In this paper, we presented a computational approach and linguistic analysis of empathic language in a large corpus of premed student essays of narrated simulated patient-doctor interactions. Specifically, we showed that semantic information at the sentence level can be very useful not only in empathy identification but provides details on the differences among the three main types of empathy: cognitive, affective, and prosocial. We presented novel and performant FrameNet-based transformer models for empathy classification. In future work, we will expand this analysis by considering discourse-level context. We will also integrate other resources like WordNet (Miller, 1995), VerbNet (Kipper et al., 2000), and take advantage of larger discourse.

6 Ethical Considerations

Despite the clear benefits that such empathy detection systems can bring, there are also ethical issues that arise from their use. First, machine learning models are susceptible to design biases that may result in systematic errors, in addition to lower transparency, loss of control, and potential lack of trust by human users (Wambsganss et al., 2021a). Moreover, such models are data-driven – and most of the time such data is potentially biased, highly sensitive, where user privacy becomes an even more important concern. For instance, although we followed the ethical protocols put forward in academia for data collection and annotation, our data is imbalanced demographically (for both pre-med students and the hypothetical patient) and limited to only one clinical scenario (i.e., breaking bad news). Furthermore, special attention should be given to models designed to empathize with vulnerable population like children and people of various abilities. Moreover, focusing only on one hypothetical medical scenario, resulted in a dataset with limited diversity. Another aspect to consider in future research is the use of self-assessment vs. third-party empathy reports. Although most of our pre-med students were highly confident in their empathetic abilities, more thorough research is needed in this direction. AI research on empathy should compare against and even integrate qualitative metrics like the Jefferson Scale of Physician Empathy (Hojat et al., 2001) or the Consultation and Relational Empathy (CARE) Measure (Mercer et al., 2004).

Obviously, we are currently far from being able to deploy such models to help in medical student training. However, our annotated corpus and experiments help shed new light on the empathy classification task and show what kind of linguistic (semantic) knowledge can contribute to it. We also hope such work will encourage future research and collaboration between AI practitioners and clinicians. Overall, developers and providers alike need to increasingly follow ethical considerations in the human-value sensitive design of these systems to ensure the well-being of their users.

References

Elizabeth G Allan and Dana Lynn Driscoll. 2014. The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. *Assessing Writing*, 21:37–55.

Giovanna Artioli, Laura Deiana, Francesco De Vincenzo, Margherita Raucci, Giovanna Amaducci, Maria Chiara Bassi, Silvia Di Leo, Mark Hayter, and Luca Ghirotto. 2021. Health professionals and students' experiences of reflective writing in learning: A qualitative meta-synthesis. *BMC medical education*, 21(1):1–14.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Hilary L Bekker, Anna E Winterbottom, Phyllis Butow, Amanda J Dillard, Deb Feldman-Stewart, Floyd J Fowler, Maria L Jibaja-Weiss, Victoria A Shaffer, and Robert J Volk. 2013. Using personal stories. *BMC Medical Informatics and Decision Making*, 13(59).

Erica Borgstrom, Rachel Morris, Diana Wood, Simon Cohn, and Stephen Barclay. 2016. Learning to care: medical students' reported value and evaluation of palliative care teaching involving meeting patients and reflective writing. *BMC medical education*, 16(1):1–9.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

Crystal Burkhardt, Ashley Crowl, Margaret Ramirez, Brianna Long, and Sarah Shrader. 2019. A reflective assignment assessing pharmacy students' interprofessional collaborative practice exposure during introductory pharmacy practice experiences. *American Journal of Pharmaceutical Education*, 83(6).

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022a. EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074, Seattle, United States. Association for Computational Linguistics.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022b. Iucl at wassa 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 228–232.

M. Cordella and S. Musgrave. 2009. Oral communication skills of international medical graduates: Assessing empathy in discourse. *Communication and Medicine*, 6(2):129–142.

Melissa Craft. 2005. Reflective writing and nursing education. *Journal of nursing education*, 44(2):53–57.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review*, 8(2):144–153.

Frédérique De Vignemont and Pierre Jacob. 2012. What is it like to feel another's pain? *Philosophy of science*, 79(2):295–316.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nancy Eisenberg, Richard A Fabes, and Tracy L Spinrad. 2006. Prosocial development. In *Volume III. Social, Emotional, and Personality Development*. John Wiley & Sons, Inc.

Y. Fan, Duncan NW, de Greck M, and Northoff G. 2011. Is there a core neural network in empathy? an fmri based quantitative meta-analysis. *Neuroscience Biobehavioral Review*, 35(3):903–911.

Shaun Gallagher. 2012. Empathy, simulation, and narrative. *Science in Context*, 25(3):355–381.

Roxana Girju. 2021. Adaptive multimodal and multisensory empathic technologies for enhanced human communication. In *Rethinking the Senses: A Workshop on Multisensory Embodied Experiences and Disability Interactions, the ACM CHI Conference on Human Factors in Computing Systems*. arXiv preprint arXiv:2110.15054.

Roxana Girju and Marina Girju. 2022. Design considerations for an NLP-driven empathy and emotion interface for clinician training via telemedicine. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27, Seattle, Washington. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. In *EACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohammadreza Hojat, Salvatore Mangione, Thomas J Nasca, Mitchell JM Cohen, Joseph S Gonnella, James B Erdmann, Jon Veloski, and Mike Magee. 2001. The jefferson scale of physician empathy: development and preliminary psychometric data. *Educational and psychological measurement*, 61(2):349–365.

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rick Iedema and Carmen Rosa Caldas-Coulthard. 2008. Introduction: Identity trouble: Critical discourse and contested identities. In *Identity trouble*, pages 1–14. Springer.

Melanie Jasper, Megan Rosser, and Gail Mooney. 2013. *Professional development, reflection and decision-making in nursing and healthcare*. John Wiley & Sons.

Geoff F Kaufman and Lisa K Libby. 2012. Changing beliefs and behavior through experience-taking. *Journal of personality and social psychology*, 103(1):1—19.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. 2000. Class-based construction of a verb lexicon. *AAAI/IAAI*, 691:696.

Allison Lahnala, Charles Welch, Béla Neuendorf, and Lucie Flek. 2022. Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4926–4938, Seattle, United States. Association for Computational Linguistics.

Geoffrey Z Liu, Oliver K Jawitz, Daniel Zheng, Richard J Gusberg, and Anthony W Kim. 2016. Reflective writing for medical students on the surgical clerkship: oxymoron or antidote? *Journal of surgical education*, 73(2):296–304.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

M. Matheny, S. Thadaney Israni, M. Ahmed, and D. Whicher (editors). 2019. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. National Academy of Medicine, Washington, DC.

Linda C Garro; Cheryl Mattingly. 2000. *Narrative and the cultural construction of illness and healing*. Univ. of California Press, Berkeley, California.

Alestair McIntyre. 1981. *After Virtue*. South Bend: University of Notre Dame Press, Notre Dame, IN.

Stewart W Mercer, Margaret Maxwell, David Heaney, and Graham Watt. 2004. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family practice*, 21(6):699–705.

George A Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

R Mintz-Binder, MM Jones, et al. 2019. When a clinical crisis strikes: Lessons learned from the reflective writings of nursing students. In *Nursing Forum*, volume 54, pages 345–351.

Jay Mundra, Rohan Gupta, and Sagnik Mukherjee. 2021. WASSA@IITK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 112–116, Online. Association for Computational Linguistics.

Jahna Otterbacher, Chee Siang Ang, Marina Litvak, and David Atkins. 2017. Show me you care: Trait empathy, linguistic style, and mimicry on facebook. *ACM Transactions on Internet Technology (TOIT)*, 17(1):1–22.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

William J Peterson, Joseph B House, Cemal B Sozener, and Sally A Santen. 2018. Understanding the struggles to be a medical provider: view through medical student essays. *The Journal of Emergency Medicine*, 54(1):102–108.

R. M. Frankel. 2000. *The (socio) linguistic turn in physician-patient communication research*. Georgetown University Press, Boston, MA.

Lian T Rameson, Sylvia A Morelli, and Matthew D Lieberman. 2012. The neural correlates of empathy: experience, automaticity, and prosocial behavior. *Journal of cognitive neuroscience*, 24(1):235–245.

Matthew Ratcliffe. 2017. Empathy without simulation. In *Imagination and Social Perspectives*, page 22. Routledge.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673, Marseille, France. European Language Resources Association.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Shuju Shi, Yinglun Sun, Jose Zavala, Jeffrey Moore, and Roxana Girju. 2021. Modeling clinical empathy in narrative essays. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 215–220.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *CoRR*, abs/1706.09528.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264, Dublin, Ireland. Association for Computational Linguistics.

Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. 2021a. Ethical design of conversational agents: towards principles for a value-sensitive design. In *International Conference on Wirtschaftsinformatik*, pages 539–557. Springer.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021b. Supporting cognitive and emotional empathic writing of students. *arXiv preprint arXiv:2105.14815*.

Sally G Warmington. 2019. *Storytelling encounters as medical education: crafting relational identity*. Routledge.

Sally G. Warmington, May-Lill Johansen, and Hamish Wilson. 2022. Identity construction in medical student stories about experiences of disgust in early nursing home placements: a dialogical narrative analysis. *BMJ open*, 12(2):e051900.

David R Williams and Ronald Wyatt. 2015. Racial bias in health care and health: challenges and opportunities. *Jama*, 314(6):555–556.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Yanmengqian Zhou, Michelle L Acevedo Callejas, Yuwei Li, and Erina L MacGeorge. 2021. What does patient-centered communication look like?: Linguistic markers of provider compassionate care and shared decision-making and their impacts on patient outcomes. *Health Communication*, pages 1–11.