Investigating Stylistic Profiles for the Task of 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

One important aspect of language is how speakers generate utterances and texts to convey their intended meanings. In this paper, we bring various aspects of the Construction Grammar (CxG) and the Systemic Functional Grammar (SFG) theories in a deep learning computational framework to model empathic language. Our corpus consists of 440 essays written by premed students as narrated simulated patient–doctor interactions. We start with baseline classifiers (state-of-the-art recurrent neural networks and transformer models). Then, we enrich these models with a set of linguistic constructions proving the importance of this novel approach to the task of empathy classification for this dataset. Our results indicate the potential of such constructions to contribute to the overall empathy profile of first-person narrative essays.

1 Introduction

Much of our everyday experience is shaped and defined by actions and events, thoughts and perceptions which can be accounted for in different ways in the system of language. The grammatical choices we make when writing an essay (i.e., pronoun use, active or passive verb phrases, sentence construction) differ from those we use to email someone, or those we utter in a keynote speech. "Word choice and sentence structure are an expression of the way we attend to the words of others, the way we position ourselves in relation to others" (Micciche, 2004). Such choices allow us to compare not only the various options available in the grammar, but also what is expressed in discourse with what is suppressed (Menéndez, 2017).

Given the great variability in the modes of expression of languages, the search for an adequate design of grammar has long motivated research in linguistic theory. One such approach is CxG (Kay and et al., 1999; Goldberg, 1995; Fillmore et al., 2006) which prioritizes the role of constructions, conventional form-meaning pairs, in the continuum between lexis and syntax (Van Valin, 2007). As such, these constructions form a structured inventory of speakers' knowledge of the conventions of their language (Langacker, 1987).

Another particular grammatical facility for capturing experience in language is Halliday's system of transitivity as part of the Systemic Functional Grammar (SFG) (Halliday, 1994; Halliday et al., 2014), a theory of language centred around the notion of language function. SFG pays great attention to how speakers generate utterances and texts to convey their intended meanings. This can make our writing effective, but also give the audience a sense of our own personality. However, unlike CxG, Halliday's system of transitivity describes the way in which the world of our experience is divided by grammar into a 'manageable set of process types' (Halliday et al., 2014) each offering not only a form-meaning mapping, but also a range of stylistic options for the construal of any given experience through language. In stylistics, researchers have used this model to uncover and study the grammatical patterns through which texts can enact a particular ideology, or an individual's distinctive 'mind style' of language (Fowler, 1996).

The idea of 'style as choice' in Halliday's transitivity system can be best understood as experiential strategies (like avoiding material processes or repeating passive voice constructions) such as those identified as contributing to a reduced sense of awareness, intentionality or control in the human agent responsible (Fowler, 2013; Simpson and Canning, 2014). Such an individual is often said to appear 'helpless' and 'detached' (Halliday, 2019; Simpson, 2003), or 'disembodied' (Hoover, 2004). Take for instance, construction choices like 'I reassured her' vs. 'She was reassured', or "I greeted her upon entrance" vs. "The nurse greeted her upon entrance" vs. "She was greeted upon entrance" – which show the degree of agency and

intended involvement on the part of the agent in the action. Such linguistic choices often occur together in stylistic profiling exercises to showcase the techniques contributing to 'passivity', or the degree of suppression of agency and power in characterisation (Kies, 1992).

In this paper, we try to bring CxG and SFG closer together in the study of discourse level construction of arguments for the analysis of empathic content of narrative essays. Specifically, inspired by research in critical discourse analysis, we are taking a step further to show ways in which such construction choices can manipulate (and even reduce) the attention we give to the agency and moral responsibility of individuals (Jeffries, 2017; Van Dijk, 2017). Specifically, such form-meaning-style mappings can be used to capture the point of view as an aspect of narrative organization and the perspective through which a story is told, the way the characters are portrayed in terms of their understanding of the processes they are involved in, as well as their own participation in the story. In this respect, "narratives seem necessary for empathy [..] they give us access to contexts that are broader than our own contexts and that allow us to understand a broad variety of situations" (Gallagher, 2012). They provide a form/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.

Drawing on Halliday's transitivity framework rooted in Systemic Functional Linguistics, this paper attempts to reveal the (dis)engaged style of empathic student essays from a semantic-grammatical point of view. Specifically, we want to investigate how certain types of processes (i.e., verbs) and constructions (i.e., passive voice) function to cast the essay writers (as main protagonists and agents) as perhaps rather ineffectual, passive, and detached observers of the events around them and of the patient's emotional states.

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 consisting of a corpus of first-person essays written by them as narrated simulated patient-doctor interactions. The corpus has been previously annotated and organized (Shi et al., 2021; Michalski and Girju, 2022) following established practices and theoretical conceptualizations in psychology (Cuff et al., 2016; Eisenberg et al., 2006; Rameson et al., 2012). Computationally, 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 measure the presence of several grammatical structures, leveraging Halliday's theory of transitivity, and its correlation with the essays' overall empathy scores. 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 linguistic mechanisms through which narratives achieve such an effect, especially applied to empathy, 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 limited to empathy recognition, empathetic response generation, or empathic language analysis in counselling sessions. Investigations of empathy as it relates to clinical practice have received even less attention given the inherent data and privacy concerns.

Most of the research on empathy detection has focused on spoken conversations or interactions, some in online platforms (e.g. (Pérez-Rosas et al., 2017; Khanpour et al., 2017; Otterbacher et al., 2017; Sharma et al., 2021; Hosseini and Caragea, 2021), very little on narrative genre (Buechel et al., 2018; Wambsganss et al., 2021), and even less in clinical settings. Buechel et al. (2018) used crowdsourced workers to self-report their empathy and distress levels and to write empathic reactions to news stories. Wambsganss et al. (2021) built a text corpus of student peer reviews collected from a German business innovation class annotated for cognitive and affective empathy levels. 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"). However, empathy is not always negative (Fan et al., 2011). A dataset reflecting empahatic language should ideally allow for expressions of

empathy that encompass a variety of emotions, and even distinguish between sympathy and empathy.[1]

Following a multimodal approach to empathy prediction, R. M. Frankel (2000) and Cordella and Musgrave (2009) identify sequential patterns of empathy in video-recorded exchanges between medical graduates and cancer patients. Sharma et al. (2020) analyzed the discourse of conversations in online peer-to-peer support platforms. Novice writers were trained to improve low-empathy responses and provided writers with adequate feedback on how to recognize and interpret others' feelings or experiences. In follow-up research, they performed a set of experiments (Sharma et al., 2021) whose results seemed to indicate that empathic written discourse should be coherent, specific to the conversation at hand, and lexically diverse.

To our knowledge, no previous research has investigated the contribution of grammatical constructions like Halliday's transitivity system to the task of empathy detection in any genre, let alone in clinical education.[2]

3 Self-reflective Narrative Essays in Medical Training

Simulation-based education (SBE) is an important and accepted practice of teaching, educating, training, and coaching health-care professionals in simulated environments (Bearman et al., 2019). Four decades-worth of SBE research has shown that "simulation technology, used under the right conditions ... can have large and sustained effects on knowledge and skill acquisition and maintenance among medical learners" (McGaghie et al., 2014). In fact, simulation-based education, an umbrella term that covers a very broad spectrum of learning activities from communication skill role-playing to teamwork simulations, is known to contribute to shaping experiences in undergraduate and postgraduate medical, nursing and other health education. In all these activities, learners contextually enact a task which evokes a real-world situation allowing them to undertake it as if it were real, even though they know it is not (Dieckmann et al., 2007; Bearman, 2003).

Personal narratives and storytelling can be viewed as central to social existence (Bruner, 1991), as stories of lived experience (Van Manen, 2016), or as a way in which one constructs notions of self (Ezzy, 1998). In this research, we focus on self-reflective narratives written by premed students given a simulated scenario. Simulation is strongly based on our first-person experiences since it relies on resources that are available to the simulator. In a simulation process, the writer puts themselves in the other's situation and asks "what would I do if I were in that situation?" Perspective taking 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. Some evidence shows that we can take such reliance on narrative resources to open up the process toward a more enriched and non-simulationist narrative practice (i.e., real doctor-patients interactions in clinical context) (Gallagher, 2012).

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[3]. They were instructed to recount (using first person voice) the hypothetical doctor-patient interaction where they explained the diagnosis and prescribed medical treatment to the patient using layman terms and language they believed would comfort as well as persuade the hypothetical patient to adhere to their prescription. Prior to writing, students completed a standard empathic training reading assignment (Baile et al., 2000). They received the following prompt instructions and scenario information.[4]

Prompt Instructions: Imagine yourself as a physician breaking bad news to a patient. Describe the dialogue between the patient and you, as their primary care physician. In your own words, write an essay reporting your recollection of the interaction as it happened (write in past tense). Think of how you would break this news if you were in this scenario in real life. In your essay, you should be reflecting on (1) how the patient felt during this scenario and (2) how you responded to your patient's

[1] Some studies don't seem to differentiate between sympathy and empathy (Rashkin et al., 2018; Lin et al., 2019).

[2] Besides our own research (Shi et al., 2021; Michalski and Girju, 2022; Dey and Girju, 2022; Girju and Girju, 2022).

[3] The patient was referred to as Betty, initially. Later in the data collection, students could also identify the patient as John.

[4] All data collected for this study adheres to the approved Institutional Review Board protocol.

questions in the scenario below.

Scenario: Betty is 32 years old, has a spouse, and two young children (age 3 and 5). You became Betty's general practitioner last year. Betty has no family history of heart disease. In the past 6 months, she has begun experiencing left-side chest pain. Betty's bloodwork has revealed that her cholesterol is dangerously high. Betty will require statin therapy and may benefit from a healthier diet and exercise.

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)[5] labeled a subset of 440 randomly selected essays at sentences level following established practices in psychology (Cuff et al., 2016; Eisenberg et al., 2006; Rameson et al., 2012). The labels are: *cognitive empathy* (the drive and ability to identify and understand another's emotional or mental states; e.g., "She looked tired"); *affective empathy* (the capacity to experience an appropriate emotion in response to another's emotional or mental state; e.g.: "I felt the pain"); and *prosocial behavior* (a response to having identified the perspective of another with the intention of acting upon the other's mental and/or emotional state; e.g.: "I reassured her this was the best way"). Everything else was "no empathy". 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[6]. 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.

In this paper, we collapsed all the affective, cognitive, and prosocial empathy labels into one *Empathy Language* label – since we are interested here only in emphatic vs. non-empathic sentences. After integrating the annotations and storing the data for efficient search (Michalski and Girju, 2022), our corpus consisted of 10,120 data points (i.e., sentences) highlighted or not with empathy. Each essay was also rated by our annotators with a score on a scale from 1-5 (one being the lowest) to reflect overall empathy content at essay level.

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

[6]10 essays per week

4 Constructions and Stylistic Profiles in Empathic Narrative Essays

In CxG, constructions can vary in size and complexity – i.e., morphemes, words, idioms, phrases, sentences. In this paper, we focus mainly on simple sentence-level constructions[7], which, since we work with English, are typically of the form S V [O], where S is the subject, V is the verb, and O is the object (e.g., a thing, a location, an attribute). For instance, "Betty took my hand" matches the construction S V O with the semantics <Agent Predicate Goal>. SFG and CxG give the same semantic analysis, modulo some terminological differences (Lin and Peng, 2006). Specifically, they agree that the sentence above describes a process (or a predicate), which involves two participant roles providing the same linking relationship between the semantic and the syntactic structures: an Actor (or Agent) / Subject, and a Goal (Patient) / Object.

We start by checking whether the subject of a sentence consists of a human or a non-human agent. After identifying the grammatical subjects in the dataset's sentences with the Python Spacy package, we manually checked the list of human agents (the five most frequent being *I* (24.56%), *She* (5.76%), *Betty* (18.43%), *John* (6.24%), *Patient* (4.86%)).[8]

Halliday's transitivity model describes the way in which the world of our experience can be divided by grammar into a manageable set of process types, the most basic of which are: *material processes* (external actions or events in the world around us; e.g., verbs like "write", "walk", "kick") and *mental processes* (internal events; e.g., verbs of thinking, feeling, perceiving). We first identify sentences containing material and mental processes by extracting the verbs in each sentence (Table 1). About 75% of the dataset contains such processes, with material processes appearing more frequently than mental ones (by a small margin: 0.9%).

Inspired by the success of Halliday's transitivity system on cognitive effects of linguistic constructions in literary texts (Nuttall, 2019), we also examine a set of construction choices which seem

[7]We also consider constructions at word level - i.e., verbs.

[8]Other subjects: *Nurse*, *Doctor*, *Family*, *Children*, *Wife*, *Husband*, and *Spouse*

to co-occur in texts as material and mental actions or events. In our quest of understanding empathy expression in student narrative essays, we want to test if such contributions lead to a reduced sense of intentionality, awareness or control for the agentive individual represented (i.e., the essay writer in the role of the doctor), and thus, identifying the stylistic profile of the narrative. Specifically, these constructions are: *Human Actor + Process (HA+P); Body Part + Process (BP+P); Other Inanimate Actor + Process (IA+P); Goal + Process (G+P)* (see Table 1). We identify HA+P to be the most common construction within our dataset, appearing in just less than half of the sentences (49.82%). The remaining constructions are much rarer with G+P being the least frequent (12.54%).

Drawing from (Langacker, 1987), Nuttall (2019) also notes that these experiences can vary in force-dynamic (energetic) quality and thus sentences exhibiting an energetic tone are linked with 'high' transitivity and those with lower or static energy can be linked to 'low' transitivity. In order to identify energetic sentences, we leverage the IBM Watson Tone Analyzer API (Yin et al., 2017) which assesses the emotions, social propensities, and language styles of a sentence. We denote sentences containing high extroversion and high confidence (values > 0.8) as energetic. Sentences with low scores are marked as static. 61.77% of the sentences exhibit a static tone, energetic tone being less frequent.

In SFG, active and passive voice plays an important role as well. Nuttall (2019) shows that, in some genres, text indicating a lower degree of agentive control tends to use more passive voice constructions. As this is also relevant to our task, we test whether voice contributes indeed to a reduced sense of intentionality, awareness or control for the Agent (in particular the essay writer playing the doctor's role) and how these features correlate with the overall empathy score at essay level. Using an in-house grammatical-role extraction tool developed on top of Spacy's dependency parser, we find that 66% of sentences use active voice and 34% passive voice.[9] 77.92% of active-voice sentences exhibit human actor subjects and only 22.08% include non-human actors. Similarly for passive voice, the majority (83.09%) of sentences had human actors. Comparing frequencies of active and passive voice across various essay empathy score ranges (Figure 1), we notice that higher empathy essays (scores >3) seem to rely more on active voice (65-70% of the sentences in active voice) as opposed to lower empathy essays (scores < 3) which have less than 65% of sentences in active voice.

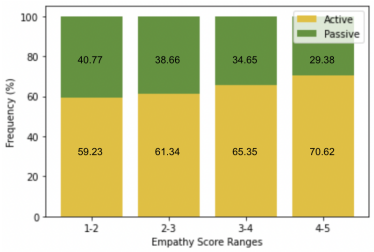

Figure 1: Frequency distribution (%) of voice in essays for various overall empathy score ranges

Stylistic research has also shown (Nuttall, 2019) the importance of movement of body parts as non-human agents. We, too, parsed sentences for the use of body parts, i.e. *eyes*, *arms*, *head* and curated a list based on anatomical terminology as defined by wiktionary.org (2022) resulting in about 18.61% of the dataset sentences (statistics for top 5 most common bodyparts are in Table 2).

Table 1 summarize all the identified constructions and stylistic features discussed in this section.

5 Empathy Classification Task

Our ultimate goal is to build an informed and performant classifier able to determine the degree of empathetic content of a medical essay overall and at sentence level. Taking advantage of form-meaning-style mappings in the language system, in this paper, we built and test a number of state-of-the-art classifiers enriched with varied constructions and stylistic features (Table 1) which are described next.

5.1 Identification of Sentence Themes

In medical training, students learn not only how to diagnose and treat patients' medical conditions, but also how to witness the patient's illness experience. In fact, in practical interactions with patients, they often switch between these positions: empathizing with the patient's situation (i.e., witnessing what it is like for the patient), and providing medical care (i.e., understanding what they need medically).

As such, we wanted to capture the distribution of such emphatic content and medical information in

[9]The active/passive voice ratio varies per genre (Strunk Jr and White, 2007). Note that in a sentence using passive voice, the subject is acted upon, which shows the main character's degree of detachment, which is of interest here.

Feature	Frequency	Definition	Example
Active	62.12%	the subject of the sentence is the one doing the action expressed by the verb	"I watched as the patient slowly sat down in the chair."
Passive	37.88%	the subject is the person or thing acted on or affected by the verb's action	"The patient I just had an appointment with is named Betty."
Material	37.39%	external actions or events in the world around us	"The nurse had already retrieved the bloodwork reports and handed them to me before I entered the room."
Mental	36.49%	events/feelings expressed by a user	" 'I can imagine that you have several questions, so I am happy to answer any questions or clear any doubts you might have.' I said to her. "
HA+P	49.82%	consists of a human actor and a material/mental process	"I calmly started explaining the treatment options."
BP+P	15.85%	consists of a non-human actor related to body parts in material/mental process	"Her shoulders started shaking when she heard the news, and I could tell she would need some time to process the news."
IE+P	18.34%	consists of an inanimate actor in material/mental process	"The file was already in the room when I arrived."
G+P	12.54%	consists of the passivisation of material/mental process and deletion of actor	"The effects of her lifestyle had already started to affect her physical strength."
Energetic	38.23%	e.g., high extroversion and confidence	"I could see Betty fidgeting with her fingers as she began to process the news."
Static	61.77%	e.g., low extroversion and confidence	"The nurse brought in the file quickly."

Table 1: Our set of SFG's transitivity constructions with their distribution and examples. Note that the total distribution should not add to 100%, as these are not mutually exclusive features.

Body Part	POS Used	Frequency	Example
Eye	subject, indirect object, prepositional object	42.96%	"I saw in her eyes tears forming as she realized the gravity of the issue at hand."
Hand	subject, prepositional object, indirect object, direct object	16.14%	"John began clasping his hands."
Head	direct object, indirect object	8.60%	"John shook his head as he sat down across from me."
Shoulder	subject, prepositional object, direct object	5.47%	"The patient shrugged his shoulders."
Body	subject, prepositional object, direct object	4.99%	"The vitals showed that the patient's body was not in its healthiest form."

Table 2: Most common body parts in the empathy essay dataset

our narrative essays of hypothetical doctor-patient interactions. Specifically, we looked at recurring topics within sentences and identified the following themes in our dataset at the sentence level: *Medical Procedural Information; Empathetic Language; Both* (Medical and Empathetic Language); and *Neither*. Sentences referring to *Medical Procedural Information* were identified based on keyword matching following established medical term vocabulary generated from Dr. Kavita Ganesan's work on clinical concepts (Ganesan et al., 2016). Sentences containing *Empathetic Language* were already annotated manually by our annotators for each essay at the sentence level (see Section 3). Sentences containing both medical procedural info and empathetic content were marked as *Both*, while remaining sentences are marked as *Neither*. Table 3 shows these categories, their definitions, examples and counts per category (10,120 sentences overall). We also give examples of two essays highlighted with these themes in the Appendix (Section 7).

In the next sections we present the classification results of various multi-class machine learning models (for each of the 4 themes: *Medical Procedural Information*, *Empathetic Language*, *Both*, and *Neither*).

5.2 Baseline Models and Analysis

In evaluating several state-of-the-art machine learning algorithms, we started with two representative baseline models: support vector machines (SVM) and logistic regression (logR). As we are interested in observing the performance of deep learning methods, we also experiment with long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997), bidirectional long-short term memory (bi-LSTM) (Graves and Schmidhuber, 2005), and convolutional neural network (CNN) (Kim, 2014) models; additionally, we use the transformer models BERT (Devlin et al., 2018) and roBERTa.

Theme	Freq.	Example
Medical Procedural Information	37.39%	"The patient's vitals showed that his body was not healthy and it was necessary to make some diet and lifestyle changes."
Empathetic Language	36.49%	"I noticed Betty looked confused and so I tried to reassure her we would do everything possible to make the changes in her lifestyle."
Both	21.28%	"I knew the statin treatment could be difficult, so I wanted to make sure Betty felt comfortable and understood the procedure."
Neither	4.84%	"The file was left on the counter, and I picked it up before going in to see Betty."

Table 3: Examples and distribution of identified themes in sentences

Classifier	Medical Procedural Information			Empathetic Language			Both			Neither		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
SVM	0.70	0.68	0.69	0.52	0.61	0.56	0.49	0.47	0.48	0.78	0.39	0.51
LogR	0.62	0.67	0.64	0.49	0.54	0.51	0.51	0.53	0.52	0.68	0.61	0.64
LSTM	0.64	0.69	0.67	0.51	0.54	0.52	0.59	0.53	0.56	0.66	0.61	0.63
biLSTM	0.65	0.7	0.68	0.51	0.54	0.52	0.56	0.53	0.54	0.68	0.62	0.65
CNN	0.70	0.71	0.70	0.52	0.54	0.53	0.64	0.53	0.57	0.71	0.63	0.66
BERT	0.69	0.72	0.70	0.55	0.61	0.58	0.57	0.63	0.60	0.68	0.65	0.66
constructionBERT	0.71	0.73	0.72	0.64	0.67	0.65	0.76	0.58	0.66	0.78	0.72	0.75
constructionBERT-*Voice:Active*	0.71	0.73	0.72	0.58	0.63	0.65	0.64	0.64	0.62	0.77	0.72	0.74
constructionBERT-*Voice:Passive*	0.71	0.73	0.72	0.65	0.67	0.66	0.76	0.61	0.67	0.78	0.72	0.75
constructionBERT-*Process:Material*	0.70	0.72	0.71	0.61	0.65	0.63	0.68	0.58	0.63	0.78	0.72	0.75
constructionBERT-*Process:Mental*	0.70	0.72	0.71	0.59	0.63	0.61	0.66	0.58	0.62	0.78	0.71	0.74
constructionBERT-*HA+P*	0.69	0.72	0.70	0.59	0.64	0.62	0.66	0.58	0.62	0.68	0.69	0.68
constructionBERT-*BP+P*	0.71	0.73	0.72	0.55	0.64	0.59	0.61	0.63	0.62	0.71	0.72	0.71
constructionBERT-*IE+P*	0.70	0.73	0.71	0.61	0.64	0.62	0.73	0.57	0.64	0.76	0.72	0.74
constructionBERT-*G+P*	0.71	0.73	0.72	0.64	0.66	0.65	0.74	0.56	0.64	0.78	0.72	0.75
constructionBERT-*Tone:Energetic*	0.71	0.73	0.72	0.58	0.62	0.60	0.66	0.57	0.61	0.78	0.72	0.75
constructionBERT-*Tone:Static*	0.71	0.73	0.72	0.64	0.62	0.63	0.71	0.58	0.64	0.78	0.73	0.75

Table 4: Precision, recall and F1 scores of all baseline classifiers on the imbalanced test dataset: 770 *Medical Procedural Information*, 722 *Empathetic Language*, 433 *Both*, 98 *Neither* sentences

As we are performing sentence classification, our features are unigrams (single words). For the logistic regression models, we used a L2 regularization and for the SVM models, a linear kernel function. We initialized the embedding layers in our neural models (LSTM, bi-LSTM, CNN) with GloVe embeddings since the expression of empathy involves larger units than words, and embeddings are known to better capture contextual information. We further decided to apply an attention layer to these models to learn patterns that may improve the classification. For the transformer BERT and roBERTa models, we use the default embeddings and 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 each type of theme, we reserve an 80/20 training/test ratio, with 5-fold cross validation. As our dataset is imbalanced, we report the precision, recall, and F1-score (harmonic mean of the precision and recall) as shown in Table 4.

We observe that the classification of *Empathetic Language* is particularly difficult. The best model is the transformer BERT model which achieves an F-1 score of 0.58. On the other hand, sentences with *Medical Procedural Information* are much easier to identify with most classifiers achieving an F-1 score above 0.65. Sentences labeled *Both* are increasingly difficult (best classifier score of 0.6 F-1). Classification scores for sentences containing *Neither* fall just short of scores from *Medical Procedural Information* sentences.

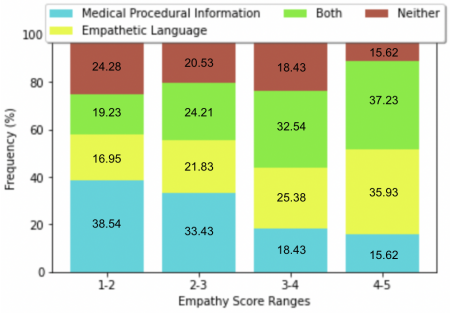

Figure 2: Frequency distribution (%) of themes in essays for various empathy score ranges

To better understand how these themes correlate with the overall empathy score at essay level, we compare frequencies and distribution of each theme for various essay empathy score ranges (Figure 2) across the entire dataset. High empathy essays

(scores >3) tend to show a large amount of *Empathetic Language* and *Both*, while low empathy essays (scores < 3) seem to favor *Medical Procedural Information* language.

Heatmaps of Medical Narrative Essays. It is also interesting to visually analyze the distribution of these themes in the layout of the narrative essays. Thus, for each essay, we highlight the sentences containing each theme and generate heat maps that might highlight high theme concentrations. We standardized the format of each essay to an A4 paper,[10] generating a 42 x 14 matrix. [11] For each essay and position – i.e., (row, column) – we note the occurrence of each theme. We then build a heat map from these counts, thus generating 3 heatmaps, one for each theme along the following overall empathy score ranges: (1-2), (2-3), (3-4), and (4-5) (Figure 3).

The heatmaps for theme *Medical Procedural Information* for low empathy score essays show darker colors (purple) indicating a higher frequency of use at the beginning and middle of the essay. Lighter colors (orange and yellow) showcasing lower concentrations of the theme seems to be more prevalent in higher empathy score essays. *Empathetic Language* tends to increase in coverage (i.e., darker color portions) from low to high-score empathy essays, with a preference toward the end of the essay.[12] *Both* themes seem to concentrate, specifically towards the top and middle of the essays for high empathy scores (darker colors). Low empathy essays also show some shades of purple (i.e. some concentration) towards the bottom and lower third of the essays.

5.3 Incorporating Halliday Features into the Theme Classifier

In this section, we seek to improve our sentence theme classifier by incorporating the constructions and stylistic features identified in Section 4. For each sentence, we append a Boolean value indicating whether each feature is present in the given sentence – e.g., if a sentence is in active voice (feature *Active* is 1; feature *Passive* is 0); if the sentence contains a HA+P (feature value is 1), and so on. Since in our baseline experiments the BERT model gave the best results across all 4 themes, we extend it here with all the features (construction-BERT) and report new scores (see bottom part of Table 4). Indeed, the inclusion of these features yields better performance, with a large increase for most of our themes including, *Empathetic Language*, *Both*, and *Neither*, and smaller performance increases in *Medical Procedural Information*.

Leave-one-out feature contribution experiments (see bottom of Table 4) show that removing *Voice: Active* and *Voice: Passive* slightly decreases performance in *Empathetic Language* and *Both* (with *Voice: Active* providing the highest decrease).

Removing *Processes* also shows a fair decrease in all themes except *Neither* which shows no change in performance. A deeper analysis indicates that *Processes: Material* helps with *Medical Procedural Information* but hurts performance on *Empathetic Language*.

The constructions *HA+P* and *BP+P* are most important for classification; the removal of *BP+P* yields the lowest F-1 score measure for detecting empathy. This shows the doctor (i.e., the student writer) paid particular attention to the patient's emotional state (thus showing empathy). Body parts in this type of discourse are particularly associated with non-verbal emotional language, which is highly indicative of empathy. *HA+P* is also an important feature for the theme *Neither*. Removal of *IE+P* gives a slight decrease in performance, while *G+P* has almost no effect on the classification results. Finally, the *Tone: Energetic* and *Tone: Static* features (constructionBERT-*Tone*) show to be important for the themes *Medical Procedural Information*, *Empathetic Language*, and *Both*. For *Tone: Energetic*, there is a 0.02 decrease in F-1 for medical procedural information, and a 0.05 for *Empathetic Language* and *Both*. For *Tone: Static*, we observe a decrease in performance for *Empathetic Language* by 0.02 and *Both* by 0.01.

With our binary classification task, we see similar patterns as constructionBERT-Tone yields much lower performances. The energetic and static tones yield 0.004 and 0.01 increases in F-1 scores for *Medical Procedural Information* and *Empathetic Language*. Our analysis also showed that G+P (Goal+Process), Processes (Mental and Material), and HA+P (Human Actor+Process) were also increasingly important for score improvements.

Interested in directly comparing the *Medical Pro-*

[10] Times New Roman, size 12: 42 lines of 14 words each

[11] We generated a separate heatmap (size: 81 x 14) for 24 essays since these were much longer and didn't fit on a standard A4 paper. These showed similar position patterns.

[12] A closer look indicates that students who wrote low-empathy essays showed a tendency to use some emotional language in the last paragraph - which appeared rather rushed and forced.

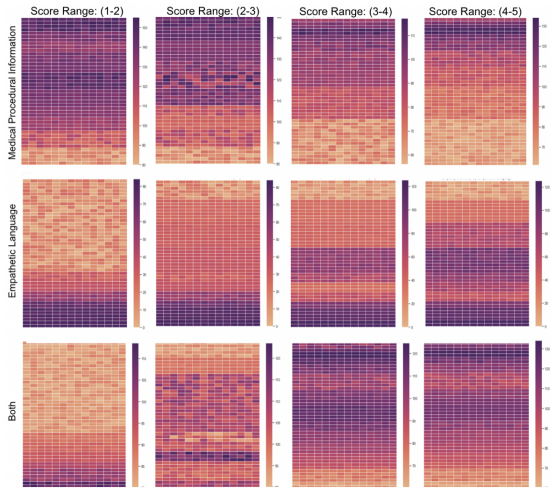

Figure 3: Heatmaps for themes in sentences of narrative essays across all overall empathy score ranges: Row#1 shows heatmaps for *Medical Procedural Information*; Row#2 for *Empathetic Language*; Row#3 for *Both*. Dark colors (purple) indicate that many essays exhibit the theme in the respective position of the essay. Light colors (yellow) indicate a small number of essays have occurrences of the theme for the given position.

cedural Information and *Empathetic Language* sentences, we further built a binary version of the simple BERT model, and another of constructionBERT, and found these tasks to be slightly easier. The binary BERT model achieved an F-1 score of 0.75 for *Medical Procedural Information* and a 0.62 for *Empathetic Language*. After adding the generated features (i.e., the binary constructionBERT), we see a small increase in F-1 scores (+0.01 for *Medical Procedural Information* and +0.03 for *Empathetic Language*).

Overall, the results of the effects of transitivity features on meaning, perceived agency and involvement of the Agent are in line with those obtained for literary genre texts by Nuttall (2019) through manual inspection. More specifically, the stylistic choices given by such linguistic constructions seem to be good indicators of the degree of perceived agency an Agent has in relation to others and the environment, as tested here for the empathy task on our dataset. In research on stylistics, the set and usage of such stylistic constructions and features in a text is known as the stylistic profile of the text. Encouraged by the correlations between Halliday's features with our essay level empathy scores, we would like to extrapolate and maintain that a set of rich stylistic constructions (like those tested in this research) can ultimately lead to informative **Empathy Profiles** – essay level form-meaning-style structures that can give an indication of the degree of social and empathetic detachment of the doctor toward the patient. Of course, while more research is needed in this direction, we believe we showed here the potential of such an approach to the task of empathy detection classification overall, and to clinical context in particular.

6 Conclusions

Medical education incorporates guided self-reflective practices that show how important it is for 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 relation to others brings together a specific set of values, attitudes, and competencies that can be supported through ongoing self-reflection. Such interactions can be captured in language via constructions as part of CxG and Halliday's transitivity system.

In this paper, we bring various aspects of these theories in a deep learning computational framework to model empathetic language in a corpus of essays written by premed students as narrated simulated patient–doctor interactions. We start with baseline classifiers (state-of-the-art recurrent neural networks and transformer models). Then, we enrich these models with a set of linguistic constructions proving the importance of this novel approach to the task of empathy classification for this dataset. Our results indicate the potential of such constructions to contribute to the overall empathy profile of first-person narrative essays.

References

Walter F Baile, Robert Buckman, Renato Lenzi, Gary Glober, Estela A Beale, and Andrzej P Kudelka. 2000. Spikes—a six-step protocol for delivering bad news: application to the patient with cancer.

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.

Margaret Bearman. 2003. Is virtual the same as real? medical students' experiences of a virtual patient. *Academic Medicine*, 78(5):538–545.

Margaret Bearman, Jennene Greenhill, and Debra Nestel. 2019. The power of simulation: a large-scale narrative analysis of learners' experiences. *Medical education*, 53(4):369–379.

Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry*, 18(1):1–21.

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*.

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

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*.

Priyanka Dey and Roxana Girju. 2022. Enriching deep learning with frame semantics for empathy classification in medical narrative essays. In *Proceedings of the 2022 Workshop on Health Text Mining and Information Analysis (LouHI) collocated with the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, hybrid. Association for Computational Linguistics.

Peter Dieckmann, David Gaba, and Marcus Rall. 2007. Deepening the theoretical foundations of patient simulation as social practice. *Simulation in Healthcare*, 2(3):183–193.

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

Douglas Ezzy. 1998. Theorizing narrative identity: Symbolic interactionism and hermeneutics. *Sociological quarterly*, 39(2):239–252.

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.

Charles J Fillmore, Paul Kay, and Laura A Michaelis. 2006. *Construction grammar*. Center for the Study of Language and Information.

Roger Fowler. 1996. *Linguistic Criticism*. Oxford: Oxford University Press, 2nd edition.

Roger Fowler. 2013. *Linguistics and Novel*. Routledge.

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

Kavita Ganesan, Shane Lloyd, and Vikren Sarkar. 2016. Discovering related clinical concepts using large amounts of clinical notes. *Biomed Eng Comput Biol*, 7(Suppl 2):27–33. Becb-suppl.2-2016-027[PII], 27656096[pmid].

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.

Adele Goldberg. 1995. Constructions: a construction grammar approach to argument structure. *Chicago: The University of Chicago*.

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

Michael AK Halliday. 1994. An introduction to functional grammar, london: Edward arnold.———& ruqaiya hasan. 1976. *Cohesion in English. London & New York: Longman. SHELL NOUNS*, 131.

Michael AK Halliday. 2019. Linguistic function and literary style: an inquiry into the language of william golding's' the inheritors'. In *Essays in modern stylistics*, pages 325–360. Routledge.

Michael Alexander Kirkwood Halliday, Christian MIM Matthiessen, Michael Halliday, and Christian Matthiessen. 2014. *An introduction to functional grammar*. Routledge.

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

David L Hoover. 2004. Altered texts, altered worlds, altered styles. *Language and Literature*, 13(2):99–118.

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.

Lesley Jeffries. 2017. *Critical stylistics: The power of English*. Bloomsbury Publishing.

Paul Kay and et al. 1999. Grammatical constructions and linguistic generalizations: the what's x doing y? construction. *Language*, 75(1):1–33.

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.

Daniel Kies. 1992. The uses of passivity: suppressing agency in nineteen eighty-four. *Advances in systemic linguistics: Recent theory and practice*, pages 229–250.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

FY Lin and AX Peng. 2006. Systemic functional grammar and construction grammar. In *Presented during the 33rd International Systemic Functional Congress*, pages 331–347.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

William C McGaghie, Saul B Issenberg, Jeffrey H Barsuk, and Diane B Wayne. 2014. A critical review of simulation-based mastery learning with translational outcomes. *Medical education*, 48(4):375–385.

Enrique Menéndez. 2017. Christopher hart: Discourse, grammar and ideology. *Pragmática Sociocultural/Sociocultural Pragmatics*, 5(2):259–262.

Laura R Micciche. 2004. Making a case for rhetorical grammar. *College Composition and Communication*, pages 716–737.

Martin Michalski and Roxana Girju. 2022. An empathy account of premed students' narrative essays. *OSF Preprints*.

Louise Nuttall. 2019. Transitivity, agency, mind style: What's the lowest common denominator? *Language and Literature*, 28(2):159–179.

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.

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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

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.

Paul Simpson. 2003. *Language, ideology and point of view*. Routledge.

Paul Simpson and Patricia Canning. 2014. Action and event. In *The Cambridge handbook of stylistics*, pages 281–299. Cambridge University Press.

William Strunk Jr and Elwyn Brooks White. 2007. *The Elements of Style Illustrated*. Penguin.

Teun A Van Dijk. 2017. *Discourse and power*. Bloomsbury Publishing.

Max Van Manen. 2016. *Researching lived experience: Human science for an action sensitive pedagogy*. Routledge.

Robert D Van Valin. 2007. Adele e. goldberg, constructions at work: the nature of generalization in language. oxford: Oxford university press, 2006. pp. vii+ 280. *Journal of Linguistics*, 43(1):234–240.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. 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.

wiktionary.org. 2022. Appendix:visual dictionary/human body - body parts. [Online; accessed 29-October-2022].

Peifeng Yin, Zhe Liu, Anbang Xu, and Taiga Nakamura. 2017. Tone analyzer for online customer service: An unsupervised model with interfered training. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1887–1895, New York, NY, USA. Association for Computing Machinery.

7 Appendix

Figure 4 shows two examples of essays, one with low empathy and one with high empathy, highlighted with the themes: *Medical Procedural Information* (cyan), *Empathetic Language* (yellow), and *Both* (green). *Neither* sentences are not highlighted. It is interesting to see that in Essay (a), the sentences mentioning diet and exercise were not identified as *Medical Procedural Information* given that they were not found in Dr. Kavita Ganesan's work on clinical concepts (Ganesan et al., 2016).

Betty is 32 years old, has a spouse, and two young children (age 3 and 5). Betty has no family history of heart disease. In the past 6 months, Betty has begun experiencing left-side chest pain. Betty's bloodwork has revealed that her cholesterol is dangerously high. Betty will require statin therapy and may benefit from a healthier diet and exercise.

I asked what medical concerns that she has or any past medical history. I confirmed with her that concerns are valid and that the chest pain is something I am concerned about also. I told her that the blood work panel that was performed confirmed her concern. I asked her what she knows about her latest blood cholesterol levels history. I informed her that the cholesterol levels have reached an actionable level and we will need to decide which plan will be the best plan to implement. I asked her what she knows about LDL and HDL and how they work and what affects each of them. I explained which parts she did not quite understand on her level of need. I explained about the change in diet and exercise that will be an important part of the plan moving forward. I explained how the other part of the plan will include the use of statins as a means of lowering the LDL levels in her body. I explained the side effects of statins in how type 2 diabetes may be one possible side effect. I explained that once on the cholesterol drug that she should keep taking them because it is keeping an important balance in the body's cholesterol levels. I explained the symptoms that she could experience with drugs that should be reported to me immediately and if there are any other concerns that she should call me with those concerns. She seemed receptive to the treatment plan. I will follow-up with her in a month to evaluate the progress and review her bloodwork. She will have a consult for diet planning within a week.

(a) Example of Essay with Empathy Score: 1

What a whirlwind that was. Everything was a surprise; the diagnosis, her reaction, her response. Heck, even my response was a bit startling. High cholesterol? But how? She has no history of this kind of thing and at such a young age too. I mean, the chest pain is a clear sign, but even that is surprising to see in a woman like Betty. It's terrible knowing that I had to tell someone that not only did they have a condition that could kill them, but it also was their fault too. I understand Betty's reaction; anyone in her situation would've thrown a tantrum. I just hated experiencing that, but I guess it's a part of the job description. After she calmed down, I told her that she could call her husband to inform him of the situation, but before calling, I figured it was a good idea to go over the next steps.

The treatment that was prescribed to Betty is called Statin Therapy. What the statin does is block the enzyme HMG CoA Reductase that is essentially what makes cholesterol. This reduces the amount of cholesterol produced in the liver. Of course, most patients don't care much for how the treatments work unless they have to do something, so naturally, Betty was more concerned with the idea of eating healthy and exercising.

I'm sure in the back of her mind, she knew that those two things would be somewhat required, but like any person who is comfortable in their lifestyle, she fought the changes. "Is there any other way?" she said. I could tell she was distraught over this idea that she had to choose between comfort and health. I find it easiest to soothe the patient by going over the plan in small steps. It's also helpful to show them the benefits in the long-term and the final goal. She seemed reluctant at first, but I think by the end of our conversation, I was able to coax her into believing that this is a good thing for her. It's good to note that it's always positive to identify these conditions as early as possible, so, at the very least, I can say with confidence that I'm glad Betty came in.

(b) Example of Essay with Empathy Score: 5

Figure 4: Two Sample Essays from the Dataset Highlighted by Sentence Themes