Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

DocuSigned by:
Signature: Nlofar Vafaie
75BB564D2ACE4A7...

Nilofar Vafaie
Name

5/15/2023 | 3:07 PM EDT
Date

Title Hurting BERT's Feelings: Toward a Computational Model of Emotion Regulation in Context

Author Nilofar Vafaie

Degree Master of Arts

Program Psychology

Approved by the Committee

DocuSigned by:
Philip Kragel
E39E7EB16272405...

Philip Kragel
Advisor

DocuSigned by:
Patricia Brennan
685971CA57FC4EC...

Patricia Brennan
Committee Member

DocuSigned by:
Phillip Wolff
A906AD9BE70D49C...

Phillip Wolff
Committee Member

Accepted by the Laney Graduate School

Kimberly Jacob Arriola, PhD, MPH
Dean, James T. Laney School of Graduate Studies

Hurting BERT's Feelings:

Toward a Computational Model of Emotion Regulation in Context

Nilofar Vafaie

Department of Psychology

Emory University

An abstract of a thesis

submitted to the Department of Psychology Laney Graduate School,

Emory University In Fulfillment of the Requirements

For the Degree of Master of Arts

June 2023

Abstract

Emotion regulation is a complex phenomenon, typically operationalized using ontologies of different strategies (e.g., cognitive reappraisal) and assessed using low dimensional assays, with affect ratings serving as the sole measure of regulation success. Such low-dimensional conceptualizations of emotion regulation limit the development computationally explicit accounts of the cognitive and neural processes involved in different regulation strategies. Here we take an alternative approach, using deep language models to test the hypothesis that emotion regulation changes the meaning of events as reflected in a high-dimensional semantic space. We conducted an online study in which participants either used reappraisal, mindfulness, or provided an objective description of stimuli after viewing short affective videos. Fine-tuned Bidirectional Encoder Representations from Transformers (BERT) were able to classify regulation strategy as well as emotional situation based on text descriptions of events. Fine-tuning in regulation specific language changed representations across layers of a BERT model, with a main effect of situation in the last layer and an interaction between strategy condition and situation. These findings suggest that regulation systematically alters the language produced when describing emotional events. The nature of these changes depends both on the type of emotional situation and the emotion regulation strategy employed. Importantly, these changes increase deeper into a language model, with implications for brain mappings of early vs late cortical areas. We further assessed generalizability of our models to independent archival data and found that model classifications better explained variations in affect compared to the experimental labels. This work paves the way for objective modeling of emotion regulation, with applications in a variety of settings in order to facilitate a deeper understanding of how emotion and emotion regulation are operationalized behaviorally and in the brain.

Keywords: emotion, emotion regulation, language, deep language models

Hurting BERT's Feelings:

Toward a Computational Model of Emotion Regulation in Context

By

NILOFAR VAFAIE

THESIS

Submitted to the Department of Psychology

Laney Graduate School, Emory University

In Fulfillment of the Requirements

For the Degree of Master of Arts

June 2023

“Oh, who can hold a fire in his hand
By thinking on the frosty Caucasus?
Or cloy the hungry edge of appetite
By bare imagination of a feast?
Or wallow naked in December snow
By thinking on fantastic summer's heat?
O, no! the apprehension of the good
Gives but the greater feeling to the worse.”

– An exiled son’s rejoinder to a father urging him to use imagination to reconfigure his suffering.

William Shakespeare, *Richard II*

Perhaps not for situations as extreme as the ones in the verses above, but when can one use imagination to reconfigure suffering – or in modern psychological parlance, cognitively reappraise a terrible situation? After all, reappraisal is a cornerstone of affective and clinical science, and one of the most studied regulatory strategies. Indeed, not only for reappraisal, but on the whole, a wealth of psychological research has shown emotion and emotion regulation to be part and parcel of daily life, and essential to healthy adaptation and functioning (Berking & Wupperman, 2012; Chervonsky & Hunt, 2019; Hu et al., 2014; Inwood & Ferrari, 2018).

Our current understanding of emotion generation and regulation points to a complex, recursive and recurrent interplay of different processes such as valuation, intro/exteroceptive and language systems (Barrett, 2017; Gross, 2015; Satpute & Lindquist, 2021). Regulation potentially overlaps with language at multiple levels, as indicated by the brain regions that are implicated both in language and emotion processing (e.g. ventrolateral prefrontal cortex, inferior frontal gyrus, and precuneus), or theoretical accounts that pose an integral role for language in the emotion generation and regulation processes (Barrett et al., 2016; Brooks et al., 2017; Gendron et al., 2012; Satpute & Lindquist, 2021). Constructivist accounts of emotion hold that there can be no emotion without language, as emotion is constructed by language concepts, taking into account biological precepts elicited along dimensions of valence and arousal (Barrett, 2017; Lindquist et al., 2015). Semantic space theory of emotion posits that emotion concepts are embedded in a high-dimensional semantic space (Cowen & Keltner, 2021). Basic emotion views too suggest the existence of semantically distinct categories of emotion (Johnson-Laird & Oatley, 1989). And while appraisal theories of emotion do not attribute a causal role to language itself, they maintain that core relational themes with deep meaning are derived with the use of language and are shared across instances of the same emotion, despite having different surface features (Moors et al., 2013). Given the purported importance of language to emotion, whether viewed as a constituent part or a reflection of aspects of the underlying cognitive processes, studies have sought various methos of delving deeper into its phenomenological role, for example by employing lexical analysis - e.g. profile of pronoun and verb tense use (Brooks et al., 2017; Nook et al., 2021; Orvell et al., 2019). However, this research has focused on specific linguistic signatures specified a-priori and has not been able to holistically look at language and assess how words represent semantic content related to emotion. Consequently, such approaches do not allow for building an explicit model of how regulation could influence emotional responding.

Here, we take an alternative approach by leveraging advanced computational and deep language models to examine the role that contextual semantic mappings play in emotion generation and its regulation. To do so, we take advantage of word embeddings – high dimensional numerical vectors that represent meaning in semantic space, and are used as proxies for human semantic representation (Chang & Chen, 2019; Günther et al., 2019; Rudkowsky et al., 2018; Schuster et al., 2019). These model spaces are built on the distributional hypothesis, positing that language reflects distributional information in the environment, such that meanings are arrived at from co-occurrence of words in similar contexts (Harris, 1954; Lopez et al., 2020; McDonald & Ramscar, 2001). The high-dimensional semantic space in different embeddings is built on large language corpora, taking into account the contextual distribution of words to arrive at a geometric embedding representing meaning. Some language models (e.g. Global Vectors for Word Representation, or word2vec) constitute static representations of meanings, whereby once a semantic space is formed, each word, regardless of its placement in any particular context will be represented with the same vector embedding (Miaschi & Dell'Orletta, 2020). Contextual word embeddings, on the other hand, take all other words in a given sentence into account in order to arrive at a contextual embedding for that word or sentence (Chang & Chen, 2019).

Recent work has shown that word embeddings represent world knowledge about object categories (Arana et al., 2023; Grand et al., 2022). In the domain of emotion, word embeddings could reflect knowledge about the emotional significance of an object in context (e.g., that a lion in the open savannah is more of a threat than a lion at a zoo). Further, context-dependent word embeddings may reflect the variable nature of emotional meaning. If emotion regulation alters the meaning of sensory information, then linguistic descriptions of emotional events can serve as a proxy into the underlying cognitive processes by reflecting the change that occurs on a semantic level.

One approach for identifying word embeddings that are context sensitive is to fit large language models using Bidirectional Encoder Representations from Transformers (BERT) (Biggiogera et al., 2021; Devlin et al., 2018; Hollis & Westbury, 2016; Rudkowsky et al., 2018; Tanana et al., 2021). Because BERT embeddings represent meaning as high-dimensional numerical vectors, they can be used to evaluate whether descriptions of emotional events systematically differ across contexts—in terms of the regulation strategy used or the type of the emotional situation itself. Another defining property of BERT embeddings are their sensitivity to context in the training set, and the ability to fine-tune embeddings on domain specific language (Chang & Chen, 2019; Miaschi & Dell'Orletta, 2020). Thus, BERT embeddings pre-trained on large text corpora can be viewed as normative accounts of emotional meaning established across a wide array of situations, whereas fine-tuned embeddings are more sensitive to differences in context in order to increase fit and model accuracy for a specific task – in our case, specific emotion regulation strategies.

Viewed from the lenses of the diverse theoretical perspectives of basic, appraisal, constructive or the semantic space theory, emotions can be characterized on the basis of semantic features captured by word embeddings. Whether emotions are best explained as abstract concepts embedded in a high-dimensional semantic space, or differentiated in terms of core relational themes in which varied instances of emotion share the same underlying meaning, there should be systematic mappings between descriptions of emotional events and specific emotion categories. Further, different patterns of appraisal should lead to systematic differences in emotional meaning. That is, different regulation strategies should be associated with distinct word embeddings. Lastly, the third family of theories characterizing emotions as situated or ad-hoc categories that are determined by context, language, and culture, among other factors (Barrett, 2017; Lindquist et al., 2015), would suggest that mappings between language and emotion categories are more variable and change depending on how they are conceptualized in the moment (Gross & Feldman Barrett, 2011).

Decoding emotional situations from context-sensitive BERT embeddings.

We evaluated these competing accounts by conducting a behavioral study in which participants either used positive reappraisal (N = 30), mindfulness (N = 29), or simple objective description of the stimuli (N = 30) to write a caption after viewing affective clips across eight varieties of emotional situations (number of stimuli in each condition, total number of stimuli). We validated emotion induction using a continuous measure of affective valence and an unconstrained text description of feelings. To test for systematicity in mappings between word embeddings and emotion categories, we extracted pre-trained BERT embeddings for each caption and used Partial Least Squares discriminant analysis (Wold et al., 2001) to build two

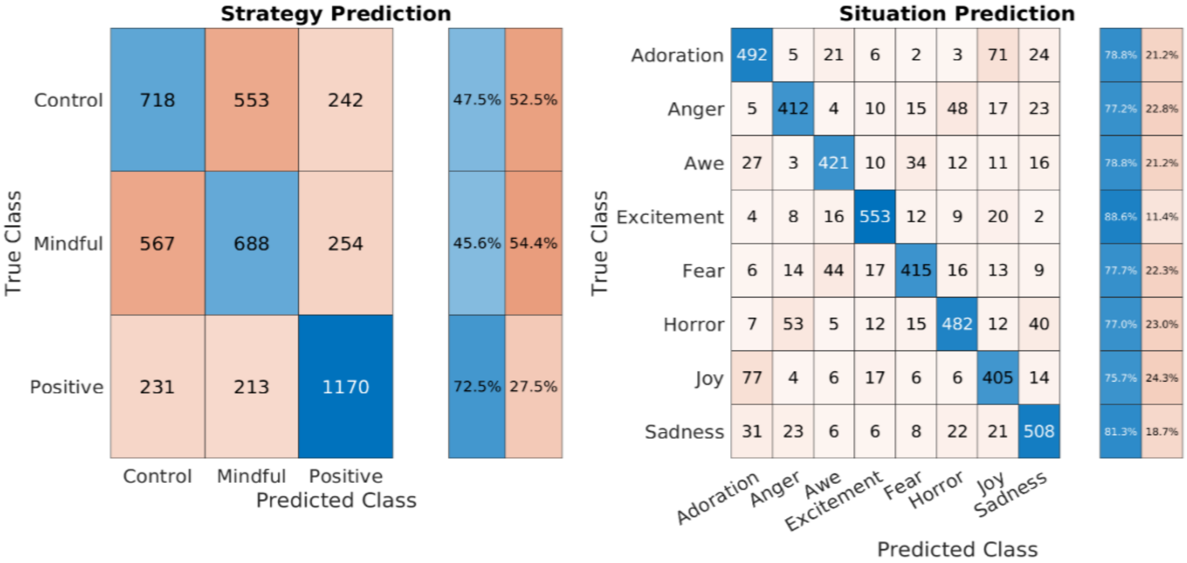


Figure 1 Pre-trained BERT embeddings predict regulation strategy (left) and emotional situation (right). Marginal cells indicate the hit rate and false positive rate for each class. Cool colors denote accurate predictions whereas warm colors denote errors.

separate decoders, one for classifying regulation strategy (3-way classification) and another for classifying emotional situation (8-way classification).

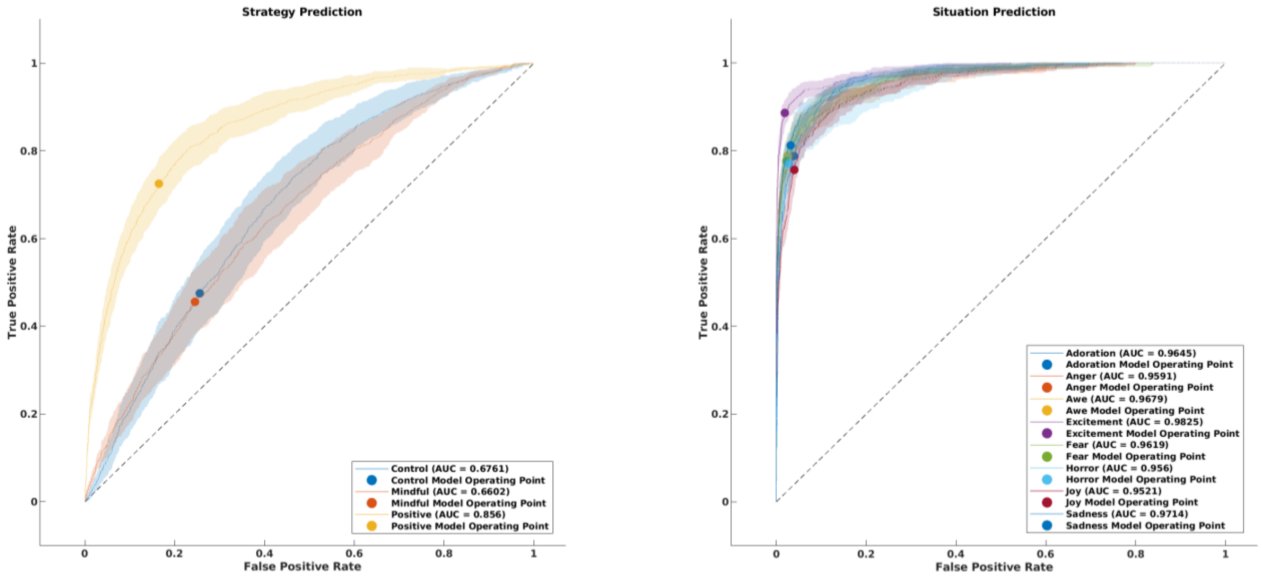


Figure 2 - Receiver operating characteristic (ROC) for multiclass classifiers. Left shows the ROC curves for each strategy prediction, and right shows the ROC curves for each situation prediction in our models. For a perfect classifier, Area Under the Curve (AUC) would be equal to 1, and for a classifier that randomly assigns observations to classes, AUC would be equal to 0.5. The shaded area around the ROC curves indicates the confidence intervals calculated from cross-validated data. The confidence intervals represent the uncertainty of the curve due to the variance in the test set for the trained model.

Decoders classified word embeddings from the pre-trained BERT model along eight emotional situations (multi-class Area Under the Curve (AUC): Adoration .97, Anger .96, Awe .97, Excitement .98; Fear .96, Horror .96, Joy .95, Sadness .97, Overall AUC=.96) and three regulation strategies with high levels of accuracy (multi-class AUC: reappraisal .86, mindfulness .66, control .68, Overall AUC =.73). See Figure 1 for confusion charts for each decoder, and Figure 2 for the corresponding Receiver operating characteristic (ROC) plots for each classifier. Table 1 shows sensitivity, specificity, AUC, and confidence intervals for each classification.

Table 1

ROC Metrics for situation and strategy PLS-DA classifiers

Category	Sensitivity	Specificity	AUC
Adoration	90.22 95% CI [89.06,91.38]	56 95% CI [54.15,57.84]	0.96 95% CI [0.96,0.97]
Anger	90.66 95% CI [89.23,92.08]	55.06 95% CI [53.26,56.87]	0.96 95% CI [0.95,0.97]
Awe	91.41 95% CI [90.18,92.64]	55.15 95% CI [53.45,56.85]	0.97 95% CI [0.96,0.97]
Excitement	91.73 95% CI [90.74,92.71]	56.26 95% CI [54.21,58.31]	0.98 95% CI [0.98,0.99]
Fear	90.89 95% CI [89.68,92.1]	55.06 95% CI [53.29,56.83]	0.96 95% CI [0.95,0.97]
Horror	89.46 95% CI [87.47,91.46]	55.96 95% CI [54.26,57.65]	0.96 95% CI [0.94,0.97]
Joy	90.04 95% CI [88.58,91.5]	54.96 95% CI [53.18,56.74]	0.95 95% CI [0.94,0.96]
Sadness	90.77 95% CI [89.63,91.91]	56.15 95% CI [54.27,58.03]	0.97 95% CI [0.96,0.98]
Control	61.89 95% CI [56.72,67.07]	55.69 95% CI [51.47,59.91]	0.68 95% CI [0.61,0.74]
Mindful	60.8 95% CI [56.08,65.51]	55.21 95% CI [51.25,59.18]	0.66 95% CI [0.6,0.72]
Positive	73.18 95% CI [70.01,76.35]	62.31 95% CI [59.38,65.25]	0.86 95% CI [0.81,0.9]

Linear readout of BERT embeddings exhibited a high ability to predict emotional situation (AUC = .95), indicating a clear delineation in language corresponding to the type of emotional situation experienced. This is consistent with semantic accounts of emotion, positing the existence of a high-dimensional semantic space wherein emotional experiences and expressions cluster along blends and gradients (Cowen & Keltner, 2021). Further, classifiers accurately predicted positive reappraisal (AUC = .86) and modestly predicted mindfulness (AUC = .66) and control (AUC = .68) conditions. These results point to the existence of information about both the regulation strategy as well as the type of emotional situations in the language produced when describing events – consistent with basic and appraisal views suggesting that there are stable mappings between different situations, cognitive appraisals, and emotion categories (Moors et al., 2013).

Emotion regulation alters the meaning conveyed by BERT embeddings

Constructionist views of emotion predict that the emotional significance of an event varies depending on how it is conceptualized in the moment. This suggests that there is not a single word embedding that should characterize the relationship between the meaning of words and emotions, but rather that mappings should depend on the nature of regulation employed. We used BERT embeddings to test this prediction by comparing BERT embeddings trained on large text corpora to embeddings from models fine-tuned to predict emotional situations from descriptions produced while participants employed different regulation strategies. Fine-tuning allows the model to retain much of its pre-trained semantic and syntactic information, however, it results in alterations of activations and representations across the layers that transform tokens into context-dependent word embeddings (Durrani et al., 2022; Merchant et al., 2020; Sun et al., 2019; Zhang et al., 2020; Zhou & Srikumar, 2021).

If emotion regulation concerns changes to representations in a high dimensional space—effectively warping a semantic field—restricting training to language generated under a specific type of regulation should differentially alter representations across layers of a BERT model. Consequently, the word embeddings from models trained on different regulation strategies should convey distinct alterations as a result of regulation. Furthermore, if different antecedents are linked to different core meanings, suggesting they can most effectively be regulated by altering different meaning structures (e.g., ‘the man took a bite out of the dog’ may have a different meaning when produced during positive reappraisal compared to mindfulness), then this change should depend on both the regulation strategy as well as the context of the emotional situation.

To assess how regulation changes word meanings in a high-dimensional semantic space, and how the interplay of emotional situation and regulation strategy impact the resulting change in meaning, we fine-tuned three BERT models, each trained to classify eight emotional situations from captions produced by participants using a different regulation strategies. To quantify deviations from pretrained BERT embeddings, we performed a representational similarity analysis (Kriegeskorte et al., 2008; Wu et al., 2020; Zhou & Srikumar, 2021) by calculating the similarity (Pearson's correlation coefficient) of embeddings from different layers of the fine-tuned models with the corresponding layers in the pre-trained BERT model. We assessed how these semantic representations for each caption differed in terms of regulation strategy, network layer, and situational context using a linear mixed effects ANOVA.

This analysis showed that compared to pretrained embeddings, regulation-specific fine-tuning differentially changed word embeddings across layers ($F_{11,89} = 1591.3$, $p < .001$; Figure 3). Furthermore, these differences were qualified by an interaction between layer depth, and the regulation condition the strategy condition ($F_{22,89} = 1.82$, $p =.01$; Figure 4). Consistent with prior studies examining the effect of fine-tuning on BERT embeddings (Merchant et al., 2020; Sun et al., 2019; Zhang et al., 2020; Zhou & Srikumar, 2021), similarity decreases with increased model depth and was lowest at layer 12 for all strategy conditions and situations (Figure 5). This is in line both with the computational literature examining the effects of fine-tuning in deep language models (Durrani et al., 2022; Merchant et al., 2020; Zhang et al., 2020; Zhou & Srikumar, 2021), and recent neuroscientific work looking at mappings between layers of deep learning models and brain activity (Caucheteux et al., 2022; Caucheteux & King, 2022; Goldstein et al., 2022; Heilbron et al., 2022; Schrimpf et al., 2021).

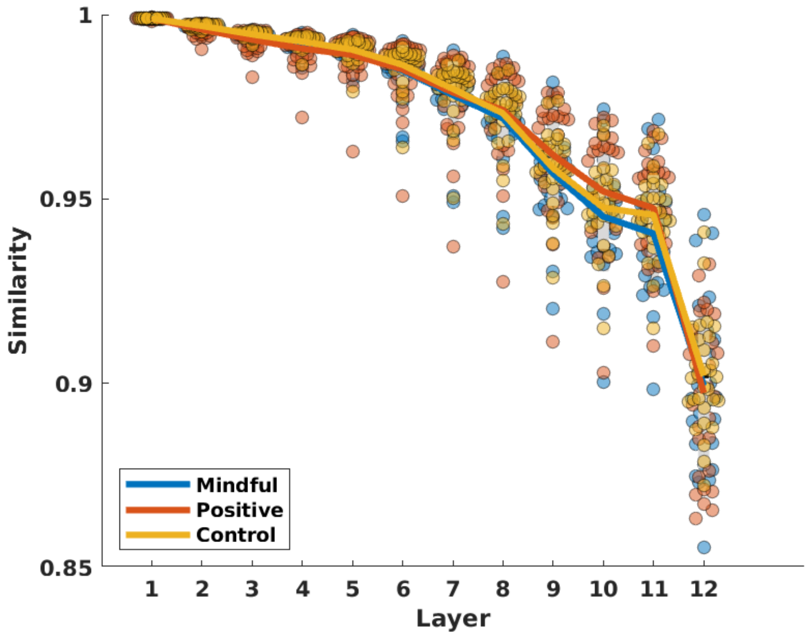


Figure 3 The average similarity for each strategy condition and all situations across layers. Solid lines indicate average similarity per layer for each strategy condition, with colors representing different strategy conditions. Colored dots are subject averages

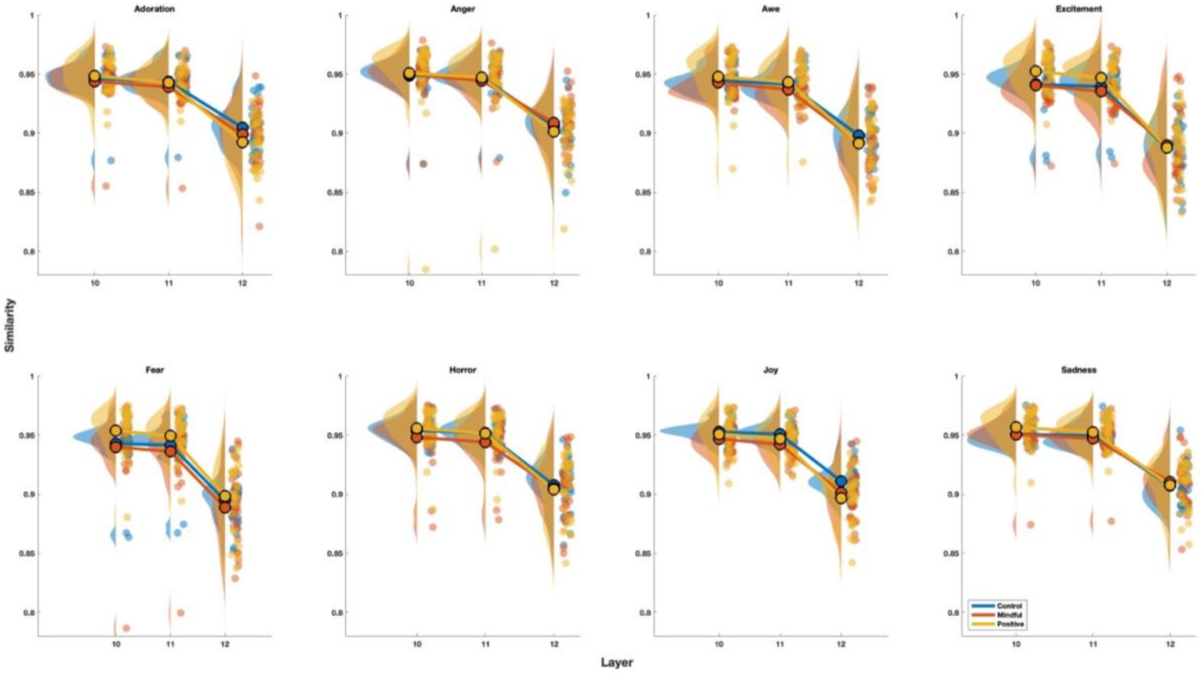


Figure 4 Interplay of strategy condition and emotional situation at the last 3 layers of fine-tuned BERT models. Solid lines indicate average similarity per layer (10 to 12) for each strategy condition, with colors representing different strategy conditions. Colored dots are subject averages for each layer across all situations. Similarity decreases as the model depth increases for all strategy conditions.

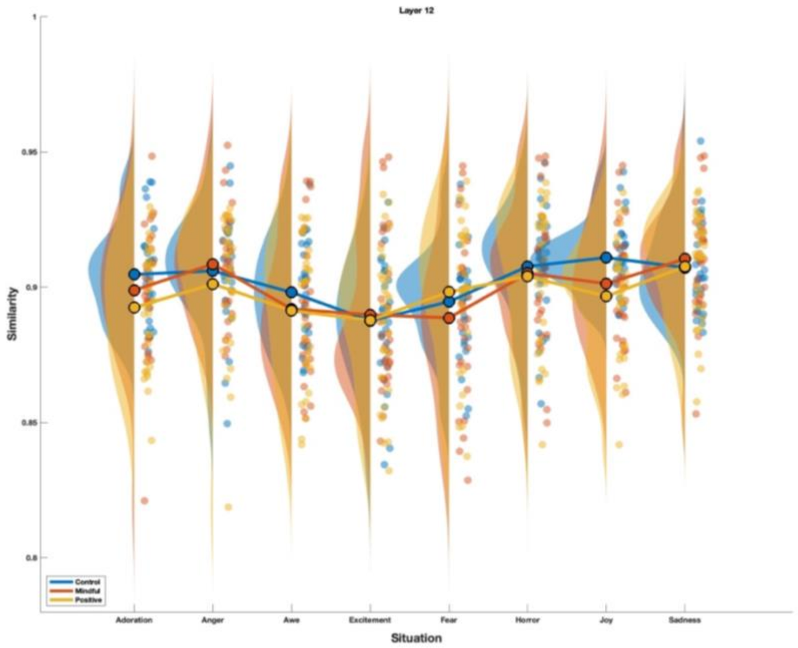


Figure 5 Average similarity of embeddings from layer 12 of pretrained and fine-tuned BERT for each strategy and emotional situation. The colored lines represent the average layer 12 similarity for each situation in each of our three strategy conditions.

We further assessed the interplay of situation and strategy in this our final layer, and found a main effect of situation in ($F_{7,89} = 25.4$, $p < .001$; Figure 5), and an interaction effect between strategy condition and situation ($F_{14,89} = 2.45$, $p <.01$; Figure 5). Post hoc tests revealed that compared to our control condition, the layer 12 embeddings for the Joy emotional category were significantly different for both Positive (Cohen’s D= 0.77 95% CI [0.258, 1.210]; p=.003) and Mindful conditions (Cohen’s D= 0.54 95% CI [0.004, 1.084]; p=.04). The embeddings for the Adoration emotional category were significantly different compared to control embeddings for the positive condition only (Cohen’s D= 0.61 95% CI [0.103, 1.064; p=.019). Only the difference between Positive and Control for the Joy category remained significantly different after correction for multiple comparisons (q=.04). These results suggest that the context of the situation and the regulation strategy are differentially altering meaning of words as indicated by differential mappings of word embeddings in the fine-tuned semantic space, consistent with constructionist theories of emotion positing that the linguistically derived meaning of an emotional situation is dependent on contextual conceptualizations based on internal and external cues.

BERT embeddings predict regulation strategy and self-reported affect in independent studies

Thus far, our findings demonstrate that BERT embeddings capture relations between emotional situations under different regulation strategies in ways that generalize across individuals. However, if regulation systematically changes the meaning of emotional events, we would expect BERT models of regulation strategy to explain variations in self-reported affect to different affective content in varied types of cognitive regulation. To test this hypothesis, we used archival data from two studies using an emotion regulation paradigm with two regulation conditions (Reappraisal and No Regulation) and linguistic descriptions (Nook et al., 2017). We fine-tuned a BERT model trained on text descriptions from our reappraisal and control conditions, and applied the resulting fine-tuned model to the archival data in order to predict regulation strategy from the provided text descriptions. We then compared affect ratings across based on BERT classifications and based on experimental conditions to see their unique contribution to affect ratings.

Our fine-tuned model classified regulation and no regulation in both the original (Binary AUC=.64) and replication study (Binary AUC=.66). Table 2 shows the ROC metrics of the classification model for each study. A repeated measure ANOVA using BERT categories (Regulation and No Regulation) and experimental conditions as factors revealed an interaction between predicted and actual regulation label for both the original ($F_{1,109} = 1106.4$, $p < .001$) and the replication study ($F_{1,109} = 1116.3$, $p < .001$).

Table 2

ROC Metrics for strategy (Regulation vs No Regulation) prediction in original and replication studies

	Sensitivity	Specificity	AUC
Original Study	54.67 95% CI [53.54,55.73]	59.42 95% CI [57.85,60.88]	0.64 95% CI [0.63,0.65]
Replication Study	54.88 95% CI [53.83,55.96]	60.67 95% CI [59.13,62.24]	0.66 95% CI [0.64,0.67]

Post-hoc tests revealed that no regulation trials classified as reappraisal had significantly lower negative affect ratings compared with no regulation trials classified as no regulation in both the original (Robust Cohen's D=-0.801 95% CI [-1.003,-0.613]; t_{118}= -10.52, p < .001), and replication (Robust Cohen's D= -0.752 95% CI [-1.010,-0.541]; t_{118}= -7.80, p < .001) studies. Similarly, reappraisal trials classified as no regulation by our model exhibited significantly higher negative affect ratings compared to with reappraisal trials classified as reappraisal in both the original (Robust Cohen's D= 0.153 95% CI [0.008, .307]; t_{110} = 2.80, p < .01) and replication studies (Robust Cohen's D= 0.267 95% CI [0.138, .459] ; t_{109} = 4.21, p < .001; Figure 6).

These results show that language provides a richer window in the process of regulation than affect ratings alone. Importantly, and in accordance with our overarching hypotheses, they show that the linguistic predictors present in description of events are stable and generalizable to different datasets and stimuli, and that BERT models trained on regulation specific language can use this information to predict the type of strategy used with meaningful consequences for self-reported affect.

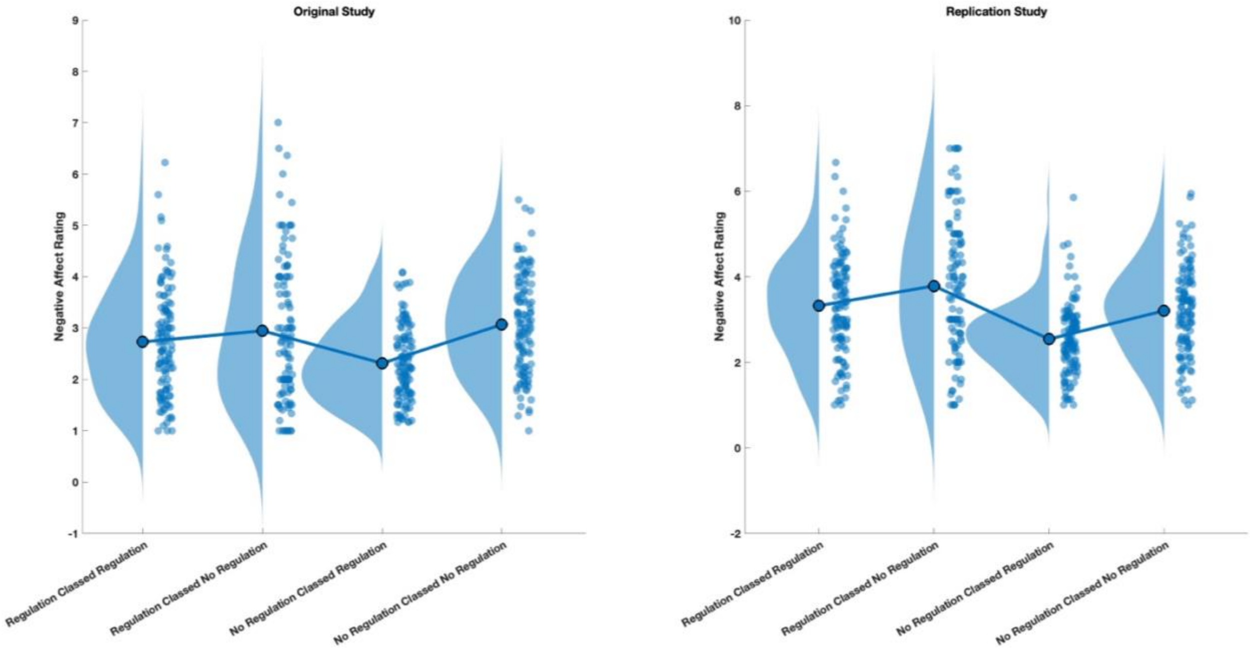


Figure 6 BERT embeddings uniquely predict variation in self-reported negative affect. Our model was sensitive to variation in affect that differed from the a priori labels, such that when our model classified a trial as regulation, even when it occurred during the experimental label of no regulation, negative affect rating was significantly lower, Conversely, when our model classified a trial as no regulation even when it occurred during the experimental label of regulation, negative affect rating was significantly higher.

Discussion

Multiple theoretical perspectives suggest that emotions have distinct positions in these high-dimensional semantic spaces (Cowen & Keltner, 2021; Johnson-Laird & Oatley, 1989). Appraisal theories suggest that core relational themes have deep meaning that is shared across instances of the same emotion, despite having different surface features (Moors et al., 2013). Based on these theoretical perspectives, we predicted that emotional situations should be conveyed in context sensitive word embeddings. Constructionists accounts, on the other hand, suggest language constitutes emotion, and that the meaning of emotion words varies depending on context (Barrett, 2017). In line with this perspective, we hypothesized that fine-tuning BERT using descriptions of emotional events while employing different types of regulation strategies should produce word embeddings that systematically differ from one another.

We leveraged deep language models to test the hypothesis that emotion regulation changes the meaning of events as reflected in a high-dimensional semantic space. We found that language reflects both the type of emotion regulation strategy used as well as the type of the emotional situation experienced. These results lend credence to semantic accounts of emotion, providing impetus to move away from limiting affective research to six basic emotions alone; working from the perspective of a high dimensional semantic space allows researchers to leverage machine learning methods to study the dynamic interplay of emotional processes at ever greater levels of complexity spanning modalities and measurements. These results are also in line with appraisal accounts of emotion, positing that deep-seated conceptualizations remain stable within emotion categories.

We further examined how training in regulation specific language alters representations across layers of a deep language model. We showed that the nature of these changes depends both on the type of emotional situation and the emotion regulation strategy employed. This is consistent with constructionist theories of emotion, denoting that emotion is evoked by a contextual conceptualization of a situation in conjunction with interoceptive cues, wherein the linguistically derived conceptualization is malleable to the context of the situation and regulation strategies. Accordingly, the dependency of the observed changes to the fine-tuned BERT embeddings suggest that regulation is working at a semantic level to alter the meaning of an experienced situation in systematic ways. Importantly, and in line with a wealth of recent literature looking at mappings between deep learning language models and brain activity (Caucheteux et al., 2022; Caucheteux & King, 2022; Goldstein et al., 2022; Kumar et al., 2022;

Russo et al., 2022), we found that the changes to fine-tuned BERT layers increase deeper into a deep language model, with implications for brain mappings of early vs late cortical areas. Furthermore, the interaction between model depth and strategy would suggest varying patterns of brain activity involving multiple regions, and commensurate with specific regulation use. Further work in conjunction with neuroimaging is required to understand how emotion regulation works at a semantic and neural level to alter meaning of a situational experience.

Lastly, we showed that deep language models trained on regulation specific language can generalize to other emotion regulation tasks that incorporate language into their regulation paradigm, and that language gives a richer, more contextualized window into the process of regulation than just affect alone. Importantly, we found that experimental trial labels of a given strategy might not provide the full picture of what type of regulation strategies participants are implementing. This could be particularly true in the case of what is typically termed a ‘no regulation’ condition, in which participants are not instructed to regulate in particular ways, or are instructed against regulating in any particular way. The fact that the subset of no regulation trials that our model is classifying as regulation also have significantly lower negative affect compared to their non-regulation counterparts takes us to the entanglement of emotion regulation and emotion generation, and the recursive automaticity of appraisal vs. reappraisal (Gross & Feldman Barrett, 2011; Zhang et al., 2023). Reacting naturally, or just looking, might not be as devoid of regulation as is commonly assumed, whether implicit or explicit; instead, the root of difference might lie in effortful vs automatic use of regulation strategies. While testing this hypothesis is beyond the scope of our current work, our models present a more holistic approach that combined with other methods such as neuroimaging and Ecological Momentary Assessment (EMA) can aide in delineating what lies at this intersection.

There are several limitations with our work. Our online study had a relatively small training sample, which could limit the scope BERT fine-tuning. Secondly, we only had two regulation strategies, one based on reappraisal and one on mindful acceptance. This could impact the generalizability of these models to different datasets and unconstrained event descriptions such as with daily-diary and EMA methods. Further, two of our chosen strategies – mindfulness and control (i.e. objective description of the experienced event) – share several features, such as a distancing component. While the fact that similar strategies are not so easily discriminated by our model further validates our hypothesis that language reflects regulation, future work could benefit from training language models on a wide range of possible regulation strategies.

Importantly, in order to get closer to shedding light on potential causal mechanisms, future work should incorporate neuroimaging to build both encoding and decoding models that are able to identify the brain processes by which the contextual interplay of emotional situation, regulation strategy as represented by systematic alteration in language are in turn operationalized in the brain. On a computational front, future studies could benefit from looking at not just the embeddings from deep language models, but also other computations performed by transformer models across layers, as well as using models other than BERT that are more biological grounded (Kumar et al., 2022).

In conjunction with neuroimaging methods, the approach presented in this paper can facilitate a deeper understanding of how emotion and emotion regulation are operationalized in the brain. Importantly, this work draws on multiple theoretical accounts of emotions, and these findings provide evidence to differing aspects of these diverse theories using a data-driven approach. In line with older basic emotion views and the more modern semantic space theory of emotion, our results show that language reflects the categorical context of emotional situations, as well as systematic semantic alterations that occur as consequence of regulation, in accordance with appraisal theories. Lastly, and in line with constructionist views of emotion, we found an interaction effect between situational category and regulation strategy, which point to a more malleable mapping between experience and semantic representations of emotion derived from the context of the situation and strategy. Taken together, this work paves the way for objective modeling of emotion regulation using language as a window into the underlying cognitive processes, with applications across a wide range of settings and implications along diverse theoretical frameworks.

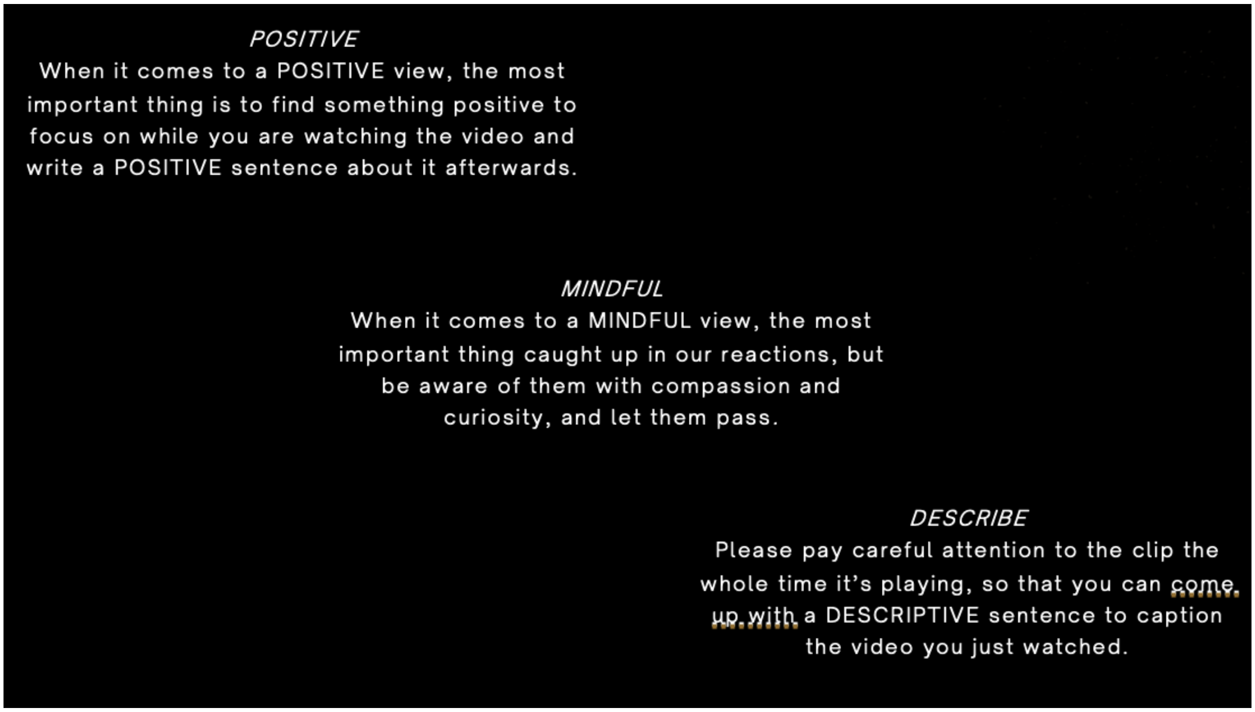
Methods

We used deep language models to see whether language varies systematically as a function of regulation strategy and/or situation, and whether regulation changes the meaning of events as reflected in the high-dimensional semantic space of fine-tuned BERT models.

Online Study

Participants

89 Participants were recruited from Amazon Mechanical Turk, and randomized into one of three conditions: (a) *POSITIVE (30)* or the cognitive reappraisal condition where participants were instructed to find something positive in the situation; (b) *MINDFUL (29)* condition where participants were instructed to notice and accept their thoughts and emotions without getting caught up in them; and (c) *DESCRIBE (30)* or the control condition where participants were instructed to passively view affective clips. Figure 7 shows snapshots of the instructions in each strategy condition.



POSITIVE
When it comes to a POSITIVE view, the most important thing is to find something positive to focus on while you are watching the video and write a POSITIVE sentence about it afterwards.

MINDFUL
When it comes to a MINDFUL view, the most important thing caught up in our reactions, but be aware of them with compassion and curiosity, and let them pass.

DESCRIBE
Please pay careful attention to the clip the whole time it's playing, so that you can come up with a DESCRIPTIVE sentence to caption the video you just watched.

Figure 2 Snapshots of regulation instructions in each strategy condition.

Stimuli

Affective clips were selected from a normed database of emotionally evocative short video clips (Cowen & Keltner, 2017). We chose clips within the extremes of positive and negative valence across 8 categories of emotional situations (Anger, Fear, Horror, Sadness, Adoration, Awe, Excitement, Joy).

Procedure

Participants received instruction in their assigned emotion regulation strategy condition and were told to apply that strategy for the entire duration of each clip. Post clip, participants were prompted to provide a written description of the content of each video. Emotion induction was assessed using a continuous measure of affect and an unconstrained text description of feelings. All participants finished a short training before proceeding to the main task. Figure 8. shows a schematic of the task paradigm.

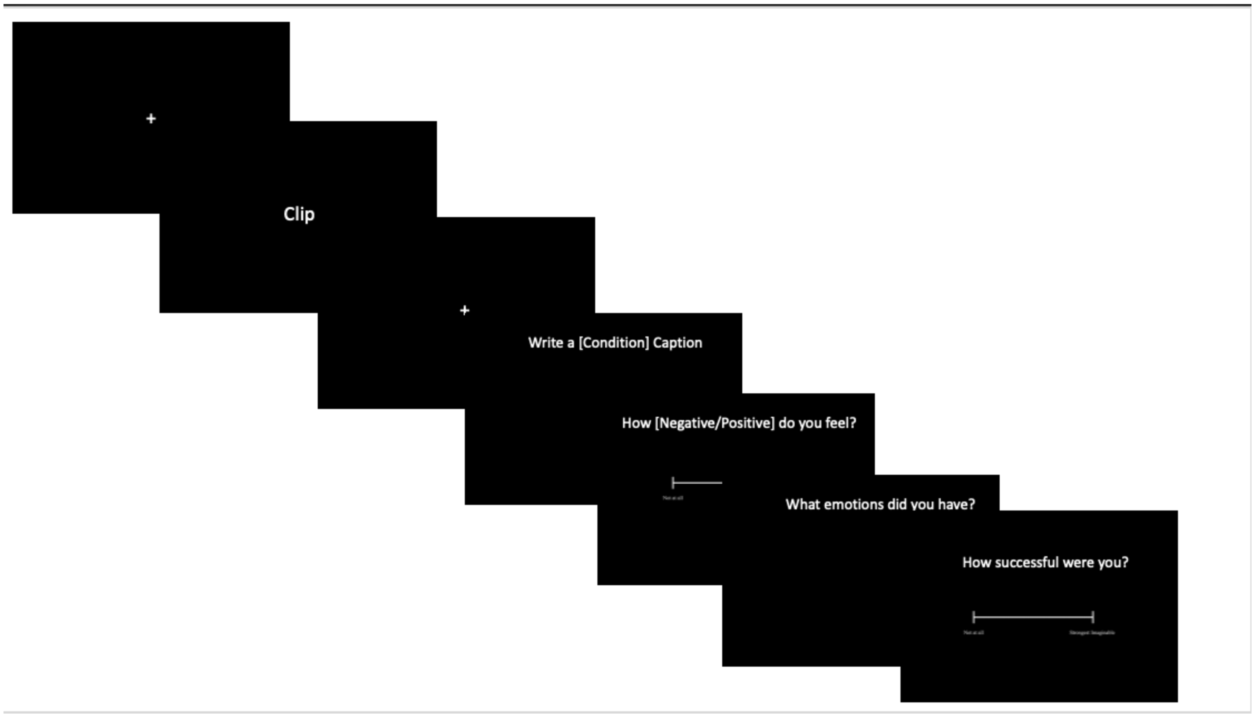

+
Clip
+
Write a [Condition] Caption
How [Negative/Positive] do you feel?
Not at all
What emotions did you have?
How successful were you?
Not at all
Strongest Imaginable

Figure 3 Experimental Paradigm for Emotion Regulation in Context (ERiC) online study.

Analysis Plan

Situation and Strategy Classification

In order to assess whether the language produced when describing emotional event differed systematically as a function of emotion regulation strategy employed or the context of the situation itself, we extracted pre-trained BERT embeddings for each caption. We used Partial least squares discriminant analysis (PLS-DA) to build two separate decoders, going from caption embeddings to regulation strategy (3-way classification) and emotional situation (8-way classification) separately. Each decoder was trained using 5-fold cross validation to predict strategy and situation respectively. Performance was assessed using Area under the Receiver Operating Characteristic curve (AUC), as well as sensitivity and specificity measures with cross validated confidence intervals.

Out of Sample Model Generalization

To test generalizability and applicability of our strategy classification model, we selected archival data from a study using an emotion regulation paradigm with a language component (Nook et al., 2017). This study asked participants to view affective images and provide text descriptions after implementing cognitive reappraisal (POSITIVE) or look (CONTROL) strategies. Accordingly, we fine-tuned a BERT model trained on text descriptions from our POSITIVE and CONTROL conditions, and applied our fine-tuned model to the archival data in order to predict regulation strategy from the provided text descriptions. We then used a repeated measure ANOVA to compare how reported affect differed across model strategy classification based on language and actual strategy label based on the experiment.

Effects of Model Training on Regulation Specific Language

We were interested not only in building generalizable models capable of classifying regulation strategy and/or emotional situation, but also understanding how regulation changes word representations and alters word meanings in a high-dimensional semantic space, and how the interplay of emotional situation and regulation strategy impact that change in meaning. Accordingly, we first fine-tuned three BERT models, each trained to predict emotional situation from captions that corresponded to one regulation strategy condition. To see how training in regulation specific language altered BERT representations, we calculated the similarity of embeddings from different layers of the fine-tuned models with corresponding layers in the pre-trained counterpart (defined as the correlation between two embedding vectors in semantic space). We then assessed how these semantic representations for each caption differed in terms of regulation strategy, network layer, and situational context using a linear mixed effects model.

References

Arana, S., Lerousseau, J. P., & Hagoort, P. (2023). Deep Learning Models to Study Sentence Comprehension in the Human Brain. *arXiv preprint arXiv:2301.06340*.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (2016). *Handbook of emotions*. Guilford Publications.

Berking, M., & Wupperman, P. (2012). Emotion regulation and mental health: recent findings, current challenges, and future directions. *Current opinion in psychiatry*, *25*(2), 128-134.

Biggiogera, J., Boateng, G., Hilpert, P., Vowels, M., Bodenmann, G., Neysari, M., Nussbeck, F., & Kowatsch, T. (2021). BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. *arXiv preprint arXiv:2106.01536*.

Brooks, J. A., Shablack, H., Gendron, M., Satpute, A. B., Parrish, M. H., & Lindquist, K. A. (2017). The role of language in the experience and perception of emotion: A neuroimaging meta-analysis. *Social cognitive and affective neuroscience*, *12*(2), 169-183.

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, *12*(1), 16327.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Chang, T.-Y., & Chen, Y.-N. (2019). What does this word mean? explaining contextualized embeddings with natural language definition. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),

Chervonsky, E., & Hunt, C. (2019). Emotion regulation, mental health, and social wellbeing in a young adolescent sample: A concurrent and longitudinal investigation. *Emotion*, *19*(2), 270.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900-E7909.

Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in cognitive sciences*, *25*(2), 124-136.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Durrani, N., Sajjad, H., Dalvi, F., & Alam, F. (2022). On the Transformation of Latent Space in Fine-Tuned NLP Models. *arXiv preprint arXiv:2210.12696*.

Gendron, M., Lindquist, K. A., Barsalou, L., & Barrett, L. F. (2012). Emotion words shape emotion percepts. *Emotion*, *12*(2), 314.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., & Cohen, A. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369-380.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, *6*(7), 975-987.

Gross, J. J. (2015). Emotion Regulation: Current Status and Future Prospects. *Psychological inquiry*, *26*(1), 1-26. https://doi.org/10.1080/1047840X.2014.940781

Gross, J. J., & Feldman Barrett, L. (2011). Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*, *3*(1), 8-16.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on psychological science*, *14*(6), 1006-1033.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*, 1744-1756.

Hu, T., Zhang, D., Wang, J., Mistry, R., Ran, G., & Wang, X. (2014). Relation between emotion regulation and mental health: a meta-analysis review. *Psychological reports*, *114*(2), 341-362.

Inwood, E., & Ferrari, M. (2018). Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: A systematic review. *Applied Psychology: Health and Well-Being*, *10*(2), 215-235.

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81-123.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, 2022.2006. 2008.495348.

Lindquist, K. A., Satpute, A. B., & Gendron, M. (2015). Does language do more than communicate emotion? *Current Directions in Psychological Science*, *24*(2), 99-108.

Lopez, W., Merlino, J., & Rodriguez-Bocca, P. (2020). Learning semantic information from Internet Domain Names using word embeddings. *Engineering Applications of Artificial Intelligence*, *94*, 103823.

McDonald, S., & Ramscar, M. (2001). Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. Proceedings of the Annual Meeting of the Cognitive Science Society,

Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*.

Miaschi, A., & Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. Proceedings of the 5th Workshop on Representation Learning for NLP,

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119-124.

Nook, E. C., Satpute, A. B., & Ochsner, K. N. (2021). Emotion naming impedes both cognitive reappraisal and mindful acceptance strategies of emotion regulation. *Affective Science*, *2*(2), 187-198.

Nook, E. C., Schleider, J. L., & Somerville, L. H. (2017). A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology: General*, *146*(3), 337.

Orvell, A., Ayduk, Ö., Moser, J. S., Gelman, S. A., & Kross, E. (2019). Linguistic shifts: A relatively effortless route to emotion regulation? *Current Directions in Psychological Science*, *28*(6), 567-573.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140-157.

Russo, A. G., Ciarlo, A., Ponticorvo, S., Di Salle, F., Tedeschi, G., & Esposito, F. (2022). Explaining neural activity in human listeners with deep learning via natural language processing of narrative text. *Scientific reports*, *12*(1), 17838.

Satpute, A. B., & Lindquist, K. A. (2021). At the neural intersection between language and emotion. *Affective Science*, *2*(2), 207-220.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Schuster, T., Ram, O., Barzilay, R., & Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18,

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods*, 1-14.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

Wu, J. M., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2020). Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*.

Zhang, J.-X., Dixon, M. L., Goldin, P. R., Spiegel, D., & Gross, J. (2023). Testing the neural separability of emotion reactivity and regulation.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020). Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987*.

Zhou, Y., & Srikumar, V. (2021). A closer look at how fine-tuning changes BERT. *arXiv preprint arXiv:2106.14282*.