*Original Article*

Sociological Methodology
2025, Vol. 55(1) 25–58

DOI: 10.1177/00811750241260729
http://sm.sagepub.com

# Contextual Embeddings in Sociological Research: Expanding the Analysis of Sentiment and Social Dynamics

**Moeen Mostafavi[1], Michael D. Porter[2], and Dawn T. Robinson[3]**

## Abstract

The authors introduce BERTNN (Bidirectional Encoder Representations from Transformers Neural Network), a novel methodology designed to expand affective lexicons, a critical component in sociological research. BERTNN estimates the affective meanings and their distribution for new concepts, bypassing the need for extensive surveys by leveraging their contextual usage in language. The cornerstone of BERTNN is the use of nuanced word embeddings from Bidirectional Encoder Representations from Transformers. BERTNN uniquely encodes words within the framework of synthesized social event sentences, preserving their meaning across actor-behavior-object positions. The model is fine-tuned on the basis of the implied sentiment changes, providing a more refined estimation of affective meanings. BERTNN outperforms previous approaches, setting a new standard in deriving multidimensional affective meanings for novel concepts. It efficiently replicates sentiment ratings that traditionally require extensive survey hours, demonstrating the power of automated modeling in sociological research. The expanded affective lexicons that can be produced with BERTNN cater to shifting cultural meanings and diverse subgroups, demonstrating the potential of computational linguistics to enrich the measurement tools in sociological research. This article underscores the novelty and significance of BERTNN in the broader context of sociological methodology.

## Keywords

affect control theory, sentiment mining, sentence embedding, computational social science, affective lexicon, Bidirectional Encoder Representations from Transformers, BERT, contextual word embeddings

The challenge of inferring the subtle affective meanings of new concepts is pivotal in both sociological research and the broader domain of sentiment analysis. Whereas sentiment analysis traditionally focuses on evaluating textual data to determine its polarity (good vs. bad), sociological research, particularly in the context of affect control

[1]Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA
[2]Systems and Information Engineering & Data Science, University of Virginia, Charlottesville, VA, USA
[3]Department of Sociology, University of Georgia, Athens, GA, USA

**Corresponding Author:**
Moeen Mostafavi, University of Virginia, Department of Systems and Information Engineering, 351 McCormick Road, Thornton Hall, Room D221, Charlottesville, VA, 22901, USA
Email:moeen@virginia.edu

theory, demands a deeper understanding that encompasses additional dimensions. Affect control theory uses affective dictionaries to map concepts onto a three-dimensional affective space known as the EPA dimensions: evaluation (good vs. bad), potency (powerful vs. powerless), and activity (active vs. passive), as introduced by Osgood, Suci, and Tannenbaum (1957).

The application of the three-dimensional EPA model significantly enriches social science research. Unlike traditional sentiment analysis that primarily focuses on the evaluative dimension, this approach incorporates potency and activity dimensions to offer deeper insights. In political science, for instance, this triadic approach could extend beyond simple approval or disapproval metrics to capture the intensity and dynamism of public opinion. In organizational behavior, it can aid in understanding workplace dynamics, not just in terms of employee satisfaction (evaluation) but also in perceived empowerment (potency) and engagement (activity). Cultural studies stand to benefit by gaining a nuanced understanding of cultural artifacts and practices, where the EPA model can elucidate complex emotional and social underpinnings. For sociology, this method can offer a more holistic view of societal trends and group behaviors, capturing the subtle interplay of emotions, power dynamics, and action tendencies within social structures. Overall, the expansion of sentiment analysis to include potency and activity dimensions can vastly enhance the depth and breadth of research across social sciences, providing a more robust, three-dimensional representation of social concepts and interactions. This approach aligns well with the evolving needs of diverse research fields, encouraging interdisciplinary collaboration and offering a more comprehensive toolkit for social scientists, even those who may not use affect control theory as their primary framework.

This multidimensional approach transcends the conventional binary evaluation of sentiment analysis, offering a more nuanced understanding of emotional significance in various sociological contexts, from interpersonal dynamics to broader societal trends. However, the traditional methodology, heavily reliant on extensive surveys, poses significant challenges in terms of time and adaptability, often struggling to capture the nuanced meanings of words in different contexts. Such limitations underscore the necessity for more sophisticated approaches that can efficiently encompass the richness of affective meanings in sociological theories across diverse settings. The introduction of computational techniques has opened new avenues for addressing this challenge (Alhothali and Hoey 2017; Li et al. 2017; Mostafavi and Porter 2021). In particular, use of contextual language to estimate affective meanings of new concepts has shown promise. However, traditional methods, often referred to as shallow word embeddings, exhibit several limitations.

Shallow word embeddings, by design, represent each word as a stand-alone, static vector. This singular representation does not change on the basis of the word's usage or context. As a result, these methods overlook the importance of context in understanding language. They fail to capture the nuanced shifts in meaning that words can take in different contexts. This inability to adapt words' representations on the basis of context also causes issues for polysemy. *Polysemy* refers to words that have multiple potential meanings depending on context. For instance, the word "bank" could refer to a financial institution or the side of a river. Shallow embeddings have no mechanism for distinguishing between these different meanings. Using a single static vector for polysemous words thus introduces inaccuracies when estimating affective meanings.

Additionally, shallow embeddings struggle with capturing complex semantic relationships. Although they can handle some basic relationships, such as analogies (e.g., king is to queen as man is to woman), they often stumble when faced with more intricate relationships, such as understanding nuanced differences between synonyms (e.g., "slim" vs. "skinny"). The limitations of shallow embeddings can result in imprecise interpretations of affective meanings. One significant challenge is their handling of words not seen during training. This is particularly problematic when trying to infer the affective meanings of less common or novel terms. Additionally, these methods do not account for the position of a word within a sentence, which is essential for grasping context-dependent meaning. For example, the word "judge" can convey different sentiments depending on whether it is used as a noun or a verb.

Deep embedding methods such as Bidirectional Encoder Representations from Transformers (BERT) represent a significant advancement in the field of language modeling, offering a more nuanced approach to contextual embeddings (Devlin et al. 2018). Unlike shallow word embeddings, BERT generates dynamic word embeddings, meaning the representation of a word changes on the basis of its context. This capability allows BERT to capture the nuanced meanings of words on the basis of their usage in a sentence. Furthermore, BERT's contextual understanding enables it to handle polysemy effectively. It can differentiate between different meanings of a word on the basis of its context.

BERT also excels in capturing deep semantic relationships between words. It considers all words in a sentence simultaneously, rather than looking at them in isolation or in a linear sequence. This approach allows BERT to understand complex relationships and dependencies between words. Additionally, BERT uses a technique called WordPiece tokenization, which breaks words into smaller subwords (Devlin et al. 2018). This technique enables BERT to handle rare words or words not seen during training, as it can understand these words in terms of their subword components. BERT also considers the position of a word within a sentence, which is crucial for understanding its meaning. It does this through the use of positional encodings, which add information about the position of a word in a sentence to its embedding (Vaswani et al. 2017).

In the context of inferring affective meanings of new concepts, BERT's capabilities make it a powerful tool. A recent study by van Loon and Freese (2023) made significant strides in using BERT for inferring affective meanings. However, their approach also exhibits certain limitations that our BERTNN algorithm addresses.[1] Their method effectively leverages context to understand the part of speech a word is used in, but it does not fully capture the broader social interaction context in which the concept is situated. Also, their method excludes compound words or words not present in the BERT vocabulary, and it focuses on extracting meanings for word "roots." This method, despite extracting context embeddings for each root, does not optimally use the broader contextual information that is crucial for understanding social interactions.

To address all these challenges, we present the BERTNN algorithm, an innovative methodology that leverages the power of contextual usage in language to efficiently estimate the affective meanings of new concepts. BERTNN is designed to construct contextual aspects in a way that captures the essence of social interactions, offering a more holistic and nuanced understanding of concepts within their societal context. This approach goes beyond simply recognizing the grammatical role of a word, delving into the deeper layers of meaning as shaped by the interplay of social dynamics. BERTNN harnesses the bidirectional transformer BERT to generate nuanced word embeddings sensitive to words' position within synthesized social event sentences. This unique encoding allows the model to capture actor-behavior-object (ABO) meaning variations, and through fine-tuning on the basis of culturally implied sentiment changes, BERTNN provides refined estimates of affective meanings.

Our work with the BERTNN algorithm introduces several innovative techniques to improve the inference of affective meanings. One key innovation is the fine-tuning of BERT on synthetic social events. By engaging with work from both social psychology and sociology, we aim to bridge the gap and highlight what might be overlooked solely from a computer-science/natural language processing perspective. This interdisciplinary approach allows a richer understanding and application of language models in social-science tasks. Instead of using BERT's pretrained model as is, we adapt it to our specific task by training it on a range of synthesized social interactions. This approach ensures the model captures the nuances of words' meanings on the basis of their grammatical positions within these events, effectively leveraging BERT's contextual understanding.

Furthermore, we aggregate the context-dependent embeddings generated by BERT to estimate the affective meanings of new concepts. This means the affective meaning of a concept is derived from its usage across multiple contexts. By aggregating these context-dependent embeddings, BERTNN captures a more comprehensive and nuanced understanding of the concept's affective meaning.

These innovations allow BERTNN to outperform previous methods in inferring the affective meanings of new concepts. BERTNN efficiently replicates sentiment ratings that traditionally rely on extensive sentiment surveys, demonstrating the power of automated modeling in sociological research. Those surveys, detailed in Heise's (2010) *Surveying Cultures*, involve collecting affective responses from participants to build a comprehensive understanding of societal sentiments.[2]

BERTNN signifies a notable progression in sociological research methodologies, facilitating enhanced application of affect-based sociological theories within the dynamic landscape of cultural contexts. By significantly broadening the scope of fundamental sentiment lexicons, this innovative computational approach revitalizes the toolkit for measuring and analyzing societal sentiments. BERTNN stands out not only for its efficiency in replicating traditionally labor-intensive sentiment ratings, but also for its adaptability to the nuanced shifts in cultural meanings. BERTNN, using artificial intelligence (AI) technology, is set to significantly contribute to the standards of sociological methodology, enriching contextual and computational analysis with its novel approach to interpreting societal sentiments.

Although large neural language models such as BERT demonstrate capabilities in manipulating linguistic form, it is important to clarify that they do not possess human-like language understanding or learning abilities (Bender et al. 2021). This distinction should be kept in mind throughout discussions of model performance.

## BACKGROUND

Consider talking to your mentor for some advice about how to behave with your colleague. Your mentor starts by asking you questions about the *culture* in the workspace and may continue asking about the *identity* of your colleague. These questions might elicit responses about *institutional constraints*, such as being the manager, or they may be about *dispositional characteristics*, such as being nice or active. On the basis of this information, your mentor may offer some initial recommendations, but you *adapt* your behavior after observing reactions from the colleague. This is a descriptive scenario for daily interaction. Affect control theory is a sociological theory that formalizes the process described in this scenario. This formal theory is both grounded and generative.

The empirically "grounded" part of this theory historically rests on research-intensive data collection methods. Affect control theory offers insights into understanding emotional changes during various social interactions, but its real-world applications have been constrained by the limited vocabulary size of affective dictionaries. Recent computational advancements provide avenues to efficiently expand this grounding, facilitating quicker adaptation to evolving cultures and cultural shifts. Building on these recent innovations, we introduce a novel methodology that not only enhances accuracy but also broadens the scope to encompass new concepts.

## AFFECT CONTROL THEORY

Affect control theory is a mathematical model of culture-based action that uses a set of event-processing equations to describe how affective meanings shift during social interactions. The theory uses culturally grounded model specifications to predict behaviors, emotions, and new cultural meanings that arise from local interactions. The theory has three core components: a scheme for representing *cultural sentiments*, a system of *event-processing* equations that characterize cultural rules, and a *control system* logic that animates the meaning-preserving assumption of the theory and makes the theory generative.

### *Cultural Sentiments*

The theory is grounded in the representation of cultural sentiments within a three-dimensional affective space, often referred to as EPA space (Heise 1977). The EPA space was originally introduced by Osgood et al. (1957). In the context of affect control theory, this space is used to systematically index various social concepts, including *identities*, *behaviors*, *emotions*, *settings*, and more. By doing so, it offers a unified metric to understand and compare these concepts in terms of their affective meanings.

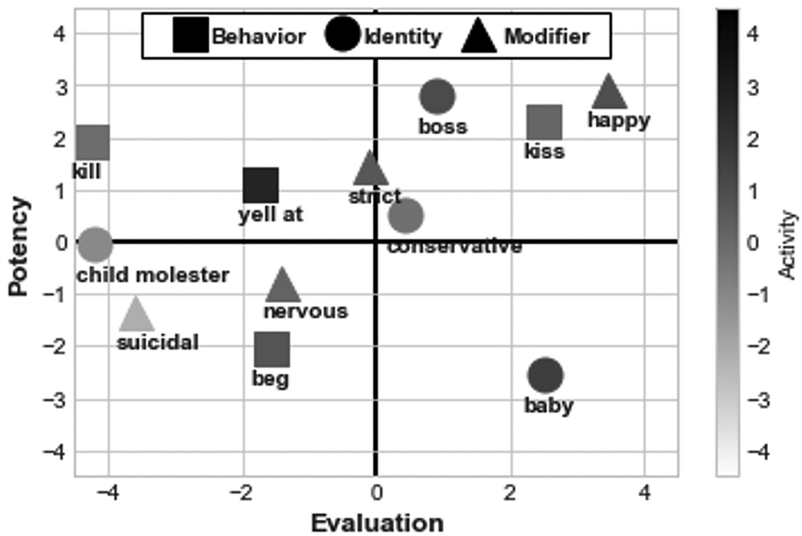


**Figure 1.** Sentiments of sample concepts in evaluation, potency, and activity space (Robinson et al. 2016).

*Note*: Color represents the activity dimension. Circles, squares, and triangles indicate identities, behaviors, and modifiers.

The EPA dimensions have been instrumental in capturing significant variations in the affective interpretations of lexicons across more than 20 national cultures (Osgood, May, and Miron 1975; Osgood et al. 1957). Further supporting the significance of EPA dimensions, Fontaine et al. (2007) demonstrated that when dimensionality is reduced on 144 features representing core components of emotions, EPA scores emerge as the first three principal components.

Affect control theory considers social interactions or events that involve an *actor* who *behaves* toward an object. Extracting the ABO components of an event is the first step in modeling interactions (Heise 2010). The *actor* and *object* are *identities* represented in the affective space, and the *behavior* connects them in the social event. Modifiers can be used to enrich the descriptions.

Figure 1 depicts U.S. cultural sentiments (Robinson et al. 2016) toward 12 concepts in EPA space. In this plot, we see that "suicidal" and "nervous" both have bad, powerless, and passive meanings, but "suicidal" is more negative in all three dimensions. On the other hand, "happy" is a pleasant, powerful, and mildly active concept. EPA ratings are typically scored on a scale of −4.3 to 4.3. This range was chosen to capture the full range of affective meanings people can experience, while still allowing a relatively fine-grained measurement of affective meaning (Heise 1979).[3]

## *Impression Change Equation*

In affect control theory, people interpret elements of their interactions (events) through labels, and these labels evoke specific meanings as indexed in EPA sentiment dictionaries. Take, for instance, an observed interaction in which "a bossy employer argues with an employee." This event leads observers to modify their affective meanings or impressions of both the *actor* (employer) and the *object* (employee) in the interaction. Specifically, the employer, as the *actor* in this scenario, may be perceived as more powerful after the interaction because of the dominant behavior of arguing, thus increasing their *potency* rating. In contrast, the employee, as the *object* of this action, may be seen as more powerless, decreasing their *potency* rating. Additionally, the act of arguing may lead to a perception of both the employer and the employee as less pleasant, thereby lowering their *evaluation* ratings. These transient meanings, shaped by the context of the event and the participants' roles, are termed "impressions" in affect control theory. Impressions are essentially contextualized affective meanings that are evoked by symbolic labels (e.g., "bossy employer" or "employee") within specific social events.

Impression change equations predict how sentiments combine to form impressions. These equations are grounded with data from respondents within a language culture. The core impression change equations describe event dynamics from basic interactions of the form "the actor behaves toward the object-person" (called ABO events). Affect control theory also has impression change equations to predict emotional responses to events (Averett and Heise 1987), interactions between emotions and identities on situational meanings (Heise and Thomas 1989; Smith 2002), nonverbal behaviors (Rashotte 2002), self-directed behaviors (Britt and Heise 1992; Smith and Francis 2005), and social settings (Smith 2002; Smith-Lovin 1979).

### *Control System*

If the *actor* behaves as expected, then the *impression* of their identity will not change far from baseline, but if the *actor* does something unexpected, then a large change from the baseline is expected. *Deflection* is the squared Euclidean distance between the baseline sentiments of ABO characters and their impressions following an event. If the impression of an ABO event is close to the initial sentiment, *deflection* is small; it grows bigger as the impression of the event drifts from the initial sentiment. Affect control theory suggests that minimizing *deflection* is the driving force in human activity. Highly deflecting events create social and physiological stress (Goldstein 1989). For example, if a grandmother fights with her grandchild, they both feel stressed. One side is likely to take an action, such as apologizing, to bring the impression of their identities back to where they view themselves in society.

The impression change equations are designed to uphold the *affect control principle*, which assumes individuals act to preserve or reinstate the cultural affective meanings linked to activated labels. These equations take EPA sentiment profiles as inputs and produce impressions as outputs. When there is a discrepancy between sentiments and impressions, it indicates how closely interactions align with cultural norms. The control system in affect control theory models the assumption that individuals strive to maintain their cultural understanding of situations. After a disruptive event, solving for the behavior profile can predict the expected response to rectify the situation or provide a new normative understanding of the observed events.

Heise's (1979) INTERACT software can simulate interactions. It uses impression change equations to solve for behaviors that minimize the deflection or predict attributes and emotions during the interaction (Heise 2007, 2013, 2020).[4] This software offers researchers a practical and rigorous way to apply the underlying theory without requiring in-depth technical expertise. INTERACT's natural language input feature enhances its accessibility, allowing theory application in a reproducible manner without the need to rederive complex mathematical equations. Consider the following set of events and interactions that we simulate using INTERACT: (1) employee greets bossy employer, (2) bossy employer asks employee, (3) employee replies to bossy employer, (4) bossy employer argues with employee, and (5) employee listens to or disobeys bossy employer.

To illustrate the dynamics of social interactions and their effect on perceived identities, we concentrate on the evaluation dimension in a sequence of interactions between an employee and a bossy employer. This focus allows clearer insight into how each action influences the evaluation impressions of the involved characters.

First, the employer and employee have baseline evaluation values of 1.49 and 1.03, respectively, with the "bossy" attribute adjusting the employer's evaluation to −1.41. During the first interaction, "employee greets bossy employer," the employer's evaluation positively shifts to 1.30, reflecting a more favorable perception due to the employee's friendly gesture, and the employee's evaluation increases to 1.22.

Second, as the sequence continues with "bossy employer asks employee," the employer's evaluation experiences another positive change, this time increasing to 3.41, suggesting a further improvement in perception despite the bossy characteristic. The employee's evaluation also adjusts to 0.93, indicating a slight decrease but still maintaining a positive perception overall.

Third, in the subsequent action, "employee replies to bossy employer," the employer's evaluation remains relatively stable, changing marginally from 3.41 to 3.44. This minor shift indicates the employer's perception is not significantly altered by the employee's cooperative behavior. The employee's evaluation rises to 1.64, reflecting a more positive perception due to the response to the employer.

Fourth, in the penultimate interaction, "bossy employer argues with employee," the employer's evaluation drops further to −0.04, marking a significant shift toward a negative perception, attributable to the confrontational behavior. Simultaneously, the employee's evaluation decreases to 0.40, indicating a more neutral or slightly negative perception due to being involved in an argument.

Finally, in "employee listens to or disobeys bossy employer," the employer's evaluation slightly recovers to 1.28 or 0.30, suggesting a modest improvement in perception. The employee's evaluation adjusts to 0.88 or −0.44, showing a rise or fall in positive perception.

Figure 2 depicts these events in all EPA dimensions, enhancing comprehension of how each interaction influences the perception of the actor and object in the social event, in line with affect control theory principles. The sequential interactions

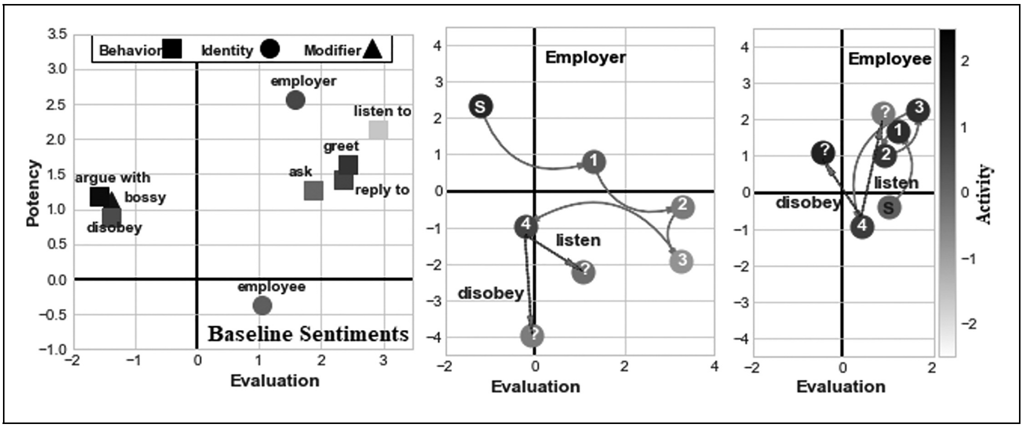


**Figure 2.** Simulating sequential events in an interaction between an employee and an employer.
*Note*: The initial sentiments for both characters are shown by *s*; the impression after each of four events are shown by numbers. After the fourth event, on the basis of what the employee does, the final sentiment for the two characters would be one of the places shown by the question mark.

discussed here are similar to our mentorship example discussed earlier. These examples show how understanding the interaction dynamics can help predict the consequences of behaviors.

The dynamic nature of social interactions plays a crucial role in shaping the affective meanings of identities, behaviors, and modifiers in affect control theory. Unlike static mappings, perception of these concepts in social events is influenced by the context and sequence of interactions. This variance is not only pivotal in understanding the fluidity of social norms, but it also reflects participants' individual experiences and interpretations. Personal experiences can significantly influence how concepts are perceived and rated in terms of the EPA dimensions. Therefore, it is essential to consider the distributional aspects of these affective meanings, acknowledging that perceptions are not deterministic but shaped by the ebb and flow of social dynamics. This understanding underscores the importance of capturing the essence of social interactions in computational models, such as BERTNN, which aims to reflect the rich, context-dependent nature of social meanings in a society.

## *Empirical Testing and Application*

Affect control theory was introduced in the 1970s (Heise 1977). Validations and applications of the theory appear in more than 100 published articles, book chapters, and books (MacKinnon and Robinson 2014; Robinson and Smith-Lovin 2018). Observational studies have revealed dynamics predicted by affect control theory equations (e.g., Francis 1997a, 1997b; Hunt 2008). Experimental studies have validated affect control theory predictions about emotional experiences (e.g., Robinson and Smith-Lovin 1992; Robinson et al. 2012), identity attributions (e.g., Heise and Calhan

1995; Robinson and Smith-Lovin 1999; Robinson, Smith-Lovin, and Tsoudis 1994; Tsoudis and Smith-Lovin 1998), and social behavior and perform(Robinson and Smith-Lovin 1992; Schröder and Scholl 2009; Youngreen et al. 2009). Survey studies have validated the theory's predictions about subcultural variation (e.g., Smith-Lovin and Douglass 1992).

Affect control theory has been used in interdisciplinary applications such as human-computer interactions (Robillard and Hoey 2018), finding how language cultures affect social response (Kriegel et al. 2017), and modeling identities and behaviors within groups (Rogers and Smith-Lovin 2019). More recently, Mostafavi (2021) proposed using affect control theory to model emotional shifts during online conversations. This approach could allow chatbots to estimate customer sentiment trajectories on the basis of messaging interactions. The chatbot could then adjust its responses accordingly, without necessarily achieving human-level "understanding" of the conversation semantics.

### *Empirical Grounding*

To formulate the process mathematically, we briefly review the quantification process using surveys. The first step is quantifying the *sentiments* that are introduced as *identity*, *behavior*, and *modifier*. For this purpose, at least 25 participants rate words of interest in EPA space (Heise 2010). In the survey, participants rate how they feel about an *identity*, *behavior*, or *modifier* such as "employer."

The sentiment ratings for each concept are aggregated and indexed in a sentiment "dictionary" for each language culture. These supply the baseline meanings for the affect control theory based model. The next step is identifying the event-processing dynamics that lead to social impressions (Robinson and Smith-Lovin 1999). For this purpose, research participants rate ABO characters again after observing a set of events. For example, participants rate the affective meaning of "employee," "greet," and "employer" after observing "the employee greets the employer." As discussed earlier, the ratings of ABO (impressions) could be different from the initial baselines (sentiments). Affect control theory uses nonlinear regression models, known as impression change equations, to estimate these changes (Heise 2013). Modifiers can be incorporated prior to impression change by changing the baseline values or sentiments (e.g., bossy employer) (Averett and Heise 1987).

### *Challenges*

Current state-of-the-art practice for generating affect control theory models for new cultures require (1) assessing the meanings of 1,000 to 2,000 commonly used labels for social event descriptors (*behaviors*, *identities*, *settings*, *emotions*) in three dimensions of meaning (*evaluation*, *potency*, and *activity*) and (2) conducting a 512-condition (partial repeated measures) survey experiment to generate data for the event-processing models (Heise 2010; Kriegel et al. 2017; Rogers 2018). These techniques are relatively efficient: they require input from only a few thousand people to produce robust, generative models capable of predicting millions of events. The impression change equations are relatively stable across time, but the sentiment dictionaries vary more in response to social change and by subculture (Heise 2010; MacKinnon and Luke 2002).

Updating sentiment dictionaries is thus a relatively efficient way to tune the theory to a new social context or time. Nonetheless, collecting a new sentiment dictionary in advance of every investigation into a new substantive domain reduces the theory's utility. To compensate for measurement error between respondents, most EPA surveys are designed so that each word is scored by at least 25 different participants (culture experts). Finding the affective meaning for 5,000 words thus requires over 125,000 ratings and 400 hours of respondent time (Heise 2010). Because of the high cost and time required, most EPA data collections have been limited to relatively small (1,000–2,000 words) dictionaries, which has restricted the applicability of affect control theory tools.

### *Toward a Computational Solution*

An alternative to conducting surveys for a large set of new concepts is expanding the current dictionaries to include those concepts. Researchers have tried different methods to build affective lexicons, which can be broadly categorized into supervised (Li et al. 2017; Mostafavi and Porter 2021) and semisupervised methods (Alhothali and Hoey 2017). In supervised methods, the model is trained on a data set where the correct outcomes are known, and it learns to predict these outcomes for new data. Semisupervised methods use a combination of labeled and unlabeled data for training, which can be beneficial when the amount of labeled data are limited.

These prior approaches use a simple form of artificial intelligence, known as a neural network, to convert words or phrases into numerical vectors known as word embeddings. These vectors, which can be thought of as points in a high-dimensional space, capture the meanings of the words or phrases they represent. The idea is that words with similar meanings will have similar vectors, and the "distance" between the vectors can be used to quantify similarity between the meanings. For example, a well-known method called "Word2vec" creates 300-dimensional vectors for 3 million unique words and phrases (Mikolov et al. 2013). The vectors are created by training the neural network on a large data set, such as the Google News data set, which includes about 100 billion words.

As discussed earlier, early practices to expand dictionaries, particularly through the use of "shallow word embedding," have inherent limitations due to their single representation for each word. This can lead to inaccuracies in estimating affective meanings, especially when considering the context-dependent nature of words. Referring to recent advancements, van Loon and Freese (2023) introduced an approach that uses BERT to estimate affective meanings. We align with van Loon and Freese's (2023) perspective, who suggest the BERT model might "bridge the gap between the 'words' of embedding models and the 'concepts' pursued by affect control theory" (p. 191). We now delve into the technical details of these alternative approaches.

Alhothali and Hoey (2017) used graph-based sentiment lexicon induction methods to find affective sentiments associated with words. Knowing the affective meaning of some seed words, the algorithm estimates the affective meaning of new words by propagating the meaning from neighbor words. Their label propagation algorithm finds a similarity measure between a large set of words. In this algorithm, the weights used to find the meaning of a new word are estimated from the similarity of the new word with words in the affective dictionary. For example, to estimate the affective meaning of *dean*, we may use a weighted mean of values from affective dictionaries for words such as *boss*, *professor*, and *faculty*. Alhothali and Hoey (2017) used similarity graphs to expand affective meanings to neighbor words in four different embedding spaces: (1) semantic lexicon–based label propagation, (2) a distribution-based approach that uses co-occurrence metrics in a corpus, (3) a neural word-embeddings method, and (4) a combination of semantic and distributional methods. They found that using both semantics and distribution-based approaches gives the best semisupervised result. Li et al. (2017), however, argued that word embedding can represent words' general meaning, including denotative, connotative, social, affective, reflected, collocative, and thematic meanings. Word-embedding graph propagation that uses general meaning similarities may thus reduce accuracy for finding affective meaning.

Joseph et al.'s (2016) novel approach hinges on an event-based affect control theory framework for deducing the EPA meanings of words. This method is distinct in its use of social events as a basis for sentiment analysis, specifically targeting the analysis of identities and behaviors within text. The primary innovation of this method lies in its ability to extract and analyze affective meanings in a manner that is deeply rooted in social contexts and interactions, thereby offering a fresh perspective in the realm of sentiment analysis.

The core of their methodology involves a sophisticated system for imputing EPA values to words on the basis of their contextual usage in social events (Joseph et al. 2016). This system not only accounts for the affective meanings of words, but it also integrates these meanings with their social and behavioral implications. The results of their study demonstrate the effectiveness of this event-based approach, showcasing its capacity to provide a more nuanced understanding of word meanings in a social context. Their article highlights the superiority of their method in extracting affective meanings through the lens of social behavior and interactions, a perspective that was relatively underexplored prior to their work. Their contributions to sentiment analysis and the understanding of affective meanings in social contexts remain significant and foundational to subsequent advancements in the field.

Li et al. (2017) and Mostafavi and Porter (2021) used supervised methods on shallow word embeddings to find affective meanings of words. To put it simply, they use supervised methods to find a mapping from higher dimensional embedding to affective space. For example, Li et al. (2017) used a regression model to estimate the three-dimensional affective meaning from a 300-dimensional Word2vec embedding. All three of these computational approaches yield good results for deriving *evaluation* sentiments for new concepts, but their performance on the *activity* and *potency* dimensions is not very strong.

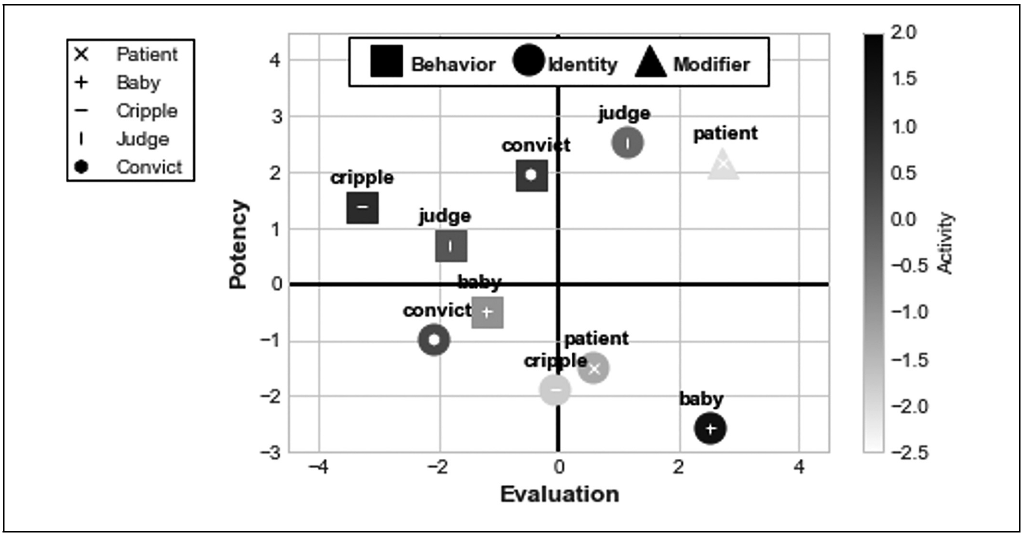


**Figure 3.** Visualization of words with different affective meaning in evaluation, potency, and activity space.

*Note*: Circles, squares, and triangles show identities, behaviors, and modifiers. The color represents the activity dimension. We can observe affective meaning of words such as "judge" is very different as an *identity* and a *behavior*.

Shallow word embeddings have only one representation for every word. On the other hand, the affective meanings of a word in *identity*, *modifier*, and *behavior* categories are different. For example, "mother," "coach," and "fool" have very different affective meanings when they are *behavior* or *identity*, but these words have only one representation in shallow embedding space. In fact, all the shallow word-embedding approaches are limited by using one representation for different meanings of a word and not considering the context. Figure 3 shows the EPA values for some words that appear in different categories. Some words are mapped very differently on the basis of their category. For example, "baby" as an *identity* is a pleasant and active character but is unpleasant and passive as a *behavior*.

Affective meaning in language is a complex task, especially when words cross categorical boundaries such as *identity*, *behavior*, and *modifier*. Shallow word-embedding approaches often overlook this nuance, using a singular representation for concepts irrespective of their categorical context. Our analysis demonstrates that such a universal approach may not accurately capture the varied affective meanings words can have depending on their usage (see Table 1). For instance, correlations across categories, especially in the dimensions of potency and activity, are not consistently high, indicating a divergence in affective meaning. This observation underscores the need for more context-sensitive models in estimating affective meanings, a need our study addresses with the advanced BERTNN algorithm. By integrating contextual understanding into our model, we aim to more accurately represent the dynamic nature of language and its affective nuances, moving beyond the limitations of traditional embedding methods.

**Table 1.** Correlation among the Affective Meanings of Identities, Behaviors, and Modifiers from a Sentiment Dictionary Collected in 2014 (Smith-Lovin et al. 2016).

| | (a) Identity-Modifier | | | (b) Behavior-Modifier | | | (c) Identity-Behavior | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | P | A | E | P | A | E | P | A |
| E | .926*** | .486 | −.587 | .983*** | .925*** | −.318 | .730*** | .354 | .400 |
| P | .772* | .624 | −.389 | .847** | .806* | −.249 | .291 | .552* | .018 |
| A | −.451 | .335 | .982*** | −.278 | −.303 | .673 | −.118 | .296 | .403 |

*Note*: A = activity; E = evaluation; P = potency.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 1 provides a detailed look at the correlation among the affective meanings of *identities*, *behaviors*, and *modifiers* from an affective dictionary collected in 2014 (Smith-Lovin et al. 2016). The table is divided into three sections, each representing the correlation between two categories: (a) identity-modifier, (b) behavior-modifier, and (c) identity-behavior. Each section further breaks down the correlation into the three dimensions of affective meanings (EPA). The correlations in Table 1 reveal that the affective meanings of words in different categories are not always highly correlated. For example, there is a correlation of .93 between the *evaluation* dimensions of words that appear in both the *identity* (a patient) and *modifier* (a patient person) categories. However, the *potency* dimension is less correlated ($r$ = .62). There is a correlation of only .4 between the *activity* dimensions of words that appear in both the *identity* and *behavior* categories.

Some categories maintain high association across categories, but other categories, such as *identity* and *behavior* in the *activity* and *potency* dimensions, have low associations implying that words used in different categories represent different affective meanings. This is consistent with the examples displayed in Figure 3. To illustrate this point, consider the word "judge." When used as an identity (e.g., "The judge asked the victim"), it might carry a different affective meaning than when it is used as a behavior (e.g., "The audiences judge the player"). This difference in affective meanings is reflected in the low correlation between the *identity* and *behavior* categories in the *activity* and *potency* dimensions.

From Table 1, it is clear that the *activity* and *potency* sentiments associated with common words that can serve as either an *identity* or a *behavior* are too small to assume they represent the same affective meanings. This highlights the need for models that can represent contextual aspects and differentiate between different meanings of a word. In other words, the low correlation values in Table 1 underscore the importance of considering the context in which a word is used when estimating its affective meaning. This is a key motivation for our work with the BERTNN algorithm, which leverages the power of contextual language usage to efficiently estimate the affective meanings of new concepts.

Van Loon and Freese (2023) applied both shallow embeddings and BERT models in an attempt to use context while representing concepts of a sentence. However, their BERT implementation did not exhibit significant gains. We posit that this is because

(1) they directly used BERT's pretrained capabilities without customization for affective modeling; (2) their particular BERT implementation did not fully use its contextual capacities; (3) their approach relied solely on "roots" word vectors rather than complete event sequences; and (4) they excluded compound terms, such as "neighborly," absent from BERT's vocabulary. While still leveraging context embeddings, their approach did not optimize the event information as we propose.

Our BERTNN methodology aims to address these limitations and unlock BERT's full potential for affective modeling. We generate full synthetic event sentences to capture nuanced ABO variations. Our training process teaches the model cultural sentiment patterns through impression formations. And we aggregate across varied contextual embeddings rather than extracting static singular root representations.

## *BERT*

In 2018, Google open-sourced a language representation model named BERT as the state-of-the-art model for a wide range of natural language processing tasks. This deep neural network model pretrained the bidirectional representation from a large set of unlabeled text. The model was pretrained on the BookCorpus (which includes 11,038 books with about 1 billion words) and English Wikipedia (which has about 800 million words) (Zhu et al. 2015).

BERT first transforms a sentence into a collection of tokens. It adds special tokens during tokenization. The [CLS] token indicates the start of a sentence, and [SEP] denotes the end of the first sentence when we have two sentences. Although we use only single sentences, these special tokens are still required by the model. If a word is in BERT's vocabulary, it becomes a single token. For uncommon or compound words, BERT breaks them into subword tokens. For example, consider the sentence "The aggressive employee yelled at the timid intern." If the word "aggressive" is not in BERT's core vocabulary of 30,000 tokens, BERT represents it as two tokens "aggress" and "##ive," where ## indicates a subword. The subword tokens get unique IDs matched to entries in BERT's vocabulary index. So "aggress" and "##ive" are mapped to specific integer IDs known as "token IDs." After tokenization, the sentences are represented by a sequence of token IDs. In this way, BERT tokenizes sentences into numerical representations, even for complex vocabulary. This enables richer contextual understanding compared with methods that operate only on common predefined words. The subword approach allows BERT to ingest a wide vocabulary critical for capturing nuanced social meanings.

Van Loon and Freese (2023) used BERT to obtain embeddings for specific words by feeding the model sentences such as "to be a judge is" and then extracting the embedding for the token "judge." However, this approach has limitations. First, if a word is not present in BERT's vocabulary, it gets split into multiple tokens, leading to the dilemma of how to aggregate these multiple embeddings into a single representation. Second, this problem is exacerbated with compound words, which often get tokenized into multiple parts. Because of these challenges, van Loon and Freese excluded compound words from their method.

Our methodology uses the BERT model more effectively. Instead of extracting embeddings for individual tokens, we focus on the output associated with the [CLS] token. In BERT's architecture, the [CLS] token is more than just a marker for the beginning of a sentence; it provides a comprehensive representation of the entire sentence. This feature of BERT is crucial to our approach. By analyzing the [CLS] token's output, we can capture the holistic meaning of the sentence, encompassing all its components in their specific context. This strategy eliminates the need to aggregate embeddings from multiple tokens, which can be complex and may not fully capture the nuances of a sentence. We enhance the contextuality of our model by feeding it with sentences that describe dynamic social events, rather than static, isolated phrases such as "to be a judge is." This use of rich, event-based sentences enables our model to derive more nuanced and context-aware representations, capturing the subtleties of social interactions and their affective implications more accurately.

BERT was trained on two primary objectives: the masked language model and next sentence prediction. The masked language model involves predicting the values of randomly masked tokens within a sentence. It uses the surrounding words on both sides of the target word in all layers to predict the masked word. For instance, consider the sentence "The student asked a question about the prerequisite courses." If the word "question" is masked, we have "The student asked a [MASK] about the prerequisite courses." The model then uses the contextual words to predict the masked word. This ability to use the full context for prediction is a key differentiator of BERT from other word-embedding models, such as Word2Vec, that rely only on neighboring words for prediction.

BERT was also pretrained on a "next sentence prediction" task. This task was designed to help BERT better understand the relationships between consecutive sentences. To illustrate, consider two sentences: (1) "The student attended the lecture." and (2) "He found the topic very interesting." BERT was trained to predict whether the second sentence logically follows the first one. This type of training enables BERT to capture longer term relationships between sentences, significantly enhancing its performance in tasks such as text summarization (Liu and Lapata 2019). BERT achieves state-of-the-art performance on various natural language processing tasks when fine-tuned on task-specific data sets (McCormick and Ryan 2019).

## METHODOLOGY

The main advantage of word representation derived from BERT over shallow word embeddings is that BERT can take into account the context of a word. This means a word can be given different representations when it is used as a *behavior* or an *identity* if it is used in a proper sentence. We take full advantage of the BERT embedding by training it on synthetic data that represent concepts and their affective categories simultaneously.

We aim to generate affective meanings for concepts across three categories: *identities*, *behaviors*, and *modifiers*. Traditional affective dictionaries are derived from surveys where participants rate a concept within one of these categories. However, to harness the power of the BERT model, we need to create synthetic sentences that represent these concepts in the context of social events. This is crucial because BERT's strength lies in understanding context; a stand-alone word might not provide enough information for the model to infer its affective meaning accurately.

In the analysis that follows, we show how to generate a contextual data set describing social events to train a deep neural network, and how to use this network to predict affective meanings of new concepts. We introduce a new framework that processes synthetic data, passes it to a pretrained BERT model, and fine-tunes the result by a three-layer neural network to generate extended affective meaning dictionaries. The data used to train our model were generated from the affective dictionary described in Smith-Lovin et al. (2016). This dictionary was developed from surveys conducted between 2012 and 2014 and includes 929 identities, 814 behaviors, and 660 modifiers. The data used in this study are publicly available on the Internet (Smith-Lovin et al. 2016).

To leverage BERT's capacity for providing context-specific sentence embeddings, we generated synthetic data to fine-tune the model. This synthetic data set comprises sentences that encompass the concepts of interest within a structured format, known as *modified actor behaves modified object* (MABMO) events. Unlike van Loon and Freese's (2023) approach, which uses token embedding of phrases such as *assailant is* to determine the embedding for *assailant*, thereby yielding a single estimation per concept, BERTNN adopts a more comprehensive method.

In BERTNN, we use synthetic sentences that describe MABMO events. An example of such a sentence could be "Happy doctor helps anxious patient." Here, "Happy" and "anxious" are the modifiers (M), "doctor" is the actor (A), "helps" is the behavior (B), and "patient" is the object (O). In this framework, each word in the sentence plays a specific role, contributing to the overall sentiment conveyed by the event.

BERTNN uses the vectorized representation of the entire sentence to estimate a 15-dimensional lexicon vector, encompassing the affective meanings across the EPA dimensions for each component (M, A, B, M, and O) in the MABMO event. This approach allows different estimations of a concept's affective meaning based on the other concepts present in the event. For example, the affective meaning of "doctor" as an actor may vary depending on the modifiers and object it interacts with in different MABMO sentences.

BERTNN aggregates these multiple estimations to generate a distribution for each concept of interest. This aggregation reflects the varied contexts in which a concept can appear, providing a nuanced understanding of its affective meaning. To illustrate with our example sentence, BERTNN would provide affective meaning estimations for "happy," "doctor," "helps," "anxious," and "patient" on the basis of their respective roles in the event. The model then aggregates these context-specific estimations across numerous MABMO events to establish a comprehensive affective profile for each concept. This MABMO event-based approach ensures the training of the BERTNN model is grounded in the dynamics of social interactions, aligning with the core principles of affect control theory and allowing a more accurate and context-aware estimation of affective meanings.

It takes only a few iterations to fine-tune the BERT model for affective meaning estimation in BERTNN. As a result, event selection is critical to get a fine-tuned model that is general enough to predict affective meaning for new concepts. We defined an algorithm to prepare data for estimating affective meanings of new concepts in diverse social interactions. The preprocessed data generate synthetic data that can represent the whole affective space of the dictionaries. Algorithm 1 describes the preprocessing steps to include contextual aspects of concepts before using the BERT model.

## ALGORITHM 1

*Preprocessing for Affective Meaning Estimation*

**Data:** Affective dictionaries for identities, modifiers, and behaviors

1. **Cluster Concepts.** Group similar concepts in each of the *identity*, *behavior*, and *modifier* dictionaries. This helps in having all the broad categories or themes of concepts present in the dictionaries.
2. **Data Splitting.** Divide the data into training, testing, and validation sets. Use stratified sampling on the basis of the clusters to ensure a balanced representation of different concept themes in each set.
3. **Event Creation.** Construct all possible ABO events using the training set. These events represent different social interactions.
4. **Impression Calculation.** Use the impression change equations to determine the perceived sentiment of concepts in all events. This step helps in understanding how different events influence the perception of the involved concepts.
5. **Deflection Analysis.** Calculate the difference between the baseline sentiments (original meanings) and the impressions (perceived meanings) of the events across all nine dimensions (3 concepts in ABO grammar $\times$ 3 EPA dimensions). This step captures the variation in factors that cause deflection, which is the difference between expected and perceived sentiments.
6. **Binary Encoding.** Label the events on the basis of whether the impression increases or decreases each concept's EPA values. For instance, a code such as "111000111" indicates that the event increases the EPA values for the actor and object but decreases them for the behavior.
7. **Training Data Generation.** Create a data generator that uses the defined binary encodings to produce synthetic data, ensuring a comprehensive representation of different social interactions.
8. **MABMO Event Creation.** Enhance the complexity of the ABO events by adding two random modifiers, resulting in MABMO events. This step helps in capturing more nuanced social interactions.

Affective meanings of concepts are distributed across EPA dimensions, but not uniformly. Some regions of EPA space contain denser clusters of concepts than others. When fine-tuning a complex model such as BERT, we use relatively few training iterations. As a result, the training data used becomes even more crucial: it needs to provide broad coverage of EPA space to avoid missing sparsely populated regions. If we randomly sampled training concepts without considering their distribution, we risk selecting a nonrepresentative sample concentrated in high-density areas. The resulting fine-tuned model would then struggle to generalize to sparser areas of the space not properly represented during training.

To avoid this issue, we use a clustering technique in our data preprocessing. By first clustering the affective concepts and sampling evenly across clusters, we help ensure our small training set captures diverse regions of EPA space. This allows the fine-tuned model to estimate sentiments accurately across dense and sparse areas of the affective distribution. The clustering guides the training data selection to be maximally representative given the practical limitations on training iterations.

For example, all the training samples may have positive activity for the behavior and as a result, the model could be more accurate if we have samples with negative activity for the behavior. Every concept in identity, behavior, and modifier dictionaries is represented by a three-dimensional affective vector. We are looking for training samples that represent different regions or clusters within each dictionary.

In the first step of algorithm 1, we need to know how many different regions/clusters we have in the dictionaries and uniformly sample from each of them. For this purpose, we clustered each identity, behavior, and modifier dictionary using *k*-means clustering. We used the elbow method and determined the number of clusters for each of these three dictionaries. We concluded that five clusters or regions are enough for all three categories. This clustering is pivotal for capturing the full spectrum of affective dimensions, ensuring a representative sample of the diverse affective landscapes present in the dictionaries. The clustering results, which are intriguing in their own right, reveal distinct affective themes and patterns within and across these dictionaries.

As is typical in machine learning methods, the second step in algorithm 1 involves dividing the data into three parts: training, testing, and validation sets. In simpler terms, the training set is like the main material used for teaching the model; the testing set helps us adjust the model's learning strategy (known as "fine-tuning"); and the validation set is like a final exam, used to evaluate the model's performance. For the fine-tuning process, we use the training set to improve the BERT model and the testing set to guide us in optimizing learning settings (termed "hyperparameters"), such as determining the number of complete learning cycles (referred to as "epochs"). We chose these settings to ensure the model's predictions are as accurate as possible when evaluated against the testing set. The validation set, which the model has not seen during training, serves as an independent check to confirm the model's effectiveness. The results reported in this article are based on how well the model performed on this validation set. To ensure a balanced representation of different types of concepts such as identities, behaviors, and modifiers, we used stratified sampling across various clusters within these categories. Specifically, we selected 80 percent of the words from each cluster for training, reserved 8 percent for testing, and the remaining 12 percent formed the validation set.

The third step captures variation in factors that change *deflection*. When we select events to train the model, they should represent the variation of concept impression.

The training data should include events that increase or decrease the baseline sentiments in each dimension. For the training set, we found the impression of all possible events in ABO grammar using impression change equations. We then calculated the difference between the impressions and baseline concepts across all nine dimensions. Among all $2^9 = 512$ possible variations, 508 types of events were present in the training set, and we used binary encoding to label them.

Our primary objective with BERTNN is to accurately estimate affective meanings across various lexical categories, specifically *identities*, *behaviors*, and *modifiers*. The initial approach in affect control theory predominantly involves ABO events, which focus on identities and behaviors. However, to effectively incorporate modifiers into our model, we transition from standard ABO events to more complex MABMO event structures. This modification in the event formation is pivotal for two reasons. First, it enables our model to deduce affective meanings associated with modifiers by observing their usage in modifying the actors and objects. Second, iterative testing showed that MABMO events yield more comprehensive affective estimations than do simpler ABO structures. The increased complexity of MABMO events offers richer context and regularities, enhancing the learning process. Consequently, by formulating synthetic sentences within MABMO structures, BERTNN is equipped to learn and predict multidimensional affective lexicons that encompass the full range of identities, behaviors, and modifiers, thus providing a more holistic understanding of social interactions.

For MABMO grammar, we need to sample two modifiers and one ABO event. We used stratified sampling on the basis of modifier clusters and ABO labels developed in step 1 of algorithm 1. These MABMO events are sentences that are passed to the BERT model.

There are two main approaches to getting vector representations from BERT. One is using the token embedding for the tokens of a word. This method is not efficient when we have compound words or a word is not in the BERT dictionary. Alternatively, the embedding for the [CLS] token can represent an embedding for the whole sentence. Using fine-tuning layers, it is possible to get representations for the concepts from this sentence embedding. We used BERT outputs that correspond to [CLS] tokens as a vectorized sentence representation. The next step is finding a mapping from the [CLS] outputs to the affective space. We passed BERT output to a fine-tuning neural network with a dense input layer followed by two "Relu" hidden layers. The term "Relu" refers to a type of activation function in neural networks that introduces nonlinearity, allowing the model to capture complex relationships in the data. The last layer is a dense layer followed by a pooler that produces the 15-dimensional vector representing EPA values for all MABMO characters. The Appendix provides more details on the technicalities of this approach.

To understand the affective meaning of a concept, especially when it is not present in our affective dictionary, we leverage BERT's ability to provide context-rich representations. However, a challenge arises: BERT can return multiple representations for a single concept, depending on its context in different sentences. We get multiple representations for a word from BERTNN depending on other terms in the MABMO event. In other words, for the affective meaning of a given actor in MABMO grammar, we get different estimates depending on the modifiers, behavior, and object used in the event. So, how do we consolidate these varied representations into a singular, meaningful affective value?

Consider the task of understanding the affective meaning of the term "moderator." Rather than evaluating this term in isolation, we integrate it into diverse synthetic social scenarios, each involving the "moderator" as a key element. For example, in the MABMO-structured event "enthusiastic moderator amicably resolves dispute with irate customer," the context provided by the additional terms helps discern the nuanced affective connotations associated with "moderator." In another instance, "diligent moderator arbitrates discussion between excited debaters," the scenario offers a different contextual setting. Our model processes these richly contextualized events. Because the affective meanings of the individual terms in these events, including "moderator," are not definitively established, our model endeavors to infer the sentiment of each term. It does so by considering the overall contextual dynamics of the event, thereby providing a more comprehensive understanding of the affective meaning attributed to the term "moderator."

The strength of this approach lies in the aggregation. By placing "moderator" in numerous events and observing its estimated sentiment across these contexts, we can derive a more robust and representative affective meaning for the term. Each individual event provides a sample estimation for "moderator," and by averaging over these many samples, we obtain a sentiment value that captures the term's general affective essence across various contexts.

Why is this approach beneficial? Using the entire sentence as context offers a richer understanding compared with analyzing a word in isolation. It is not just about identifying the part of speech (whether it is an identity, behavior, or modifier), but about understanding the word's role and sentiment within a social interaction. For instance, a "moderator" in a peaceful discussion might have a different sentiment than one in a heated debate. By considering multiple scenarios, we capture these nuances.

Comparing BERTNN with van Loon and Freese's (2023) model, the main differences are the following:

- BERTNN generates multiple synthetic sentences that encapsulate the dynamics of social events, leveraging the contextual embedding capabilities of the BERT model. These sentences are constructed on the basis of the MABMO grammar from the affect control theory literature. This approach allows BERTNN to account for deflection and regions of concepts during training, providing a more comprehensive representation of social interactions than do static sentences.
- Whereas van Loon and Freese's (2023) added fine-tuning layers and trained only those layers, BERTNN goes a step further. It fine-tunes the entire BERT model in conjunction with the additional layers for several epochs, ensuring a more integrated optimization process.
- BERTNN uses sentence embeddings as opposed to the token embeddings of core concepts. This design choice enables BERTNN to accurately estimate the affective meaning

of compound words and multitoken words, overcoming the limitations of solely relying on token embeddings.

- BERTNN provides a distribution for affective meaning, which is derived from the variety of synthetic data used. This distributional approach offers a richer understanding of affective meanings by considering the diverse contexts in which a term can appear.

We made BERTNN a publicly available tool that researchers with minimal coding experience can use to estimate the affective meaning of their words of interest.[5]

## RESULTS

Comparing the performance of BERTNN with that of prior approaches requires analyzing the models across several quantitative evaluation criteria. We use three standard metrics that provide complementary perspectives on estimation accuracy and association with ground truth affective ratings.

The first two metrics—mean absolute error (MAE) and root mean squared error (RMSE)—assess the absolute deviation between a model's predicted EPA values and the true sentiments from the survey dictionary. MAE simply averages the absolute differences across all dimensions and test-set concepts. RMSE also averages errors but places more weight on larger discrepancies by squaring before taking the mean. For both, lower values indicate better estimation performance.

The third metric we use is the correlation coefficient between a model's outputs and the survey EPA ratings. Correlation measures the linear relationship between predicted and actual values, with a maximum possible value of 1 indicating perfect linear association. Higher correlation signifies the model more consistently aligns with the ground truth affective spectrum, even if offset by a constant amount.

To evaluate the efficacy of our model in determining affective meanings, we juxtaposed its performance against several established methodologies from prior research: word analogy methods as described by Kozlowski, Taddy, and Evans (2019), regression techniques highlighted by Li et al. (2017), translation matrix methods introduced by Mostafavi and Porter (2021), and application of the BERT model to core concepts as presented by van Loon and Freese (2023). The validation set used for this comparative analysis was meticulously curated, encompassing 139 identities, 122 behaviors, and 99 modifiers, all chosen via the stratified sampling technique discussed in prior sections.

Table 2 shows the outcomes of this comparative study. BERTNN achieves the lowest MAE and RMSE and highest correlation values overall for identities and modifiers, demonstrating lower estimation error across the EPA dimensions. For behaviors, it attains the best MAE and RMSE for the evaluation and activity sentiments. For behaviors, BERTNN also has the minimum error for evaluation and activity dimensions and second lowest for potency.

In essence, BERTNN produces affective ratings with lower error compared with previous approaches for the majority of lexicon categories and dimensions. It also exhibits higher correlation with ground truth for evaluation universally, and for the majority of categories on potency and activity.

**Table 2.** Performance of Several Models on Smith-Lovin et al.'s (2016) Data.

| | | MAE | | | RMSE | | | Correlation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | E | P | A | E | P | A | E | P | A |
| Identity | Analogy stepW. | .92 | .92 | .74 | 1.15 | 1.14 | .94 | .744*** | .538*** | .299*** |
| | Analogy_regression | .93 | .94 | .76 | 1.16 | 1.17 | .96 | .748*** | .505*** | .249*** |
| | StepW Translation (Mostafavi and Porter 2021) | .80 | .78 | .69 | .97 | 1.04 | .87 | .824*** | .633*** | .509*** |
| | CoreBERT (van Loon and Freese 2023) | .65 | .69 | .66 | .88 | .89 | .79 | .860*** | .638*** | .485*** |
| | BERTNN | **.53** | **.54** | **.56** | **.73** | **.70** | **.72** | **.893***** | **.817***** | **.611***** |
| Behavior | Analogy stepW. | 1.24 | .69 | .67 | 1.54 | .90 | .85 | .662*** | .387*** | .352*** |
| | Analogy_regression | 1.25 | .72 | .70 | 1.56 | .92 | .88 | .653*** | .388*** | .295*** |
| | StepW Translation (Mostafavi and Porter 2021) | 1.07 | .69 | .58 | 1.21 | .84 | .82 | .767*** | .456*** | .512*** |
| | CoreBERT (van Loon and Freese 2023) | .74 | **.44** | .51 | .97 | **.55** | .64 | .893*** | **.746***** | **.747***** |
| | BERTNN | **.69** | .47 | **.48** | **.87** | .62 | **.61** | **.911***** | .707*** | .730*** |
| Modifier | Analogy stepW. | .98 | .70 | .89 | 1.23 | .88 | 1.10 | .827*** | .832*** | .537*** |
| | Analogy_regression | 1.01 | .74 | .94 | 1.27 | .91 | 1.13 | .816*** | .834*** | .504*** |
| | StepW Translation (Mostafavi and Porter 2021) | .85 | .59 | .73 | .98 | .67 | .84 | .869*** | .840*** | .650*** |
| | CoreBERT (van Loon and Freese 2023) | .63 | .50 | .65 | .80 | .64 | .79 | .932*** | .883*** | .752*** |
| | BERTNN | **.60** | **.44** | **.54** | **.78** | **.58** | **.71** | **.936***** | **.912***** | **.812***** |

*Note*: Boldface type indicates the best model. Our model, BERTNN, performed best in most categories. A = activity; BERTNN = Bidirectional Encoder Representations from Transformers Neural Network; E = evaluation; MAE = mean absolute error; P = potency; RMSE = root mean squared error.
$^{***}p < .001$.

All the similar works (except van Loon and Freese 2023) used shallow embedding (Kozlowski et al. 2019; Li et al. 2017; Mostafavi and Porter 2021). We used the publicly available code of Mostafavi and Porter (2021) and Kozlowski et al. (2019) to replicate their works and compare the results. For Li et al. (2017) we had to implement their method in Python. To further improve their methods, we added tuning layers, such as adding step-wise regression to the analogy method. We used a pretrained model in van Loon and Freese (2023) to compare with their work. Table 2 shows the best result we could get from other methods. We see that in some metrics, such as the correlation metric for activity, the improvement over shallow-embedding methods is about 20 percent for behaviors.

In terms of error rates, BERTNN achieves values that are comparable with differences between different affective dictionaries. For example, the MAE value that BERTNN achieves for modifiers in Table 2 is smaller than that for three of the dictionaries shown in Appendix Table A1. In other words, BERTNN estimation error on the validation set is small enough to say this model estimates EPA values similar to running a new survey.

**Table 3.** Comparing Estimated Affective Meanings for "Judge" as an Identity and a Behavior.

| Judge | Evaluation | Potency | Activity | Estimated Evaluation | Estimated Potency | Estimated Activity |
|---|---|---|---|---|---|---|
| Behavior | −1.83 | .71 | .07 | −1.40 | .88 | −.03 |
| Identity | 1.15 | 2.53 | −.22 | 1.95 | 2.27 | .67 |

**Table 4.** Correlation Analysis for Identities.

| (a) | E | P | A | (b) | E | P | A | (c) | EE | EP | EA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EE | .894*** | .423*** | .094 | E | 1.00*** | .462*** | .133 | EE | 1.00*** | .448*** | .044 |
| EP | .394*** | .816*** | .186 | P | .452*** | 1.0*** | .334*** | EP | .448*** | 1.0*** | .242* |
| EA | .009 | .269** | .611*** | A | .133 | .334*** | 1.0*** | EA | .044 | .242* | 1.0*** |

*Note*: (a) Correlation between estimated values (EE, EP, and EA) and the affective dictionary value (E, P, and A). We can also compare correlation of the words in the three dimensions shown in (b) and correlation of the estimated values of the three dimensions shown in (c) to find how close the off-diagonal entries are in the estimation compared with dictionary values. A = activity; E = evaluation; P = potency.
*$p < .05$; **$p < .01$; ***$p < .001$.

One problem with shallow embeddings is they cannot distinguish between words used as identities versus behaviors. Using BERTNN, we can differentiate between these two cases. For example, Table 3 shows that estimated values for "judge" are different when the word is considered as an identity or a behavior.

Tables 4 and 5 present a correlation analysis between the EPA values estimated by BERTNN and the original survey dictionary for identities and behaviors. The high diagonal values indicate BERTNN's outputs are strongly linearly associated with the ground truth ratings, validating its accuracy. But diagonal correlation terms alone do not fully validate a model. We also want estimated dimensions to have the proper relationships with each other, mirroring the survey data. The off-diagonal terms assess this, that is, whether changes in one estimated dimension correlate with changes in another dimension similarly to the original data. For instance, in Table 4c, the correlation of .448 between estimated evaluation and potency matches the correlation of .462 in the survey data (Table 4b). This alignment shows BERTNN preserves the interrelationships between dimensions seen in human ratings.

The proximity between the off-diagonal terms in Tables 4b and 5b and Tables 4c and 5c demonstrates that BERTNN replicates not just absolute values but also the relative multidimensional patterns in the seed lexicon. Capturing these subtle relationships is critical for modeling the nuances of cultural sentiments.

Variations in estimated affective meanings from the model can arise from a variety of sources, including, first, splits of the original data to make training, testing, and validation sets. We conducted a series of robustness checks, and we obtain similar results using the test set for model selection and increasing or decreasing the number of

**Table 5.** Correlation Analysis for Behaviors.

| (a) | E | P | A | (b) | E | P | A | (c) | EE | EP | EA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EE | .910*** | .396*** | −.279** | E | 1.0*** | .521*** | −.237* | EE | 1.0*** | .521*** | −.288** |
| EP | .507*** | .707*** | .141 | P | .521*** | 1.0*** | .191 | EP | .521*** | 1.0*** | 0.105 |
| EA | −.276** | .059 | .731*** | A | −.237* | .191 | 1.0*** | EA | −.288** | 0.105 | 1.0*** |

*Note*: (a) Correlation between estimated values (EA, EP, and EA) and the affective dictionary value (E, P, and A). We can also compare correlation of the words in the three dimensions shown in (b) and correlation of the estimated values of the three dimensions shown in (c) to find how close the off-diagonal entries are in the estimation compared with dictionary values. A = activity; E = evaluation; P = potency.
*$p < .05$; **$p < .01$; ***$p < .001$.

**Table 6.** Using BERTNN to Represent One Participant Evaluating Concepts in MABMO Grammar.

| Happy [Modifier] | | | Doctor [Actor] | | | Help [Behavior] | | | Wonderful [Modifier] | | | Mother [Object] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | P | A | E | P | A | E | P | A | E | P | A | E | P | A |
| .28 | 2.61 | .82 | 2.55 | 2.75 | .47 | 1.91 | 2.16 | .96 | 2.81 | 2.81 | 1.13 | 2.81 | 2.48 | .66 |

*Note*: A = activity; BERTNN = Bidirectional Encoder Representations from Transformers Neural Network; E = evaluation; MABMO = modified actor behaves modified object; P = potency.

iterations. We also simulated this process with various seeds, and the final results are similar. Second, mean, median, or similar statistics such as trim-mean of multiple sample estimations for a concept can lead to different results. In this study, we used the mean values across all dimensions. Finally, whether an identity is used as an actor or object in the MABMO grammar, or a modifier is modifying an actor or object, can affect the results. The values we get for a concept in these cases vary but they are close. We concatenated the results and aggregated for the concept in general. It is also possible to distinguish different affective meanings for concepts operating in the role of an actor or object. One can run studies to differentiate the values one gets for identity as the actor or the object.

## *Model Outputs*

BERTNN is designed to represent the evaluation of affective meanings of concepts by simulating a participant's perspective. For instance, consider the phrase "happy doctor help wonderful mother." Each word in this phrase carries a specific affective meaning, and BERTNN estimates these meanings on the basis of the context in which they appear. Table 6 provides a detailed breakdown of the estimated affective values for each concept in the phrase. The table shows the three primary affective dimensions (EPA). For instance, the word "happy" as a modifier has an evaluation score of 3.28, indicating a positive sentiment. Similarly, the word "doctor" as an actor has an evaluation score of 2.55, suggesting a moderately positive sentiment.

**Table 7.** Estimated Affective Meaning of "Phone" from Different MABMO Events.

| MABMO Event | E | P | A |
|---|---|---|---|
| Respectful neurotic phones grumpy golfer | .86 | .79 | .99 |
| Nostalgic warden phones ambitious tutor | 1.02 | .89 | 1.11 |
| Hurt fiancé phones conventional intimate | 1.17 | 1.04 | 1.15 |
| Rigid scapegoat phones egotistical magician | .48 | .72 | 1.04 |
| Straightforward single parent phones opportunistic hippie | 1.16 | .94 | 1.09 |
| Relentless golfer phones withdrawn spouse | 1.23 | 1.01 | 1.09 |

*Note*: A = activity; E = evaluation; MABMO = modified actor behaves modified object; P = potency.

**Table 8.** Estimating Affective Meaning of "Phone" from 300 MABMO Events.

| | E | P | A |
|---|---|---|---|
| Mean | 1.03 | .98 | 1.06 |
| Standard deviation | .25 | .09 | .10 |
| Minimum | −.52 | .64 | .79 |
| Maximum | 1.61 | 1.21 | 1.47 |

*Note*: A = activity; E = evaluation; MABMO = modified actor behaves modified object; P = potency.

A standout feature of BERTNN is its ability to provide contextual representations of words. This means the same word can have varied affective meanings on the basis of the context in which it appears. To illustrate, consider the word "phone" as a behavior. Traditional affective dictionaries might not have evaluated this concept. However, with BERTNN, we can create multiple MABMO events with the word "phone" to estimate its affective meaning in different scenarios. Table 7 provides a snapshot of the estimated affective meanings of the word "phone" in various MABMO events.

For a more comprehensive understanding, we aggregated the results from 300 such events to derive a reliable estimate for the word "phone" as a behavior. Table 8 summarizes the statistics from these events, providing insights into the mean, standard deviation, minimum, and maximum values for the affective dimensions of the word "phone."

These cases illustrate how BERTNN can replicate human-level sentiment inferences for new concepts on the basis of contextual usage across observed interactions. The examples validate BERTNN's ability to effectively incorporate term positions and event information when making affective judgments.

BERTNN offers a dynamic way to estimate affective meanings of concepts on the basis of their context. By examining words in varied MABMO events, we can derive a more nuanced understanding of their affective values, which might not be possible with traditional dictionaries. The tables presented here serve as concrete examples of BERTNN's capabilities, providing specific estimations for individual words and aggregated insights for a holistic view.

## DISCUSSION

Creating a synthetic data set of events corresponding to the core grammar of affect control theory, preprocessing these data, and then sending them through the pretrained BERT word-embedding model yielded predictions that seem to have great promise for generating the cultural sentiments required by affect control theory. We argue that this approach produces sentiments that are closer to the concept sentiments required for affect control theory, rather than sentiments for words or root words. We also note that the learning approach used by BERTNN is not unlike how humans acquire social meaning through observation of social interactions. When we encounter a new culture or subculture with an identity we do not recognize, we observe interactions to gather information (some of the form MABMO), and each interaction leaves an impression about the likely meaning of that identity. If someone always does good things, for example, we will come to infer a positive evaluation for their identity. If we do not know the identity of an actor (in this context, the identity of a person is masked in MABMO interactions), we would have something similar to M*BMO. On the basis of multiple interactions, an affective meaning for the unknown identity is formed in our minds. Like the impression change equations described previously, we expect a person to be nice if we observe lots of nice behaviors in prior interactions, or if the person routinely interacts with other positively identified actors. In this way, our method of learning the affective meanings associated with known or unknown identities is very similar to the masked language model.

To further bolster our discussion on the analogy between human cultural learning and how language models learn linguistic representations, we draw insights from Arseniev-Koehler and Foster (2022). Their work provides valuable perspective on how distributed representations in machine learning, such as Word2vec, aim to predict incoming words and complete patterns, emphasizing the importance of understanding the connotations shared between word vectors. They delve into the cognitive lens of these methods, illustrating how word vectors can be seen as activations of lower level units or artificial neurons. This perspective offers a nuanced understanding of the connotations shared between word vectors, emphasizing the importance of shared patterns of activation in determining semantic relationships.

BERTNN's capability to estimate the distribution of affective meanings marks a significant advancement in computational social science. This feature not only enriches our understanding of the variability with which concepts are used in natural language, but it also sheds light on the degree of cultural consensus regarding concepts' meanings in EPA space. By analyzing diverse contexts in which a term appears, BERTNN captures the breadth of its semantic landscape, reflecting the multifaceted nature of cultural sentiments. This approach parallels human cultural learning, where exposure to a variety of contexts leads to a nuanced understanding of social meanings. Such distributional analysis is vital for sociological theories that hinge on the fluidity and dynamism of cultural norms and values. By enabling deeper insight into how sentiments are distributed and perceived across different scenarios, BERTNN paves the way for more sophisticated models of social interaction and cultural analysis, reinforcing the synergy between computational linguistics and sociological inquiry.

The potential of using such computational methods to advance sociological research is significant. Many sociological theories rely on understanding norms, values, and sentiments to model social interactions and dynamics. However, manually surveying populations to build representations of these intangible attributes can be extremely resource and time intensive. Our proposed modeling framework demonstrates how machine learning can be leveraged to bypass this constraint by rapidly inferring latent attributes from language.

As one example, affect control theory is one of sociology's most enduring and developed mathematical theories. Nonetheless, application and extensions of the theory are seriously limited by the need for resource-intensive data collection to incorporate concepts from new substantive domains. Traditional methods require time, expense, and significant burden on human research participants and researchers. Machine learning approaches, of the sort described here, promise to minimize this burden by sidestepping the need for human raters. This approach could greatly expand researchers' ability to apply affect control theory to new substantive domains at greater speed and at a lower burden for respondents, researchers, and funders. Using machine learning approaches pretrained on very large data sets should rapidly expand the existing sentiment dictionaries "on the fly," rather than requiring researchers to design, implement, analyze, and interpret new cultural data. Moreover, the ability to train these context-embedding models on new data sets opens the possibility of moving into novel cultural and subcultural domains more quickly (skipping years of fieldwork or thousands of surveys).

Our work relates broadly to research in natural language processing on constructing representations of social meanings. Beyond affective lexicons, natural language processing researchers have proposed various methods for inducing social lexicons related to power dynamics, agency, and bias (Sap et al. 2017, 2019). These approaches apply social psychology theories, presenting opportunities for cross-disciplinary work to bring in sociological perspectives. In particular, incorporating contextual information seems crucial for properly representing how social meanings vary across settings and roles, as discussed in our work. Engaging with this related natural language processing literature could further strengthen the connections we draw between computational representation learning and traditional methods for quantifying cultural knowledge. Our contextual affective lexicon expansion methodology offers potential for advancing social lexicon induction.

Our primary focus has been on leveraging BERTNN within the affect control theory framework, but the implications of our methodology extend far beyond this paradigm. Our approach, grounded in the rich contextual understanding of language afforded by machine learning, opens new avenues for exploring cultural meanings across various sociological domains. By efficiently inferring latent attributes from linguistic patterns, our model can significantly reduce the time and resources required for traditional data collection methods, which often rely on extensive fieldwork or surveys. BERTNN's s utility is not confined to affect control theory alone. Its capacity to rapidly adapt to new cultural and subcultural domains, coupled with its ability to handle a wide range of sociological concepts, make it a powerful tool for a broad spectrum of sociological research. By bridging computational representation learning with traditional methods of quantifying cultural knowledge, our methodology stands to contribute significantly to the advancement of social lexicon induction and the broader study of cultural meanings.

# APPENDIX

We use a large pretrained BERT[6] and fine-tune it for finding affective lexicons. Our approach uses the vectorized representation of the whole sentence. The pipeline for our method (BERTNN) is shown in Figure A1. The $L_2$ loss used in the neural networks minimized the squared error between estimated affective meaning and the target values across all 15 dimensions. Figure A1 shows how a sentence representing a social interaction is tokenized and passed to the model (McCormick and Ryan 2019). The output layer of the BERT model gives the embeddings for all the tokens shown as $C, T_1, ..., T_N, T_{[SEP]}$ where $T_k$ is the vector representation of $k^{th}$ token and $C$ corresponds to the $[CLS]$ token and can represent the sentence embedding.

We implemented the neural network in Python using Torch and Transformers packages. We used "AdamW" with a learning rate of $2 \times 10^{-5}$ and batch size of 64. The BERT model and neural network are tuned for a few epochs. The data in the testing set are used to decide how many iterations result in a good model. The neural network output is a 15-dimension vector that is highly correlated with the affective meaning variables.

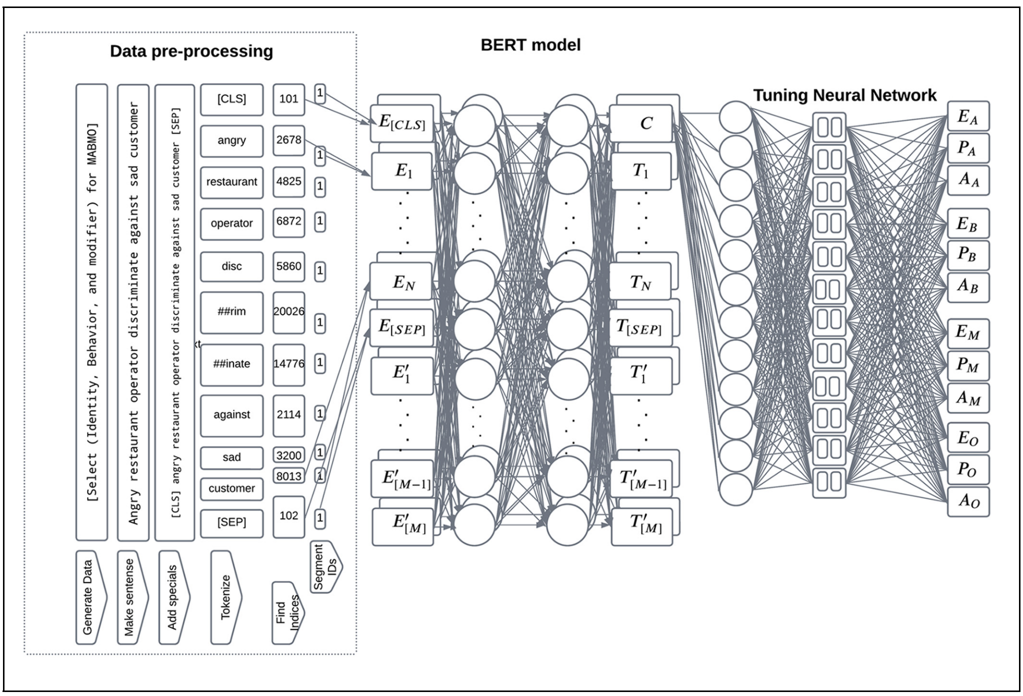


**Figure A1.** BERTNN pipeline.

*Note*: Sentences describing an event are generated in the first step and preprocessed to pass to a BERT model. Then outputs corresponding to the [CLS] token are passed to a three-layer neural network to find affective meaning.

**Table A1.** Comparing Four Different Affective Dictionaries.

| | | MAD | | | RMSD | | | Correlation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | E | P | A | E | P | A | E | P | A | Count |
| Identity | Indiana | .41 | .48 | .44 | .53 | .63 | .57 | .956*** | .905*** | .807*** | 317 |
| | North Carolina | .57 | .58 | .69 | .71 | .71 | .85 | .927*** | .906*** | .653*** | 453 |
| | Texas | .45 | .52 | .77 | .59 | .64 | .95 | .936*** | .910*** | .539*** | 319 |
| Behavior | Indiana | .54 | .61 | .59 | .67 | .74 | .73 | .967*** | .801*** | .665*** | 288 |
| | North Carolina | .68 | .5 | .52 | .82 | .63 | .65 | .954*** | .798*** | .697*** | 500 |
| | Texas | .53 | .56 | .76 | .65 | .69 | .94 | .965*** | .808*** | .321*** | 225 |
| Modifier | Indiana | .61 | .52 | .58 | .71 | .65 | .70 | .969*** | .923*** | .857*** | 276 |
| | North Carolina | .78 | .72 | .63 | .92 | .89 | .81 | .958*** | .907*** | .794*** | 407 |
| | Texas | .78 | .6 | .78 | .88 | .73 | .92 | .970*** | .909*** | .794*** | 58 |

*Note*: We compare Indiana (Francis and Heise 2006), Texas (Schneider 2006), and North Carolina (Smith-Lovin and Heise 1978) with the affective dictionary used in this study (Smith-Lovin et al. 2016). A = activity; E = evaluation; MAD = mean absolute distance; P = potency; RMSD = root mean square distance.
***$p < .001$.

The tuning layers in our framework are trained on the basis of the target affective values. In training the neural network, we used early stopping to get reasonably low errors in the training and test sets. To find reasonably low estimation errors for the model, we compared mean absolute distance, root mean square distance, and the correlation between common concepts in different affective dictionaries.

To contextualize the accuracy metrics of our BERT-based model, we compare various English affective dictionaries. This comparison is crucial for establishing a benchmark for the expected variability in sentiment analysis across different sources and contexts. Table A1 compares the affective dictionary used in this study (Smith-Lovin et al. 2016) with some other affective dictionaries. These dictionaries, developed in distinct time periods and locations, exhibit inherent variations in sentiment assessments. By juxtaposing these variations with the performance of our model, we can gauge the model's reliability in mirroring the diversity and nuances present in human sentiment interpretation. Such a comparison is pivotal, as it aligns our model's accuracy with the inherent discrepancies found in traditional sentiment analysis tools. This approach underscores the robustness of our BERTNN model, particularly in replicating and potentially surpassing the variability observed in conventional affective dictionaries. The results in Table 2 further elucidate this point, demonstrating that our model's error rates and correlation metrics are comparable with the differences observed among established dictionaries. This evidence supports the assertion that our BERTNN model can estimate EPA values with an accuracy akin to conducting a new, comprehensive sentiment survey.

We actively checked L2loss for the training and test sets and had early stopping considering both values. We stopped the process after we used 457 batches of size 64 to train the model. After convergence, the trained model can predict the affective meaning for a large number of concepts.

**Acknowledgments**

We are grateful to the anonymous reviewers and editors for helpful feedback on this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr. Robinson's work was partly supported by ARO W911NF1710509.

## ORCID iDs

Moeen Mostafavi https://orcid.org/0000-0001-8101-5025
Dawn T. Robinson https://orcid.org/0000-0002-6809-0474

## Notes

1. BERTNN stands for "BERT Neural Network," indicating that our innovative method combines the original BERT model with additional fine-tuning neural network layers.
2. Readers interested in examining concrete examples of such sentiment dictionaries can refer to the repository at http://affectcontroltheory.org.
3. This approach is not uniform across all affect control theory measurements, but it is the classic approach and was used to collect all the data used in this article. The data used in the current article were collected using Attitude (Heise and Lewis 1988), Surveyor 2.0 (Heise 2001), or Surveyor 3.0 (Morgan and Morgan 2017). In these tools, EPA ratings were collected on bipolar scales ranging from −4.3 to 4.3 marked by labels of *neutral*, *slightly*, *quite*, *extremely*, and *infinitely*. Smith-Lovin (1987) conducted successive interval scaling on these measures to confirm that the size of the intervals between measurement points is uneven, with the distance between *extremely* and *infinitely* being 1.3 times the distance between each of the other markers.
4. David R. Heise, "Java INTERACT 2," distributed at https://affectcontroltheory.org.
5. The link to the tool is available at https://github.com/moeenm/BERTNN and https://affectcontroltheory.org/bertnn/.
6. Specifically, BERT large model (uncased), which is a 24-layer neural network with 1,024 hidden dimension and 16 attention heads. This model has 336 million parameters (Devlin et al. 2018).

## References

Alhothali, Areej, and Jesse Hoey. 2017. "Semi-supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations."arXiv. Retrieved June 9, 2024. https://arxiv.org/abs/1703.09825.

Arseniev-Koehler, Alina, and Jacob G. Foster. 2022. "Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What It Means to Be Fat."*Sociological Methods & Research* 51(4): 1484–1539.

Averett, Christine, and David R. Heise. 1987. "Modified Social Identities: Amalgamations, Attributions, and Emotions."*Journal of Mathematical Sociology* 13(1–2):103–32.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Pp. 610–23 in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada*. New York: Association for Computing Machinery.

Britt, Lory, and David R. Heise. 1992. "Impressions of Self-Directed Action."*Social Psychology Quarterly* 55(4):335–50.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."arXiv. Retrieved June 9, 2024. https://arxiv.org/abs/1810.04805.

Fontaine, Johnny R., Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2007. "The World of Emotions Is Not Two-Dimensional."*Psychological Science* 18(12):1050–57.

Francis, Clare, and David R. Heise. 2006. "Mean Affective Ratings of 1,500 Concepts by Indiana University Undergraduates in 2002–3."Distributed with program INTERACT at Affect Control Theory Website. https://affectcontroltheory.org//usa-indiana-2003/.

Francis, Linda E. 1997a. "Emotion, Coping, and Therapeutic Ideologies."*Social Perspectives on Emotion* 4:71–102.

Francis, Linda E. 1997b. "Ideology and Interpersonal Emotion Management: Redefining Identity in Two Support Groups."*Social Psychology Quarterly* 60(2):153–71.

Goldstein, David M. 1989. "Control Theory Applied to Stress Management."*Advances in Psychology* 62: 481–91.

Heise, David R. 1977. "Social Action as the Control of Affect."*Behavioral Science* 22(3):163–77.

Heise, David R. 1979. *Understanding Events: Affect and the Construction of Social Action*. New York: Cambridge University Press.

Heise, David R. 2001. "Project Magellan: Collecting Cross-Cultural Affective Meanings via the Internet."*Electronic Journal of Sociology* 5(3).

Heise, David R. 2007. *Expressive Order: Confirming Sentiments in Social Actions*. Berlin, Germany: Springer.

Heise, David R. 2010. *Surveying Cultures*. Hoboken, NJ: John Wiley.

Heise, David R. 2013. "INTERACT Guide." Retrieved June 9, 2023. https://kb-research.ca/archives/ACT/recovered/InteractGuide.pdf.

Heise, David R. 2020. "Java INTERACT 2." https://affectcontroltheory.org/resources-for-researchers/tools-and-software/interact/.

Heise, David R., and Cassandra Calhan. 1995. "Emotion Norms in Interpersonal Events."*Social Psychology Quarterly* 58(4):223–40.

Heise, David R., and Elsa Lewis. 1988. *Programs Interact and Attitude: Software and Documentation*. Dubuque, IA: William C. Brown.

Heise, David R., and Lisa Thomas. 1989. "Predicting Impressions Created by Combinations of Emotion and Social Identity."*Social Psychology Quarterly* 52(2):141–48.

Hunt, Pamela M. 2008. "From Festies to Tourrats: Examining the Relationship between Jamband Subculture Involvement and Role Meanings."*Social Psychology Quarterly* 71(4):356–78.

Joseph, Kenneth, Wei Wei, Matthew Benigni, and Kathleen M. Carley. 2016. "A Social-Event Based Approach to Sentiment Analysis of Identities and Behaviors in Text."*Journal of Mathematical Sociology* 40(3):137–66.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings."*American Sociological Review* 84(5):905–49.

Kriegel, Darys J., Muhammad Abdul-Mageed, Jesse K. Clark, Robert E. Freeland, David R. Heise, Dawn T. Robinson, Kimberly B. Rogers, and Lynn Smith-Lovin. 2017. "A Multilevel Investigation of Arabic-Language Impression Change."*International Journal of Sociology* 47(4):278–95.

Li, Minglei, Qin Lu, Yunfei Long, and Lin Gui. 2017. "Inferring Affective Meanings of Words from Word Embedding."*IEEE Transactions on Affective Computing* 8(4):443–56.

Liu, Yang, and Mirella Lapata. 2019. "Text Summarization with Pretrained Encoders." . Retrieved June 9, 2024. https://arxiv.org/abs/1908.08345.

MacKinnon, Neil J., and Alison Luke. 2002. "Changes in Identity Attitudes as Reflections of Social and Cultural Change."*Canadian Journal of Sociology/Cahiers Canadiens de Sociologie* 27(3):299–338.

MacKinnon, Neil J., and Dawn T. Robinson. 2014. "Back to the Future: 25 Years of Research in Affect Control Theory."*Advances in Group Processes* 31:139–73.

McCormick, Chris, and Nick Ryan. 2019. "BERT Word Embeddings Tutorial." Retrieved June 9, 2024. https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."arXiv. Retrieved June 9, 2024. https://arxiv.org/abs/1301.3781.

Morgan, Jonathan H., and Geoffrey P. Morgan. 2017. "New Surveyor 3.0: A Light-Weight Data Driven Approach for Capturing Affective Meaning." Retrieved June 9, 2024. https://affectcontroltheory.org/resources-for-researchers/tools-and-software/surveyor-3-0/.

Mostafavi, Moeen. 2021. "Adapting Online Messaging Based on Emotional State." In *Proceedings of the 29th Conference on User Modeling, Adaptation, and Personalization, Utrecht, Netherlands*. New York: Association for Computing Machinery.

Mostafavi, Moeen, and Michael D. Porter. 2021. "How Emoji and Word Embedding Helps to Unveil Emotional Transitions during Online Messaging." In *2021 IEEE International Systems Conference (SysCon)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. Retrieved June 9, 2024. https://ieeexplore.ieee.org/document/9447137.

Osgood, Charles E., William H. May, and Murray S. Miron. 1975. *Cross-Cultural Universals of Affective Meaning*. Urbana: University of Illinois Press.

Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Rashotte, Lisa Slattery. 2002. "Incorporating Nonverbal Behaviors into Affect Control Theory."*Electronic Journal of Sociology* 6(3).

Robillard, Julie M., and Jesse Hoey. 2018. "Emotion and Motivation in Cognitive Assistive Technologies for Dementia."*Computer* 51(3):24–34.

Robinson, Dawn T., Jody Clay-Warner, Christopher D. Moore, Tiffani Everett, Alexander Watts, Traci N. Tucker, and Chi Thai. 2012. "Toward an Unobtrusive Measure of Emotion during Interaction: Thermal Imaging Techniques." Pp. 225–66 in *Biosociology and Neurosociology*, Vol. 29: Advances in Group Processes, edited by W. Kalkhoff, S. R. Thye, and E. J. Lawler. Leeds, UK: Emerald.

Robinson, Dawn T., and Lynn Smith-Lovin. 1992. "Selective Interaction as a Strategy for Identity Maintenance: An Affect Control Model."*Social Psychology Quarterly* 55(1):12–28.

Robinson, Dawn T., and Lynn Smith-Lovin. 1999. "Emotion Display as a Strategy for Identity Negotiation."*Motivation and Emotion* 23(2):73–104.

Robinson, Dawn T., and Lynn Smith-Lovin. 2018. "Affect Control Theories of Social Interaction and Self." Pp. 139–65 in *Contemporary Social Psychological Theories*, 2nd ed., edited by P. J. Burke. Stanford, CA: Stanford University Press.

Robinson, Dawn T., Lynn Smith-Lovin, Bryan C. Cannon, Jesse K. Clark, Robert Freeland, Jonathan H. Morgan, and Kimberly B. Rogers. 2016. "Mean Affective Ratings of 932 Identities, 810 Behaviors, and 660 Modifiers by University of Georgia Undergraduates in 2012–2014." https://affectcontroltheory.org/usa-georgia-dictionary-2015/.

Robinson, Dawn T., Lynn Smith-Lovin, and Olga Tsoudis. 1994. "Heinous Crime or Unfortunate Accident? The Effects of Remorse on Responses to Mock Criminal Confessions."*Social Forces* 73(1): 175–90.

Rogers, Kimberly B. 2018. "Do You See What I See? Testing for Individual Differences in Impressions of Events."*Social Psychology Quarterly* 81(2):149–72.

Rogers, Kimberly B., and Lynn Smith-Lovin. 2019. "Action, Interaction, and Groups." Pp. 67–86 in *The Wiley Blackwell Companion to Sociology*, edited by G. Ritzer, and W. W. Murphy. Hoboken, NJ: Wiley Blackwell.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Daniel Jurafsky, Noah A. Smith, and Yejin Choi. 2019. "Social Bias Frames: Reasoning about Social and Power Implications of Language." . Retrieved June 9, 2024. https://arxiv.org/abs/1911.03891.

Sap, Maarten, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, Yejin Choi, Martha Palmer, Rebecca Hwa, and Sebastian Riedel. 2017. "Connotation Frames of Power and Agency in Modern Films." Pp. 2329–34 in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Stroudsburg, PA: Association for Computational Linguistics.

Schneider, Andreas. 2006. "Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998." https://affectcontroltheory.org/usa-texas-1998/.

Schröder, Tobias, and Wolfgang Scholl. 2009. "Affective Dynamics of Leadership: An Experimental Test of Affect Control Theory."*Social Psychology Quarterly* 72(2):180–97.

Smith, Herman W. 2002. "The Dynamics of Japanese and American Interpersonal Events: Behavioral Settings versus Personality Traits."*Journal of Mathematical Sociology* 26(1–2):71–92.

Smith, Herman W., and Linda E. Francis. 2005. “Social vs. Self-Directed Events among Japanese and Americans.”*Social Forces* 84(2):821–30.

Smith-Lovin, Lynn. 1979. “Behavior Settings and Impressions Formed from Social Scenarios.”*Social Psychology Quarterly* 42(1):31–43.

Smith-Lovin, Lynn. 1987. “Impressions from Events.”*Journal of Mathematical Sociology* 13(1–2):35–70.

Smith-Lovin, Lynn, and William Douglass. 1992. “An Affect Control Analysis of Two Religious Subcultures.”*Social Perspectives in Emotions* 1:217–47.

Smith-Lovin, Lynn, and David R. Heise. 1978. “Mean Affective Ratings of 2,106 Concepts by University of North Carolina Undergraduates in 1978” [computer file]. *Distributed with program INTERACT at Affect Control Theory Website*. https://affectcontroltheory.org/usa-north-carolina-1978/.

Smith-Lovin, Lynn, Dawn T. Robinson, Bryan C. Cannon, Jesse K. Clark, Robert Freeland, Jonathan H. Morgan, and Kimberly B. Rogers. 2016. “Mean Affective Ratings of 929 Identities, 814 Behaviors, and 660 Modifiers by University of Georgia and Duke University Undergraduates and by Community Members in Durham, NC, in 2012–2014.” https://affectcontroltheory.org/usa-combined-surveyor-dictionary-2015/.

Tsoudis, Olga, and Lynn Smith-Lovin. 1998. “How Bad Was It? The Effects of Victim and Perpetrator Emotion on Responses to Criminal Court Vignettes.”*Social Forces* 77(2):695–722.

van Loon, Austin, and Jeremy Freese. 2023. “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language.”*American Behavioral Scientist* 67(2):175–200.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” Pp. 5998–6008 in *Proceedings of the Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA. Red Hook, NY: Curran Associates.

Youngreen, Reef, Bridget Conlon, Dawn T. Robinson, and Michael J. Lovaglia. 2009. “Identity Maintenance and Cognitive Test Performance.”*Social Science Research* 38(2):438–46.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books.” Pp. 19–27 in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

## Author Biographies

**Moeen Mostafavi** received a PhD in systems engineering from the University of Virginia. He is interested in computational social science research questions pertaining to decision making. Understanding emotions from textual conversations was the focus of his dissertation.

**Michael D. Porter** is an associate professor of data science and systems and information engineering at the University of Virginia. His research focuses on modeling and forecasting human behavior and social dynamics, identifying patterns and anomalies in sequential data, and data linkage.

**Dawn T. Robinson** is a professor of sociology at the University of Georgia. Her recent research includes cross-cultural work on affective-linguistic culture, computational models of cross-cultural interactions, developing thermographic measures of emotion during social interaction, studying emotional responses to injustice, and analyzing the social networks of children and adolescents.