LEARNING AFFECTIVE MEANINGS THAT DERIVES THE SOCIAL BEHAVIOR USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cultural sentiments of a society characterize social behaviors, but modeling sentiments to manifest every potential interaction remains an immense challenge. Affect Control Theory (ACT) offers a solution to this problem. ACT is a generative theory of culture and behavior based on a three-dimensional sentiment lexicon. Traditionally, the sentiments are quantified using survey data which is fed into a regression model to explain social behavior. The lexicons used in the survey are limited due to prohibitive cost. This paper uses a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for developing a replacement for these surveys. This model achieves state-of-the-art accuracy in estimating affective meanings, expanding the affective lexicon, and allowing more behaviors to be explained.

1 INTRODUCTION

Consider talking to your mentor for some advice about how to *behave* with your colleague. Your mentor probably starts asking questions about the *culture* in the workspace and may continue asking about the *identity* of the person. These questions could be about *institutional constraints* such as being the manager, or they could be personality *sentiment* such as being nice or active. Knowing all this information, your mentor may have some initial recommendations, but you *adapt* your behavior after observing the reactions from the colleague. This is a descriptive scenario for a daily interaction. Affect Control Theory (ACT) quantifies the variables in this scenario and produces equations for simulating human behavior in different scenarios.

ACT was introduced in the 1970s (Heise, 1977) and has been validated in more than 100 social science research (Robinson & Smith-Lovin, 2018). Recently ACT has been used in applications such as Human-Computer Interactions (Robillard & Hoey, 2018). ACT uses a three-dimensional affective meaning space as a quantized form of meaning (Heise, 1977).

ACT uses Evaluation [good vs. bad], Potency [powerful vs. powerless], and Activity [active vs. passive] (EPA) space introduced by Osgood et al. (1957) as a semantic differential form of affective meaning. These affective lexicons represent the word of interest within cultural and social boundaries (Robinson & Smith-Lovin, 1992). Fontaine et al. (2007) reduced dimensionality from 144 features representing the main components of emotions, and the first three principal components were EPA space.

A survey study can measure these values within a cultural group. To compensate for unreliability in the survey, each rating should be replicated at least 25 times. Other than baseline sentiments, ACT needs a sentiment rating after observing social interactions to model the behavior. Finding these affective meanings is a challenging task, and it limits the applicability of ACT in many applications. As an alternative approach, researchers used supervised (Mostafavi & Porter, 2021; Li et al., 2017) and semi-supervised (Alhothali & Hoey, 2017) methods on shallow word-embeddings to find affective lexicons. However, their result specifically in *activity* and *potency* dimensions needs improvement. In this paper, we discuss the main limitations for the applicability of shallow word-embeddings. Then we discuss how to generate a data-set to pass to a deep neural network. We use Bidirectional Encoder Representations from Transformers (BERT) embedding and fine-tune it for

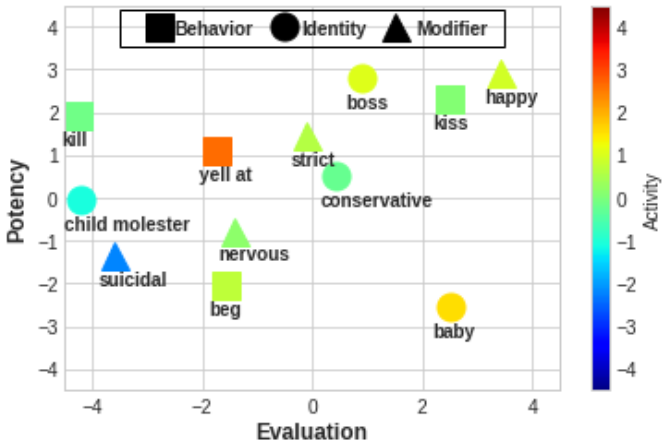

Figure 1: Sentiments of sample words in EPA space. Circles, squares, and triangles show identities, behaviors, and modifiers. The color represents the activity dimension.

finding affective lexicons. The result of this paper achieves state-of-the-art accuracy in estimating affective lexicons. This result is within the accepted range of uncertainty in the ACT literature.

2 Affect Control Theory

Social interactions or events at least include an *actor* that *behaves* an *object*. These Actor-Behavior-Object (ABO) are the first pieces of modeling interactions (Heise, 2010). The actor/object in an event may have a modifier specifying their characteristics. The baseline EPA representation of words is known as *sentiments*. Sentiments for some sample words are plotted in Figure 1. In this plot, we can observe 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 active word. Note the range of EPA ratings is from 4.3 to 4.3.

Social events have impressions on ABO characters. Experimental tests explain what contributing facts derive the impressions of events (Robinson & Smith-Lovin, 1999). Consider an example of observing "a bossy employer argues with an employee". This observation makes people evaluate both the *actor* and the *object* of the interaction are less pleasant than initial thoughts. On the other hand, they may feel the employer is more powerful, and the employee is more powerless than their baseline sentiments after this event.

Deflection is the 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 initial sentiments, deflection is small, and it gets bigger when the impression of the event leaves the initial sentiments. Minimizing *deflection* is the driving force in human activity. Highly deflecting events create social and physiological distress (Goldstein, 1989). For example, if a grandmother fights with her grandchild, they feel distressed and prefer to do something. We expect one side to take an action, like apologize, to bring the impression of their identities back to where society views them. This event is very different from two soldiers fighting in a battle. The soldiers are supposed to fight with enemies in battle, so they do not feel social pressure.

Heise (2013) developed a software called INTERACT. It simulates these types of interaction. For a simple simulation, we simulated the following scenario using INTERACT,

1. Employee greets bossy employer.
2. Bossy employer asks employee.
3. Employee replies to bossy employer.
4. Bossy employer argues with employee.
5. Employee listens to / disobeys bossy employer.

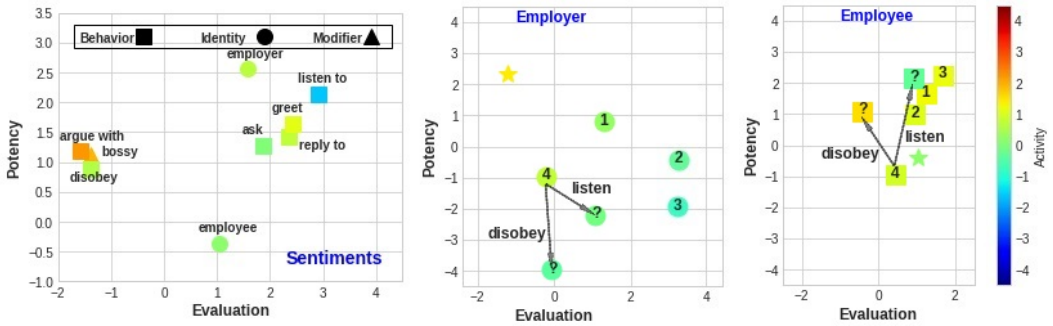

Figure 2: Simulating sequential events in an interaction between an employee and employer. The initial sentiments for both characters are shown by star signs and the sentiments after each of four events are shown by numbers. After the fourth event, based on what the employee does, the final sentiment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

The above scenario is visualized in Figure 2. For the fifth event, we have shown how the impression of different actions by the employer has significantly moved the states for both the actor and the object. This is a simplified scenario of our mentorship example in the beginning. It shows how understanding the interaction dynamic can make us predict the consequences of our behavior.

We can explain impression changes in a descriptive form such as the following two ones,

- *Actors* seem nice when they behave in a positive way toward others. This describes *morality* fact in ACT literature. Observing the *evaluation* dimension for the *actors* after he greets the *object*, we can find this *behavior* resulted in an impression of being nicer (getting larger evaluation) comparing to the state in the last step.
- Active *behaviors* make the *actors* seem more active. Observing the boss's activity, he is considered more active after he argues with [active behavior] the employee.

As we have seen in the descriptive analyses, events change the impression of ABO in EPA space. They move them toward or away from their sentiments. ACT uses regression models, known as impression change equations, to estimate these changes. These models resulted from surveys that asked people to evaluate ABO in EPA space after events (Heise, 2013). Let $X = [A_e, A_p, A_a, B_e, B_p, B_a, O_e, O_p, O_a]^T$, $X^2 = [A_eB_e, A_eO_e, A_eB_a, \ldots A_aO_a]^T$, and $X^3 = [A_eB_eO_e, A_eB_eO_p, A_eB_eO_a, \ldots A_aB_aO_a]^T$ then we can write basic impression change equations in the following format,

$$X' = \alpha X + \beta X^2 + \gamma X^3 \tag{1}$$

where α, β, and γ are constant matrices, $\{A, B, O\}$, represent the ABO characters, and $\{e, p, a\}$ represent EPA components before the event, and X' represent the resulting impression after the event. Equation (1) models the basic interaction scenario. When we have modified or the interaction happens in a specific social institution, we can use similar approaches to model it. When we have modified identities, a bossy employer for example, *amalgamation equations* similar to (2) gives the sentiments for an identity with a modifier (Averett & Heise, 1987).

$$C_e = -17 + 0.62M_e - 0.14M_p - 0.18M_a + 0.5I_e \tag{2}$$

where, C_e, M_e, M_p, M_a, I_e represent evaluation of the modifier-identity, evaluation of the modifier, potency of the modifier, activity of the modifier, and evaluation of the identity respectively. Equation (2) is a weighted average of the evaluation for the modifier and the identity. Similarly, *potency* and *activity* of the modifier affect the evaluation of the modifier-identity (Averett & Heise, 1987).

3 Related Works

Alhothali & Hoey (2017) used graph-based sentiment lexicon induction methods to find affective lexicons associated with words. They used similarity graphs to expand affective meanings

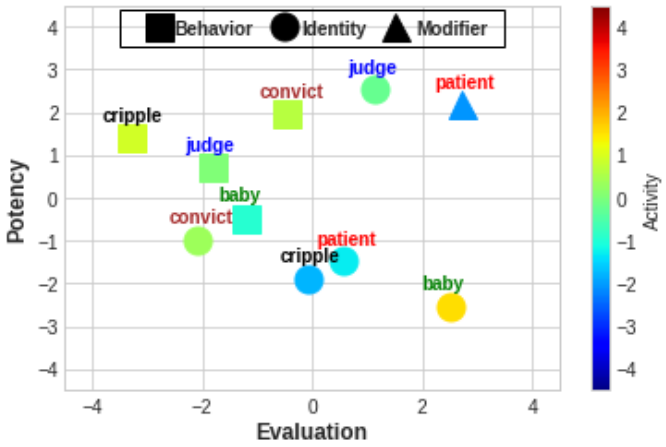

Figure 3: Visualization of words with different affective meaning in EPA space. Circles, squares, and triangles show identities, behaviors, and modifiers. The color represents the activity dimension.

to neighbor words in four different embedding spaces. They found that using both semantics and distributional-based approaches gives the best semi-supervised result close to supervised algorithms.

Li et al. (2017) argued that word-embedding represents the words' general meaning, including denotative meaning, connotative meaning, social meaning, affective meaning, reflected meaning, collocative meaning, and thematic meaning. However, ACT is developed only based on affective meanings of the words. So word-embedding graph propagation considers similarities that reduce accuracy for finding affective meaning.

Kozlowski et al. (2019) used word analogy to find cultural dimensions from embedding. Mostafavi & Porter (2021); Li et al. (2017) used supervised methods on shallow word-embeddings to find affective meanings of the words. However, all these approaches are limited by using one representation for different meanings of a word.

The words in affective dictionaries are categorized as identity, modifier, and behavior. The words are mapped to different points in EPA space when they are evaluated in each category. On the other hand, if a word has two or more meanings or different parts of speech, it still has only one representation in the embedding space. So, for example, "mother", "coach", and "fool" appear in both categories of behavior and identities, but they have only one representation in embedding space. In Figure 3 we can observe how some words are mapped very differently based on their category. For example, "baby" as an *identity* is a pleasant and active character but is unpleasant and passive as a *behavior*.

To find how the affective meaning of words in different categories are related, we find the correlation between EPA values of shared words across various categories in Table 1. From these tables, we can have observations such as, *activity* and *potency* dimensions of common words between *identity* and *behavior* are too small to assume they represent the same affective meanings. So we need a model that can represent contextual aspects and differentiate between different meanings of a word.

Table 1: Correlation between the affective meaning of words in identity, behavior, and modifier category of a dictionary collected in 2014 (Smith-Lovin et al., 2016).

Identity-Modifier			
	E	P	A
E	0.93	0.49	-0.58
P	0.77	0.62	-0.39
A	-0.45	0.33	0.98

Modifier-Behavior			
	E	P	A
E	0.98	0.92	-0.32
P	0.85	0.8	-0.25
A	-0.27	-0.3	0.67

Identity-Behavior			
	E	P	A
E	0.73	0.35	0.4
P	0.29	0.55	0.02
A	-0.11	0.3	0.4

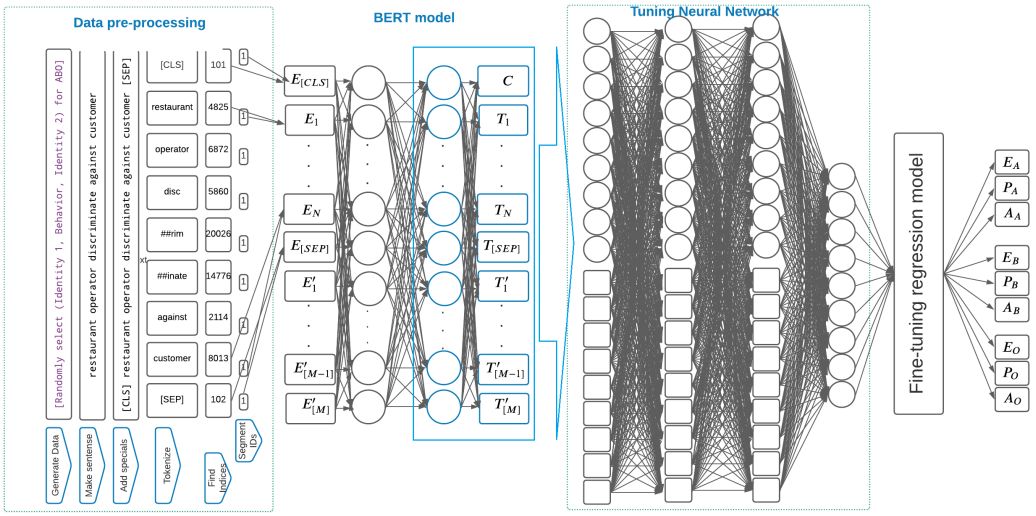

Figure 4: BERTNN pipeline. Sentences describing an event are generated in the first step and pre-processed to pass to a BERT model. Then the last two layers of the BERT model are concatenated to make contextual word-embedding. This embedding is passed to a three-layer neural network to find affective meaning, and finally, the output of the neural network is fine-tuned using a regression model.

We can observe in Table 1 that affective meanings of the words in different categories are not necessarily highly correlated. For example, there is only a 0.4 correlation between *activity* dimensions of words that appear both in *identity* and *behavior* categories.

4 Bidirectional Encoder Representations from Transformers

In 2018, Google open-sourced BERT as the state-of-the-art pre-training model for Natural Language Processing (NLP) (Devlin et al., 2018). This model is designed to be pre-trained on unlabeled data and uses both left and right context in all layers. BERT gives state of the art in many challenging NLP tasks (McCormick & Ryan, 2019). BERT has been pre-trained on Wikipedia and book corpora that includes more than 10000 books. We use the base version of BERT in this study.

The main advantage of word-embeddings derived from BERT over shallow word-embeddings is being context-based embeddings. So we can get different representations when we have a word as a *behavior* or an identity. But we have to feet the dataset so that it can understand in which category we are working.

To use BERT as a contextual word-embedding we should have the following processing (McCormick & Ryan, 2019),

1. BERT requires two special tokens in a sentence. First, [CLS] represents the start of a sentence. Special token [SEP] comes at the end of the sentence when we are only passing one sentence.
2. BERT tokenizer tokenizes the sentence. This tokenizer assigns multiple tokens to a word if the word is not in its vocabulary. In that case, tokens are WordPieces, prefix, and suffix are sub-words.
3. Token IDs are assigned from the tokenizer.
4. BERT always expects two sentences. Since we are passing only one sentence, the sentence ID is set to 1 for all the tokens in the sentence.

In Figure 4 you can find how a sentence is processed before passing to a BERT model.

Table 2: Comparing the result of this paper with commonly used approaches in the literature.

	Identity									Behavior								
	MAE			RMSE			Cor.			MAE			RMSE			Cor.		
	E	P	A	E	P	A	E	P	A	E	P	A	E	P	A	E	P	A
Word analogy 1&2 + Stepwise Regression	0.94	0.92	0.82	1.22	1.11	1.02	0.63	0.55	0.12	1.13	0.71	0.68	1.42	0.88	0.84	0.74	0.46	0.5
Word analogy + Linear Regression	0.95	0.95	0.81	1.22	1.16	1	0.64	0.48	0.15	1.2	0.75	0.73	1.48	0.93	0.9	0.71	0.44	0.38
Word analogy + Stepwise Regression	0.93	0.92	0.81	1.2	1.13	0.99	0.65	0.53	0.18	1.17	0.71	0.68	1.44	0.9	0.83	0.73	0.45	0.51
Translation matrix + Stepwise Regression	0.58	0.63	0.57	0.8	0.83	0.72	0.85	0.78	0.68	0.8	0.56	0.58	1.02	0.74	0.73	0.87	0.66	0.67
BERTNN	**0.42**	**0.38**	**0.28**	**0.51**	**0.47**	**0.37**	**0.93**	**0.93**	**0.93**	**0.59**	**0.41**	**0.27**	**0.75**	**0.5**	**0.36**	**0.94**	**0.86**	**0.93**

Table 3: Comparing estimated affective meanings for "judge" as an *identity* and *behavior*.

Beh	EEB	EPB	EAB	EB	PB	AB	Ident	EEA	EPA	EAA	EA	PA	AA
judge	-2.03	0.98	0.24	-1.83	0.71	0.07	judge	1.57	1.56	-0.42	1.15	2.53	-0.22

5 METHOD

As we discussed earlier, BERT can be a contextual embedding space. There are a couple of different approaches to find word and sentence embedding from the BERT model. We discovered that concatenating the last two layers gives a sentence embedding that results in our task's state of the art.

We should process the data to use the BERT for extending affective dictionaries. The following steps show how we use BERT model to extend affective dictionaries,

1. One affective dictionary that includes both *identities* and *behaviors* is selected. 85% of the data is used for the training and 15% for the test.
2. Two identities and one behavior are randomly selected from the train set. They are combined to make ABO form.
3. ABO sentences are pre-processed and passed to the pre-trained BERT model.
4. The last two layers of the BERT model are concatenated to make a sentence embedding.
5. A four-layer neural network is used to find the 9-dimensional vector of EPA values for ABO characters. The first three layers have 9 *dense* and 9 *Tanh* parts. The last layer is a dense layer that makes the 9-dimensional vector.
6. A regression model is used to fine-tune the neural network.

The pipeline for our method (BERTNN) is shown in Figure 4. This pipeline was finalized after we iterated on hyperparameters with different scenarios.

6 RESULTS

We implemented different word analogy, regressions, and translation matrix methods to find affective meanings, and their results are shown in Table 2. We used Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation analysis to compare the result.

From Table 2, we can observe that our approach reached state of the art across all the metrics. To evaluate how close our extended dictionaries are to the baseline affective meanings, we did a correlation analysis in Tables 4 and 5. In these figures, we find the correlation (A) between the result of the test set in our method and the EPA values from the surveys and (B) the correlation for the estimation and survey results.

One problem with estimation from shallow-embeddings was having the same embedding for words as *identity* or *behavior*. Using the last two layers of the BERT model as embedding, we can differentiate between these two cases. In Table 3 we can observe that estimated values for "judge" are different when it is considered as *identity* or *behavior*.

Table 4: Correlation analysis for identities. (a) Correlation between estimated values and the affective dictionary value. We can also compare (b) and (c) to find how close the off-diagonal entries are in the estimation comparing to dictionary values.

	EA	PA	AA
EEA	0.93	0.57	0.06
EPA	0.57	0.93	0.31
EAA	0.04	0.25	0.93

(a)

	EA	PA	AA
EA	1.00	0.55	0.05
PA	0.55	1.00	0.25
AA	0.05	0.25	1.00

(b)

	EEA	EPA	EAA
EEA	1.00	0.62	0.06
EPA	0.62	1.00	0.33
EAA	0.06	0.33	1.00

(c)

Table 5: Correlation analysis for behaviors. (a) Correlation between estimated values and the affective dictionary value. We can also compare (b) and (c) to find how close the off-diagonal entries are in the estimation comparing to dictionary values.

	EB	PB	AB
EEB	0.94	0.55	-0.23
EPB	0.59	0.86	0.13
EAB	-0.25	0.11	0.93

(a)

	EB	PB	AB
EB	1.00	0.59	-0.21
PB	0.59	1.00	0.12
AB	-0.21	0.12	1.00

(b)

	EEB	EPB	EAB
EEB	1.00	0.63	-0.26
EPB	0.63	1.00	0.12
EAB	-0.26	0.12	1.00

(c)

Tables 4 and 5 show that diagonal terms in the correlation of estimated values and values from the survey dictionary are reasonably large. On the other hand, the off-diagonal entries from the estimation are close to the ones from the survey.

7 Concluding Remarks

A big limitation in the applicability of ACT to different contexts such as NLP tasks was the limitation of affective vocabulary collected from surveys. In this paper, an approach to make training and test sentences that describe an ABO event. Then we used BERT as an embedding and fine-tuned it using a neural network and regression model. Our approach, BERTNN, resulted in state of the art in estimating affective meaning.

References

Areej Alhothali and Jesse Hoey. Semi-supervised affective meaning lexicon expansion using semantic and distributed word representations. *arXiv preprint arXiv:1703.09825*, 2017.

Christine Averett and David R Heise. Modified social identities: Amalgamations, attributions, and emotions. *Journal of Mathematical Sociology*, 13(1-2):103–132, 1987.

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

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

David M Goldstein. Control theory applied to stress management. In *Advances in Psychology*, volume 62, pp. 481–491. Elsevier, 1989.

David R Heise. Social action as the control of affect. *Behavioral Science*, 22(3):163–177, 1977.

David R Heise. *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons, 2010.

David R Heise. Interact guide. *Department of Sociology, Indiana University*, 2013.

Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456, 2017.

Chris McCormick and Nick Ryan. Bert word embeddings tutorial, May 2019. URL `http://www.mccormickml.com`.

Moeen Mostafavi and Michael Porter. How emoji and word embedding helps to unveil emotional transitions during online messaging. In *2021 IEEE International Systems Conference (SysCon)*. IEEE, 2021.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

Julie M Robillard and Jesse Hoey. Emotion and motivation in cognitive assistive technologies for dementia. *Computer*, 51(3):24–34, 2018.

Dawn T Robinson and Lynn Smith-Lovin. Selective interaction as a strategy for identity maintenance: An affect control model. *Social Psychology Quarterly*, pp. 12–28, 1992.

Dawn T Robinson and Lynn Smith-Lovin. Emotion display as a strategy for identity negotiation. *Motivation and Emotion*, 23(2):73–104, 1999.

Dawn T Robinson and Lynn Smith-Lovin. Affect control theories of social interaction and self. 2018.

Lynn Smith-Lovin, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, Robert Freeland, Jonathan H Morgan, and Kimberly B Rogers. Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014. *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*, 2016.