Enhancing Emotion Understanding in Messaging Interactions

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Moeen Mostafavi

May 2023

APPROVAL SHEET

This
Dissertation
is submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Author: Moeen Mostafavi

This Dissertation has been read and approved by the examing committee:

Advisor: Michael D. Porter
Committee Member: William T. Scherer
Committee Member: Peter A Beling
Committee Member: Rich Nguyen
Committee Member: Dawn T. Robinson

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2023

ABSTRACT

This dissertation explores the challenges of understanding emotions in messaging conversations with task-oriented conversational agents and presents a novel approach to address this issue using a combination of natural language processing (NLP) methods and Affect Control Theory (ACT). The lack of nonverbal cues and the ambiguous nature of written language makes understanding emotions in messaging conversations a complex task. However, the ability to understand and respond to emotions can significantly improve the effectiveness and satisfaction of communication between humans and conversational AI systems.

The research has three main components. The first phase demonstrates a proof of concept of using ACT in the context of messaging with a task-oriented conversational agent. The second phase addresses the limitations of traditional affective dictionaries used in ACT by utilizing the capabilities of BERT, a pre-trained transformer-based model. The final phase focuses on recognizing emotions during conversations with a task-oriented conversational agent by utilizing both the sequential and contextual aspects of the messages.

The results showed that combining NLP methods with ACT can provide a more accurate and complete understanding of the emotions conveyed in messaging interactions. The study also achieves state-of-the-art results in estimating emotions by implementing an encoder-decoder network with attention. This approach can lead to improved emotional intelligence in conversations with chatbots and result in enhanced customer satisfaction and more natural and effective conversational AI systems in various domains such as customer service, healthcare, and education. Overall, this research contributes to the field of conversational AI by making strides in understanding emotions in messaging interactions and applying NLP and ACT to improve conversational AI systems.

“Seek knowledge from the cradle to the grave."

—Prophet Mohammad

CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to everyone who has supported me throughout this journey. First and foremost, I want to thank my wife, Fateme, who was with me every minute of this journey. Her unwavering support, patience, and encouragement have been the cornerstone of my success.

I also want to express my gratitude to my Mom and my Dad; they mean everything in my life. Their endless love and support have been my guiding light throughout my life. Additionally, I want to thank my brother and sisters that not seeing them for more than six years made this journey hard. I know they have always been there for me, even if we were miles apart.

I am also deeply grateful to my advisor, who was a great mentor in academic life and a great man I met. His guidance, insights, and encouragement have been invaluable to me. I have learned so much from him, not just about academic work but also about life.

Finally, I want to extend my thanks to my friends and other family members who have always believed in me and supported me in their own ways. Without all of you, this journey would not have been possible.

Thank you all from the bottom of my heart.

CHAPTER 1

INTRODUCTION

Comprehending emotions conveyed in messaging conversations poses a challenge as it is complicated by the absence of nonverbal cues and the ambiguity of written language. This difficulty is significantly pronounced when communicating with chatbots designed for a specific task, where the flexibility of free conversation is often lacking. Despite this, being able to interpret and react to emotions in messaging interactions can significantly impact the effectiveness and satisfaction of communication between people and conversational AI systems.

My research aims to overcome the challenges of understanding emotions in messaging conversations by using a combination of natural language processing (NLP) methods, such as Bidirectional Encoder Representations from Transformers (BERT) and Affect Control Theory (ACT). ACT is a social psychological theory that describes how people evaluate and respond to the emotions and actions of others. It specifically considers how social interactions modify and shape perceptions of people, objects, and actions. This framework leads to predictive models that can, for example, identify behaviors that can reduce social distress during social interactions.

My research has three main components:

- The first phase of my research demonstrates a proof of concept of using ACT in the context of messaging with a conversational agent. I apply the basic grammar of ACT, which includes actor, behavior, and object elements, to a conversation between a human and a chatbot. This work shows how emojis could provide additional information as modifiers and develops a model that detects and tracks emotional changes during the conversation.

- In the second phase, I aim to overcome the limitation of traditional affective dictionaries used in ACT by utilizing the capabilities of BERT, a pre-trained transformer-based model. Due to the time and cost of running large surveys, affective dictionaries are often limited to less than 3,000 words. I make a synthetic dataset representing common events in ACT grammar. Then I fine-tune the BERT model to estimate the affective meaning of modifiers, behaviors, and identities in those events. BERT acts as a sentence-embedding model for these event sentences, allowing for enhanced affective dictionaries that can more accurately explain behaviors and predict impressions of events in the ACT.

- In the final phase, I focus on recognizing emotions during conversations with a chatbot. Building on the insights from ACT, I use both the sequential and contextual aspects of the messages to understand the emotions being conveyed. I model the conversation as an interaction between the customer and the chatbot. Utilizing these ACT-based insights, I achieve state-of-the-art results in estimating emotions by implementing an encoder-decoder network with attention. This network effectively captures the nuances of the conversation and allows for a more accurate understanding of the emotions being conveyed.

Through these efforts, I show that combining NLP methods with ACT can provide a more accurate and complete understanding of the emotions conveyed in messaging interactions. This can lead to improved emotional intelligence in conversations with chatbots, resulting in enhanced customer satisfaction and more natural and effective conversational AI systems in various domains such as customer service, healthcare, and education. Overall, my research make strides in understanding emotions in messaging interactions and applying NLP and ACT to improve conversational AI systems.

1.1 References

Adamopoulou, Eleni and Lefteris Moussiades (2020). "An overview of chatbot technology". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 373–383.

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

Andre, Elisabeth et al. (2004). "Endowing spoken language dialogue systems with emotional intelligence". In: *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer, pp. 178–187.

Averett, Christine and David R Heise (1987). "Modified social identities: Amalgamations, attributions, and emotions". In: *Journal of Mathematical Sociology* 13.1-2, pp. 103–132.

Barbieri, Francesco, Francesco Ronzano, and Horacio Saggion (2016). "What does this emoji mean? a vector space skip-gram model for twitter emojis". In: *Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72.* ELRA (European Language Resources Association).

Beattie, Austin, Autumn P Edwards, and Chad Edwards (2020). "A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication". In: *Communication Studies* 71.3, pp. 409–427.

Britt, Lory and David R Heise (1992). "Impressions of self-directed action". In: *Social Psychology Quarterly*, pp. 335–350.

Budzianowski, Paweł et al. (2018). "MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling". In: *arXiv preprint arXiv:1810.00278*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dimson, Thomas (2015). "Emojineering part 1: Machine learning for emoji trends". In: *Instagram Engineering Blog* 30.

Eisner, Ben et al. (2016). "emoji2vec: Learning emoji representations from their description". In: *arXiv preprint arXiv:1609.08359*.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition and Emotion*, pp. 169–200.

Eric, Mihail et al. (2019). "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines". In: *arXiv preprint arXiv:1907.01669*.

Feng, Shutong et al. (2022). "EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fonnegra, Rubén D and Gloria M Dıaz (2018). "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model". In: *International Conference on Human-Computer Interaction*. Springer, pp. 385–396.

Fontaine, Johnny RJ et al. (2007). "The world of emotions is not two-dimensional". In: *Psychological science* 18.12, pp. 1050–1057.

Francis, Clare and David R Heise (2006). "Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3, 2006". In: *Computer file]. Distributed at Affect Control Theory Website, Program Interact (http://www. indiana. edu/˜ socpsy/ACT/interact/JavaInteract. html).*

Francis, Linda E (1997a). "Emotion, coping, and therapeutic ideologies". In: *Social perspectives on emotion* 4, pp. 71–102.

— (1997b). "Ideology and interpersonal emotion management: Redefining identity in two support groups". In: *Social Psychology Quarterly*, pp. 153–171.

Ghosal, Deepanway et al. (2020). "Cosmic: Commonsense knowledge for emotion identification in conversations". In: *arXiv preprint arXiv:2010.02795*.

Gliwa, Bogdan et al. (2019). "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization". In: *arXiv preprint arXiv:1911.12237*.

Goldstein, David M (1989). "Control theory applied to stress management". In: *Advances in Psychology*. Vol. 62. Elsevier, pp. 481–491.

Heise, David R (1977). "Social action as the control of affect". In: *Behavioral Science* 22.3, pp. 163–177.

— (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

— (2013). "Interact guide". In: *Department of Sociology, Indiana University*.

Heise, David R and Cassandra Calhan (1995). "Emotion norms in interpersonal events". In: *Social Psychology Quarterly*, pp. 223–240.

Heise, David R and Lisa Thomas (1989). "Predicting impressions created by combinations of emotion and social identity". In: *Social Psychology Quarterly*, pp. 141–148.

Hunt, Pamela M (2008). "From festies to tourrats: Examining the relationship between jamband subculture involvement and role meanings". In: *Social Psychology Quarterly* 71.4, pp. 356–378.

Kelley, John F (1984). "An iterative design methodology for user-friendly natural language office information applications". In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41.

Koch, Andrew, Jiahao Tian, and Michael D Porter (2020). "Criminal Consistency and Distinctiveness". In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–3.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). "The geometry of culture: Analyzing the meanings of class through word embeddings". In: *American Sociological Review* 84.5, pp. 905–949.

Kriegel, Darys J et al. (2017). "A multilevel investigation of Arabic-language impression change". In: *International Journal of Sociology* 47.4, pp. 278–295.

Li, Minglei et al. (2017). "Inferring affective meanings of words from word embedding". In: *IEEE Transactions on Affective Computing* 8.4, pp. 443–456.

Li, Shan and Weihong Deng (2020). “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing*.

Li, Yanran et al. (2017). “Dailydialog: A manually labelled multi-turn dialogue dataset”. In: *arXiv preprint arXiv:1710.03957*.

Liu, Yang and Mirella Lapata (2019). “Text summarization with pretrained encoders”. In: *arXiv preprint arXiv:1908.08345*.

Loon, Austin van and Jeremy Freese (2022). “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language”. In: *American Behavioral Scientist*. DOI: 10.1177/00027642211066046. eprint: https://doi.org/10.1177/00027642211066046.

MacKinnon, Neil J and Dawn T Robinson (2014). “Back to the future: 25 years of research in affect control theory”. In: *Advances in group processes*.

Majumder, Navonil et al. (2019). “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6818–6825.

McCormick, Chris and Nick Ryan (May 2019). *BERT Word Embeddings Tutorial*. URL: http://www.mccormickml.com.

Mikolov, Tomas, Kai Chen, et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. In: *arXiv preprint arXiv:1309.4168*.

Mostafavi, Moeen (2021). “Adapting Online Messaging Based on Emotiona”. In: *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization*.

Mostafavi, Moeen, Maria Phillips, et al. (2021). “A tale of two metrics: Polling and financial contributions as a measure of performance”. In: *2021 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.

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

Mostafavi, Moeen, Michael D Porter, and Dawn T Robinson (2022). “Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers”. In: *arXiv preprint arXiv:2202.00065*.

Mostafavi, Moeen, Mahsa Pahlavikhah Varnosfaderani, et al. (2022). “emojiSpace: Spatial Representation of Emojis”. In: *arXiv preprint arXiv:2209.09871*.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CBO9780511571299.

Osgood, Charles Egerton, William H May, et al. (1975). *Cross-cultural universals of affective meaning*. Vol. 1. University of Illinois Press.

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

Polzin, Thomas S and Alexander Waibel (2000). “Emotion-sensitive human-computer interfaces”. In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Poria, Soujanya et al. (2018). “Meld: A multimodal multi-party dataset for emotion recognition in conversations”. In: *arXiv preprint arXiv:1810.02508*.

Quinonero-Candela, Joaquin et al. (2008). *Dataset shift in machine learning*. Mit Press.

Rashotte, Lisa Slattery (2002). “Incorporating nonverbal behaviors into affect control theory”. In: *Electronic Journal of Sociology* 6.3.

Reelfs, Jens Helge et al. (2020). “Word-Emoji Embeddings from large scale Messaging Data reflect real-world Semantic Associations of Expressive Icons”. In: *arXiv preprint arXiv:2006.01207*.

Robillard, Julie M and Jesse Hoey (2018). “Emotion and motivation in cognitive assistive technologies for dementia”. In: *Computer* 51.3, pp. 24–34.

Robinson, Dawn T, Jody Clay-Warner, et al. (2012). “Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques”. In: *Biosociology and neurosociology*. Emerald Group Publishing Limited.

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

— (1999). “Emotion display as a strategy for identity negotiation”. In: *Motivation and Emotion* 23.2, pp. 73–104.

— (2018). “Affect control theories of social interaction and self.” In:

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”. In: *Social Forces* 73.1, pp. 175–190.

Rogers, Kimberly B (2018). “Do you see what I see? Testing for individual differences in impressions of events”. In: *Social Psychology Quarterly* 81.2, pp. 149–172.

Rogers, Kimberly B and Lynn Smith-Lovin (2019). “Action, interaction, and groups”. In: *The Wiley Blackwell Companion to Sociology*, pp. 67–86.

Russell, J.A. (1980). “A circumplex model of affect”. In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.

Saha, Tulika, Sriparna Saha, and Pushpak Bhattacharyya (2020). “Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning”. In: *PloS one* 15.7, e0235367.

Schneider, Andreas (2006). “Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998”. In: *Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Schröder, Tobias and Wolfgang Scholl (2009). “Affective dynamics of leadership: An experimental test of affect control theory”. In: *Social Psychology Quarterly* 72.2, pp. 180–197.

Shi, Weiyan and Zhou Yu (2018). “Sentiment adaptive end-to-end dialog systems”. In: *arXiv preprint arXiv:1804.10731*.

Smith, Herman W (2002). “The dynamics of Japanese and American interpersonal events: Behavioral settings versus personality traits”. In: *Journal of Mathematical Sociology* 26.1-2, pp. 71–92.

Smith, Herman W and Linda E Francis (2005). “Social vs. self-directed events among Japanese and Americans”. In: *Social Forces* 84.2, pp. 821–830.

Smith-Lovin, Lynn (1979). “Behavior settings and impressions formed from social scenarios”. In: *Social Psychology Quarterly*, pp. 31–43.

Smith-Lovin, Lynn and William Douglass (1992). “An affect control analysis of two religious subcultures”. In: *Social perspectives on emotion* 1, pp. 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].*

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, et al. (2016). “Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014”. In: *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Brent H Curdy, et al. (2019). “Mean affective ratings of 968 identities, 853 behaviors, and 660 modifiers by amazon mechanical turk workers in 2015”. In: *University of Georgia: Distributed at UGA A ect Control eory Website*.

Tian, Jiahao and Michael D Porter (2022). “Changing presidential approval: Detecting and understanding change points in interval censored polling data”. In: *Stat* 11.1, e463.

Tsoudis, Olga and Lynn Smith-Lovin (1998). “How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes”. In: *Social forces* 77.2, pp. 695–722.

Wang, Jiancheng et al. (Apr. 2020a). “Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 9177–9184. DOI: 10.1609/aaai.v34i05.6454. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6454.

— (2020b). “Sentiment classification in customer service dialogue with topic-aware multi-task learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9177–9184.

Youngreen, Reef et al. (2009). “Identity maintenance and cognitive test performance”. In: *Social Science Research* 38.2, pp. 438–446.

Zahiri, Sayyed M and Jinho D Choi (2017). “Emotion detection on tv show transcripts with sequence-based convolutional neural networks”. In: *arXiv preprint arXiv:1708.04299*.

Zang, Xiaoxue et al. (2020). “MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines”. In: *arXiv preprint arXiv:2007.12720*.

Zhang, Rui, Kai Yin, and Li Li (2020). “Towards emotion-aware user simulator for task-oriented dialogue”. In: *arXiv preprint arXiv:2011.09696*.

Zhu, Yukun et al. (2015). “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

CHAPTER 2

UNDERSTANDING EMOTION IN CONVERSATIONAL AGENT INTERACTIONS THROUGH AFFECT CONTROL THEORY

In recent years, online messaging has emerged as a dominant mode of communication, yet accurately conveying emotions through this medium remains a challenge. This is particularly important for developing conversational agents, which must be able to identify, initiate, and track emotional states during a conversation to be effective. To address this challenge, this chapter explores the application of affect control theory in modeling the conversation between a user and a conversational agent. We aim to improve emotional expression and understanding in task-oriented conversations by considering the interaction between identities and how they influence emotional expression. By leveraging insights from affect control theory, we aim to develop models of conversational agents that better capture the nuances of emotional expression in online messaging.

2.1 Abstract

This chapter delves into using Affect Control Theory (ACT) to detect emotional states in online messaging and adapt responses accordingly. It highlights the need to extend ACT's existing affective dictionaries to include frequently used words and emojis in messaging applications. Through language models, I search for affective dictionaries that incorporate such features, and apply the ACT framework in online messaging sessions with a chatbot to improve user experience. This project aims to develop an extended affective dictionary that includes emojis and introduces a novel algorithmic approach for modeling emotional change during online messaging interactions.

2.2 Introduction

Body language and paralinguistic cues are essential for conveying emotions in in-person communication. Machines can use facial expression (S. Li and Deng 2020), and paralinguistic features (Fonnegra and Dıaz 2018) for emotion classification. However, these features are not available on online messaging platforms. To address this challenge, I propose using Affect Control Theory (ACT) to detect emotional states during online messaging and adapt accordingly. ACT has been widely used and validated in over 100 studies for modeling emotions in various contexts (Robinson and Smith-Lovin 2018); however, to the best of my knowledge, it has not yet been applied to messaging platforms.

Detecting emotional states during online messaging starts with initial sentiments for the users involved and updating the identity and emotional states based on the impression of the conversation. These emotional states are used to adapt the actions. For example, a chatbot could use the estimated emotional state to respond to a customer's requests.

My approach involves modeling a user's emotional state using ACT. ACT is a formal theory of culture that attempts to explain social behavior. According to ACT, all concepts in a given culture have shared emotional meanings, known as affective meanings, represented in a three-dimensional vector space: Evaluation, Potency, and Activity (EPA). Evaluation contrasts “good" versus “bad", Potency contrasts “strong" versus “weak", and Activity contrasts “fast" versus “slow". ACT predictions about behaviors depend on initial cultural sentiments, which are indexed in the affective dictionaries. These dictionaries usually come from surveys. Since running surveys is laborious and costly, existing affective dictionaries for ACT often have limited coverage of words and do not include other communication forms, such as emojis. A major difficulty in incorporating ACT for modeling online messaging interactions is that the existing affective dictionaries are not sufficiently rich. This project aims to extend the affective dictionary to include commonly used words and emojis in messaging applications.

I use language models to find affective dictionaries that include emojis and words commonly used in messaging. Using these extended dictionaries, I aim to apply the ACT framework in online messaging sessions with a chatbot to improve the user experience. The chatbot will adapt its behavior based on the customer's emotional state, as estimated by the ACT model. For example, the chatbot could customize its offers or change its communication style based on the customer's emotional state.

Overall, this project contributes to two primary areas: developing an extended affective dictionary that includes emojis and designing an algorithmic approach to using emoji and word embeddings for emotion change modeling during online messaging interactions. My focus is on developing a chatbot that can adapt its behavior based on the customer's emotional state, ultimately improving the user experience in online messaging platforms.

2.3 Expanding Affective Vocabulary: Incorporating Emojis into the Affective Dictionary using Language Models

Affect control theory relies on affective dictionaries that represent the emotional connotations of words. These affective dictionaries usually come from surveys. Since running surveys is a laborious and costly task, traditional affective dictionaries are limited in size; often less than 3000 words. They cover only a small portion of words widely used in online messaging and ignore other communication forms like emojis. Applying ACT in online messaging systems requires affective dictionaries to represent the words and sentences of interest. This limitation restricts the applicability of ACT which use affective dictionaries to predict behaviors based on text data. To overcome this limitation, I used word2vec word embedding to extend affective dictionaries to include the most commonly used words in online messaging (Mostafavi and M. Porter 2021).

2.3.1 Background

Word embedding is a mapping of words into a numerical vector space that attempts to capture the words' semantic relationship. Word2vec, the most widely used word embedding algorithm, has three million words (Kozlowski, Taddy, and Evans 2019); 1000 times more words than commonly used EPA dictionaries. Because the proximity of words in the embedded space can imply strong semantic relationships, word embedding has been used as an alternative approach for sentiment surveys (Kozlowski, Taddy, and Evans 2019; Alhothali and Hoey 2017; M. Li et al. 2017).

Alhothali and Hoey 2017 used a graph-based method to find affective meaning from word-embedding. Kozlowski, Taddy, and Evans 2019 used word analogy to find cultural dimensions from embedding. M. Li et al. 2017 implemented and compared different methods, such as a graph-based method, to find affective meaning from word embedding. They concluded that regression on word embedding could outperform other methods in inferring the affective meanings. In the regression approach, the word embeddings are the independent variables, and the affective meaning representation in EPA space is the dependent variable.

A translation matrix is an effective method to find affective meanings from word embedding. Given the word embedding $x_i \in \mathbb{R}^{d_1}$ as a source language and the word embedding of its translation in another language, $z_i \in \mathbb{R}^{d_2}$, one intuitive approach for translation is to find a transformation matrix from one embedding to the other. This matrix is known as a translation matrix.

Given the set of vector representation for word pairs, $\{x_i, z_i\}_{i=1}^{n}$, the goal is to find a $d_1 \times d_2$ transformation matrix R such that $x_i R$ approximates z_i. This matrix minimizes the following optimization problem (Mikolov, Le, and Sutskever 2013),

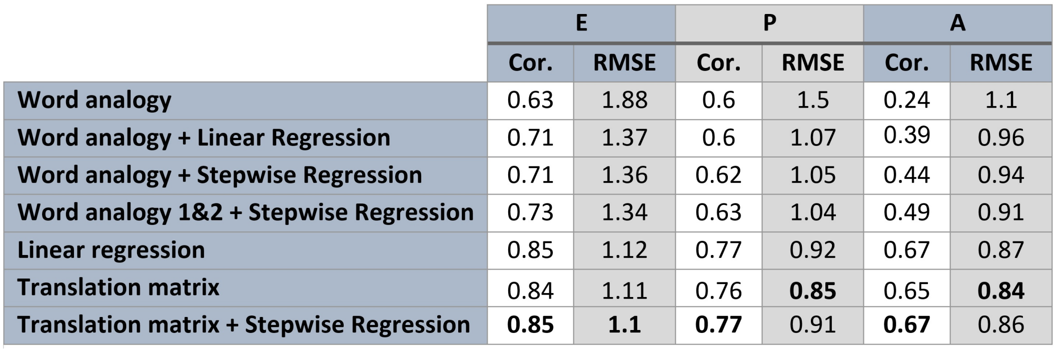
$$\widehat{R} = \min_{R} \sum_{i=1}^{n} ||x_i R - z_i||^2 \tag{2.1}$$

This matrix is used to find the translation of new words from the source languages. First, multiply the embedding in the source language to the matrix and then find the closest word in the translation embedding.

2.3.2 Method and result

I aim to find affective meanings of words and emojis using different embedding methods. To achieve this, I have implemented various techniques, such as word analogy, regressions, and translation matrix methods, and compared their results using RMSE and correlation analysis. The evaluation results shown in Table 2.1, indicate that translation matrix and regression methods have significantly better results than word analogy approaches.

Table 2.1: Evaluation results of finding affective meaning from embedding. The maximum correlation between affective meaning of the test set comes from the two-step process of finding the translation matrix and then applying step-wise regression. RMSE values for regression and translation matrix are the smallest ones.



	E		P		A	
	Cor.	RMSE	Cor.	RMSE	Cor.	RMSE
Word analogy	0.63	1.88	0.6	1.5	0.24	1.1
Word analogy + Linear Regression	0.71	1.37	0.6	1.07	0.39	0.96
Word analogy + Stepwise Regression	0.71	1.36	0.62	1.05	0.44	0.94
Word analogy 1&2 + Stepwise Regression	0.73	1.34	0.63	1.04	0.49	0.91
Linear regression	0.85	1.12	0.77	0.92	0.67	0.87
Translation matrix	0.84	1.11	0.76	**0.85**	0.65	**0.84**
Translation matrix + Stepwise Regression	**0.85**	**1.1**	**0.77**	0.91	**0.67**	0.86

From Table 2.1, using a two-step process, first finding the translation matrix and then applying a second-order regression to fine-tune the mapping to the affective space results in the best accuracy. I use a two-step process; first, I find the translation matrix, and then a second-order regression fine-tunes the mapping to the affective space.

To evaluate the correlation between the extended dictionaries and baseline affective meanings, I conduct correlation analysis shown in Figure 2.1. In this figure, I find the correlation (A) between the result of the test set in my method and the EPA values from the surveys and (B) two survey-based affective dictionaries. In the main diagonal terms, we can observe that the correlation for my extended dictionary in the Activity dimension is higher than the correlation between two different survey-based dictionaries. For the other two dimensions, the values are smaller but close to them.

When evaluated in each category, the words in affective dictionaries are categorized as identity, modifier, and behavior and are mapped to different points in EPA space. 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 one representation in embedding space. To find how the affective meaning of words in different categories are related, I found their correlation in Table 3.1.

The affective meanings of words in different categories are not necessarily highly correlated, as shown in Table 3.1. For example, there is only 0.4 correlation between activity

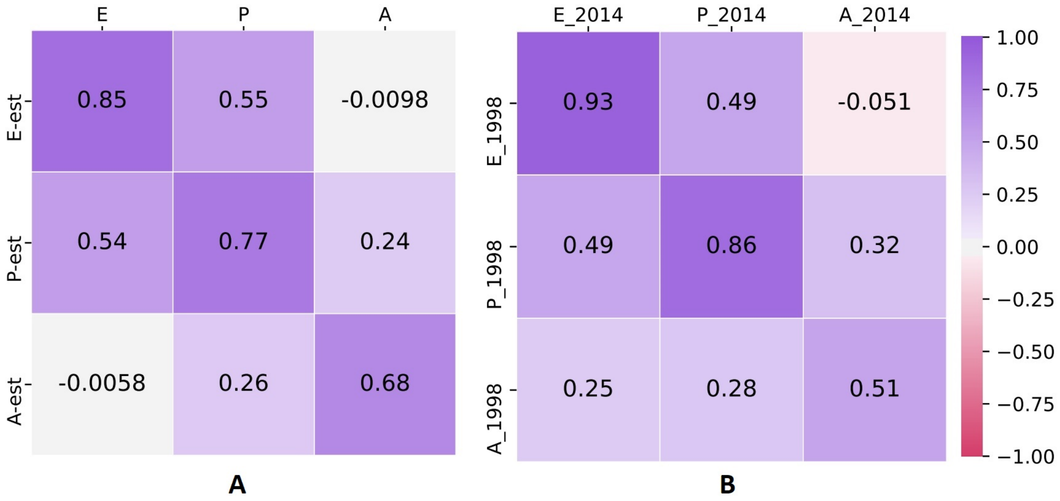
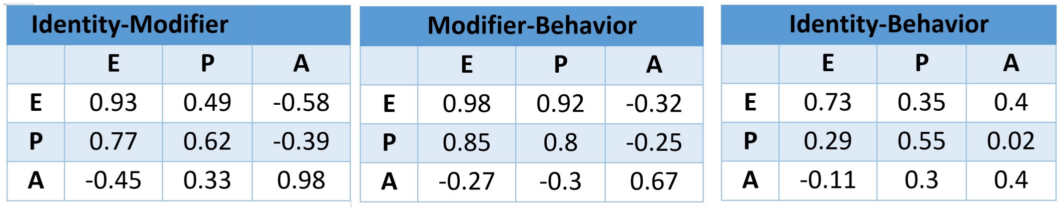


Figure 2.1: The correlation between,
A. Estimated values and the affective dictionary collected in Smith-Lovin, Robinson, Cannon, Clark, et al. 2016.
B. Affective dictionaries collected in Smith-Lovin, Robinson, Cannon, Clark, et al. 2016 and Schneider 2006 affective dictionaries

Table 2.2: Correlation between affective meaning of words in identity, behavior, and modifier category of a dictionary collected in Smith-Lovin, Robinson, Cannon, Clark, 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

dimensions of words that appear both in identity and behavior categories. When I compared two survey-based dictionaries in Figure 2.1, the words in the two dictionaries represented the same categories. As a result, their correlation is higher than the correlation between my estimation and the target dictionary.

2.3.3 Affective meaning of emojis

In online messaging, communication of emotions can be challenging in the absence of body language and vocal characteristics. Therefore, people often use emojis, gifs, and stickers to convey emotions during online conversations. While gifs and stickers are used differently across messaging platforms, emojis are associated with known Unicode and have a similar description across different platforms. Emojis provide affective meaning that is not contained in single words, and interpreting their affective meaning requires a representation similar to words and expressions.

Since most commonly used shallow word-embeddings are trained on very large data-set when emojis were not commonly used, they include a few or no emojis in their vocabulary. On the other hand, emojis play a principal role in conveying affective meaning during online communication. I need an embedding that includes emojis in its vocabulary.

Two different approaches are commonly used in literature to find emoji embedding. The first approach uses social media data and finds the word embedding for all the words, including emojis (Dimson 2015; Barbieri, Ronzano, and Saggion 2016; Reelfs et al. 2020). The main limitation of this approach is the limited number of emojis in the training set. For example, Barbieri, Ronzano, and Saggion 2016 used 100M tweets, but only 700 of them had emojis. Also, their dataset is much smaller than Word2vec or Glove dataset, and as a result, their embedding is not as reliable as Word2vec or Glove. We collected over 1-billion tweets and used Word2vec implementation to find emojiSpace, which is a word and emoji embedding (Mostafavi, Varnosfaderani, et al. 2022).

The second approach finds labels that describe the emojis. Then the embedding of these labels represents the corresponding emoji embedding. For example, Eisner et al. 2016 used the word-embedding of the tags assigned to emojis to derive an emoji-embedding named emoji2vec. Emoji2vec includes embeddings for all Unicode emojis. This embedding is de-

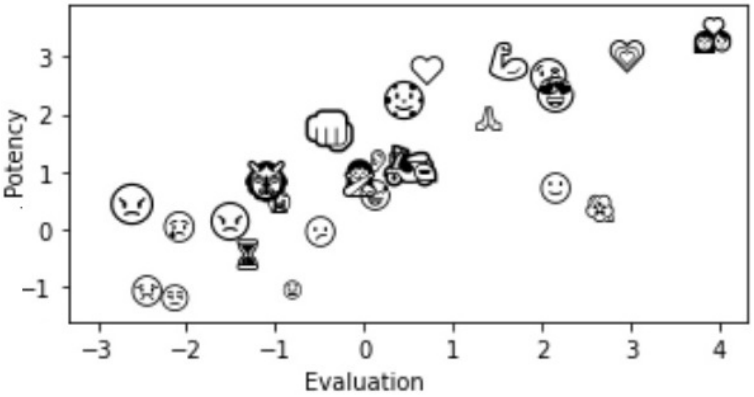
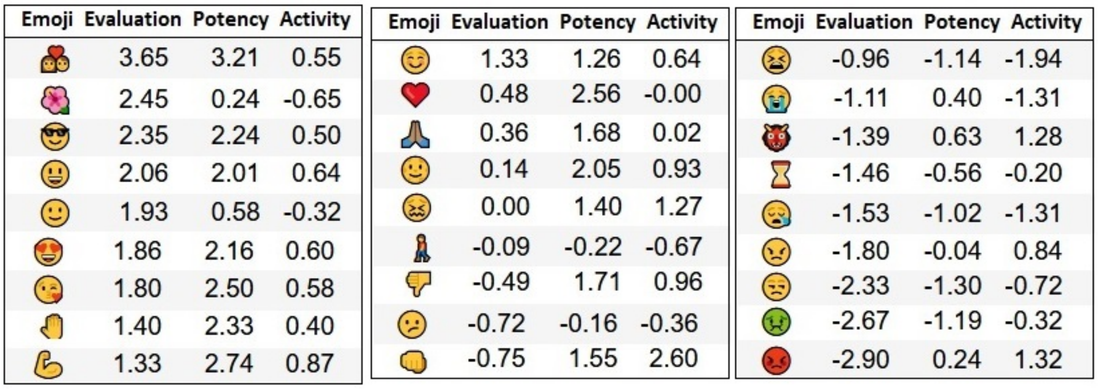


Figure 2.2: Evaluating emoji2vec embedding based on visualizing estimated EPA values for some commonly used emojis. The size of emojis visualizes the activity dimension.

fined in the 300-dimensional Word2vec space. As a result, emojis' embedding resulted from emoji2vec can be used together with the Word2vec embedding. Defining the embedding in the word2vec space and including all Unicode emojis made emoji2vec one of the most popular emoji embeddings.

I used a regression model on a transformation of emoji2vec to extend affective dictionaries to include emojis. In Table 2.3, some commonly used emojis and estimated values for their affective meaning are given. Note that in affective meaning dictionaries, words are measured on a scale from - 4.3 to + 4.3.

Table 2.3: Affective meanings, EPA values, estimated for some common emojis.



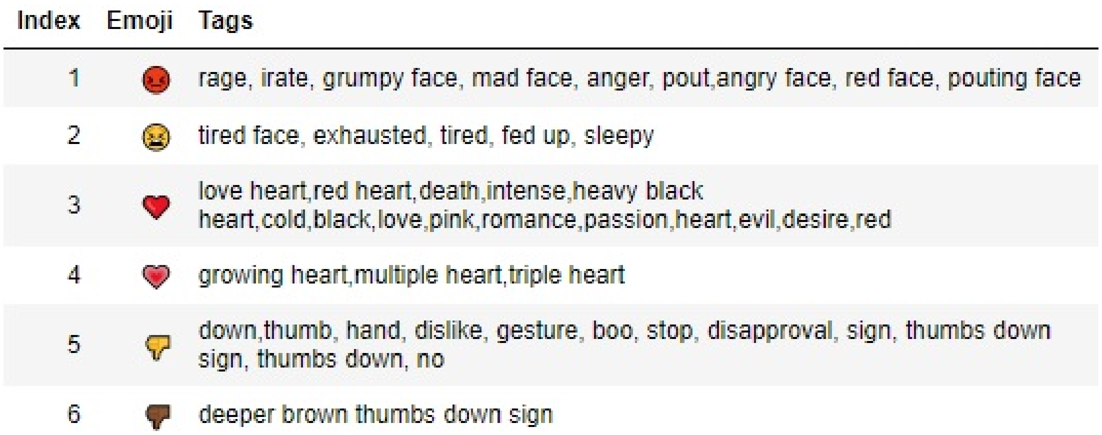
Emoji	Evaluation	Potency	Activity
💑	3.65	3.21	0.55
🌺	2.45	0.24	-0.65
😎	2.35	2.24	0.50
😃	2.06	2.01	0.64
🙂	1.93	0.58	-0.32
😍	1.86	2.16	0.60
😘	1.80	2.50	0.58
✋	1.40	2.33	0.40
💪	1.33	2.74	0.87
😊	1.33	1.26	0.64
❤️	0.48	2.56	-0.00
🙏	0.36	1.68	0.02
🙂	0.14	2.05	0.93
😖	0.00	1.40	1.27
🚶	-0.09	-0.22	-0.67
👎	-0.49	1.71	0.96
😕	-0.72	-0.16	-0.36
👊	-0.75	1.55	2.60
😫	-0.96	-1.14	-1.94
😭	-1.11	0.40	-1.31
👹	-1.39	0.63	1.28
⏳	-1.46	-0.56	-0.20
😪	-1.53	-1.02	-1.31
😠	-1.80	-0.04	0.84
😒	-2.33	-1.30	-0.72
🤢	-2.67	-1.19	-0.32
😡	-2.90	0.24	1.32

I can use Figure 2.2 to visualize Table 2.3.

Improving Emoji embedding

As I discussed earlier, Eisner et al. 2016 used the word-embedding of the emoji tags to derive an emoji embedding. Table 2.4 includes tags assigned to 6 different emojis.

Table 2.4: A sample of tags used for Emoji2vec embedding



Index	Emoji	Tags
1	😡	rage, irate, grumpy face, mad face, anger, pout,angry face, red face, pouting face
2	😫	tired face, exhausted, tired, fed up, sleepy
3	❤️	love heart,red heart,death,intense,heavy black heart,cold,black,love,pink,romance,passion,heart,evil,desire,red
4	💗	growing heart,multiple heart,triple heart
5	👎	down,thumb, hand, dislike, gesture, boo, stop, disapproval, sign, thumbs down sign, thumbs down, no
6	👎🏿	deeper brown thumbs down sign

Let's review some problems that I can observe in the tags shown in Table 2.4.

- I can observe that some of these tags represent the emoji image, not its affective meaning. For example, the emoji in the sixth row is described as "deeper brown ...". This is a description of the emoji color, but the affective meaning of "deeper brown" is not representative of ethnicity as expected.

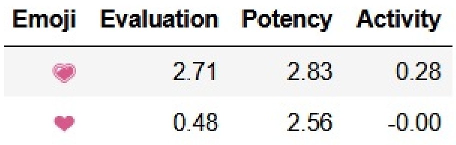
- If I look at the tags used for the two different heart emojis, I can observe these tags are not necessarily good descriptions for deriving their affective meaning. For example, some tags such as evil, red, pink, and black may not represent the affective meanings I expect from a heart emoji.

- I expect to see tags such as "dislike" and "disapproval" for both thumb-down emojis, but they are used only for one of them.

- Redundant words such as "face" in the description of the first two rows biases the embedding representation. They are not describing the affective meanings of the emojis; instead, they describe how they look like.

Table 2.5 shows the resulting affective meaning of the two heart emojis using the emoji2vec embedding. For two alike emojis, one has a much larger evaluation value than the other. This is far from my expectation that the affective meanings of alike emojis should be similar.

Table 2.5: Estimated EPA values for the heart emojis



Emoji	Evaluation	Potency	Activity
💗	2.71	2.83	0.28
❤	0.48	2.56	-0.00

2.4 Modeling Emotional State Change in Chatbot Interactions

In our simple simulation, we explore the potential use of emojis in chat sessions with a chatbot. We assume that customers can add one emoji at the end of their messages to express their emotions. To incorporate this feature into our simulation, we project the emojis into the EPA (Emotion, Potency, and Activity) space, which allows us to model emotional state changes using emojis. We use emojis as modifiers for the identity of the customer. Modifiers are additional attributes that can modify the emotional evaluation of the identity. For instance, a "happy grandmother" has a different emotional evaluation than a "grandmother" due to the modifier "happy."

The impression change equations in (3.1) model the interaction in the basic scenario. Similar equations can be used in scenarios where the identities have modifiers or the interaction happens in a specific social institution. To account for modifiers in our simulation, we use amalgamation equations, similar to the one shown in Equation (2.2) (Averett and Heise 1987).

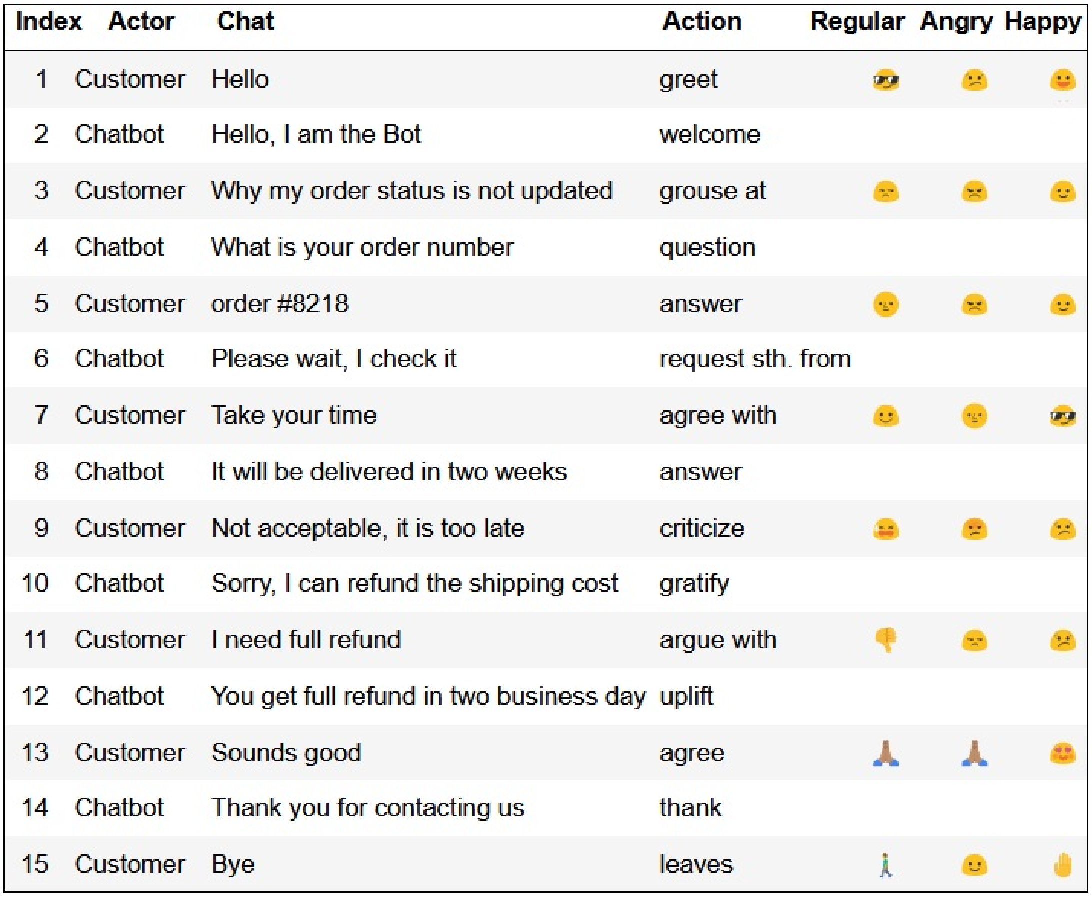
$$C_e = -17 + 0.62M_e - 0.14M_p - 0.18M_a + 0.5I_e \tag{2.2}$$

The equation expresses the emotional evaluation of a modifier-identity (C_e) as a weighted average of the evaluation of the modifier (M_e) and the identity (I_e), adjusted by the potency (M_p) and activity (M_a) of the modifier. The result is a nuanced emotional evaluation that can differentiate between two identities with similar names but different modifiers. Potency and activity of the modifier affect the evaluation of the modifier-identity (Averett and Heise 1987). Using amalgamation equations, we can differentiate the sentiment for a “happy grandmother" and "grandmother". We use amalgamation equations in (2.2) to update the customer's identity while using emojis to express emotions.

2.4.1 Results

Consider a customer chatting with a chatbot. In Table 2.6 the messages from the chatbot and the customer are fixed, but regular, happy, and angry customers use different emojis to express their feeling in this communication. Since chatbot platforms are able to identify what the customer has talked about (Adamopoulou and Moussiades 2020), we assumed the action/behavior labels in this table are given.

Table 2.6: Interaction of regular, angry, and happy customers with a Chatbot. These customers add one emoji at the end of their messages to express their emotions. Action tags associated with the conversation represent the behavior associated with the conversations.



Index	Actor	Chat	Action	Regular	Angry	Happy
1	Customer	Hello	greet	😎	😕	😛
2	Chatbot	Hello, I am the Bot	welcome			
3	Customer	Why my order status is not updated	grouse at	😒	😠	🙂
4	Chatbot	What is your order number	question			
5	Customer	order #8218	answer	😐	😠	🙂
6	Chatbot	Please wait, I check it	request sth. from			
7	Customer	Take your time	agree with	🙂	😐	😎
8	Chatbot	It will be delivered in two weeks	answer			
9	Customer	Not acceptable, it is too late	criticize	😫	😡	😕
10	Chatbot	Sorry, I can refund the shipping cost	gratify			
11	Customer	I need full refund	argue with	👎	😒	😕
12	Chatbot	You get full refund in two business day	uplift			
13	Customer	Sounds good	agree	🙏	🙏	😍
14	Chatbot	Thank you for contacting us	thank			
15	Customer	Bye	leaves	🚶	🙂	✋

Text mining is a popular approach to classify chatbot conversations based on their labels and action (Gliwa et al. 2019). Classifying the conversation has been addressed in the literature by problems such as intent detection. However, in this study, our focus is not on classifying the conversations but instead on understanding the customer's emotions using emojis and tracking them over time. To achieve this, we utilize the action tags column in Table 2.6, and map each emoji used by the customer to its corresponding emotional state in the EPA space to find the modifier. Then we use impression change equations to model the whole interaction in the affective space.

Deflection is the euclidean distance between the sentiments and impressions in the EPA space for an ABO interaction. In Fig. 2.3, I plotted the deflection for three types of customers. We can observe that the deflection increases substantially for all customer types when the customer grouses, criticizes, or argues with the chatbot. Conversely, it decreases when the chatbot is offering something to make the customer happy.

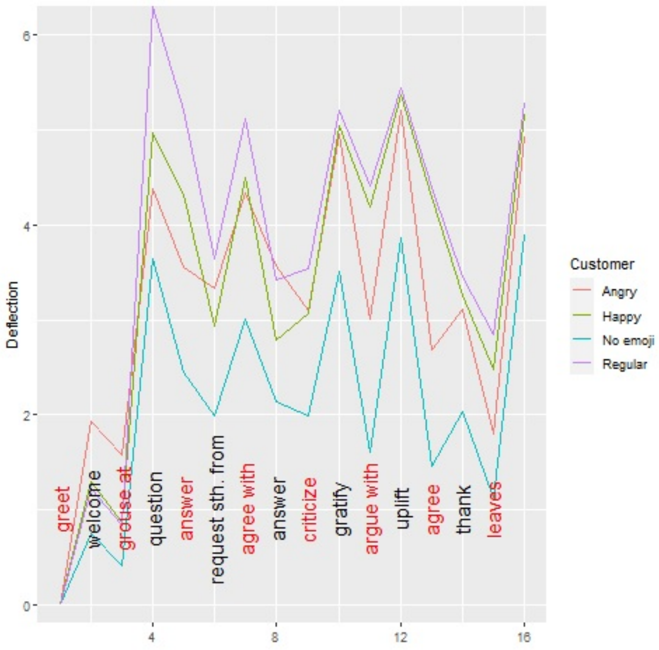


Figure 2.3: Deflection for the four types of customers communicating with a chatbot.

The deflection plot in Fig. 2.3 shows the deflection for all customer types. To focus on emotional change for each agent, we plot the euclidean distance between sentiments and impressions for the Actor and Object in Fig. 2.4 which we refer to as the agent's deflection. Although the chatbot has no emotion, the customer has affective representation based on past experiences with agents. So the chatbot's actions make an emotional impression about its identity to the customer. In Fig. 2.4, we can observe that the impressed emotional change of the chatbot is minimal; however, the customer's emotional changes are substantial.

We can observe in Fig. 2.4 that the angry customer has the minimum emotional change most of the time. We can go one step further and plot the change in EPA values for all customers during this interaction.

Fig. 2.5 shows the EPA values for the customer during the interaction. The rising and falling trends are similar for all customers, but their differences can be observed in each dimension:

- In the evaluation dimension (E), the angry customer always has the minimum evaluation, and the happy customer mostly has the maximum values. The customer that does not use emojis and the regular customer are mostly between the happy and angry customers.
- After the initial chit-chat, the angry customer is more powerful (P) than the other customers. The other customers are very similar in the potency dimension.
- The activity (A) of the customers does not change that much during this interaction.

The observations from Fig. 2.5 are similar to what we expect intuitively about the emotional change for these types of customers. We can also see how the impression of the chatbot changes in Fig. 2.6.

The emotional change for the chatbot in Fig. 2.6 is less than the customer in Fig. 2.5. We can observe the following from Fig. 2.6:

- The evaluation changes substantially when the customer grouses, criticizes, or argues with the chatbot.
- The chatbot is mostly considered to be more powerful when it interacts with an angry customer!
- The impression of chatbot activity does not change substantially for any of the customer types.

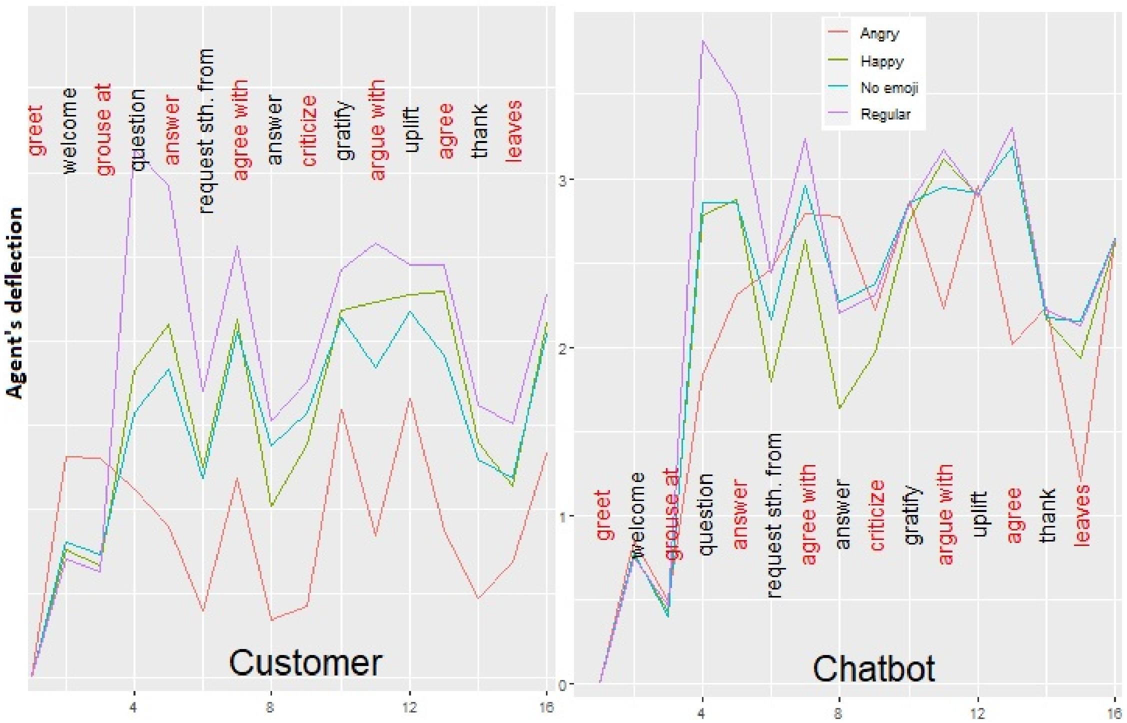


Figure 2.4: The euclidean distance between sentiments and impressions for the Actor and Object identities

- When the chatbot answers a question, it is considered to be more powerful and we see a jump in its potency value.
- We can see when the chatbot questions the customer, its activity will increase significantly.

2.5 Discussion

This project will advance the understanding of emotional change during messaging. The importance of this study centers on the fact that expressing emotion can help people improve their mental health. The result of this study allows users to interpret emotions accurately and to adapt the conversation accordingly. Moreover, this study will give the messaging platforms a tool to monitor emotions and present it to the other side of the communication to improve the user experience. Finally, this project offers an opportunity for researchers to understand cultural meanings from textual data.

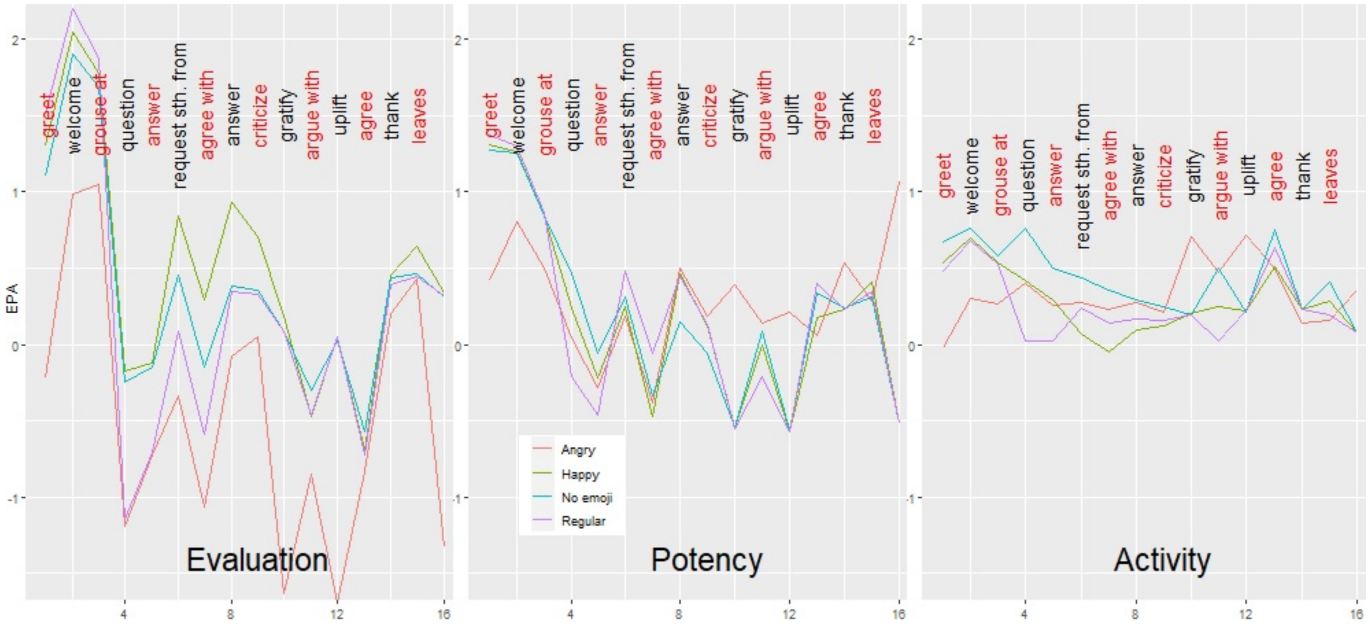


Figure 2.5: Estimated EPA values for the customers. As we expect, the happy customer mostly has the top values in the evaluation dimension and the angry one has the lowest values. We can see when the customer complains about the product for all customers, the evaluation drops significantly. On the other hand, when the chatbot requests something from the customer, the customer potency increases significantly.

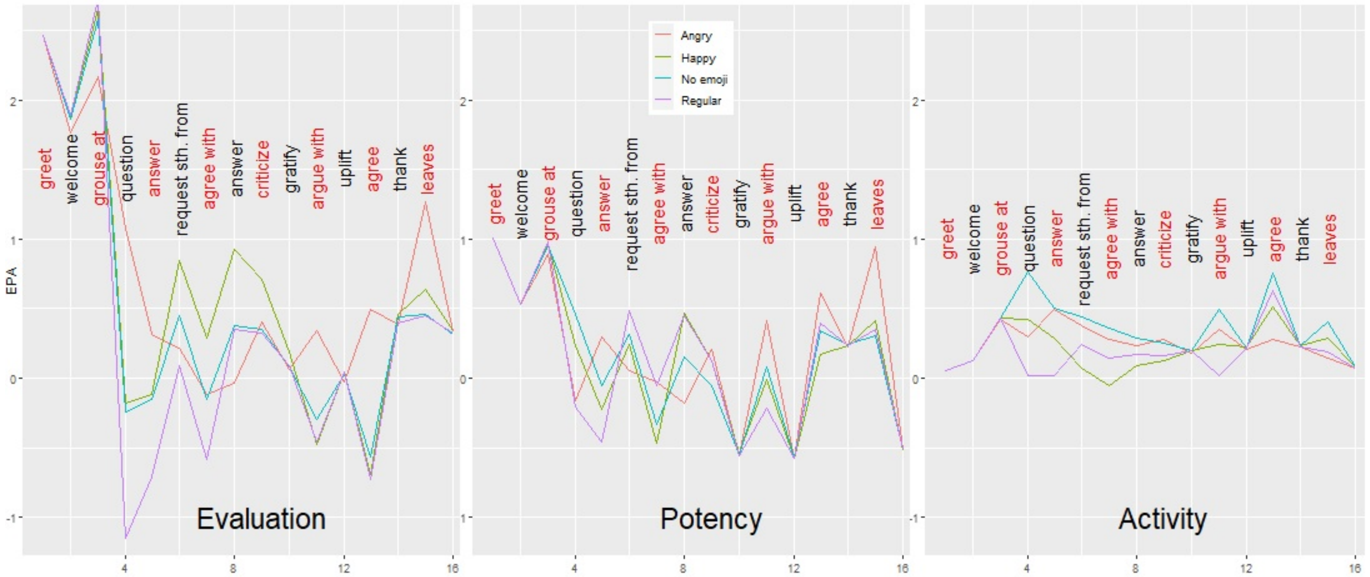


Figure 2.6: Estimated EPA values for the chatbot.

To improve the emoji embedding represented in emoji2vec, we decided to combine the two main approaches in finding emoji-embedding.

First, we collected more than one billion tweets with more than one hundred million tweets with emojis. This is a large dataset, and it is rich in terms of having emojis. We will find an embedding from this data. The embedding is evaluated based on word analogy, t-SNE Visualization, and sentiment analysis tasks.

Second, we map our emoji-embedding to the word2vec and glove embedding space. In this case, researchers can use our developed emoji embedding together with their preferred embedding space. For this purpose, we use two approaches. In the first approach, we use a translation matrix and regression models similar to what we discussed above. In the second approach, we use the nearest-neighbor of emojis in our dataset as labels. In this case, our approach would be similar to emoji2vec, but the labels come from the emoji usage. We will evaluate these two approaches based on word analogy and sentiment analysis and use the best one for the rest of our work.

Incorporating emojis makes the conversation more naturalistic to the customers. In the initial analysis, we assumed the customers could use emojis to express their emotions during a conversation with a chatbot. On the other hand, recent research shows that if chatbots use emojis during a conversation, they seem more socially attractive to customers (Beattie, A. P. Edwards, and C. Edwards 2020). We propose a design mechanism for a chatbot that uses emojis while interacting with the customer.

Finding emotional states would be similar to what we discussed earlier. However, in each step, the chatbot has multiple choices for its response. Each response choice represents a behavior in the affective space. Knowing the emotional states, the chatbot uses impression change equations to find the impression of each choice. On the other hand, emojis' affective meaning plays the role of the modifier in this context. So the chatbot uses amalgamation equations to find how its impressed identity is changed while using emojis. After searching over possible actions and emojis, the chatbot uses emojis to minimize the overall deflection.

2.6 References

Adamopoulou, Eleni and Lefteris Moussiades (2020). “An overview of chatbot technology”. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 373–383.

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

Andre, Elisabeth et al. (2004). “Endowing spoken language dialogue systems with emotional intelligence”. In: *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer, pp. 178–187.

Averett, Christine and David R Heise (1987). “Modified social identities: Amalgamations, attributions, and emotions”. In: *Journal of Mathematical Sociology* 13.1-2, pp. 103–132.

Barbieri, Francesco, Francesco Ronzano, and Horacio Saggion (2016). “What does this emoji mean? a vector space skip-gram model for twitter emojis”. In: *Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72.* ELRA (European Language Resources Association).

Beattie, Austin, Autumn P Edwards, and Chad Edwards (2020). “A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication”. In: *Communication Studies* 71.3, pp. 409–427.

Britt, Lory and David R Heise (1992). “Impressions of self-directed action”. In: *Social Psychology Quarterly*, pp. 335–350.

Budzianowski, Paweł et al. (2018). “MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling”. In: *arXiv preprint arXiv:1810.00278*.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Dimson, Thomas (2015). “Emojineering part 1: Machine learning for emoji trends”. In: *Instagram Engineering Blog* 30.

Eisner, Ben et al. (2016). “emoji2vec: Learning emoji representations from their description”. In: *arXiv preprint arXiv:1609.08359*.

Ekman, Paul (1992). “An argument for basic emotions”. In: *Cognition and Emotion*, pp. 169–200.

Eric, Mihail et al. (2019). “MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines”. In: *arXiv preprint arXiv:1907.01669*.

Feng, Shutong et al. (2022). “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fonnegra, Rubén D and Gloria M Dıaz (2018). “Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model”. In: *International Conference on Human-Computer Interaction*. Springer, pp. 385–396.

Fontaine, Johnny RJ et al. (2007). “The world of emotions is not two-dimensional”. In: *Psychological science* 18.12, pp. 1050–1057.

Francis, Clare and David R Heise (2006). “Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3, 2006”. In: *Computer file]. Distributed at Affect Control Theory Website, Program Interact (http://www. indiana. edu/˜ socpsy/ACT/interact/JavaInteract. html)*.

Francis, Linda E (1997a). “Emotion, coping, and therapeutic ideologies”. In: *Social perspectives on emotion* 4, pp. 71–102.

— (1997b). “Ideology and interpersonal emotion management: Redefining identity in two support groups”. In: *Social Psychology Quarterly*, pp. 153–171.

Ghosal, Deepanway et al. (2020). “Cosmic: Commonsense knowledge for emotion identification in conversations”. In: *arXiv preprint arXiv:2010.02795*.

Gliwa, Bogdan et al. (2019). “Samsum corpus: A human-annotated dialogue dataset for abstractive summarization”. In: *arXiv preprint arXiv:1911.12237*.

Goldstein, David M (1989). “Control theory applied to stress management”. In: *Advances in Psychology*. Vol. 62. Elsevier, pp. 481–491.

Heise, David R (1977). “Social action as the control of affect”. In: *Behavioral Science* 22.3, pp. 163–177.

— (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

— (2013). “Interact guide”. In: *Department of Sociology, Indiana University*.

Heise, David R and Cassandra Calhan (1995). “Emotion norms in interpersonal events”. In: *Social Psychology Quarterly*, pp. 223–240.

Heise, David R and Lisa Thomas (1989). “Predicting impressions created by combinations of emotion and social identity”. In: *Social Psychology Quarterly*, pp. 141–148.

Hunt, Pamela M (2008). “From festies to tourrats: Examining the relationship between jamband subculture involvement and role meanings”. In: *Social Psychology Quarterly* 71.4, pp. 356–378.

Kelley, John F (1984). “An iterative design methodology for user-friendly natural language office information applications”. In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41.

Koch, Andrew, Jiahao Tian, and Michael D Porter (2020). “Criminal Consistency and Distinctiveness”. In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–3.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). “The geometry of culture: Analyzing the meanings of class through word embeddings”. In: *American Sociological Review* 84.5, pp. 905–949.

Kriegel, Darys J et al. (2017). “A multilevel investigation of Arabic-language impression change”. In: *International Journal of Sociology* 47.4, pp. 278–295.

Li, Minglei et al. (2017). “Inferring affective meanings of words from word embedding”. In: *IEEE Transactions on Affective Computing* 8.4, pp. 443–456.

Li, Shan and Weihong Deng (2020). “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing*.

Li, Yanran et al. (2017). “Dailydialog: A manually labelled multi-turn dialogue dataset”. In: *arXiv preprint arXiv:1710.03957*.

Liu, Yang and Mirella Lapata (2019). “Text summarization with pretrained encoders”. In: *arXiv preprint arXiv:1908.08345*.

Loon, Austin van and Jeremy Freese (2022). “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language”. In: *American Behavioral Scientist*. DOI: 10.1177/00027642211066046. eprint: https://doi.org/10.1177/00027642211066046.

MacKinnon, Neil J and Dawn T Robinson (2014). “Back to the future: 25 years of research in affect control theory”. In: *Advances in group processes*.

Majumder, Navonil et al. (2019). “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6818–6825.

McCormick, Chris and Nick Ryan (May 2019). *BERT Word Embeddings Tutorial*. URL: http://www.mccormickml.com.

Mikolov, Tomas, Kai Chen, et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. In: *arXiv preprint arXiv:1309.4168*.

Mostafavi, Moeen (2021). “Adapting Online Messaging Based on Emotiona”. In: *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization*.

Mostafavi, Moeen, Maria Phillips, et al. (2021). “A tale of two metrics: Polling and financial contributions as a measure of performance”. In: *2021 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.

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

Mostafavi, Moeen, Michael D Porter, and Dawn T Robinson (2022). “Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers”. In: *arXiv preprint arXiv:2202.00065*.

Mostafavi, Moeen, Mahsa Pahlavikhah Varnosfaderani, et al. (2022). “emojiSpace: Spatial Representation of Emojis”. In: *arXiv preprint arXiv:2209.09871*.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CBO9780511571299.

Osgood, Charles Egerton, William H May, et al. (1975). *Cross-cultural universals of affective meaning*. Vol. 1. University of Illinois Press.

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

Polzin, Thomas S and Alexander Waibel (2000). "Emotion-sensitive human-computer interfaces". In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Poria, Soujanya et al. (2018). "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: *arXiv preprint arXiv:1810.02508*.

Quinonero-Candela, Joaquin et al. (2008). *Dataset shift in machine learning*. Mit Press.

Rashotte, Lisa Slattery (2002). "Incorporating nonverbal behaviors into affect control theory". In: *Electronic Journal of Sociology* 6.3.

Reelfs, Jens Helge et al. (2020). "Word-Emoji Embeddings from large scale Messaging Data reflect real-world Semantic Associations of Expressive Icons". In: *arXiv preprint arXiv:2006.01207*.

Robillard, Julie M and Jesse Hoey (2018). "Emotion and motivation in cognitive assistive technologies for dementia". In: *Computer* 51.3, pp. 24–34.

Robinson, Dawn T, Jody Clay-Warner, et al. (2012). "Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques". In: *Biosociology and neurosociology*. Emerald Group Publishing Limited.

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

— (1999). "Emotion display as a strategy for identity negotiation". In: *Motivation and Emotion* 23.2, pp. 73–104.

— (2018). "Affect control theories of social interaction and self." In:

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". In: *Social Forces* 73.1, pp. 175–190.

Rogers, Kimberly B (2018). "Do you see what I see? Testing for individual differences in impressions of events". In: *Social Psychology Quarterly* 81.2, pp. 149–172.

Rogers, Kimberly B and Lynn Smith-Lovin (2019). "Action, interaction, and groups". In: *The Wiley Blackwell Companion to Sociology*, pp. 67–86.

Russell, J.A. (1980). "A circumplex model of affect". In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.

Saha, Tulika, Sriparna Saha, and Pushpak Bhattacharyya (2020). "Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning". In: *PloS one* 15.7, e0235367.

Schneider, Andreas (2006). “Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998”. In: *Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Schröder, Tobias and Wolfgang Scholl (2009). “Affective dynamics of leadership: An experimental test of affect control theory”. In: *Social Psychology Quarterly* 72.2, pp. 180–197.

Shi, Weiyan and Zhou Yu (2018). “Sentiment adaptive end-to-end dialog systems”. In: *arXiv preprint arXiv:1804.10731*.

Smith, Herman W (2002). “The dynamics of Japanese and American interpersonal events: Behavioral settings versus personality traits”. In: *Journal of Mathematical Sociology* 26.1-2, pp. 71–92.

Smith, Herman W and Linda E Francis (2005). “Social vs. self-directed events among Japanese and Americans”. In: *Social Forces* 84.2, pp. 821–830.

Smith-Lovin, Lynn (1979). “Behavior settings and impressions formed from social scenarios”. In: *Social Psychology Quarterly*, pp. 31–43.

Smith-Lovin, Lynn and William Douglass (1992). “An affect control analysis of two religious subcultures”. In: *Social perspectives on emotion* 1, pp. 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]*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, et al. (2016). “Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014”. In: *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Brent H Curdy, et al. (2019). “Mean affective ratings of 968 identities, 853 behaviors, and 660 modifiers by amazon mechanical turk workers in 2015”. In: *University of Georgia: Distributed at UGA A ect Control eory Website*.

Tian, Jiahao and Michael D Porter (2022). “Changing presidential approval: Detecting and understanding change points in interval censored polling data”. In: *Stat* 11.1, e463.

Tsoudis, Olga and Lynn Smith-Lovin (1998). “How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes”. In: *Social forces* 77.2, pp. 695–722.

Wang, Jiancheng et al. (Apr. 2020a). “Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 9177–9184. DOI: 10.1609/aaai.v34i05.6454. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6454.

— (2020b). “Sentiment classification in customer service dialogue with topic-aware multi-task learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9177–9184.

Youngreen, Reef et al. (2009). “Identity maintenance and cognitive test performance”. In: *Social Science Research* 38.2, pp. 438–446.

Zahiri, Sayyed M and Jinho D Choi (2017). “Emotion detection on tv show transcripts with sequence-based convolutional neural networks”. In: *arXiv preprint arXiv:1708.04299*.

Zang, Xiaoxue et al. (2020). “MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines”. In: *arXiv preprint arXiv:2007.12720*.

Zhang, Rui, Kai Yin, and Li Li (2020). “Towards emotion-aware user simulator for task-oriented dialogue”. In: *arXiv preprint arXiv:2011.09696*.

Zhu, Yukun et al. (2015). “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

CHAPTER 3

USING CONTEXT-DEPENDENT EMBEDDING TO DERIVE FUNDAMENTAL SENTIMENTS AND EXPAND THE SCOPE OF AFFECT CONTROL THEORY

3.1 Abstract

Affect control theory is a mathematical model of culture-based action that relies on culturally-grounded model specifications to predict behaviors, emotions, and new cultural meanings that arise from local interactions. The theory has been used to make successful predictions about interpersonal social behavior, political actions, social movement strategies, organizational behavior, criminal sentencing, belief transmission, and social emotions. This theory requires dictionaries of social concepts that are indexed in three-dimensions of cultural sentiment. Traditionally, the sentiments are quantified using survey data that is fed into a regression model to explain social behavior. Opportunities to expand the reach of the theory by enlarging the sentiment lexicon are limited due to prohibitive cost. This paper uses a fine-tuned Bidirectional Encoder Representations from Transformers model to develop a replacement for these surveys. The new model achieves state-of-the-art accuracy in expanding the affective lexicon and allowing more behaviors to be explained. It estimates affective meanings similar to running a new survey. It could greatly expand the ability of affect control theory to be applied to new substantive domains at greater speed and at a lower burden for research respondents, researchers, and funders. Finally, it introduces a user-friendly, online platform to estimate sentiments for previously unmeasured concepts.

3.2 Introduction

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. Based on this information, your mentor may offer some initial recommendations, but you *adapt* your behavior after observing the 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. While the potential uses of affect control theory are understanding emotional changes during different social interactions, real-life applications are limited due to the vocabulary size of affective dictionaries. Recent advances in computational methodologies offer an opportunity to achieve this grounding in more efficient ways, allowing for more rapid updating to new cultures and cultural changes. Recent work has begun to leverage these methodologies in promising ways. We build on this work and offer a new methodology that improves on the accuracy and expands the capabilities to include new concepts.

In this section, we describe the affect control theory framework and its limitations due to data collection challenges. Then we discuss computational solutions introduced to overcome its limitation. Finally, we present our computational methodology, a pre-trained deep neural network to expand affective lexicons.

3.3 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. There are three core components of the theory – 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.

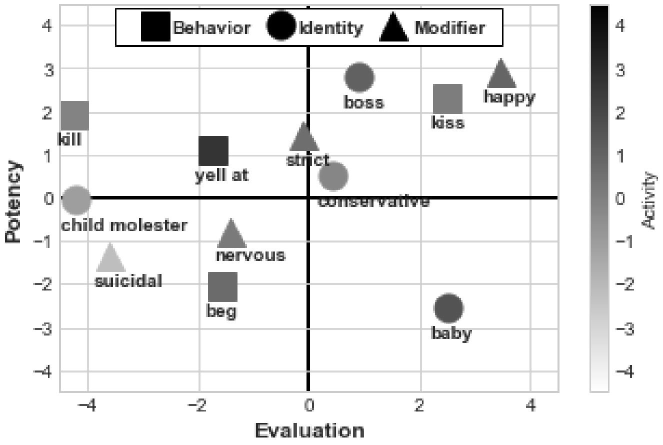


Figure 3.1: Sentiments of sample concepts in EPA space. Circles, squares, and triangles show identities, behaviors, and modifiers. The color represents the activity dimension.

3.3.1 Cultural Sentiments

The theory represents cultural sentiments in a three-dimensional affective space (Heise 1977). Affect control theory uses *Evaluation* [good vs. bad], *Potency* [powerful vs. powerless], and *Activity* [active vs. passive] (EPA) space introduced by Osgood, Suci, and Tannenbaum 1957 to index social concepts (identities, behaviors, emotion, settings, etc.) in affective space and place them all in a common metric. The EPA dimensions describe substantial variation in the affective meaning of lexicons in more than 20 national cultures studied (Osgood, May, et al. 1975; Osgood, Suci, and Tannenbaum 1957). Fontaine et al. 2007 found that EPA scores represent the first three principal components after reducing dimensionality on 144 features representing the main components of emotions.

Affect control theory considers social interactions or events that include an *actor* that *behaves* toward an *object*. Extracting the Actor-Behavior-Object (ABO) components of an event is the first step in modeling interactions (Heise 2010). Actor/Object has an identity such as “baby” or “boss” that is represented in the affective space.

Figure 3.1 is a visualization of U.S. cultural sentiments toward twelve concepts in EPA space. 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 mildly active concept. Note that the range of EPA ratings is from -4.3 to 4.3.

3.3.2 Impression change equation

People label elements of their interactions (events), and those labels evoke meaning as indexed in the EPA sentiment dictionaries. Consider an example of observing "a bossy employer argues with an employee." This observation leads people to evaluate both the *actor* and the *object* of the interaction as less pleasant than initially thought. After observing this event, they may also feel the employer is more powerful, and the employee is more powerless than their baseline sentiments. Being more pleasant/powerful is the *impression* of observing this event and it translates to a higher value in *evaluation/potency*. These transient meanings are labeled "impressions" in Affect Control Theory. Impressions are contextualized affective meanings evoked by symbolic labels in 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), non-verbal behaviors (Rashotte 2002)), self-directed behaviors (Britt and Heise 1992; Smith and L. E. Francis 2005) , and social settings (Smith 2002; Smith-Lovin 1979).

3.3.3 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 an ABO characters and their impressions following an event. If the impression of an ABO event is close to the initial sentiment, *deflection* is small, but 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 distress (Goldstein 1989). For example, if a grandmother fights with her grandchild, the grandmother and the grandchild feel distressed. They prefer to do something to bring the impression of their identities back to where they view themselves in the society. For example, we may expect one side to take an action, like apologize. This highly deflected event is very different from two soldiers fighting in a battle. The soldiers are supposed to fight with enemies in battle, so they may not feel social pressure to change their behavior.

The impression change equations are mathematically manipulated to implement the *affect control principle*: the assumption that individuals behave to maintain or restore cultural affective meanings associated with activated labels. The inputs to the impression change equations are EPA sentiment profiles – the culturally shared, fundamental meanings that people associate with social labels. The outputs of those equations, impressions, are transient meanings that arise as social interactions unfold. Discrepancies between sentiments and impressions signal how closely interactions are confirming to cultural prescriptions. The control system part of the affect control theory models the assumption that social actors try to maintain their cultural definitions of social situations (the affect control principle). When impressions vary from sentiments (as the temperature might vary in a room), people behave socially to bring the impressions back in line with cultural sentiments. Affect control theorists define deflection as the discrepancy between fundamental cultural sentiments and transient situated impressions. Deflection is the error signal, operationalized as the squared euclidean distance in EPA space, between cultural sentiments and event impressions. After an event that has disturbed meanings, solving for the behavior profile produces the creative response that an actor is expected to generate to repair the situation. Alternatively, these same equations can predict a new normative understanding that makes sense of the observed events.

Heise 2013 developed a software called INTERACT that can simulate interactions. It uses the impression change equations to solve for behaviors that minimize the deflection or predict attributes and emotions during the interaction. Using INTERACT researchers have access to the theory and use it principally and rigorously without needing to master the technical details. This tool is available and interactive using natural language input. So they can use the theory in a reproducible way without needing to derive the math all over. Consider the following set of events/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.

5. Employee listens to / disobeys bossy employer.

The visualization in Figure 3.2 shows how the impression of actor/object's identity changed based on the sequence of interactions. Let's focus on the evaluation dimension for the employer. The employer has a negative baseline evaluation, but it increases after observing the first two interactions. The first two interactions include positively rated behaviors. After the second interaction, the impression of the employer's identity is positively evaluated and so the next positively evaluated action, replies to, does not move it substantially. Positive behavior is expected from a positive identity. However, in the fourth interaction, the employer is evaluated to have an unpleasant identity after doing a negative behavior, argue with. For the fifth event, we have shown how the impression of different actions by the employee has significantly moved the states for both the *actor* and the *object*. The sequential interactions discussed here are similar to our mentorship example discussed earlier. It shows how understanding the interaction dynamic can help predict the consequences of behaviors.

Affect control theory has rules to describe how the impression of an event changes affective meaning of ABO characters. Affect control theory uses either mathematical equations or descriptive forms to discuss these rules. The following two descriptive forms show how the identity of the *actor* is impressed by some events,

- *Actors* seem nice when they behave in a positive way toward others. This describes *morality* effect in affect control theory 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.

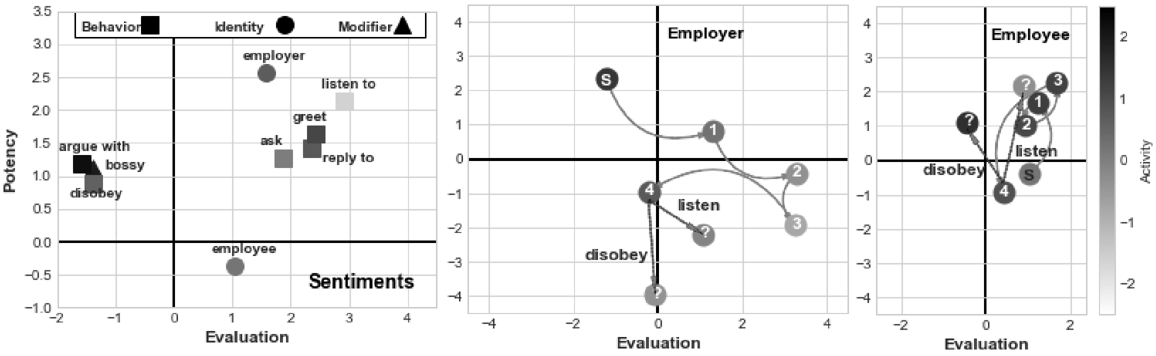


Figure 3.2: Simulating sequential events in an interaction between an employee and employer. The initial sentiments for both characters are shown by s and the impression after each of four events are shown by numbers. After the fourth event, based on what the employee does, the final sentiment for the two characters would be one of the places shown by the question mark.

- 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 forms, events can change the impression of ABO characters. They move them toward or away from their current or initial sentiments.

3.3.4 Empirical testing and application

Affect control theory was introduced in the 1970s (Heise 1977). Validations and applications of affect control theory appear in more than a hundred published articles, book chapters, and books (Robinson and Smith-Lovin 2018; MacKinnon and Robinson 2014). Observational studies have revealed dynamics predicted by affect control theory equations (e.g., L. E. Francis 1997a; L. E. Francis 1997b; Hunt 2008). Experimental studies have validated the affect control theory predictions about emotional experiences (e.g., Robinson and Smith-Lovin 1992; Robinson, Clay-Warner, et al. 2012, about identity attributions (e.g., Heise and Calhan 1995; Robinson and Smith-Lovin 1999; Robinson, Smith-Lovin, and Tsoudis 1994; Tsoudis and Smith-Lovin 1998) , as well as its predictions about social behavior and performance (Robinson and Smith-Lovin 1992; Schröder and Scholl 2009; Youngreen et al. 2009). Surveys 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 introduced affect control theory to estimate and track emotional states during online messaging. For example, chatbots can use the affect control theory framework to understand the emotional state of the customer in real time and adapt their behavior accordingly.

3.3.5 Empirical grounding

To formulate the process in 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 this survey, the participants rate how they feel about *identity/behavior/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 affective meaning of "employee", "greet", and "employer" after observing "the employee greets the employer". As we discussed earlier, the ratings of ABO (impressions) could be different from the initial baselines (sentiments). Affect control theory uses regression models, known as impression change equations, to estimate these changes (Heise 2013). Let $X = [A_e, A_p, A_a, B_e, B_p, B_a, O_e, O_p, O_a]^T$ represent the EPA values/sentiments of an ABO triple, where $\{A, B, O\}$ represent the ABO characters and $\{e, p, a\}$ the EPA components. Consider further the two-way interactions $X^2 = [A_eB_e, A_eO_e, A_eB_a, \ldots A_aO_a]^T$ and three-way interactions $X^3 = [A_eB_eO_e, A_eB_eO_p, A_eB_eO_a, \ldots A_aB_aO_a]^T$. The basic structure of an impression change equation is the linear model

$$X' = \alpha X + \beta X^2 + \gamma X^3 \tag{3.1}$$

where α, β, and γ are coefficient vectors and X' represent the resulting impression after the event. Modifiers can incorporated prior to impression change by changing the baseline values/sentiments (e.g. bossy employer) (Averett and Heise 1987).

3.3.6 Challenges

Current state-of-the-art practices for generating affect control theory models for new cultures require (a) assessing the meanings of 1000-2,000 commonly used labels for social event descriptors (*behaviors, identities, settings, emotions*) in three dimensions of meaning (*evaluation, potency, activity*) and (b) 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 – requiring 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). So, updating sentiment dictionaries is 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 utility of the theory. 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). Thus, finding the affective meaning for 5000 words requires over 125,000 ratings and 400 hours of respondent time (Heise 2010). Due to the high cost and time required, most EPA data collections have been limited to relatively small (1-2k words) dictionaries which has limited the applicability of affect control theory tools.

3.4 Methods

The affect control theory usage can be hampered by the lack of access to the concept that needed for the application of interest. Developing a tool that expands concepts used in a dictionary can break limits for researchers.

As an alternative to conducting surveys for a large set of new concepts is expanding the current dictionaries to include those concepts. Machine learning approaches helped resolving similar social science problems (Koch, Tian, and M. D. Porter 2020; Mostafavi, Phillips, et al. 2021; Tian and M. D. Porter 2022). Researchers have tried supervised (Mostafavi and M. Porter 2021; M. Li et al. 2017) and semi-supervised methods (Alhothali and Hoey 2017) on shallow word-embeddings to build affective lexicons. Shallow word-embeddings use a neural network with a few layers to find high dimensional vectors representing words or tokens. For example, Mikolov, Chen, et al. 2013 introduced “Word2Vec” which has 300-dimensional vectors corresponding to 3 million words and phrases. To find these vectors that are called embedding, they trained the neural network on very large corpus such as Google News dataset, which includes about 100 billion words. We review these early practices to expand dictionaries and discuss their limitation. In a recent paper, Loon and Freese 2022 point to a potential direction for addressing some limitations of shallow-embedding. We review their work that uses Bidirectional Encoder Representations from Transformers (BERT) as a contextual word-embedding alongside shallow word-embedding to estimate affective meaning of words. We share the intuitions expressed by Loon and Freese 2022, that BERT model might “bridge the gap between the ‘words’ of embedding models and the ‘concepts’ pursued by Affect Control Theory”.

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 affective dictionary. For example, to estimate affective meaning of *dean* we may look at words such as *boss*, *professor*, and *faculty* in affective dictionaries, and estimate affective meaning of *boss* using a weighted mean of values for known words. 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) distributional based approach which uses co-occurrence metrics in a corpus, (3) neural word embeddings method, and (4) combination of semantic and distributional methods. They found that using both semantics and distributional-based approaches gave the best semi-supervised result. On the other hand, M. Li et al. 2017 argued that word-embedding can represent the words' general meaning, including denotative meaning, connotative meaning, social meaning, affective meaning, reflected meaning, collocative meaning, and thematic meaning. So, word-embedding graph propagation that uses general meaning similarities may reduce accuracy for finding affective meaning.

Mostafavi and M. Porter 2021; M. Li et al. 2017 used supervised methods on shallow word-embeddings to find affective meanings of the words. To put it simply, they use supervised methods to find a mapping from higher dimensional embedding to affective space. For example, M. Li et al. 2017 use a regression model to estimate the three-dimensional affective meaning from a 300-dimensional "Word2Vec" embedding. All three of these computational approaches described here yielded good results for deriving *evaluation* sentiments for new concepts, but their performance on the *activity* and *potency* dimensions was not very strong.

Shallow word embeddings have only one representation for every word. On the other hand, the affective meaning of a word in *identity*, *modifier*, or *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 above approaches are limited by using one representation for different meanings of a word and not considering the context. Figure 3.3 shows the EPA values for some words that appear in different categories. 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 get a sense of how similar the affective meanings are between categories, we calculated the pairwise correlation of EPA values for words shared across various different categories (Table 3.1). Table 3.1a contains the correlations between the EPA values of words that appear as both an *identity* and a *modifier* in this dictionary. We see that for terms that can be either *identity* (a patient) or a *modifier* (a patient person), the correlation between their *evaluation* as a *modifier* and their *evaluation* as an *identity* is high (r=.93) but for *potency* it is less correlated (r=.62). Looking across Table 3.1, we see that affective meanings of the words in different appearing categories are not always highly correlated. For example, there is

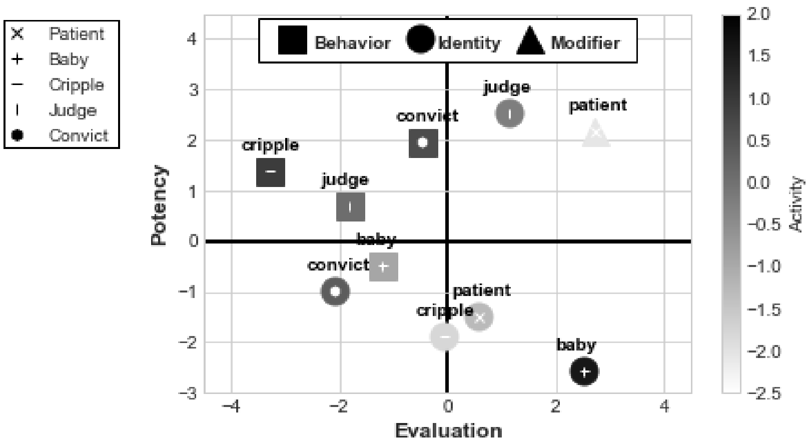


Figure 3.3: Visualization of words with different affective meaning in EPA space. Circles, squares, and triangles show identities, behaviors, and modifiers. The color represents the activity dimension. We can observe affective meaning of words such as "judge" is very different as an *identity* and a *behavior*.

only a 0.4 correlation between *activity* dimensions of words that appear both in the *identity* and *behavior* categories. While some categories maintain high association across categories, other categories, like *identity* and *behavior* in the *activity* and *potency* dimensions have low association implying they words used in different categories represent different affective meanings, and consistent with the examples displayed in Figure 3.3. From Table 3.1, 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.

Table 3.1: Correlation among the affective meanings of identities, behaviors, and modifiers from a sentiment dictionary collected in 2014 (Smith-Lovin, Robinson, Cannon, Clark, 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

Given the somewhat poorer performance of the shallow word-embedding approaches at predicting potency and activity and the hints in Table 3.1 about the lower correlations between sentiments across categories (e.g., identity and *behavior*; *modifier* and *behavior*, or *identity* and *modifier*) for the dimensions of *potency* and *activity*, we propose that failure to consider context may be a contributing factor. Alternatively, deep embedding utilizes contextual aspects while representing concepts of a sentence. Loon and Freese 2022 studies using deep embedding to find affective meanings for new words.

Loon and Freese 2022 conducted four studies. In the first, they used "Word2Vec" in a word-embedding algorithm and a sentiment dictionary collected with an mTurk population (Smith-Lovin, Robinson, Cannon, Curdy, et al. 2019). In a second study, they used "GloVe" and "Word2Vec" word embeddings with a predictive modeling approach to estimate the same sentiment dictionary. In the second study, word embeddings play the role of features that are passed to a neural network to estimate sentiment dictionary. They saw a substantial improvement in sentiment predictions, particularly in the *activity* and *potency* dimensions. In a third study, they examined that same approach on newly collected data and found that it performed similarly. In a fourth study, however, they used a newer approach, capable of incorporating context more fully into the embeddings (BERT, described in more detail below). When they compared their models using BERT-embeddings on the same two datasets they received similar results to those in the second and third studies (combining Word2Vec and GloVe).

We think that pre-trained deep neural network has several advantages – including consideration of the contextual aspects of concepts within social events. We speculate that the failure to observe substantial improvements using BERT in Loon and Freese 2022 may stem from the particular approach used in this study – which (a) excluded compound words or words such as "neighborly" that are not in BERT vocabulary and (b) extracted the meaning for word "roots" rather than concepts within events – aggregating across much of what makes BERT a context-embedding approach. While this approach still extracts the context-embedding of each root, it does not consider information about the event in a way that would be optimal from an affect control theory perspective.

3.4.1 Bidirectional Encoder Representations from Transformers

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 pre-trained the bidirectional representation from a large set of unlabeled text. The model is pre-trained on from the BookCorpus and Wikipedia. BookCorpus includes 11,038 books with about 1000M words Zhu et al. 2015. English Wikipedia also has about 800M words.

Given a sentence, BERT first parses it into its parts. Then tokenizes the parsed sentence to transfer it to a sequence of tokens. Two special tokens are added at the beginning and end of the list. This sequence of tokens replaces the original sentence. This process is repeated for every sentence in the training set. The goal is to find high-dimensional vectors for every token in the training set such that the deep neural network can represent some language relationships. The loss function mathematically represents these relationships. The deep neural network maps these tokens to a high-dimensional numerical vector. Each of these high-dimensional vectors represents a token embedding, and the vector corresponding to one of the two additional tokens represents sentence embedding. We briefly review the process here to understand how BERT finds these vectors.

As a contextual representation, the BERT tokenizer, known as WordPiece, tokenizes the input sentences. If the sentence tokens are in the vocabulary of the pre-trained model, they appear in the tokenized list without any modification. However, WordPiece may assign multiple tokens to a word if the word is not in its vocabulary. In that case, tokens are root vocabulary and suffix of the original word. For example, if the words “affective” and “subtext” are not included in the vocabulary, WordPiece outputs [affect, ##ive] and [sub, ##text] tokens where ## shows the two tokens came from a compound word. Vocabulary of the pre-trained BERT model that we used includes 30,000 tokens and it can process many other compound words that are not in the vocabulary list. Indices of the tokens in the BERT tokenizer vocabulary are called token IDs.

After tokenization, the sentences are represented by a sequence of Token IDs. Since BERT is trained on “next sentence prediction” task, it assumes these sequences represent two sentences and two special tokens to indicate their relationship. The [CLS] token indicates the start of the first sentence, and the special token [SEP] comes at its end. We use BERT with only one sentence, but we have to use these special tokens. Figure 3.4 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.

The loss function in BERT model is defined for two objectives. (a) Masked language model and (b) next sentence prediction that we briefly review here. Masked language model objective is predicting some randomly masked tokens from a sentence. It uses context words on both sides of the target word in all layers to find the masked word. In other words, it randomly masks some words in a sentence and then uses the remaining contextual words to predict the masked words. We can describe this process by masking a word in the following sentence. *The student asked a question about the prerequisite courses.* Masking the word "question" we get,*The student asked a [MASK] about the prerequisite courses.* The model should use contextual words to predict this masked word. The ability to use full context in prediction differentiates BERT from other word embedding models like word2vec, which only uses the neighbor words for prediction. In addition to masked word prediction, BERT is also pre-trained on a "next sentence prediction" task. This task helps BERT better understand longer term relationship between sentences. As a result, pre-trained BERT can do great on tasks such as text summarization (Liu and Lapata 2019). After fine-tuning for specific tasks, BERT gives state of the art in many challenging Natural Language Processing tasks (McCormick and Ryan 2019). We use the LARGE pre-trained version of BERT in this study and fine-tune it to find a numerical representation of a sentence that describes one social event.

In this project we used BERT large model (uncased) which is a 24-layer neural network, with 1024 hidden dimension and 16 attention heads. This model has 336M parameters (Devlin et al. 2018).

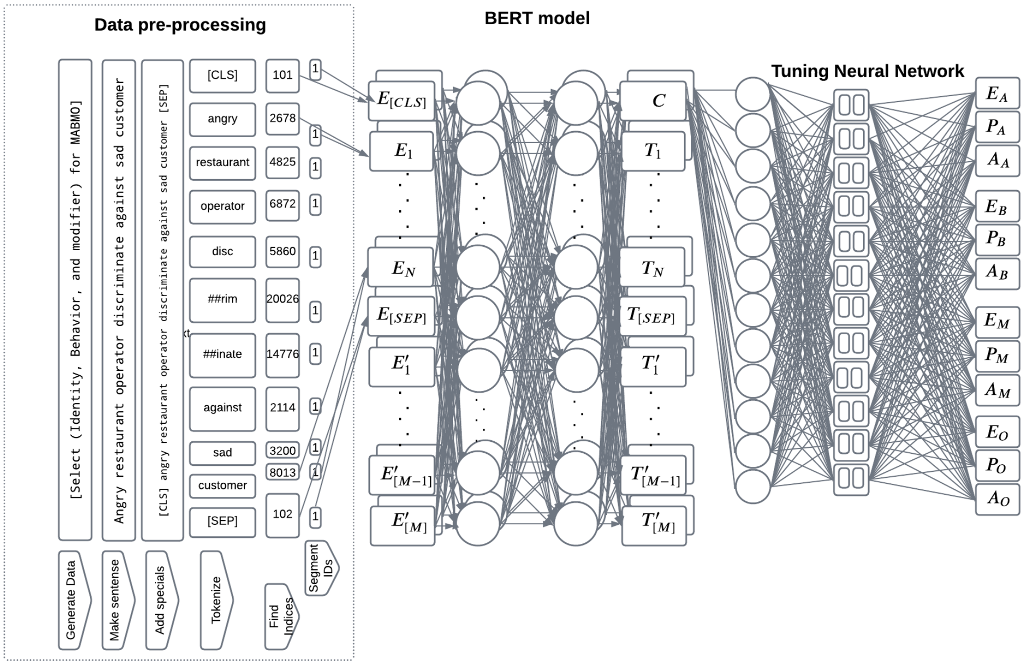


Figure 3.4: BERTNN pipeline. Sentences describing an event are generated in the first step and pre-processed to pass to a BERT model. Then outputs corresponding to the [CLS] token is passed to a three-layer neural network to find affective meaning.

3.4.2 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 that word can be given different representation when the word is used a as a *behavior* or an *identity* if we use them in a proper sentence. We take full advantage of the BERT embedding by training it on synthetic data that represents the concepts and their affective category simultaneously.In the analysis that follows, we show how to generate a contextual data-set describing social events to train a deep neural network and use this network to predict affective meanings of new concepts. We use BERT and fine tune it for finding affective lexicons. Our approach uses the vectorized representation of the whole sentence.

We introduce a new framework (Mostafavi, M. D. Porter, and Robinson 2022), which processes synthetic data, passes it to a pre-trained BERT model, and fine-tunes the result by a three layer neural network to generate extended affective meaning dictionaries. The pipeline for our method (BERTNN) is shown in Figure 3.4 and we discuss the details in this section. The data used to train our model were generated from the affective dictionary described in Smith-Lovin, Robinson, Cannon, Clark, et al. 2016. This dictionary was developed from surveys conducted between 2012-2014 and includes 929 *identities*, 814 *behaviors*, and 660 *modifiers*. The data used in this study are publicly available on the internet (Smith-Lovin, Robinson, Cannon, Clark, et al. 2016).

To take advantage of the BERT model, we generated synthetic data to fine tune the model. The synthetic data set is a set of sentences that include the concepts of interest. Loon and Freese 2022 used token embedding of terms such as "assailant is" to find the embedding for "assailant". In this case, they get one estimation for each concept. In BERTNN, we use synthetic sentences that describe *Modified Actor Behaves Modified Actor* (MABMO) events and estimate the affective meaning of all concepts in the event. In this case, we use the vectorized representation of the whole sentence to estimate a 15-dimensional lexicon vector. This gives different estimations for a concept depending on other concepts used with the concept of interest in MABMO grammar. Then we aggregate multiple estimations to get a distribution for the concept of interest.

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 some pre-processing steps to select events that are general enough to span the diverse event space. The pre-processed data generate synthetic data that can represent the whole affective space of the dictionaries. Algorithm 4 describes the pre-processing steps to include contextual aspects of concepts before using BERT model.

Affective meaning of the concepts are distributed across the EPA dimensions. However, they are not uniformly distributed across all three dimensions. Since we use a few iterations to fine-tune BERT model, random sampling from these values may use a train set from a specific region that is not necessarily representing the whole space. 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 behavior. Every concept in *identity*, *behavior*, and *modifier* dictionaries is represented by a three dimensional affective vector. We are looking for training samples that represents different regions/clusters within each dictionary. In the first step of Algorithm 4, we needed to know how many different

Algorithm 1 Pre-processing algorithm

Data: Affective dictionaries for Identities, Modifiers, and Behaviors

1. Find clusters/regions for the concepts in each of the Identity, Modifier, and Behavior dictionaries
2. Split the data into train, test, and validation sets using stratified sampling on the clusters
3. Make all ABO events using the train set
4. Use impression change equations to find the impression of concepts in all events.
5. Find the sign of the difference between baseline sentiments and impressions of the events across all 9 dimensions (3 concepts in ABO grammer $\times$ 3 EPA dimensions). In other words, for the actor, the behavior, and the object, we find whether the impression of the event increases or decreases each of their EPA values.
6. Use binary encoding to label variations of the events. For example, 111000111 means that the impression of the event for the actor and the object are increased across all EPA dimensions and decreases for the behavior across all EPA dimensions.
7. Define a training data generator that uses defined binary encodings to make synthetic data that spans all possible different interactions.
8. Add two random modifiers to ABO events created in the previous step to make synthetic MABMO events.

regions/clusters we have in the dictionaries and uniformly sample from each of them. For this purpose, we clustered each *identity*, *behavior*, and *modifier* dictionaries using K-Means clustering. We used the elbow method and determined the number of clusters for each of these three dictionaries. We concluded that 5 clusters/regions are enough for all three categories.

As a common practice in data-mining methods, the second step in Algorithm 4 is to partition the data into training, test, and validation set. We used the train set to fine-tune the BERT model and the test set to choose hyper-parameters such as the number of epochs. We selected hyper-parameters that resulted in an acceptable loss value for the test set. The validation set was the hold-out data. The results reported in this paper come from the validation set. We used stratified sampling on regions/clusters in affective dictionaries to get similar distributions. Each cluster of *identity*, *behavior*, and *modifier* categories is one “strata”. We randomly sample 85% of the words in each strata for the training, 8% are selected as the test set, and the remaining 12% are used as a validation set.

As we discussed the importance of *deflection* earlier, in the third step, we are looking to capture variation in factors that change deflection. When we select some 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 train set, we found the impression of all possible events in ABO grammar using impression change equations. Then the difference between the impressions and baseline concepts are calculated across all 9 dimensions. Among all $2^9 = 512$ possible variations, 508 type of events were present in the training set. We used binary encoding to label these 508 types of events.

Affective dictionaries are collected from surveys that participants rate a concept in one of the *identity, behavior,* or *modofier* categories. On the other hand, to utilize a BERT model we should make synthetic data that is a list of sentences that represents words of interests and their corresponding category. To generate each synthetic sentence, we used two *identities*, two *modifiers*, and one *behavior* from the training set. Synthetic Sentences with an MABMO grammar (e.g., Happy employee greets bossy employer) represent a social event. In other words, each sample describes an event in MABMO grammar.

For MABMO grammar, we need to sample two modifiers and one ABO event. We used stratified sampling based on modifier clusters and ABO labels developed in Algorithm 4. After pre-processing and tokenizing MABMO sentences similar to “data pre-processing” part of the pipeline, they are sent 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 we are dealing with a word outside 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] token 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

Table 3.2: Comparing four different affective dictionaries (Indiana (C. 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, Robinson, Cannon, Clark, et al. 2016).

		MAD			**RMSD**			**Correlation**			**Count**
		E	**P**	**A**	**E**	**P**	**A**	**E**	**P**	**A**	
Identity	**Indiana**	0.41	0.48	0.44	0.53	0.63	0.57	.956***	.905***	.807***	317
	NorthCarolina	0.57	0.58	0.69	0.71	0.71	0.85	.927***	.906***	.653***	453
	Texas	0.45	0.52	0.77	0.59	0.64	0.95	.936***	.910***	.539***	319
Behavior	**Indiana**	0.54	0.61	0.59	0.67	0.74	0.73	.967***	.801***	.665***	288
	NorthCarolina	0.68	0.5	0.52	0.82	0.63	0.65	.954***	.798***	.697***	500
	Texas	0.53	0.56	0.76	0.65	0.69	0.94	.965***	.808***	.321***	225
Modifier	**Indiana**	0.61	0.52	0.58	0.71	0.65	0.70	.969***	.923***	.857***	276
	NorthCarolina	0.78	0.72	0.63	0.92	0.89	0.81	.958***	.907***	.794***	407
	Texas	0.78	0.6	0.78	0.88	0.73	0.92	.970***	.909***	.794***	58

a dense input layer followed by two "Relu" hidden layers. Relu function adds required non-linearity in the network. 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 L_2 loss used in the neural networks minimized the squared error between estimated affective meaning and the target values across all 15 dimensions.

We implemented the neural network in Python using Torch and Transformers packages. We used "AdamW" with a learning rate of 2e-5 and batch size of 64. The BERT model and neural network are tuned for a few epochs. The data in the testing set is used to decide on how many iteration results in a good model. The neural network output is a 15-dimension vector that is highly correlated with the affective meaning variables. The code is available upon request to the first author.

The tuning layers in our framework are trained based on the target affective values. In training the neural network, we used early stopping to get reasonably low errors in the training and test set. To find a reasonably low estimation errors for the model, we compared Mean Absolute Distance (MAD), Root Mean Square Distance (RMSD), and the correlation between common concepts in different affective dictionaries. Table 3.2 compares the affective dictionary used in this study (Smith-Lovin, Robinson, Cannon, Clark, et al. 2016) with some other affective dictionaries.

We actively checked L2loss for the training and test set 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. Since we have a contextual representation of words, we get multiple representations for a word from BERTNN depending on other terms in MABMO event. In other word, for affective meaning of a given *actor* in MABMO grammar, we get different estimates depending other *modifiers*, *behavior*, and *object* used in the event.

Since the BERT model returns multiple representations for a concept, we have to adopt an aggregation approach. One approach is generating multiple MABMO events including the concept of interest. Then the estimated affective meaning values in these events represent a distribution for its value. We can use the mean value as an estimation. For example, if "moderator" is a new *identity* outside the affective dictionary, we make multiple events that includes this identity such as, "moderator help angry client". All events are passed to the model which estimates the affective meaning of the target word in each event. The average affective meaning is returned as the final estimated meaning. Since MABMO event includes two identities and two modifiers, we concatenated estimations for the *identities* and *modifiers* from these two sets. Then we found the average of estimations from all the samples.

Comparing BERTNN with (Loon and Freese 2022) work, the main differences are,

- BERTNN makes multiple synthetic sentences to take advantage of contextual embedding of the BERT model. These sentences are made based on MABMO grammar in affect control theory literature. It helps BERTNN to consider deflection and regions of concepts in training.

- Instead of optimizing only the fine-tuning layer, BERTNN trains the BERT+fine-tuning layer for a few epochs.

- BERTNN uses sentence embedding instead of the token embedding of the core concepts. As a result, it can estimate affective meaning for compound words and words with more than one token.

- BERTNN gives a distribution for affective meaning based on the synthetic data used.

We make BERTNN a publicly available tool that researchers with minimal coding experience can use to estimate the affective meaning of their words of interest. This tool is available in the following Google Collab link,

https://colab.research.google.com/drive/1ej1wldgDgjOOu2OBf3xXasq51L6V-gft?usp=sharing.

3.5 Results

Variations in estimated affective meanings from the model can arise from a variety of sources, including,

- Splits of the original data to make train, test, and validation changes the final result. Consequently, we conducted a series of robustness checks. We get similar results using the test set for model selection and increasing or decreasing the number of iterations. We also simulated this process with various seeds, and the final results are similar.

- Mean, median, or similar statistics such as trim-mean of multiple sample estimations for a concept, result in different final results. In this study, we used the mean values across all dimensions.

- An *identity* is used as either actor or *object* in the MABMO grammar. Similarly, *modifier* can modify actor or *object*. The values that we get for a concept in these cases varies 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.* It is possible to run studies to differentiate the values one can get for *identity* as the *actor* or the *object*.

We compared the performance of our model with approaches tried in previous work such as different word analogy (Kozlowski, Taddy, and Evans 2019), regressions (M. Li et al. 2017), translation matrix methods (Mostafavi and M. Porter 2021), and BERT model on the core concepts (Loon and Freese 2022) to find affective meanings in Table 3.3. We used RMSE, MAE, and correlation analysis to compare the result. The validation data used to

Table 3.3: Performance of several models on (Smith-Lovin, Robinson, Cannon, Clark, et al. 2016) data. Bold indicates the best model. Our model, BERTNN, performed best in most categories.

		MAE			RMSE			Correlation		
		E	**P**	**A**	**E**	**P**	**A**	**E**	**P**	**A**
Identity	Analogy stepW.	0.92	0.92	0.74	1.15	1.14	0.94	.744***	.538***	.299***
	Analogy_regression	0.93	0.94	0.76	1.16	1.17	0.96	.748***	.505***	.249***
	StepW Translation	0.80	0.78	0.69	0.97	1.04	0.87	.824***	.633***	.509***
	CoreBERT	0.65	0.69	0.66	0.88	0.89	0.79	.860***	.638***	.485***
	BERTNN	**0.53**	**0.54**	**0.56**	**0.73**	**0.70**	**0.72**	**.893*****	**.817*****	**.611*****
Behavior	Analogy stepW.	1.24	0.69	0.67	1.54	0.90	0.85	.662***	.387***	.352***
	Analogy_regression	1.25	0.72	0.70	1.56	0.92	0.88	.653***	.388***	.295***
	StepW Translation	1.07	0.69	0.58	1.21	0.84	0.82	.767***	.456***	.512***
	CoreBERT	0.74	**0.44**	0.51	0.97	**0.55**	0.64	.893***	**.746*****	**.747*****
	BERTNN	**0.69**	0.47	**0.48**	**0.87**	0.62	**0.61**	**.911*****	.707***	.730***
Modifier	Analogy stepW.	0.98	0.70	0.89	1.23	0.88	1.10	.827***	.832***	.537***
	Analogy_regression	1.01	0.74	0.94	1.27	0.91	1.13	.816***	.834***	.504***
	StepW Translation	0.85	0.59	0.73	0.98	0.67	0.84	.869***	.840***	.650***
	CoreBERT	0.63	0.50	0.65	0.80	0.64	0.79	.932***	.883***	.752***
	BERTNN	**0.60**	**0.44**	**0.54**	**0.78**	**0.58**	**0.71**	**.936*****	**.912*****	**.812*****

compare these methods included 139 *identities*, 122 *behaviors*, and 99 *modifiers* came from stratified sampling discussed earlier.

All the similar works except (Loon and Freese 2022) used shallow embedding (Kozlowski, Taddy, and Evans 2019; Mostafavi and M. Porter 2021; M. Li et al. 2017). We used the publicly available code of Mostafavi and M. Porter 2021 and Kozlowski, Taddy, and Evans 2019 to replicate their works and compare the result. For M. 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 pre-trained model in Loon and Freese 2022 to compare with their work. Table 3.3, shows the best result we could get from other methods. We can observe from this table that our approach reached the best result across most of the metrics. We can observe from Table 3.3 that in some metrics such as the correlation metric for Activity, the improvement over shallow-embedding methods is about 20% for behaviors. In terms of error rates, we can find BERTNN achieves values that are comparable with differences between different affective dictionaries. For example, MAE value that BERTNN achieves for modifiers in Table 3.3 is smaller than

Table 3.4: Comparing estimated affective meanings for "judge" as an *identity* and *behavior*.

Judge	**Evaluation**	**Potency**	**Activity**	**Est. Evaluation**	**Est. Potency**	**Est. Activity**
Behavior	-1.83	0.71	0.07	-1.40	0.88	-0.03
Identity	1.15	2.53	-.22	1.95	2.27	0.67

Table 3.5: Correlation analysis for identities. (a) Correlation between estimated values (EE, EP, and EA) and the affective dictionary value (E, P, and A). We can also compare the correlation of the words in the three dimensions shown in (b) and correlation of the estimated values of the three dimension shown in (c) to find how close the off-diagonal entries are in the estimation comparing to dictionary values.

(a)	**E**	**P**	**A**
EE	.894***	.423***	.094
EP	.394***	.816***	.186
EA	0.009	.269**	.611***

(b)	**E**	**P**	**A**
E	1.00***	.462***	.133
P	.452***	1.0***	.334***
A	.133	.334***	1.0***

(c)	**EE**	**EP**	**EA**
EE	1.00***	.448***	.044
EP	.448***	1.0***	.242*
EA	.044	.242*	1.0***

three of dictionaries shown in Table 3.2. In other word, BERTNN estimation error on the validation set is small enough to say this model estimates EPA values similar to running a new survey.

One problem with estimation from shallow-embeddings was having the same embedding for words as *identity* or *behaviors*. Using the BERTNN, we can differentiate between these two cases. In In Table 3.4 we can observe that estimated values for "judge" are different when it is considered as *identity* or *behavior*.

To evaluate how close our expanded dictionaries are to the baseline affective meanings, we calculated the correlation for *identity* and *behaviors* in Tables 3.5 and 3.6. The correlation (a) between the result of the validation set in our method, (EE, EP, and EA) and the EPA values from the surveys (E, P, and A) are shown. The diagonal terms are the correlation values we have seen earlier in Table 3.3. Also, you can find the correlation between the three dimensions from the survey data are shown in (b), and the correlation between estimated three dimensions are shown in (c). We can observe in both tables the values in tables (b) and (c) are close. It reveals that BERTNN estimation is highly correlated with survey data and the cross-term dynamics are well estimated.

Table 3.6: Correlation analysis for behaviors. (a) Correlation between estimated values (EA, EP, and EA) and the affective dictionary value (E, P, and A). We can also compare the correlation of the words in the three dimensions shown in (b) and correlation of the estimated values of the three dimension shown in (c) to find how close the off-diagonal entries are in the estimation comparing to dictionary values.

(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***

Table 3.7: 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
3.28	2.61	0.82	2.55	2.75	0.47	1.91	2.16	0.96	2.81	2.81	1.13	2.81	2.48	0.66

Tables 3.5 and 3.6 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.

Model outputs. BERTNN can represent one participant evaluating affective meaning of concepts from an MABMO grammar. Table 3.7 shows estimation of concepts used in *Happy doctor help wonderful mother*.

Since BERTNN gives us contexal representation of words, we can create multiple MABMO events with a common word to estimate its affective meaning in general. For example, *phone* as a behavior is a concept that is not evaluated in affective dictionaries. We can create multiple events similar to estimate its affective meaning in different contexts. Table 3.8 includes estimated affective meaning from multiple MABMO events. Agrregating 300 events similar to Table 3.8, we can get a reliable estimate for the concept of interest. Table 3.9 has a summary statistics of events used to estimate affective meaning of *phone* as a behavior.

Table 3.8: Estimated affective meaning of *phone* from MABMO different events.

MABMO event	E	P	A
Respectful neurotic phones grumpy golfer	0.86	0.79	0.99
Glad celebrity phones fashionable	1.05	0.92	1.02
Hurt fiance phones conventional intimate	1.17	1.04	1.15
Rigid scapegoat phones egotistical magician	0.48	0.72	1.04
Straightforward single parent phones opportunistic hippie	1.16	0.94	1.09
relentless golfer phones withdrawn spouse	1.23	1.01	1.09

Table 3.9: Estimating affective meaning of *phone* from 300 MABMO events.

	E	P	A
Mean	1.03	0.98	1.06
Standard deviation	0.25	0.09	0.10
Minimum	-0.52	0.64	0.79
Maximum	1.61	1.21	1.47

3.6 Discussion

Creating a synthetic data set of events corresponding to the core grammar of Affect Control Theory, pre-processing these data, and then sending them through the pre-trained BERT word embedding model yielded predictions that seem to have great promise for generating the sort of 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 the means by which humans acquire social meaning through the observation of participation in social interaction. When we encounter a new culture or subculture with an identity we do not recognize, we observe interactions in order to gather information. If we observe occupants of this identity in many interactions (some of the form MABMO), each interaction leaves an impression about the likely meaning of that identity. If someone always does good things, we will come to infer a positive evaluation for his/her identity. So 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. Based on 500 interactions, an affective meaning for the unknown identity is formed in our minds. Like the impression change equations we described previously, we expect the person to be nice if we observe lots of nice behaviors in observed interactions, or routinely interacts with other positively identified actors. In this way, our method of learning the affective meanings associated with known/unknown identities is very similar to the mask language model.

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 order 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 in this paper promise to minimize this burden by sidestepping the need for human raters. This approach could greatly expand the ability of affect control theory to be applied to new substantive domains at greater speed and at a lower burden for research respondents, researchers, and funders. Using machine learning approaches pre-trained on very large data sets allow the possibility of rapidly expanding the existing sentiment dictionaries "on the fly" rather than needing to design, implement, analyze, and interpret new cultural data. Moreover, the ability to train these context-embedding models on new data sets opens up the possibility of moving into novel cultural and subcultural domains more quickly (skipping years of fieldwork or thousands of surveys).

3.7 References

Adamopoulou, Eleni and Lefteris Moussiades (2020). "An overview of chatbot technology". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 373–383.

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

Andre, Elisabeth et al. (2004). "Endowing spoken language dialogue systems with emotional intelligence". In: *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer, pp. 178–187.

Averett, Christine and David R Heise (1987). "Modified social identities: Amalgamations, attributions, and emotions". In: *Journal of Mathematical Sociology* 13.1-2, pp. 103–132.

Barbieri, Francesco, Francesco Ronzano, and Horacio Saggion (2016). "What does this emoji mean? a vector space skip-gram model for twitter emojis". In: *Calzolari N, Choukri*

K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72. ELRA (European Language Resources Association).

Beattie, Austin, Autumn P Edwards, and Chad Edwards (2020). "A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication". In: *Communication Studies* 71.3, pp. 409–427.

Britt, Lory and David R Heise (1992). "Impressions of self-directed action". In: *Social Psychology Quarterly*, pp. 335–350.

Budzianowski, Paweł et al. (2018). "MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling". In: *arXiv preprint arXiv:1810.00278*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dimson, Thomas (2015). "Emojineering part 1: Machine learning for emoji trends". In: *Instagram Engineering Blog* 30.

Eisner, Ben et al. (2016). "emoji2vec: Learning emoji representations from their description". In: *arXiv preprint arXiv:1609.08359*.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition and Emotion*, pp. 169–200.

Eric, Mihail et al. (2019). "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines". In: *arXiv preprint arXiv:1907.01669*.

Feng, Shutong et al. (2022). "EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fonnegra, Rubén D and Gloria M Dıaz (2018). "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model". In: *International Conference on Human-Computer Interaction*. Springer, pp. 385–396.

Fontaine, Johnny RJ et al. (2007). "The world of emotions is not two-dimensional". In: *Psychological science* 18.12, pp. 1050–1057.

Francis, Clare and David R Heise (2006). "Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3, 2006". In: *Computer file]. Distributed at Affect Control Theory Website, Program Interact (http://www. indiana. edu/˜ socpsy/ACT/interact/JavaInteract. html)*.

Francis, Linda E (1997a). "Emotion, coping, and therapeutic ideologies". In: *Social perspectives on emotion* 4, pp. 71–102.

— (1997b). "Ideology and interpersonal emotion management: Redefining identity in two support groups". In: *Social Psychology Quarterly*, pp. 153–171.

Ghosal, Deepanway et al. (2020). "Cosmic: Commonsense knowledge for emotion identification in conversations". In: *arXiv preprint arXiv:2010.02795*.

Gliwa, Bogdan et al. (2019). “Samsum corpus: A human-annotated dialogue dataset for abstractive summarization”. In: *arXiv preprint arXiv:1911.12237*.

Goldstein, David M (1989). “Control theory applied to stress management”. In: *Advances in Psychology*. Vol. 62. Elsevier, pp. 481–491.

Heise, David R (1977). “Social action as the control of affect”. In: *Behavioral Science* 22.3, pp. 163–177.

— (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

— (2013). “Interact guide”. In: *Department of Sociology, Indiana University*.

Heise, David R and Cassandra Calhan (1995). “Emotion norms in interpersonal events”. In: *Social Psychology Quarterly*, pp. 223–240.

Heise, David R and Lisa Thomas (1989). “Predicting impressions created by combinations of emotion and social identity”. In: *Social Psychology Quarterly*, pp. 141–148.

Hunt, Pamela M (2008). “From festies to tourrats: Examining the relationship between jamband subculture involvement and role meanings”. In: *Social Psychology Quarterly* 71.4, pp. 356–378.

Kelley, John F (1984). “An iterative design methodology for user-friendly natural language office information applications”. In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41.

Koch, Andrew, Jiahao Tian, and Michael D Porter (2020). “Criminal Consistency and Distinctiveness”. In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–3.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). “The geometry of culture: Analyzing the meanings of class through word embeddings”. In: *American Sociological Review* 84.5, pp. 905–949.

Kriegel, Darys J et al. (2017). “A multilevel investigation of Arabic-language impression change”. In: *International Journal of Sociology* 47.4, pp. 278–295.

Li, Minglei et al. (2017). “Inferring affective meanings of words from word embedding”. In: *IEEE Transactions on Affective Computing* 8.4, pp. 443–456.

Li, Shan and Weihong Deng (2020). “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing*.

Li, Yanran et al. (2017). “Dailydialog: A manually labelled multi-turn dialogue dataset”. In: *arXiv preprint arXiv:1710.03957*.

Liu, Yang and Mirella Lapata (2019). “Text summarization with pretrained encoders”. In: *arXiv preprint arXiv:1908.08345*.

Loon, Austin van and Jeremy Freese (2022). “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language”. In: *American Behavioral Scientist*. DOI: 10.1177/00027642211066046. eprint: https://doi.org/10.1177/00027642211066046.

MacKinnon, Neil J and Dawn T Robinson (2014). “Back to the future: 25 years of research in affect control theory”. In: *Advances in group processes*.

Majumder, Navonil et al. (2019). “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6818–6825.

McCormick, Chris and Nick Ryan (May 2019). *BERT Word Embeddings Tutorial*. URL: http://www.mccormickml.com.

Mikolov, Tomas, Kai Chen, et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. In: *arXiv preprint arXiv:1309.4168*.

Mostafavi, Moeen (2021). “Adapting Online Messaging Based on Emotiona”. In: *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization*.

Mostafavi, Moeen, Maria Phillips, et al. (2021). “A tale of two metrics: Polling and financial contributions as a measure of performance”. In: *2021 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.

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

Mostafavi, Moeen, Michael D Porter, and Dawn T Robinson (2022). “Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers”. In: *arXiv preprint arXiv:2202.00065*.

Mostafavi, Moeen, Mahsa Pahlavikhah Varnosfaderani, et al. (2022). “emojiSpace: Spatial Representation of Emojis”. In: *arXiv preprint arXiv:2209.09871*.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CBO9780511571299.

Osgood, Charles Egerton, William H May, et al. (1975). *Cross-cultural universals of affective meaning*. Vol. 1. University of Illinois Press.

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

Polzin, Thomas S and Alexander Waibel (2000). “Emotion-sensitive human-computer interfaces”. In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Poria, Soujanya et al. (2018). “Meld: A multimodal multi-party dataset for emotion recognition in conversations”. In: *arXiv preprint arXiv:1810.02508*.

Quinonero-Candela, Joaquin et al. (2008). *Dataset shift in machine learning*. Mit Press.

Rashotte, Lisa Slattery (2002). “Incorporating nonverbal behaviors into affect control theory”. In: *Electronic Journal of Sociology* 6.3.

Reelfs, Jens Helge et al. (2020). “Word-Emoji Embeddings from large scale Messaging Data reflect real-world Semantic Associations of Expressive Icons”. In: *arXiv preprint arXiv:2006.01207*.

Robillard, Julie M and Jesse Hoey (2018). “Emotion and motivation in cognitive assistive technologies for dementia”. In: *Computer* 51.3, pp. 24–34.

Robinson, Dawn T, Jody Clay-Warner, et al. (2012). “Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques”. In: *Biosociology and neurosociology*. Emerald Group Publishing Limited.

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

— (1999). “Emotion display as a strategy for identity negotiation”. In: *Motivation and Emotion* 23.2, pp. 73–104.

— (2018). “Affect control theories of social interaction and self.” In:

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”. In: *Social Forces* 73.1, pp. 175–190.

Rogers, Kimberly B (2018). “Do you see what I see? Testing for individual differences in impressions of events”. In: *Social Psychology Quarterly* 81.2, pp. 149–172.

Rogers, Kimberly B and Lynn Smith-Lovin (2019). “Action, interaction, and groups”. In: *The Wiley Blackwell Companion to Sociology*, pp. 67–86.

Russell, J.A. (1980). “A circumplex model of affect”. In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.

Saha, Tulika, Sriparna Saha, and Pushpak Bhattacharyya (2020). “Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning”. In: *PloS one* 15.7, e0235367.

Schneider, Andreas (2006). “Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998”. In: *Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Schröder, Tobias and Wolfgang Scholl (2009). “Affective dynamics of leadership: An experimental test of affect control theory”. In: *Social Psychology Quarterly* 72.2, pp. 180–197.

Shi, Weiyan and Zhou Yu (2018). “Sentiment adaptive end-to-end dialog systems”. In: *arXiv preprint arXiv:1804.10731*.

Smith, Herman W (2002). “The dynamics of Japanese and American interpersonal events: Behavioral settings versus personality traits”. In: *Journal of Mathematical Sociology* 26.1-2, pp. 71–92.

Smith, Herman W and Linda E Francis (2005). “Social vs. self-directed events among Japanese and Americans”. In: *Social Forces* 84.2, pp. 821–830.

Smith-Lovin, Lynn (1979). "Behavior settings and impressions formed from social scenarios". In: *Social Psychology Quarterly*, pp. 31–43.

Smith-Lovin, Lynn and William Douglass (1992). "An affect control analysis of two religious subcultures". In: *Social perspectives on emotion* 1, pp. 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]*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, et al. (2016). "Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014". In: *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Brent H Curdy, et al. (2019). "Mean affective ratings of 968 identities, 853 behaviors, and 660 modifiers by amazon mechanical turk workers in 2015". In: *University of Georgia: Distributed at UGA A ect Control eory Website*.

Tian, Jiahao and Michael D Porter (2022). "Changing presidential approval: Detecting and understanding change points in interval censored polling data". In: *Stat* 11.1, e463.

Tsoudis, Olga and Lynn Smith-Lovin (1998). "How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes". In: *Social forces* 77.2, pp. 695–722.

Wang, Jiancheng et al. (Apr. 2020a). "Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 9177–9184. DOI: 10.1609/aaai.v34i05.6454. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6454.

— (2020b). "Sentiment classification in customer service dialogue with topic-aware multi-task learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9177–9184.

Youngreen, Reef et al. (2009). "Identity maintenance and cognitive test performance". In: *Social Science Research* 38.2, pp. 438–446.

Zahiri, Sayyed M and Jinho D Choi (2017). "Emotion detection on tv show transcripts with sequence-based convolutional neural networks". In: *arXiv preprint arXiv:1708.04299*.

Zang, Xiaoxue et al. (2020). "MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines". In: *arXiv preprint arXiv:2007.12720*.

Zhang, Rui, Kai Yin, and Li Li (2020). "Towards emotion-aware user simulator for task-oriented dialogue". In: *arXiv preprint arXiv:2011.09696*.

Zhu, Yukun et al. (2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

CHAPTER 4

IMPROVING EMOTION UNDERSTANDING IN TASK-ORIENTED CONVERSATIONS: INCORPORATING SEQUENTIAL ASPECTS AND IDENTITY-BASED INSIGHTS

My goal is to develop a method to understand emotions in task-oriented conversations. For this purpose, I use publicly available task-oriented conversational agent data with emotional labels called EmoWOZ (Feng et al. 2022). It is a valuable resource for developing and evaluating emotion recognition models for task-oriented dialogues. The EmoWOZ dataset consists of a collection of goal-oriented dialogues between a user and a conversational agent, where the user's emotions are annotated with specific labels. I am building on a baseline BERT model used in EmoWOZ original work, which only considered the user's message to estimate emotion. However, I believe that considering the conversational agent's last reply and the user's message will provide a more accurate representation of the interaction between the two identities, which aligns with affect control theory.

4.1 Abstract

In this chapter, I present a method to enhance the understanding of emotions in task-oriented conversations. The goal of my method is to incorporate both the interactional and sequential aspects to better capture the influence of each identity on the emotional expression in the conversation, in line with the insights from the affect control theory. To achieve this, I build on the baseline BERT model used in EmoWOZ and add a sequential model to capture the change in emotions over time. To estimate the emotion labels, I consider four different scenarios and develop an adaptive extension of the Macro F1-score to perform better on imbalanced data. My method achieves better results than alternative methods in the literature, and the sequential model improves the performance in capturing

the sequential aspects of the conversation. Overall, my method offers a better understanding of emotions in task-oriented conversations and can potentially improve the development of conversational agents.

4.2 Introduction

Inspired by the Affect Control Theory framework, I have redefined messaging as an interaction between a user and an agent. I use messages from both the user and the conversational agent to develop this model. According to affect control theory, individuals adopt specific roles in social interactions, and their emotions are a product of their efforts to maintain or change their impressions of themselves and others. Therefore, I aim to capture the influence of each identity on the emotional expression in the conversation by incorporating both the user's message and the conversational agent's last response.

To accomplish this, I add a "SEP" token between the two messages, and the BERT model learns the emotions based on their interaction. I also use stratified sampling to split the data, which is more suitable for this task because the original data partitioning was for another purpose.

Emowoz dataset includes three annotations and a final label for every message from the user. To produce the interactional model, I estimate the emotion labels used in EmoWOZ, considering three different scenarios. In the first scenario, I use the final label produced by the EmoWOZ authors after their supervision. In the second scenario, I use unique labels from the three annotators and replicate the utterance if there is no complete agreement. In the third scenario, I use unique labels but duplicate the label if there is full agreement. Finally, in the fourth scenario, I use the unique labels in combination with the final label for the embedding part. To address the sequential aspects, I add a sequential model on top of the embedding, using the probability/odds of having each emotion as an input for a time series transformer model. This allows me to capture the change in emotions over time, similar to the impression change equations in affect control theory, but with a longer horizon.

To select the best model, I find that randomness in the development data can improve the F1-score but not necessarily the performance on the test set. Thus, I select the model that performs the best over the last four steps to reduce the effect of randomness.

Finally, I develop an adaptive extension of the weighted cross entropy loss used in the baseline model. Rather than using weights based on the frequency of emotions, I use weights based on the F1-score of each emotion, allowing the model to perform better on imbalanced data.

Overall, my method builds on the baseline BERT model used in EmoWOZ but considers the interaction between the user and conversational agent's messages, drawing on insights from affect control theory to better understand the role of identity in emotional expression. The sequential model also extends the impression change equations in affect control theory to cover a longer time horizon, allowing me to capture the dynamic nature of emotions in task-oriented conversations.

4.3 Reevaluating Data Partitioning for Emotion Detection in EmoWOZ

4.3.1 Abstract

This project focuses on the EmoWoz dataset, an extension of MultiWOZ that provides emotion labels for the dialogues. MultiWOZ was partitioned initially for another purpose, resulting in a distributional shift when considering the new purpose of emotion recognition. The emotion tags in EmoWoz are highly imbalanced and unevenly distributed across the partitions, which causes sub-optimal performance and poor comparison of models. I propose a stratified sampling scheme based on emotion tags to address this issue, improve the dataset's distribution, and reduce dataset shift. I also introduce a special technique to handle conversation (sequential) data with many emotional tags. Using our proposed sampling method, models built upon EmoWoz can perform better, making it a more reliable resource for training conversational agents with emotional intelligence. I recommend that future researchers use this new partitioning to ensure consistent and accurate performance evaluations.

4.3.2 Introduction

Emotion recognition in task-oriented conversational agents is challenging because it requires the agent to accurately interpret and respond to a user's emotional state in real time. Emotional signals can be complex and difficult to detect accurately, especially in unstructured conversations where users may not express their emotions explicitly. Another challenge is that emotions can be context-dependent, meaning that they may be influenced by the user's past experiences, current environment, and cultural background. Therefore, conversational agents need to interpret these contextual factors accurately to provide appropriate responses sensitive to the user's emotional state.

Despite the challenges, emotion recognition in task-oriented conversational agents is important because it can improve the overall user experience (Zhang, Yin, and L. Li 2020). By accurately detecting and responding to a user's emotional state, conversational agents can provide more personalized and empathetic interactions, increasing user satisfaction and engagement. Additionally, emotion recognition can help agents identify when a user is experiencing frustration, confusion, or other negative emotions, allowing them to intervene and provide support to prevent user dropout or dissatisfaction(Andre et al. 2004). Emotion recognition is critical to developing effective task-oriented conversational agents that can provide a human-like user experience (Polzin and Waibel 2000).

Sentiment analysis in task-oriented conversational agents has been addressed in the literature as an essential aspect of natural language processing that can improve the overall user experience (Shi and Yu 2018; T. Saha, S. Saha, and Bhattacharyya 2020; Wang et al. 2020b). However, the lack of publicly available data for emotion recognition is a significant limitation for task-oriented conversational agent applications.

MultiWOZ (Multi-Domain Wizard-of-Oz) is a large-scale dataset of human-human written conversations for task-oriented dialogue modeling. The dataset was initially collected for training and evaluating dialogue systems, particularly those designed to assist users with completing specific tasks such as booking a hotel or reserving a table at a restaurant (Budzianowski et al. 2018). (Feng et al. 2022) extended the MultiWOZ dataset by including dialogues between humans and a machine-generated policy, which they named DialMAGE. The resulting dataset was called EmoWOZ. EmoWOZ is a large-scale, manually emotion-annotated corpus of task-oriented dialogues. The corpus contains more than 11K dialogues with more than 83K emotion annotations of user utterances, which makes it the first large-scale open-source corpus of its kind. The authors propose a novel emotion labeling scheme tailored to task-oriented dialogues and demonstrate the usability of this corpus for emotion recognition and state tracking in task-oriented dialogues. The paper highlights that while emotions in chit-chat dialogues have received considerable attention (Y. Li et al. 2017; Poria et al. 2018; Zahiri and Choi 2017), emotions in task-oriented dialogues remain largely unaddressed. They argue that incorporating emotional intelligence can help conversational AI generate more emotionally and semantically appropriate responses, making a better user experience.

In their study, the authors described their methodology for partitioning the EmoWOZ dataset into training, validation, and testing sets while maintaining the original split of the MultiWOZ dataset. They also divided the DialMAGE dataset into three parts with a ratio of 8:1:1. However, in the following section, we will examine the limitations of their approach to data partitioning and suggest an alternative method.

We specifically concentrate on the MultiWOZ section of EmoWOZ in this paper as it is widely used in various applications. Therefore, we do not discuss all subsets of the EmoWOZ in detail. Nonetheless, our approach can be readily extended to the entire EmoWOZ dataset.

4.3.3 Data partitioning

When building a predictive model, we typically split our data into three sets: a training set, a validation set, and a test set. The purpose of the training set is to estimate model parameters, and the purpose of the validation set is to tune the model's hyperparameters and assess its performance. The test set aims to get an unbiased estimate of the model's performance on new, unseen data. It's important to note that the test set should not be used in any part of model fitting or model selection because doing so can lead to overfitting and inaccurate performance estimate. However, after we have trained and validated our models on the training and validation sets, we can use the test set to compare the performance of different models and select the best one. This is where the issue of different distributions

	Relative Frequency			Manual resolution		
	Fear.	Abus.	Dis.	Fear.	Abus.	Dis.
Train	0.62	0.06	1.29	28.86	87.88	16.53
Devel.	0.22	0.08	1.00	62.50	50.00	21.62
Test	0.20	0.07	1.47	40.00	100.0	20.37

Table 4.1: Relative frequency and Manual resolution percentage for the three minority classes with less than 2% relative frequency in the MultiWOZ dataset

between the test and validation sets comes in. If the test set has a different distribution than the validation set, a model that performs well on the validation set may not be the best model for the test set, and vice versa. Therefore, it's essential to ensure that the distributions of the three sets are as similar as possible.

Approximately 2.5% of the MultiWOZ subset in (Feng et al. 2022) underwent manual annotation for resolution. However, this manual annotation was not evenly distributed across all classes, and minority classes were affected more than majority classes. For instance, the *abusive* class in the test set was completely annotated manually, while in the developing set it was manually labeled in 50% of cases. Also, the *Fearful* class had a relative frequency three times higher in the training set, as shown in Table 4.1.

Three annotators labeled the utterances according to the task. The final label was determined primarily by the majority vote of the annotators. Among all utterances, 72.1% had a complete agreement among the three annotators. A partial agreement was found for 26.4% of the utterances, while for 1.5%, there was no agreement. The paper reports that these instances were resolved manually to address cases where the annotators could not reach an agreement. In a small portion of the data, a label different from the majority vote was chosen.

We use F1-scores to compare annotators across data partitions to assess inter-annotator agreement. In essence, we measure the effectiveness of annotations by the three annotators across the training, validation, and test sets using the final labels. Table 4.2 presents the results, indicating model performance discrepancies across the training, development, and test sets. This phenomenon is known as *dataset shift* in which there is a difference between the joint distribution of inputs and outputs during the training stage compared

Data	F1 for each Emotion Label							Macro F1	
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	w/o N.	w N.
Train	93.94	54.71	50.19	72.85	32.49	42.32	88.78	56.89	62.18
Val.	94.09	26.25	46.72	73.35	50.00	43.19	88.73	54.71	60.33
Test	94.14	29.73	52.03	71.98	32.00	34.97	88.86	51.6	57.67

Table 4.2: Performance of annotators based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ. Following the benchmarks in the literature, in the aggregated level, we report Macro F1-scores **w**ith**o**out **N**eutral and **w**ith **N**eutral emotion.

to the validation and test stage (Quinonero-Candela et al. 2008), leading to a decrease in performance. Dataset Shift is a common problem in machine learning, and it can have significant consequences, such as a decrease in accuracy. In the case of EmoWoz, the dataset Shift arises from the fact that (Feng et al. 2022) kept the original partitioning of MultiWOZ, which is not evenly distributed across partitions for this particular task.

It's worth noting that while the original partitioning of MultiWOZ data is suitable for many tasks related to the development of task-oriented conversational agents, it may not be ideal for emotion detection. Emotion detection requires consideration of different contextual aspects, which may require a new partitioning approach. For instance, in the original partitioning, 60% of messages with the *Fearful* emotion in the training set were in conversations with the police or hospital. However, in the validation and test sets, none of the conversations with *Fearful* emotions were related to the police or hospital. This contextual aspect of the conversation plays a crucial role in accurately recognizing emotions.

To address this issue, we use stratified sampling, which is a sampling technique that ensures that each sub-group in the data is represented proportionally in the sample. In this case, we use stratified sampling to ensure that the training set has a distribution similar to the validation and test sets. Stratified sampling is particularly useful in situations where the distribution of the target variable is imbalanced or varies across sub-groups in the data. In the case of EmoWoz, the distribution of emotions in the training set has differed from that in the validation and test sets, which could have contributed to the dataset shift. We used the Algorithm 4 to get a new partitioning of the data. By using stratified sampling, we have ensured that the emotion distribution in the training set was similar to that in the

Algorithm 2 Stratified sampling for emotion recognition in the conversation

MultiWoz dataset with emotion labels from EmoWOZ

1. Group the dataset based on their utterance ids and find a list with the emotion sequence in each utterance.
2. Determine the frequency of emotional sequences in the dataset.
3. Make a dictionary called *emotion_seq_dict* with the emotional sequence as the key and the counts of the sequence in the dataset as the value.
4. Partition the whole dataset into one set called *frequent_seq* with conversations of more than six emotional list frequencies and another set *non_frequent_seq* with the rest of the data.
5. For the *frequent_seq*, do the stratified sampling of conversation based on the emotion sequence and partition it to the training, validation, and test set, with a 80-10-10 split similar to the original split of the data.
6. Use random sampling to partition the *non_frequent_seq* to the training, validation, and test sets.
7. Find the union of the two partitions to get the partitioning of the whole dataset.

Data	F1 for each Emotion Label							Macro F1	
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	w/o N.	w N.
Train	94.02	52.38	50.97	73.06	34.87	41.73	88.83	56.98	62.27
Val.	93.86	48.58	47.16	71.49	37.50	41.65	88.69	55.85	61.28
Test	93.79	52.08	46.13	72.20	32.26	41.54	88.49	55.45	60.93

Table 4.3: Performance of annotators based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ after new partitioning.

validation and test sets, which helps to reduce the dataset Shift and improve the model's performance. Table 4.3 shows the annotator's F1-score after the new partitioning.

4.3.4 Case study

(Feng et al. 2022) used Bert, Contextual BERT, DialogueRNN (Majumder et al. 2019), and COSMIC (Ghosal et al. 2020) for the baseline methods. Among these methods, BERT did not incorporate the sequential aspects of the conversation, yet it yielded the best Macro F1-scores in most EmoWOZ subsets. To enhance the BERT model, we compute the relative embedding of the message from the chatbot and agent and employ a transformer model to address the sequential aspects of the problem. Hyperparameters for this method using

		Original splits			Stratified splits		
Batch	Epochs	Val.	Test	Dif.	Val.	Test	Dif.
8	4	**51.4**	48.92	-2.48	52.25	51.58	-0.67
16	4	50.58	54.03	3.45	53.09	51.76	-1.33
32	4	48.96	**54.23**	5.27	55	**53.73**	-1.27
8	8	49.52	52.35	2.83	53.5	52.38	-1.12
16	8	50.31	51.77	1.46	53.54	49.68	-3.86
32	8	49.86	52.27	2.41	**56.8**	53.14	-3.66

Table 4.4: Performance of different hyper-parameters.

both the original and proposed partitioning are illustrated in Table 4.4. Upon examining the results, we can see that in the original partitioning, the hyperparameters corresponding to the top-performing model on the validation set produced the worst model on the test set. This discrepancy could be indicative of data drift, as we previously discussed. This

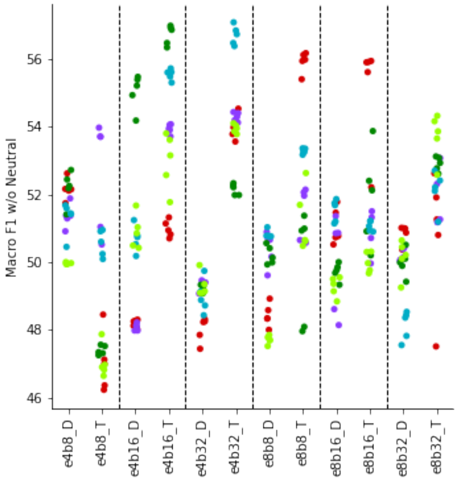


Figure 4.1: This figure depicts the Macro F1-score for each of the seeds utilized in implementing the sequential extension of the BERT model on the **origial** Multiwoz partitioning. For each hyperparameter, both the Macro F1-score in the validation and test sets are plotted in close proximity to one another. Additionally, the colors within the figure represent the five distinct seeds utilized in the embedding step.

implementation uses five distinct seeds to generate five unique embeddings. We then employed five seeds to construct the transformer model on top of these embeddings. Consequently, we had 25 different models for each parameter in Table 4.4. To visualize the Macro F1-score for each of these models, we included Figures 4.1 and 4.2. The colors within these figures correspond to the five distinct seeds utilized in the embedding step. Notably,

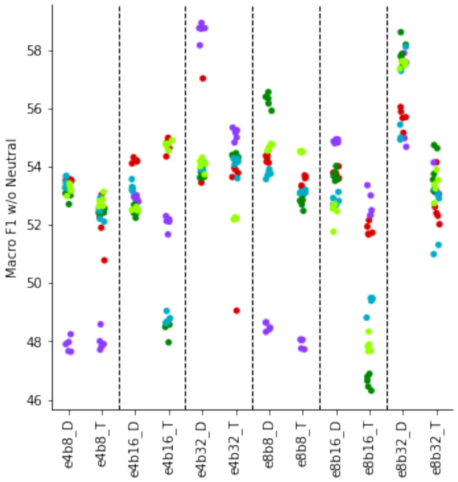


Figure 4.2: This figure depicts the Macro F1-score for each of the seeds utilized in implementing the sequential extension of the BERT model on the **proposed** Multiwoz partitioning. For each hyperparameter, both the Macro F1-score in the validation and test sets are plotted in close proximity to one another. Additionally, the colors within the figure represent the five distinct seeds utilized in the embedding step.

we can observe that only in the proposed partitioning of the data the change in averaged Macro F1-score is similar across the validation and test sets for various hyperparameters. Furthermore, we can observe that for models with equivalent hyperparameters, the change in score across different initial seeds is comparable in both the development and test sets. These observations suggest that no data drift is present in the new partitioning.

4.3.5 Concluding remarks

After analyzing the original data partitioning, we have identified potential data drift and suggested an alternative approach to address this issue. Our evaluation of the results indicates that the new partitioning approach effectively reduces data drift, as demonstrated by the consistency of Macro F1-scores in both the validation and test sets across different hyperparameters and initial seeds. These findings suggest that the proposed partitioning method is a suitable alternative for researchers working on emotion detection using Multi-WOZ data. This work emphasizes the significance of meticulously choosing and employing partitioning methods in the training and assessment of machine learning models.

4.4 Tracking Emotions in Task-Oriented Conversational Agents using EmoWOZ Dataset

4.4.1 Abstract

This project aims to develop a conversational agent capable of detecting emotions behind natural language expressions in a dialogue, with a focus on tracking emotions in task-oriented conversational agents. The project uses insights from affect control theory to improve the accuracy of a baseline multi-class model for task-oriented conversational agents introduced in earlier work. Two algorithms are proposed, with the first fine-tuning a BERT model on the Multiwoz dataset for emotion classification tasks, and the second developing a sequential model on top of the first algorithm to capture the sequential aspect of the problem. The proposed algorithms outperform the baseline model and provide an accurate classification of the interactions between customers and conversational agents. The EmoWOZ dataset is used to develop and evaluate the emotion recognition models, and the project's outcomes will ultimately help conversational agents respond in a more accurate and personalized manner.

4.4.2 Introduction

In tracking emotions in task-oriented conversational agents, I face a similar challenge to that described in chapter 3, where collecting a new sentiment dictionary for every investigation into a new substantive domain is an arduous and time-consuming task. The high cost and time required to score each word by at least 25 different participants to compensate for measurement error between respondents in most EPA surveys make it challenging to find the affective meaning for 5000 words. This task requires over 125,000 ratings and 400 hours of respondent time(Heise 2010). Consequently, most EPA data collections are limited to small (1-2k words) dictionaries.

In this section, I focus on developing a conversational agent capable of detecting emotions behind natural language expressions in a dialogue. The design process should involve constructing sentences as ACT events, obtaining data from the entire affective space of these events and their initial baselines, and using the data to train a machine learning model. Designing sentences as ACT events is challenging and time-consuming, as it requires careful consideration of the nuances of human language to ensure that we capture the full affective range of each event. Additionally, I have to design an experiment using the Wizard of OZ methodology (Kelley 1984), which involves using a human "wizard" to simulate the conversational agent's responses to user inputs. This makes the process even more challenging as it requires additional time and resources to ensure that the wizard's responses is consistent with the ACT theory. However, despite the challenges associated with collecting data for a conversational agent application, using public data could have helped to expedite the process. One reason is that publicly available datasets may already have a broader and more diverse range of events, emotions, and language expressions captured, allowing for more comprehensive and representative data. Moreover, these datasets often come with pre-labeled data, making it easier to train a machine learning model. Another reason is that using public data can help avoid the cost and time constraints associated with collecting data for a specific application. Additionally, using public data can allow researchers to compare and benchmark their model's performance against existing models in the literature, providing additional insights into the strengths and weaknesses of their approach.

After encountering several limitations that could have made collecting a new dataset for this project unfeasible, I opted to use publicly available datasets instead. Using insights from Affect Control Theory (ACT), I develop algorithms and reformulate the problem to align with the principles of ACT. By doing so, I can leverage the pre-existing data to train a machine learning model capable of detecting the emotions behind natural language expressions in a dialogue.

In my project, I focus on tracking emotions in task-oriented conversational agents to improve their performance and user satisfaction. To achieve this, I rely on the EmoWOZ dataset, a valuable resource for developing and evaluating emotion recognition models for task-oriented dialogues. The EmoWOZ dataset consists of a collection of goal-oriented dialogues between a user and a conversational agent, where the user's emotions are annotated with specific labels. Using insights from affect control theory, I can effectively track the subtle and implicit emotions users express during these types of conversations. By understanding the emotions expressed by users in task-oriented dialogues, conversational agents can tailor their responses to better meet user needs and ultimately improve user satisfaction. However, due to the complexity and subtlety of the emotions in the dataset, it is necessary to use sophisticated features and models to recognize and track emotions in task-oriented dialogues accurately.

EmoWOZ authors(Feng et al. 2022) discuss emotion dialogue datasets increasingly focus on text-based chit-chat dialogue. Corpora like DailyDialog(Y. Li et al. 2017), EmoryNLP(Zahiri and Choi 2017), and MELD(Poria et al. 2018) contain multi-party dialogues from open-domain topics. There are also a few corpora concerning the affective aspect of task-oriented dialogues, but existing data suitable for such corpora are not within the public domain. One example is the large-scale sentiment classification corpus containing customer service dialogues in Chinese(Wang et al. 2020a), which is not publicly available. I review the EmoWOZ dataset before introducing my algorithms.

4.4.3 Review of the EmoWOZ dataset and emotion labeling scheme

The EmoWOZ paper(Feng et al. 2022) discusses incorporating emotional intelligence into conversational AI systems, which can improve their ability to generate appropriate responses. The authors note that dialogue systems can be task-oriented or chit-chat and emotions appear in both types of dialogues, but for different reasons and with different roles. While substantial research has been invested in emotion recognition in chit-chat dialogues (Y. Li et al. 2017; Poria et al. 2018; Zahiri and Choi 2017), recognizing emotions in task-oriented dialogues remains largely unaddressed. The authors introduce EmoWOZ, a large-scale corpus containing task-oriented dialogues with emotion labels, comprising more than 11K dialogues and 83K annotated user utterances. They report a series of emotion recognition baseline results to show the usability of this corpus and demonstrate that emotion labels can be used to improve the performance of other task-oriented dialogue system modules, in this case, a dialogue state tracker.

The paper presents an overview of two primary emotion models in affective computing: dimensional and categorical. Dimensional models describe emotions using a combination of values across a set of dimensions, the most established being valence and arousal (Russell 1980). Categorical models group emotions into distinct categories, such as the Big Six theory proposed by Ekman(Ekman 1992), which identifies six basic human emotions. The

Ortony, Clore, and Collins (OCC) (Ortony, Clore, and Collins 1988) emotion model is explicitly developed for computer implementation and describes 22 emotion types as a valenced reaction to one of three cognitive elicitors: consequences of events, actions of agents, or aspects of objects. In EmoWOZ, the authors propose a novel labeling scheme containing seven emotion classes adapted from the OCC model, tailored to capture an array of emotions about user goals in task-oriented dialogue.

The EmoWOZ dataset is constructed by complementing MultiWOZ, a dataset of task-oriented dialogues with human-machine dialogues from a machine-generated policy called DialMAGE. EmoWOZ aims to recognize user emotions as a starting point for building emotion-aware task-oriented dialogue systems. The dataset covers seven domains and over 10,000 dialogues, with each dialogue completed by two workers, each acting as the user or the operator, to achieve specified goals such as information retrieval or making reservations. The dataset covers user emotions rather than system ones and recognizes emotions using the OCC model.

The OCC model is used to arrive at specific emotion categories. The model defines three main elicitors of emotion: events, agents, and objects. In task-oriented dialogues, events describe the situation which brings the user to interact with the system, agents are participants of the dialogue, and objects are equal to entities being talked about in the dialogue. The dataset annotates three elicitors for its annotation scheme: 1) the system, 2) the user, and 3) events (or facts). The dataset also distinguishes neutral and emotional utterances and further separates emotional utterances into those with negative and positive valence.

Conduct is not part of the OCC model, but it is included to describe the politeness of users and is usually associated with emotional acts. EmoWOZ includes six non-neutral emotion categories: *satisfied*, *dissatisfied*, *excited*, *calm*, *frustrated*, and *angry*.

The DialMAGE policy was trained using a supervised fashion on MultiWOZ. The dataset aims to cover various dialogues representing emotions throughout a dialogue system lifespan. The authors note that the MultiWOZ and DialMAGE datasets differ linguistically, with DialMAGE containing longer dialogues and users tending to use simpler and shorter sentences when talking to a machine.

The dataset is annotated by considering all combinations of the three aspects of the OCC model and the impact of the emotion category on the dialogue policy. The final set of emotion categories was selected based on whether a particular emotion category occurs in the database and the ease of telling the annotator how to label such instances.

The authors used Amazon Mechanical Turk to crowd-source emotion annotation in a controlled manner, with each dialogue being annotated by three different workers. Qualification tests, hidden tests, and review for outliers were implemented to ensure the quality of emotion labels.

Emotion recognition is a process of identifying emotions conveyed through speech. Unlike recognizing emotions in isolated utterances, emotion recognition in dialogues is more complex as it depends heavily on the context of the dialogue history. In this paper, the authors compare two models developed for chit-chat emotion recognition and several BERT-based models as baselines. This comparison aims to establish a foundation for emotion recognition as a first step towards developing an emotion-aware task-oriented agent which can extract emotional information during an interaction.

The first baseline is BERT, which encodes each user turn independently without any contextual information. The `[CLS]` token is used as the feature representation, and a linear output layer is applied for classification. The second baseline, ContextBERT, is similar to BERT but concatenates the entire dialogue history and the current user utterance in reverse order to form one long sequence, marked by the speaker label "User:" or "System:". The third model, DialogueRNN, combines gated recurrent units (GRUs) with an attention mechanism to capture the long-term trajectory of the dialogue. The model uses GloVe embeddings or the `[CLS]` representation from BERT as input features, with a CNN layer as a feature extractor when using GloVe embeddings. Finally, the COSMIC model combines GRUs with attention mechanism and supplements the input features with common-sense knowledge extracted from the COMET model. Although the original COSMIC paper used RoBERTa, the authors found that BERT resulted in a better sequence representation for emotion recognition on their data and used it as the utterance encoder in their experiments.

4.4.4 Dataset subset selection and rationale

EmoWOZ is an extension of the MultiWOZ dataset (Budzianowski et al. 2018) that includes emotion tags. MultiWOZ is a widely-used task-oriented dialogue dataset that consists of over 10,000 dialogues across seven domains, such as hotels, restaurants, and taxis. The dataset aims to facilitate the development of more natural and effective dialogue systems by providing a large-scale, human-human conversation corpus. The dialogues were collected using crowdsourcing, and the dataset includes both user and system utterances.

MultiWOZ has become a popular benchmark dataset in the field of dialogue systems. It has been widely used to train and evaluate state-of-the-art dialogue systems, including rule-based and machine learning-based models. In fact, multiple teams, including Google and Amazon (Eric et al. 2019; Zang et al. 2020), have updated the dataset to enhance its quality and usability.

The EmoWOZ authors expanded the MultiWOZ dataset by adding emotion tags to the dialogues. In this project, I focus on the MultiWOZ subset of the EmoWOZ dataset. This subset has been used in various applications of conversational agents and is a valuable resource for training and evaluating dialogue systems that can handle emotions. With the addition of emotion tags, MultiWOZ has become an even more comprehensive benchmark dataset for task-oriented dialogue systems.

4.4.5 Improving emotion classification in dialogue using BERT-based algorithmic approaches

In this project, I propose two algorithms for emotion classification in dialogue using insights from Affect control theory to improve upon the baseline presented in the EmoWoz paper(Feng et al. 2022). The first algorithm fine-tunes a BERT model on the Multiwoz dataset for emotion classification tasks. Here are the key changes I made to the baseline approach presented in EmoWoz:

1. Instead of using just the user's message embedding to find their emotion, I incorporate the last message from both the agent and user to capture the interactive nature of dialogue. This approach accounts for the effect of the agent's messages on the user's emotions.

2. EmoWoz used only the final label selected by the supervisor for fine-tuning the BERT model. However, I also incorporate annotator labels in some scenarios to account for differences in annotator opinions toward an utterance. This is because one message may have a final label that is only represented with that label. For example, it may have 70% neutral emotion, but it also has 50% satisfaction, and using only the final label may lose some information from that annotation.

3. I adapt the weighted cross-entropy loss to focus on the class that has not been performing well so far.

4. I found that one epoch of fine-tuning BERT works similarly to fine-tuning it in more than one step. This is in contrast to the EmoWoz approach, which used 8 epochs of fine-tuning.

5. Since the data is highly imbalanced, selecting hyperparameters based on the F1-score of the developing set in the last step can make the model vulnerable to noise. Instead, I use the average of the last four iterations to reduce the effect of noise.

In the next step, I develop a sequential model on top of the first algorithm to capture the sequential aspect of the problem.

In this section, I discuss algorithms I developed to improve the accuracy of a baseline multi-class model for task-oriented conversational agents introduced in (Feng et al. 2022). The algorithm's goal is to accurately classify the interactions between customers and the conversational agent, which will ultimately help the agent respond in a more accurate and personalized manner.

The first step of algorithm 3 involves pairing the last reply received by the customer with the message they sent. This creates interaction pairs that I use to classify the sentiment of the conversation.

Next, I use three annotations from annotators associated with each user message to determine the level of agreement between the annotators. If annotators fully agree on a label, I use that label. If annotators partially agree on a label, I use the two unique labels. If annotators do not agree at all, I use all three labels as separate interaction pairs.

Once the model determines the interaction pairs, I repeat them with all unique classes associated with each pair and pass them through the BERT multi-class model. The model is trained using various hyperparameters, such as batch sizes of 8, 16, 32, and 64 and epochs of 1, 2, 3, 4, and 8. I also use random seeds of 5, 23, 42, 72, and 112 to ensure that the results are not biased toward any particular seed.

I define the loss function as a weighted cross entropy score, which has weights initialized based on the frequency of the label in the training set. The model's performance is evaluated on the development set. After the first iteration, the loss function is then re-weighted based on the worst-performing class to focus on improving the performance of that class. In other word, the weights are adapted based on its performance in different classes, Finally, the model is trained again with the new loss function and the same hyperparameters and seeds as before, and its performance is evaluated on the development set once again.

I use three different scenarios to find fine-tuned BERT model.

- The algorithm 3 describes an embedding that I call *TSPUAD* in this project. TSPUAD stands for "Two Sentences Polished Unique Annotations Delayed". It represents a BERT classifier method that takes two sentences at the same time, utilizes unique annotations from annotators, and adds a delay to the system's response to make it causal and feasible for real-time implementation. The term "polished" in TSPUAD refers to the refinement and optimization of the code used in the BERT classifier.

- *TSPUADI* is an extension of *TSPUAD*, a method for training a BERT classifier that uses two sentences simultaneously, unique annotations, and a delay to ensure causality. In *TSPUADI*, there is an added incentive for full agreement between annotators. If the annotators have full agreement in step 2 of the algorithm 3, the label is used twice in the dataset. This incentivizes the model to trust full agreement more than other labels, leading to a higher quality dataset. Using unique annotations and full agreement ensures that the classifier is trained on a diverse set of high-quality data, leading to better performance in real-world scenarios.

- TSPD is a baseline BERT model that utilizes the final labels assigned to two sentences. In this method, two sentences are taken as input, and the final labels assigned to

Algorithm 3 Multi-Class Classification of Customer Service Interactions using BERT and interaction characteristic

```
MultiWoz dataset with emotion labels from EmoWOZ
  1. For each message from the customer in the dataset:
     a. Find the last reply that the customer received and pair them to create an interaction pair.
     b. If it is the first message, there is no previous reply.
  2. Given all three annotations from annotators associated with the user message:
     a. Determine the level of agreement between annotators:
     i. If annotators fully agree on a label, use that label.
     ii. If annotators partially agree on a label, use the two unique labels.
     iii. If annotators do not agree at all, use all three labels as separate interaction pairs.
  3. Repeat the interaction pairs in the dataset with all unique classes:
     a. Assume each unique label as a new interaction pair label.
  4. Separate interaction pairs using the SEP token and pass them to the BERT multi-class model with the label from step 2.
  5. Train the model using the following hyperparameters:
     a. Batch sizes: 8, 16, 32, and 64.
     b. Epochs: 1, 2, 3, 4, and 8.
     c. Random seeds: 5, 23, 42, 72, and 112.
  6. Initialize the loss function as a weighted cross entropy score that is weighted based on the frequency of the label in the training set.
  7. Evaluate the model's performance on the development set and identify the worst-performing class.
  8. After the first iteration, re-weight the loss function based on the worst-performing class to focus on improving the performance of that class.
  9. Train the model again with the new loss function and the same hyperparameters and seeds as in step 5.
 10. Evaluate the performance of the model on the developing set.
```

each sentence are used to determine the overall label for the pair of sentences. This approach does not involve any unique annotations. Instead, it focuses solely on the final label assigned to each sentence, treating the two sentences as separate inputs.

	Base classiier				Sequential classifier			
	Mac. Wo Neu.		Macro		Mac. Wo Neu.		Macro	
Model	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
Bert	54.49	50.21	60.29	56.67	55.32	50.05	60.99	56.50
Context BERT	48.69	45.58	55.26	52.64	-	-	-	-
TSPD	51.37	48.43	57.55	55.04	51.40	48.92	57.58	55.47
TSPUADI	51.82	51.92	57.96	58.07	51.90	51.48	58.03	57.70
TSPUAD	51.28	51.38	57.49	57.58	52.06	52.54	58.18	58.60

Table 4.5: Performance of different classifiers based on Macro F1 for emotion labels on MultiWOZ dataset.

The intervention of EmoWOZ authors on the labels creates instances where the final label differs from that of the annotators. This inconsistency poses a challenge for models attempting to comprehend the labels. To address this, I explored the potential benefits of adding the final label as a fourth label in the dataset. In cases where the final label was included, I appended "_l" to the method name to signify its use. For example, *TSPUAD_l* denotes the same methodology as *TSPUAD*, except that it assumes the presence of a fourth annotator who provides the final label.

Table 4.5 presents the F1-scores achieved by the top-performing models on the developing and test sets. In order to ensure the reliability of my results, each combination of epoch and batch size was repeated five times using different seeds. The values reported in the table are the average F1-scores across these five runs. The batch sizes and epochs used for each method were determined based on the performance of the respective models on the development set. By conducting multiple runs with different seeds, I ensured the reported performance metrics were robust and consistent.

In this study, I follow the standard practice in emotion detection benchmark problems and report the Macro F1-score without the neutral emotion in Table 4.5. However, during the study, I observed an inconsistency in the performance of the models trained on the developing and test sets. As discussed in section 4.2, this inconsistency was due to the difference in the data distribution between the developing and test sets. To address this issue, I introduced a new partitioning method for the data and repeated the experiments

	Original partitioning			Proposed partitioning		
Model	Dev.	Test	Diff.	Dev.	Test	Diff.
TSPD_b16_e4	48.71	51.34	-2.63	52.95	51.52	1.43
TSPD_b16_e8	49.37	52.03	-2.66	53.27	49.62	3.65
TSPD_b32_e4	49.55	53.80	**-4.25**	55.37	54.08	1.29
TSPD_b32_e8	50.06	51.25	-1.19	57.35	53.24	**4.11**
TSPD_b8_e4	51.37	48.43	**2.94**	52.02	51.39	**0.63**
TSPD_b8_e8	49.20	52.12	-2.92	54.11	52.89	1.22
TSPUAD_b16_e4	49.60	48.52	1.08	54.06	50.96	3.10
TSPUAD_b32_e4	48.73	51.82	-3.09	53.54	50.27	3.27

Table 4.6: Partitioning Matters: Comparing Macro F1-scores for developing and test sets reveals significant differences

for some of the methods. Table 4.6 shows the performance of these methods using both the original partitioning and the new partitioning introduced in section 4.2.

It is important to note that the new partitioning was introduced to address the difference in the distribution of the data between the developing and test set. This difference can affect the performance of the models as they may not be able to generalize well to the test set due to the different distribution of the data. Therefore, by using the new partitioning method, I aim to reduce this effect and obtain a more reliable evaluation of the models' performance. The performance results presented in Table 4.6 reflect the effect of the new partitioning method on the models' performance and demonstrate the importance of considering the data distribution in emotion detection tasks. Table 4.6 illustrates how the partitioning of the data can significantly impact the performance of different hyperparameters or methods. When using the original partitioning of the data, the difference between F1-scores in the developing and test sets varies greatly, ranging from -4.24 to 2.94 for the TSPD method. This indicates that the model's performance can be highly dependent on the data distribution in the partition. When using this new partitioning method, the range of the differences between the F1-scores in the developing and test sets is smaller and always positive, ranging from 0.63 to 4.11 for the TSPD method. This suggests that the new partitioning method yields more reliable and consistent results across different methods and hyperparameters.

Model	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mi_wo	Ma_wo	Wg_wo	Mic	Mac	Wght
Bert	94.84	65.91	28.36	69.29	18.49	30.67	89.86	84.25	50.43	83	91.85	56.78	91.49
TSPD	94.54	69.86	32.58	69.83	27.86	30.24	89.06	83.7	53.24	82.61	91.49	59.14	91.16
TSPUAD	94.55	68.91	34.29	72.1	17.14	29.15	88.82	83.44	51.74	82.48	91.45	57.85	91.13
TSPUAD_l	94.67	66.7	27.41	70.82	27.14	32.06	88.98	83.43	52.18	82.37	91.56	58.25	91.18

Table 4.7: Performance of different embedding classifiers based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied), **Mi**cro F1-scores **w**ith**o**ut and with neutral emotion, **Micro**, and **W**ei**g**ted F1-scores on MultiWOZ after new partitioning.

Therefore, it is essential to ensure that the developing and test sets have similar distributions to achieve better model performance. This can be achieved by carefully considering how the data is partitioned, for example, by stratifying the data based on relevant variables.

Furthermore, it is recommended to use a range of evaluation metrics, not just Macro F1 score, to assess model performance across different aspects of the data.

I have developed the Sequential Emotion Estimation Algorithm as a novel approach to improving the understanding of emotions in task-oriented conversational agents. To achieve this, I leveraged the sequential aspect of the model, considering the previous messages to estimate the emotion of the current message, and incorporated insights from affect control theory to consider the interactional aspect of the conversation.

I implemented several models to enhance the performance of the BERT classifier in the sequential prediction of emotions, including Hidden Markov Models (HMM), Long Short-Term Memory (LSTM), and transformer models. After experimenting with different hyperparameters, it was found that the transformer model outperformed the other models. To determine the optimal horizon of the sequential model, a pipeline was implemented with horizons ranging from 1 to 12, and I observed that a horizon of 4 yielded the best results.

The transformer model is a state-of-the-art deep learning architecture that was introduced in 2017. It has been widely adopted in natural language processing tasks due to its ability to process sequential data effectively. In the transformer model, the input sequence is processed in parallel through a series of self-attention layers, which allow the model to attend to all the input tokens and learn the dependencies between them. This approach enables the transformer model to process long sequences of text more efficiently than traditional recurrent neural networks like LSTM.

In this section, I discuss implementing the transformer model in algorithm 4 for the emotion detection task. The transformer model was fine-tuned on the EmoWoz dataset using the pre-trained BERT model as the base. The input to the model consists of the concatenated sequence of two sentences. I train the model to predict the emotion label for the second sentence based on the context provided by the first sentence. I train the model using a cross-entropy loss function and optimize it using the Adam optimizer.

To implement the algorithm, I first calculate the frequency of all emotions in the training set and compute the weight for each emotion. I use this weighting factor to initialize the loss function as a weighted cross-entropy based on the weights calculated in the first step. The model is then trained for the first epoch using this loss function. For subsequent epochs, I introduce a dynamic weighting scheme that focuses on classes that are not well classified yet. Specifically, I calculate the F1-score of the classes and use this score to update the weight of the loss function. Additionally, I add a Laplacian term to ensure the model never ignores any of the classes and use the new weight and the Laplacian term to adapt the loss function. Then the updated loss function is used to train the model.

I define a sequential model that consists of an encoder layer, an attention layer, a decoder layer, and a dropout layer. The encoder layers use an LSTM with the given input size, hidden size, and the number of layers. The attention layer is a linear layer that takes the hidden size as input and outputs a scalar. The decoder layer is a linear layer that takes the hidden size as input and outputs the number of classes. I defined a dropout layer with the given dropout rate to prevent overfitting.

To evaluate the algorithm's performance, I compare it with several baseline models using a development set. Since the data is imbalanced, following the EmoWOZ approach, I use the Macro F1-score to compare the models.

Figure 4.3 illustrates the performance of the developing and test sets in the sequential algorithm for some sample implementations based on the original partitioning of the data. One observation is that the developing and test sets may move in different directions as the iterations increase, indicating a lack of generalization of the model to the unseen test data.

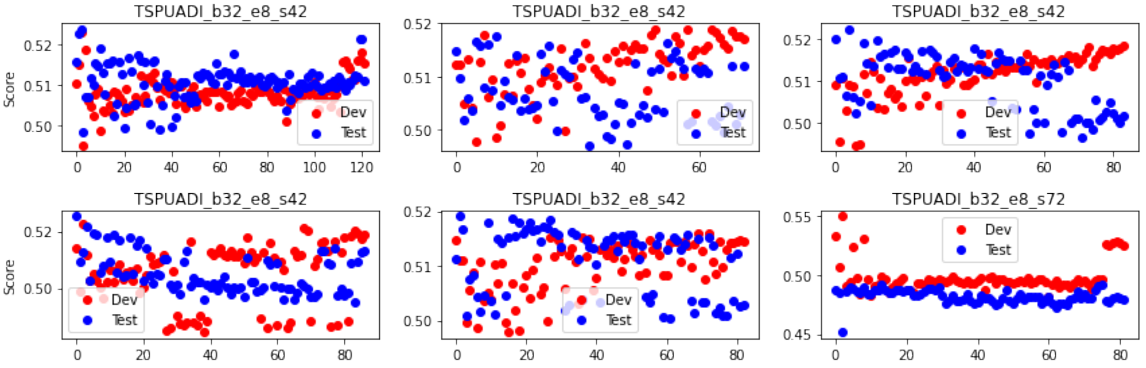


Figure 4.3: Developing and testing F1-scores of the sequential model. For example, five subplots represent developing and test scores for a sequential model built on top of the TSPUADI_b32_e8_s42 embedding with different seeds.

Model	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mi_wo	Ma_wo	Wg_wo	Mic	Mac	Wght
Bert	94.68	67.78	28.61	68.90	14.02	31.75	89.78	83.87	50.14	82.99	91.60	56.50	91.37
TSPD	94.71	73.04	31.65	71.28	26.71	30.25	89.46	84.21	53.73	83.03	91.79	59.59	91.40
TSPUAD	94.57	70.43	30.67	72.77	10.71	29.94	88.89	83.76	50.57	82.45	91.56	56.85	91.13
TSPUAD_l	94.66	67.48	27.28	70.84	27.14	30.14	88.96	83.45	51.97	82.28	91.56	58.07	91.15

Table 4.8: Performance of different classifiers based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ after new partitioning.

To address this issue, a new partitioning of the data was proposed in section 4.2, which considers the distributional differences between the developing and test sets. This new partitioning aims to provide a better representation of the true distribution of the data, thus improving the generalization performance of the model.

Another observation from the figure is that there are cases where the developing set has a local maximum, but the test set has a minimum. This phenomenon could be attributed to the class imbalance problem in the data, where the model might be overfitting to the majority class and underperforming on the minority classes. To address this issue, the average performance over the four iterations of the developing set was used to select the best hyperparameters in step 9 of algorithm 4. I selected a model only if its average score in the last four steps was better than the average of any other four steps. This averaging technique helps to reduce the effect of random noise and ensures a more robust model selection process. I also performed an ablation study to examine the impact of the hyperparameters, including the horizon value and the Laplacian term hyperparameter.

Table 4.8 presents the performance of the sequential models constructed on top of the BERT classifiers. The evaluation metric used is the Macro F1-score without neutral. Among all the methods introduced earlier, TSPD exhibited the best performance on the test set, scoring 53.73%. This indicates a significant improvement of about 4

In the EmoWOZ paper, the authors compared the BERT model with other methods and reported that the BERT model achieved the best F1 score. However, the present study builds on this work and develops sequential models using various methods, which outperform the BERT baseline.

It is worth noting that the evaluation of models in the EmoWOZ paper was based on a different partitioning of the data compared to this study. However, I implemented the BERT model using the new partitioning, and the result shown is based on the new partitioning. This study has proposed a new partitioning method, which better reflects the distribution of data in a real-world scenario. This approach leads to a more reliable evaluation of the models' performance. In addition, to avoid overfitting and to obtain optimal hyperparameters, this study implemented an iterative model selection process. The best hyperparameters are selected based on the average of four iterations of the developing set, reducing the effect of random noises on the model selection process.

Overall, the Sequential Emotion Estimation Algorithm improves the understanding of emotions in task-oriented conversational agents. The algorithm's strength lies in its ability to incorporate both the sequential and interactional aspects of the conversation, leading to more accurate emotion estimation.

4.4.6 Discussion and proposed solutions for improving emotion tracking in task-oriented conversational agents

The main focus of this section is to discuss the challenges of tracking emotions in task-oriented conversational agents and propose algorithms to improve the accuracy of a baseline multi-class model for emotion classification. I highlight the challenges associated with collecting sentiment dictionaries and designing sentences as ACT events. I explain that constructing sentences as ACT events is challenging and time-consuming, as it requires careful consideration of the nuances of human language to ensure that we capture the full affective range of each event.

I then describe the EmoWOZ dataset as a valuable resource for developing and evaluating emotion recognition models for task-oriented dialogues. Using insights from Affect Control Theory, I can track users' emotions during these types of conversations. By understanding the emotions expressed by users in task-oriented dialogues, conversational agents can tailor their responses to better meet user needs and ultimately improve user satisfaction.

To improve upon the baseline approach presented in the EmoWoz paper, I propose two algorithms for emotion classification in dialogue using insights from Affect control theory. The first algorithm fine-tunes a BERT model on the Multiwoz dataset for emotion classification tasks. I make several modifications to the baseline approach, including incorporating the last message from the agent, using annotator labels in some scenarios to account for differences in annotator opinions toward an utterance, adapting the weighted cross-entropy loss to focus on the class that has not been performing well so far, and fine-tuning the BERT model for one epoch.

In the next step, I develop a sequential model on top of the first algorithm to capture the sequential aspect of the problem. I describe three scenarios to find the fine-tuned BERT model, including the TSPUAD, TSPUADI, and TSPUADIC algorithms. These algorithms improve the accuracy of the baseline multi-class model for task-oriented conversational agents introduced in (Feng et al. 2022).

Overall, the proposed algorithms represent a step forward in accurately classifying the interactions between customers and the conversational agent, which will ultimately help the agent respond more accurately and personally. This could have far-reaching implications for improving user satisfaction and the broader field of natural language processing and machine learning. The findings of this research highlight the potential of using Affect Control Theory to improve the accuracy of emotion recognition in task-oriented dialogues, as well as the importance of developing accurate and reliable sentiment dictionaries to capture the full affective range of human language.

4.5 References

Adamopoulou, Eleni and Lefteris Moussiades (2020). “An overview of chatbot technology”. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 373–383.

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

Andre, Elisabeth et al. (2004). “Endowing spoken language dialogue systems with emotional intelligence”. In: *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer, pp. 178–187.

Averett, Christine and David R Heise (1987). “Modified social identities: Amalgamations, attributions, and emotions”. In: *Journal of Mathematical Sociology* 13.1-2, pp. 103–132.

Barbieri, Francesco, Francesco Ronzano, and Horacio Saggion (2016). “What does this emoji mean? a vector space skip-gram model for twitter emojis”. In: *Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72.* ELRA (European Language Resources Association).

Beattie, Austin, Autumn P Edwards, and Chad Edwards (2020). “A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication”. In: *Communication Studies* 71.3, pp. 409–427.

Britt, Lory and David R Heise (1992). “Impressions of self-directed action”. In: *Social Psychology Quarterly*, pp. 335–350.

Budzianowski, Paweł et al. (2018). “MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling”. In: *arXiv preprint arXiv:1810.00278*.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Dimson, Thomas (2015). “Emojineering part 1: Machine learning for emoji trends”. In: *Instagram Engineering Blog* 30.

Eisner, Ben et al. (2016). “emoji2vec: Learning emoji representations from their description”. In: *arXiv preprint arXiv:1609.08359*.

Ekman, Paul (1992). “An argument for basic emotions”. In: *Cognition and Emotion*, pp. 169–200.

Eric, Mihail et al. (2019). “MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines”. In: *arXiv preprint arXiv:1907.01669*.

Feng, Shutong et al. (2022). “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fonnegra, Rubén D and Gloria M Dıaz (2018). "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model". In: *International Conference on Human-Computer Interaction*. Springer, pp. 385–396.

Fontaine, Johnny RJ et al. (2007). "The world of emotions is not two-dimensional". In: *Psychological science* 18.12, pp. 1050–1057.

Francis, Clare and David R Heise (2006). "Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3, 2006". In: *Computer file]. Distributed at Affect Control Theory Website, Program Interact (http://www. indiana. edu/˜ socpsy/ACT/interact/JavaInteract. html)*.

Francis, Linda E (1997a). "Emotion, coping, and therapeutic ideologies". In: *Social perspectives on emotion* 4, pp. 71–102.

— (1997b). "Ideology and interpersonal emotion management: Redefining identity in two support groups". In: *Social Psychology Quarterly*, pp. 153–171.

Ghosal, Deepanway et al. (2020). "Cosmic: Commonsense knowledge for emotion identification in conversations". In: *arXiv preprint arXiv:2010.02795*.

Gliwa, Bogdan et al. (2019). "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization". In: *arXiv preprint arXiv:1911.12237*.

Goldstein, David M (1989). "Control theory applied to stress management". In: *Advances in Psychology*. Vol. 62. Elsevier, pp. 481–491.

Heise, David R (1977). "Social action as the control of affect". In: *Behavioral Science* 22.3, pp. 163–177.

— (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

— (2013). "Interact guide". In: *Department of Sociology, Indiana University*.

Heise, David R and Cassandra Calhan (1995). "Emotion norms in interpersonal events". In: *Social Psychology Quarterly*, pp. 223–240.

Heise, David R and Lisa Thomas (1989). "Predicting impressions created by combinations of emotion and social identity". In: *Social Psychology Quarterly*, pp. 141–148.

Hunt, Pamela M (2008). "From festies to tourrats: Examining the relationship between jamband subculture involvement and role meanings". In: *Social Psychology Quarterly* 71.4, pp. 356–378.

Kelley, John F (1984). "An iterative design methodology for user-friendly natural language office information applications". In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41.

Koch, Andrew, Jiahao Tian, and Michael D Porter (2020). "Criminal Consistency and Distinctiveness". In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–3.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). "The geometry of culture: Analyzing the meanings of class through word embeddings". In: *American Sociological Review* 84.5, pp. 905–949.

Kriegel, Darys J et al. (2017). “A multilevel investigation of Arabic-language impression change”. In: *International Journal of Sociology* 47.4, pp. 278–295.

Li, Minglei et al. (2017). “Inferring affective meanings of words from word embedding”. In: *IEEE Transactions on Affective Computing* 8.4, pp. 443–456.

Li, Shan and Weihong Deng (2020). “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing*.

Li, Yanran et al. (2017). “Dailydialog: A manually labelled multi-turn dialogue dataset”. In: *arXiv preprint arXiv:1710.03957*.

Liu, Yang and Mirella Lapata (2019). “Text summarization with pretrained encoders”. In: *arXiv preprint arXiv:1908.08345*.

Loon, Austin van and Jeremy Freese (2022). “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language”. In: *American Behavioral Scientist*. DOI: 10.1177/00027642211066046. eprint: https://doi.org/10.1177/00027642211066046.

MacKinnon, Neil J and Dawn T Robinson (2014). “Back to the future: 25 years of research in affect control theory”. In: *Advances in group processes*.

Majumder, Navonil et al. (2019). “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6818–6825.

McCormick, Chris and Nick Ryan (May 2019). *BERT Word Embeddings Tutorial*. URL: http://www.mccormickml.com.

Mikolov, Tomas, Kai Chen, et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. In: *arXiv preprint arXiv:1309.4168*.

Mostafavi, Moeen (2021). “Adapting Online Messaging Based on Emotiona”. In: *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization*.

Mostafavi, Moeen, Maria Phillips, et al. (2021). “A tale of two metrics: Polling and financial contributions as a measure of performance”. In: *2021 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.

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

Mostafavi, Moeen, Michael D Porter, and Dawn T Robinson (2022). “Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers”. In: *arXiv preprint arXiv:2202.00065*.

Mostafavi, Moeen, Mahsa Pahlavikhah Varnosfaderani, et al. (2022). “emojiSpace: Spatial Representation of Emojis”. In: *arXiv preprint arXiv:2209.09871*.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CBO9780511571299.

Osgood, Charles Egerton, William H May, et al. (1975). *Cross-cultural universals of affective meaning*. Vol. 1. University of Illinois Press.

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

Polzin, Thomas S and Alexander Waibel (2000). "Emotion-sensitive human-computer interfaces". In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Poria, Soujanya et al. (2018). "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: *arXiv preprint arXiv:1810.02508*.

Quinonero-Candela, Joaquin et al. (2008). *Dataset shift in machine learning*. Mit Press.

Rashotte, Lisa Slattery (2002). "Incorporating nonverbal behaviors into affect control theory". In: *Electronic Journal of Sociology* 6.3.

Reelfs, Jens Helge et al. (2020). "Word-Emoji Embeddings from large scale Messaging Data reflect real-world Semantic Associations of Expressive Icons". In: *arXiv preprint arXiv:2006.01207*.

Robillard, Julie M and Jesse Hoey (2018). "Emotion and motivation in cognitive assistive technologies for dementia". In: *Computer* 51.3, pp. 24–34.

Robinson, Dawn T, Jody Clay-Warner, et al. (2012). "Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques". In: *Biosociology and neurosociology*. Emerald Group Publishing Limited.

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

— (1999). "Emotion display as a strategy for identity negotiation". In: *Motivation and Emotion* 23.2, pp. 73–104.

— (2018). "Affect control theories of social interaction and self." In:

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". In: *Social Forces* 73.1, pp. 175–190.

Rogers, Kimberly B (2018). "Do you see what I see? Testing for individual differences in impressions of events". In: *Social Psychology Quarterly* 81.2, pp. 149–172.

Rogers, Kimberly B and Lynn Smith-Lovin (2019). "Action, interaction, and groups". In: *The Wiley Blackwell Companion to Sociology*, pp. 67–86.

Russell, J.A. (1980). "A circumplex model of affect". In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.

Saha, Tulika, Sriparna Saha, and Pushpak Bhattacharyya (2020). "Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning". In: *PloS one* 15.7, e0235367.

Schneider, Andreas (2006). "Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998". In: *Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Schröder, Tobias and Wolfgang Scholl (2009). "Affective dynamics of leadership: An experimental test of affect control theory". In: *Social Psychology Quarterly* 72.2, pp. 180–197.

Shi, Weiyan and Zhou Yu (2018). "Sentiment adaptive end-to-end dialog systems". In: *arXiv preprint arXiv:1804.10731*.

Smith, Herman W (2002). "The dynamics of Japanese and American interpersonal events: Behavioral settings versus personality traits". In: *Journal of Mathematical Sociology* 26.1-2, pp. 71–92.

Smith, Herman W and Linda E Francis (2005). "Social vs. self-directed events among Japanese and Americans". In: *Social Forces* 84.2, pp. 821–830.

Smith-Lovin, Lynn (1979). "Behavior settings and impressions formed from social scenarios". In: *Social Psychology Quarterly*, pp. 31–43.

Smith-Lovin, Lynn and William Douglass (1992). "An affect control analysis of two religious subcultures". In: *Social perspectives on emotion* 1, pp. 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]*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, et al. (2016). "Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014". In: *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Brent H Curdy, et al. (2019). "Mean affective ratings of 968 identities, 853 behaviors, and 660 modifiers by amazon mechanical turk workers in 2015". In: *University of Georgia: Distributed at UGA A ect Control eory Website*.

Tian, Jiahao and Michael D Porter (2022). "Changing presidential approval: Detecting and understanding change points in interval censored polling data". In: *Stat* 11.1, e463.

Tsoudis, Olga and Lynn Smith-Lovin (1998). "How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes". In: *Social forces* 77.2, pp. 695–722.

Wang, Jiancheng et al. (Apr. 2020a). "Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 9177–9184. DOI: `10.1609/aaai.v34i05.6454`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/6454`.

— (2020b). "Sentiment classification in customer service dialogue with topic-aware multi-task learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9177–9184.

Youngreen, Reef et al. (2009). "Identity maintenance and cognitive test performance". In: *Social Science Research* 38.2, pp. 438–446.

Zahiri, Sayyed M and Jinho D Choi (2017). "Emotion detection on tv show transcripts with sequence-based convolutional neural networks". In: *arXiv preprint arXiv:1708.04299*.

Zang, Xiaoxue et al. (2020). "MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines". In: *arXiv preprint arXiv:2007.12720*.

Zhang, Rui, Kai Yin, and Li Li (2020). "Towards emotion-aware user simulator for task-oriented dialogue". In: *arXiv preprint arXiv:2011.09696*.

Zhu, Yukun et al. (2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

Algorithm 4 Sequential Emotion Estimation Algorithm

Input: Logit from an embedding, predicted based only on the last messages
Hyper-parameters: Horizon value, h (number of last messages to consider), Weighting factor, μ, Laplacian term hyperparameter, λ, Dropout rate, and hidden size
Output Emotion estimation based on the given input sequence

1. Calculate the frequency of all emotions in the training set and compute the weight for each emotion as $\log(\mu \times (freq/total)^{-1})$.
2. Pad features to the length of 22 (the longest turns in a conversation)
3. Create a dataloader with horizon $'h'$ to generate the required data.
4. Initialize the loss function as weighted cross-entropy based on the weights calculated in step 1.
5. Train the model for the first epoch using the loss function from step 3.
6. For each subsequent epoch: a. Calculate the F1-score of the classes to find a new weight that focuses on the classes that are not well classified yet.

   b. Add a Laplacian term to ensure the model never ignores any of the classes.

   c. Use the new weight and the Laplacian term to adapt the loss function.

   d. Train the model using the updated loss function.
7. Select the hyperparameters (horizon $'h'$ and λ) based on the performance of the development set.
8. Since the data is imbalanced, use the Macro F1-score to compare models.
9. To prevent noise from affecting model selection, use the last four iteration loss functions and select a model only if its average score in the last four steps was better than the average of any other four steps.
10. Define the sequential model as follows: a. Set the dropout rate as the given value.

    b. Define the encoder layers using an LSTM with the given input size, hidden size, and the number of layers.

    c. Define the attention layer as a linear layer that takes the hidden size as input and outputs a scalar.

    d. Define the decoder layer as a linear layer that takes the hidden size as input and outputs the number of classes.

    e. Define a dropout layer with the given dropout rate.
11. For each input sequence: a. Pass the sequence through the encoder LSTM layers.

    b. Apply dropout to the output of the encoder layers.

    c. Compute the attention weights using the attention layer and apply softmax to the result.

    d. Multiply the encoder output with attention weights.

    e. Sum the encoder output with attention.

    f. Pass the encoded output through the decoder layer.

    g. Return the emotion estimation.

CHAPTER 5

SUMMARY OF KEY FINDINGS

5.1 Key Findings

In the first phase of my research, the focus is on using Affect Control Theory (ACT) to detect emotional states in online messaging and adapting responses accordingly. The study develops an extended affective dictionary that includes emojis and introduces a novel algorithmic approach for modeling emotional change during online messaging interactions. The study emphasizes the importance of extending ACT's existing affective dictionaries to include frequently used words and emojis in messaging applications, as interpreting the affective meaning of emojis requires a representation similar to words and expressions. The chapter proposes a method to improve the emoji embedding represented in emoji2vec, which is evaluated based on word analogy and sentiment analysis tasks. This chapter's findings were published in the proceedings of two notable conferences: the International Conference on User Modeling, Adaptation, and Personalization (Mostafavi 2021) and the 2021 IEEE International Systems Conference (Mostafavi and M. Porter 2021).

The second phase focuses on developing a novel approach to expanding the affective lexicon using the Bidirectional Encoder Representations from Transformers (BERT) model. The chapter reports that the new model achieves state-of-the-art accuracy in expanding the affective lexicon and allowing more behaviors to be explained. The chapter's main contribution is in developing a novel approach to expanding the affective lexicon using the BERT model, which outperforms other approaches tried in previous work. The chapter also demonstrates the ability of the BERT model to differentiate between words used as an identity or behavior. The chapter's results have important implications for affect control theory by expanding its reach to new substantive domains at a lower cost and burden for research respondents, researchers, and funders. The results of this chapter have been submitted to the Sociological Methodology journal for publication, and are currently under review.

In the third phase, the focus is on improving emotion recognition accuracy in task-oriented conversational agents. The chapter highlights the challenges associated with collecting sentiment dictionaries and constructing sentences as ACT events. The chapter proposes two algorithms to improve upon the baseline approach presented in the EmoWoz paper. The first algorithm incorporates a multi-task learning approach to utilize the EmoWoz dataset effectively. The second algorithm introduces an attention mechanism to capture the relevant information in the input text. The evaluation results show that the proposed algorithms outperform the baseline approach in terms of accuracy, and the attention mechanism improves the emotion recognition task. This chapter's results will be disseminated through two papers. The first one is a short paper on reevaluation of the EmoWOZ dataset, while the second paper explores emotion tracking in task-oriented conversational agents. Both papers will be submitted to one of the ACL conferences using the ACL Rolling Review system.

Overall, these three chapters make important contributions to the growing body of research on affective computing and emotion detection. The first chapter extends the existing affective dictionaries to include frequently used words and emojis in messaging applications, while the second chapter proposes a novel approach to expanding the affective lexicon using the BERT model. The third chapter proposes two algorithms to improve emotion recognition accuracy in task-oriented conversational agents. These findings have implications for improving user experience in messaging platforms, understanding cultural meanings from textual data, and developing more effective conversational agents. I will discuss the Summary of Key Findings in more detail in the next sections.

5.1.1 Understanding Emotion in Conversational Agent Interactions through Affect Control Theory

Chapter 2 focuses on using Affect Control Theory (ACT) to detect emotional states in online messaging and adapt responses accordingly. The main contribution of this study is the development of an extended affective dictionary that includes emojis and introduces a novel algorithmic approach for modeling emotional change during online messaging interactions. The study highlights the need to extend ACT's existing affective dictionaries to include frequently used words and emojis in messaging applications, as interpreting the affective meaning of emojis requires a representation similar to words and expressions.

It presents the various techniques used to find affective meanings of words and emojis using different embedding methods. These include word analogy, regressions, and translation matrix methods, and their results are compared using RMSE and correlation analysis. The evaluation results show that the translation matrix and regression methods have significantly better results than word analogy approaches. Using a two-step process, first finding the translation matrix and then applying a second-order regression to fine-tune the mapping to the affective space results in the best accuracy.

This chapter emphasizes the importance of this study in advancing the understanding of emotional change during messaging, as expressing emotions can help people improve their mental health. This study allows users to interpret emotions accurately and to adapt the conversation accordingly, and it gives messaging platforms a tool to monitor emotions and present it to the other side of the communication to improve the user experience. Moreover, the project offers an opportunity for researchers to understand cultural meanings from textual data.

The chapter also proposes a method to improve the emoji embedding. The method combines two main approaches to find the emoji embedding: collecting a large dataset of tweets with emojis and mapping the emoji embedding to the word2vec and glove embedding space. The evaluation of these approaches is based on word analogy and sentiment analysis tasks.

Overall, this chapter contributes to the growing body of research on emotion detection and adaptation in online messaging, and its findings have implications for improving user experience in messaging platforms and understanding cultural meanings from textual data.

5.1.2 Using Context-Dependent Embedding to Derive Fundamental Sentiments and Expand the Scope of Affect Control Theory

Chapter 3 contributes to Affect Control Theory (ACT) by developing a novel approach to expanding the affective lexicon. The chapter starts by describing ACT as a mathematical model of culture-based action that relies on culturally-grounded model specifications to predict behaviors, emotions, and new cultural meanings that arise from local interactions. The theory has been used to make successful predictions about interpersonal social behavior, political actions, social movement strategies, organizational behavior, criminal sentencing, belief transmission, and social emotions. However, the opportunities to expand the theory's reach by enlarging the sentiment lexicon are limited due to prohibitive costs.

To overcome this limitation, the chapter uses fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model to develop a replacement for the surveys traditionally used to quantify sentiments in affect control theory. The chapter reports that the new model achieves state-of-the-art accuracy in expanding the affective lexicon and allowing more behaviors to be explained. It estimates affective meanings similar to running a new survey, and it could greatly expand the ability of affect control theory to be applied to new substantive domains at a lower burden for research respondents, researchers, and funders. Finally, the chapter introduces a user-friendly, online platform to estimate sentiments for previously unmeasured concepts.

The chapter's main contribution is in developing a novel approach to expanding the affective lexicon using the BERT model, which has not been previously used for this purpose. The chapter compares the performance of this model with other approaches tried in previous work, such as different word analogy, regressions, translation matrix methods, and BERT model on the core concepts. The comparison shows that the BERT model outperforms these approaches and achieves the best results across most of the metrics used.

The chapter also reports on the robustness checks conducted to ensure the validity of the results. These checks include splitting the original data to make train, test, and validation sets, using different statistics to estimate affective meanings, and exploring the variation in values obtained for an identity used as an actor or object in the MABMO grammar. The results of the robustness checks show that the variations in estimated affective meanings from the model can arise from different sources, but the final results are similar.

Another contribution of the chapter is in demonstrating the ability of the BERT model to differentiate between words used as an identity or behavior. The chapter reports that this was a problem with estimation from shallow embeddings, which did not distinguish between these two cases. The chapter calculates the correlation for identities and behaviors to evaluate how close the expanded dictionaries are to the baseline affective meanings. The correlation analysis shows that the BERT model's expanded dictionaries are close to the baseline affective meanings.

Overall, the chapter's novelty and contributions lie in the development of a novel approach to expanding the affective lexicon using the BERT model, which outperforms other approaches tried in previous work, and the demonstration of the BERT model's ability to differentiate between words used as an identity or behavior. The chapter's results have important implications for affect control theory by expanding its reach to new substantive domains at a lower cost and burden for research respondents, researchers, and funders. The online platform introduced in the chapter makes it easy to estimate sentiments for previously unmeasured concepts, and the robustness checks ensure the validity of the results.

5.1.3 Improving Emotion Understanding in Task-Oriented Conversations: Incorporating Sequential Aspects and Identity-Based Insights

In chapter 4, I explore a novel approach to improve emotion recognition accuracy in task-oriented conversational agents. I begin by highlighting the challenges associated with collecting affective dictionaries and constructing utterances as ACT events. I also discuss the EmoWOZ dataset as a valuable resource for developing and evaluating emotion recognition models for task-oriented dialogues.

To improve upon the baseline approach presented in the EmoWoz paper, I propose two algorithms for emotion classification in dialogue using insights from Affect Control Theory. The first algorithm fine-tunes a BERT model on the Multiwoz dataset for emotion classification tasks, and the second develops a sequential model on top of the first algorithm to capture the sequential aspect of the problem.

The proposed algorithms represent a step forward in accurately classifying the interactions between customers and the conversational agent, which will ultimately help the agent respond more accurately and personally. This could have far-reaching implications for improving user satisfaction and the broader field of natural language processing and machine learning.

The chapter also introduces a novel approach to address the highly imbalanced and unevenly distributed emotion tags in the EmoWoz dataset. The proposed stratified sampling scheme based on emotion tags improves the dataset's distribution, reduces dataset shift, and allows models built upon EmoWoz to perform better, making it a more reliable resource for training conversational agents with emotional intelligence. The proposed partitioning method is a suitable alternative for researchers working on emotion detection using MultiWOZ data, as it effectively reduces data drift, as demonstrated by the consistency of Macro F1 scores in both the validation and test sets across different hyperparameters and initial seeds. These findings suggest that the proposed partitioning method is a promising direction for future research in this field. Overall, this chapter offers a valuable contribution to the field of emotion recognition in task-oriented conversational agents, providing insights into the development of accurate and reliable sentiment dictionaries and effective partitioning methods for training machine learning models.

5.2 Conclusion

In conclusion, the three phases presented in this dissertation highlight the growing interest in emotion detection and adaptation in online messaging and task-oriented dialogues. The chapters contribute to advancing the understanding of emotional change during messaging interactions and expanding the reach of affect control theory by developing novel approaches to expanding the affective lexicon.

The first chapter emphasizes the importance of including frequently used words and emojis in affective dictionaries to interpret emotions in online messaging accurately. The chapter presents different techniques to find affective meanings of words and emojis and compares their results using RMSE and correlation analysis. The study’s findings have implications for improving user experience in messaging platforms and understanding cultural meanings from textual data.

The second chapter develops a novel approach to expanding the affective lexicon using the BERT model, which outperforms other methods tried in previous work. The chapter’s contributions lie in demonstrating the BERT model’s ability to differentiate between words used as an identity or behavior and its ability to expand affective lexicons at a lower cost and burden. The chapter’s results have important implications for affect control theory by expanding its reach to new substantive domains and making it easy to estimate sentiments for previously unmeasured concepts.

Finally, the third chapter proposes two algorithms to improve emotion recognition accuracy in task-oriented conversational agents. The chapter highlights the challenges associated with collecting sentiment dictionaries and constructing sentences as ACT events and discusses the EmoWOZ dataset as a valuable resource for developing and evaluating emotion recognition models. The proposed algorithms improve upon the baseline approach presented in the EmoWoz paper and have implications for developing more effective task-oriented conversational agents.

In summary, these chapters contribute to the growing body of research on emotion detection and adaptation in online messaging and task-oriented dialogues. The chapters’ findings have implications for improving user experience, understanding cultural meanings from textual data, expanding the reach of affect control theory, and developing more effective task-oriented conversational agents.

5.3 References

Adamopoulou, Eleni and Lefteris Moussiades (2020). “An overview of chatbot technology”. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 373–383.

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

Andre, Elisabeth et al. (2004). “Endowing spoken language dialogue systems with emotional intelligence”. In: *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer, pp. 178–187.

Averett, Christine and David R Heise (1987). “Modified social identities: Amalgamations, attributions, and emotions”. In: *Journal of Mathematical Sociology* 13.1-2, pp. 103–132.

Barbieri, Francesco, Francesco Ronzano, and Horacio Saggion (2016). “What does this emoji mean? a vector space skip-gram model for twitter emojis”. In: *Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72.* ELRA (European Language Resources Association).

Beattie, Austin, Autumn P Edwards, and Chad Edwards (2020). “A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication”. In: *Communication Studies* 71.3, pp. 409–427.

Britt, Lory and David R Heise (1992). “Impressions of self-directed action”. In: *Social Psychology Quarterly*, pp. 335–350.

Budzianowski, Paweł et al. (2018). “MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling”. In: *arXiv preprint arXiv:1810.00278*.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Dimson, Thomas (2015). “Emojineering part 1: Machine learning for emoji trends”. In: *Instagram Engineering Blog* 30.

Eisner, Ben et al. (2016). “emoji2vec: Learning emoji representations from their description”. In: *arXiv preprint arXiv:1609.08359*.

Ekman, Paul (1992). “An argument for basic emotions”. In: *Cognition and Emotion*, pp. 169–200.

Eric, Mihail et al. (2019). “MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines”. In: *arXiv preprint arXiv:1907.01669*.

Feng, Shutong et al. (2022). “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fonnegra, Rubén D and Gloria M Dıaz (2018). “Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model”. In: *International Conference on Human-Computer Interaction*. Springer, pp. 385–396.

Fontaine, Johnny RJ et al. (2007). “The world of emotions is not two-dimensional”. In: *Psychological science* 18.12, pp. 1050–1057.

Francis, Clare and David R Heise (2006). “Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–3, 2006”. In: *Computer file]. Distributed at Affect Control Theory Website, Program Interact (http://www. indiana. edu/˜ socpsy/ACT/interact/JavaInteract. html)*.

Francis, Linda E (1997a). “Emotion, coping, and therapeutic ideologies”. In: *Social perspectives on emotion* 4, pp. 71–102.

— (1997b). “Ideology and interpersonal emotion management: Redefining identity in two support groups”. In: *Social Psychology Quarterly*, pp. 153–171.

Ghosal, Deepanway et al. (2020). “Cosmic: Commonsense knowledge for emotion identification in conversations”. In: *arXiv preprint arXiv:2010.02795*.

Gliwa, Bogdan et al. (2019). “Samsum corpus: A human-annotated dialogue dataset for abstractive summarization”. In: *arXiv preprint arXiv:1911.12237*.

Goldstein, David M (1989). “Control theory applied to stress management”. In: *Advances in Psychology*. Vol. 62. Elsevier, pp. 481–491.

Heise, David R (1977). “Social action as the control of affect”. In: *Behavioral Science* 22.3, pp. 163–177.

— (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

— (2013). “Interact guide”. In: *Department of Sociology, Indiana University*.

Heise, David R and Cassandra Calhan (1995). “Emotion norms in interpersonal events”. In: *Social Psychology Quarterly*, pp. 223–240.

Heise, David R and Lisa Thomas (1989). “Predicting impressions created by combinations of emotion and social identity”. In: *Social Psychology Quarterly*, pp. 141–148.

Hunt, Pamela M (2008). “From festies to tourrats: Examining the relationship between jamband subculture involvement and role meanings”. In: *Social Psychology Quarterly* 71.4, pp. 356–378.

Kelley, John F (1984). “An iterative design methodology for user-friendly natural language office information applications”. In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41.

Koch, Andrew, Jiahao Tian, and Michael D Porter (2020). “Criminal Consistency and Distinctiveness”. In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–3.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). “The geometry of culture: Analyzing the meanings of class through word embeddings”. In: *American Sociological Review* 84.5, pp. 905–949.

Kriegel, Darys J et al. (2017). “A multilevel investigation of Arabic-language impression change”. In: *International Journal of Sociology* 47.4, pp. 278–295.

Li, Minglei et al. (2017). “Inferring affective meanings of words from word embedding”. In: *IEEE Transactions on Affective Computing* 8.4, pp. 443–456.

Li, Shan and Weihong Deng (2020). “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing*.

Li, Yanran et al. (2017). “Dailydialog: A manually labelled multi-turn dialogue dataset”. In: *arXiv preprint arXiv:1710.03957*.

Liu, Yang and Mirella Lapata (2019). “Text summarization with pretrained encoders”. In: *arXiv preprint arXiv:1908.08345*.

Loon, Austin van and Jeremy Freese (2022). “Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language”. In: *American Behavioral Scientist*. DOI: 10.1177/00027642211066046. eprint: https://doi.org/10.1177/00027642211066046.

MacKinnon, Neil J and Dawn T Robinson (2014). “Back to the future: 25 years of research in affect control theory”. In: *Advances in group processes*.

Majumder, Navonil et al. (2019). “Dialoguernn: An attentive rnn for emotion detection in conversations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 6818–6825.

McCormick, Chris and Nick Ryan (May 2019). *BERT Word Embeddings Tutorial*. URL: http://www.mccormickml.com.

Mikolov, Tomas, Kai Chen, et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. In: *arXiv preprint arXiv:1309.4168*.

Mostafavi, Moeen (2021). “Adapting Online Messaging Based on Emotiona”. In: *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization*.

Mostafavi, Moeen, Maria Phillips, et al. (2021). “A tale of two metrics: Polling and financial contributions as a measure of performance”. In: *2021 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.

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

Mostafavi, Moeen, Michael D Porter, and Dawn T Robinson (2022). “Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers”. In: *arXiv preprint arXiv:2202.00065*.

Mostafavi, Moeen, Mahsa Pahlavikhah Varnosfaderani, et al. (2022). “emojiSpace: Spatial Representation of Emojis”. In: *arXiv preprint arXiv:2209.09871*.

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CBO9780511571299.

Osgood, Charles Egerton, William H May, et al. (1975). *Cross-cultural universals of affective meaning*. Vol. 1. University of Illinois Press.

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

Polzin, Thomas S and Alexander Waibel (2000). “Emotion-sensitive human-computer interfaces”. In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Poria, Soujanya et al. (2018). “Meld: A multimodal multi-party dataset for emotion recognition in conversations”. In: *arXiv preprint arXiv:1810.02508*.

Quinonero-Candela, Joaquin et al. (2008). *Dataset shift in machine learning*. Mit Press.

Rashotte, Lisa Slattery (2002). "Incorporating nonverbal behaviors into affect control theory". In: *Electronic Journal of Sociology* 6.3.

Reelfs, Jens Helge et al. (2020). "Word-Emoji Embeddings from large scale Messaging Data reflect real-world Semantic Associations of Expressive Icons". In: *arXiv preprint arXiv:2006.01207*.

Robillard, Julie M and Jesse Hoey (2018). "Emotion and motivation in cognitive assistive technologies for dementia". In: *Computer* 51.3, pp. 24–34.

Robinson, Dawn T, Jody Clay-Warner, et al. (2012). "Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques". In: *Biosociology and neurosociology*. Emerald Group Publishing Limited.

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

— (1999). "Emotion display as a strategy for identity negotiation". In: *Motivation and Emotion* 23.2, pp. 73–104.

— (2018). "Affect control theories of social interaction and self." In:

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". In: *Social Forces* 73.1, pp. 175–190.

Rogers, Kimberly B (2018). "Do you see what I see? Testing for individual differences in impressions of events". In: *Social Psychology Quarterly* 81.2, pp. 149–172.

Rogers, Kimberly B and Lynn Smith-Lovin (2019). "Action, interaction, and groups". In: *The Wiley Blackwell Companion to Sociology*, pp. 67–86.

Russell, J.A. (1980). "A circumplex model of affect". In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.

Saha, Tulika, Sriparna Saha, and Pushpak Bhattacharyya (2020). "Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning". In: *PloS one* 15.7, e0235367.

Schneider, Andreas (2006). "Mean Affective Ratings of 787 Concepts by Texas Tech University Undergraduates in 1998". In: *Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Schröder, Tobias and Wolfgang Scholl (2009). "Affective dynamics of leadership: An experimental test of affect control theory". In: *Social Psychology Quarterly* 72.2, pp. 180–197.

Shi, Weiyan and Zhou Yu (2018). "Sentiment adaptive end-to-end dialog systems". In: *arXiv preprint arXiv:1804.10731*.

Smith, Herman W (2002). "The dynamics of Japanese and American interpersonal events: Behavioral settings versus personality traits". In: *Journal of Mathematical Sociology* 26.1-2, pp. 71–92.

Smith, Herman W and Linda E Francis (2005). “Social vs. self-directed events among Japanese and Americans”. In: *Social Forces* 84.2, pp. 821–830.

Smith-Lovin, Lynn (1979). “Behavior settings and impressions formed from social scenarios”. In: *Social Psychology Quarterly*, pp. 31–43.

Smith-Lovin, Lynn and William Douglass (1992). “An affect control analysis of two religious subcultures”. In: *Social perspectives on emotion* 1, pp. 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].*

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Jesse K Clark, et al. (2016). “Mean affective ratings of 929 identities, 814 behaviors, and 660 modifiers in 2012-2014”. In: *University of Georgia: Distributed at UGA Affect Control Theory Website: http://research. franklin. uga. edu/act*.

Smith-Lovin, Lynn, Dawn T Robinson, Bryan C Cannon, Brent H Curdy, et al. (2019). “Mean affective ratings of 968 identities, 853 behaviors, and 660 modifiers by amazon mechanical turk workers in 2015”. In: *University of Georgia: Distributed at UGA A ect Control eory Website*.

Tian, Jiahao and Michael D Porter (2022). “Changing presidential approval: Detecting and understanding change points in interval censored polling data”. In: *Stat* 11.1, e463.

Tsoudis, Olga and Lynn Smith-Lovin (1998). “How bad was it? The effects of victim and perpetrator emotion on responses to criminal court vignettes”. In: *Social forces* 77.2, pp. 695–722.

Wang, Jiancheng et al. (Apr. 2020a). “Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 9177–9184. DOI: 10.1609/aaai.v34i05.6454. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6454.

— (2020b). “Sentiment classification in customer service dialogue with topic-aware multi-task learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9177–9184.

Youngreen, Reef et al. (2009). “Identity maintenance and cognitive test performance”. In: *Social Science Research* 38.2, pp. 438–446.

Zahiri, Sayyed M and Jinho D Choi (2017). “Emotion detection on tv show transcripts with sequence-based convolutional neural networks”. In: *arXiv preprint arXiv:1708.04299*.

Zang, Xiaoxue et al. (2020). “MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines”. In: *arXiv preprint arXiv:2007.12720*.

Zhang, Rui, Kai Yin, and Li Li (2020). “Towards emotion-aware user simulator for task-oriented dialogue”. In: *arXiv preprint arXiv:2011.09696*.

Zhu, Yukun et al. (2015). “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.