RECOGNIZING AFFECTIVE EVENTS AND EMBODIED EMOTIONS IN NATURAL LANGUAGE

by

Yuan Zhuang

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

August 2024

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Yuan Zhuang**
has been approved by the following supervisory committee members:

Ellen M. Riloff, Chair(s) __
Date Approved

Marina Kogan, Member __
Date Approved

Rada Mihalcea, Member __
Date Approved

Jeffrey Phillips, Member __
Date Approved

Vivek Srikumar, Member __
Date Approved

by **Mary W. Hall**, Chair/Dean of
the Department/College/School of **Computing**
and by **Darryl P. Butt**, Dean of The Graduate School.

ABSTRACT

Affective text analysis, such as sentiment analysis and emotion recognition, has long been studied in the research community but still remains challenging. One major reason is that current natural language processing systems still struggle to recognize implicit affective expressions, where affect is conveyed without any affect-bearing cues. To address this challenge, this dissertation focuses on two learning tasks to acquire two types of implicit affective expressions, which are common and critical for affective text analysis.

The first learning task is affective event recognition, which aims to classify if an event impacts most people positively (e.g., "*I watched the sunrise*"), negatively (e.g., "*I broke my leg*") or neutrally (e.g., "*I opened the door*"). This dissertation first identifies the limitations of previous approaches and introduces a deep learning classifier to mitigate these limitations. It also presents two novel semi-supervised learning methods to produce more training data to improve a classifier. The first method, *Discourse-Enhanced Self-Training*, produces new affective events by using coreference relations between events and sentiment expressions. The second method, *Multiple View Co-Prompting*, generates new affective events of high quality by prompting language models. Experiments show that the new affective events produced by these two methods substantially improve affective event classifiers.

The second learning task is to recognize expressions of embodied emotion in natural language, which refer to physical responses in our body when emotion arises (e.g., "*my legs shake due to fear*"). This dissertation first introduces a new task that aims to identify whether a body part mention is involved in any embodied emotion or not. It also presents two semi-supervised algorithms to generate weakly labeled data to improve a classifier. The first algorithm extracts weakly labeled data from text by using manner expressions with emotion, and the second algorithm generates weakly labeled data by prompting a large language model. Experiments demonstrate that the harvested weakly labeled data can train an effective classifier on its own. Furthermore, it can improve a supervised classifier when combined with gold training data.

For my family: my parents, my sister, my fiancée and my cats.

CONTENTS

ABSTRACT . . . . . **iii**

LIST OF FIGURES . . . . . **vii**

LIST OF TABLES . . . . . **viii**

ACKNOWLEDGEMENTS . . . . . **ix**

CHAPTERS

1. INTRODUCTION **1**

1.1 Improving Affective Event Recognition Systems 4
1.2 Recognizing Expressions of Embodied Emotion in Natural Language 6
1.3 Dissertation Claims and Research Contributions 7
1.4 Dissertation Outline 9

2. BACKGROUND **11**

2.1 Affective Text Analysis 11
2.2 Affective Event Recognition 18
2.3 Weakly Supervised Learning 23
2.4 Prompting Large Language Models 27
2.5 Embodied Emotion 29

3. IMPROVING AFFECTIVE EVENT RECOGNITION WITH DEEP-LEARNING MODELS **35**

3.1 Basic Concepts in Affective Event Recognition 35
3.2 Limitations of Existing Methods for Affective Event Recognition 37
3.3 Aff-BERT: A Deep-Learning Model for Affective Event Recognition 40
3.4 Evaluating Aff-BERT 41
3.5 Conclusion 42

4. IMPROVING AFFECTIVE EVENT RECOGNITION BY USING DISCOURSE-ENHANCED SELF-TRAINING **44**

4.1 Harvesting Events with Coreferent Sentiment Expressions 45
4.2 Discourse-Enhanced Self-Training 49
4.3 Gold Dataset Creation 52
4.4 Experimental Results 53
4.5 Conclusion 55

5. IMPROVING AFFECTIVE EVENT RECOGNITION BY MULTIPLE VIEW CO-PROMPTING **57**

5.1 Acquiring Affective Events with Multiple View Co-Prompting 58
5.2 Evaluation 65
5.3 Conclusions 73

6. RECOGNIZING EXPRESSIONS OF EMBODIED EMOTION IN NATURAL LANGUAGE 75

6.1 Task Formulation 76
6.2 Data Collection 78
6.3 Evaluating Emotion Classifiers 80
6.4 Producing Weakly Labeled Data for Embodied Emotions 83
6.5 Experimental Results 86
6.6 Analysis 89
6.7 Conclusion 91

7. CONCLUSIONS AND FUTURE WORK 93

7.1 Research Summary and Contributions 93
7.2 Future Research Directions 95

REFERENCES 104

LIST OF FIGURES

3.1 The architecture of Aff-BERT. 40
4.1 Illustration of affective polarity scoring in Discourse-Enhanced Self-Training. . 49
4.2 Learning curves through 10 iterations. 54
4.3 Learning curves of models with training sets of different sizes. 55
5.1 Flowchart for acquiring affective events with Multiple View Co-Prompting. . . 59
5.2 Examples of the Emotion Prompt. 63
5.3 Newly labeled events generated across iterations. 67
5.4 Learning curve of Co-Prompting. 72
5.5 Results for different training set sizes. 73
6.1 Illustration of body part references associated with or not associated with embodied emotions. 76
6.2 Dependency relation patterns. 85
6.3 Example prompt for GPT3.5. 85
6.4 F1 scores based on body part frequency. 91

LIST OF TABLES

1.1 Examples of affective states described implicitly in novels. 3
3.1 Examples of affective events correctly and incorrectly labeled in AEKB. 38
3.2 Performance of AEKB over the TWITTER dataset. 39
3.3 Performance on the BLOG test set. 42
4.1 Syntactic patterns of coreferent sentiment expressions. 45
4.2 Examples of harvested tweets and extracted events. 47
4.3 Examples of harvested tweets and new sentiment terms. 48
4.4 Results for learning from unlabeled data. 54
4.5 Recall and precision across polarities. 55
5.1 Examples of events generated by the Associated Event Prompt for seed events. 62
5.2 Experimental results for TWITTER data. 69
5.3 Experimental results for BLOG data. 69
5.4 Impact of multiple views on TWITTER data. 70
5.5 Counts of labels changed by Co-Prompting. 71
5.6 Manual analysis of polarity labels. 71
6.1 Frequencies of different body parts in the CHEER dataset. 80
6.2 Statistics of the CHEER dataset in terms of annotated body part mentions. . . . 80
6.3 Embodied Emotion examples in CHEER. 81
6.4 Evaluating emotion classification models. 83
6.5 Zero-shot prompting and gold training results. 87
6.6 Results with weakly labeled data only. 88
6.7 Using gold and weakly labeled data together. 89
6.8 The effects of removing E_{PAT} or E_{LM} from the weakly labeled data, one at a time. 90
6.9 Error cases. 91

ACKNOWLEDGEMENTS

My journey in natural language processing began 7 years ago. At that time, I thought that it was just about spending the next few years building different learning models. However, I soon realized that it was much more than that. Pursuing a PhD is an intelligence challenge, where one needs to explore research questions without standard answers and constantly keep up with new algorithms that emerge every day. Moreover, it is a mental challenge. Success usually comes only after months or even years of failures. One must learn how to balance life and research, fight against stress, and confront setbacks. This journey has been one of the most important and unforgettable journeys in my life, where I learned to be a researcher and also gained mental resilience. I would not have been able to make it without the support of many people.

I would like to express my greatest thanks to my advisor Ellen Riloff. Ellen taught me how to be a good researcher. I have learned many, many research principles and skills from her. Moreover, I was always amazed by her instincts, insights and high-level views about research. Most importantly, Ellen is a very caring advisor. She supported and encouraged me throughout my journey, ensuring that I remained focused and motivated. She is not only an excellent advisor but also a great friend to me. I enjoyed talking to Ellen about everything in our weekly meetings, especially our cats. I am grateful for her mentorship and her profound impact on my academic and personal growth.

I thank Marina Kogan, Rada Mihalcea, Jeff Phillips, and Vivek Srikumar for being on my dissertation committee and for their guidance and support. They have provided insightful feedback, helped me become aware of potential improvements in my dissertation study, and guided me to view the impacts of my study from different perspectives. I am very lucky to have them on my dissertation committee.

I have learned a lot from my collaboration with other researchers. I thank Kiri Wagstaff for her supervision during our collaboration to develop relation extraction models on planetary scientific publications. Kiri is a very thoughtful and patient advisor with great

leadership. With her help, I was fortunate to work together with a group of planetary scientists at NASA. I am also grateful for the collaboration with Marina Kogan and Di Wang. With their support, I was able to explore topics in the field of crisis informatics and worked on identifying social cues from Twitter, which I believe would be a very important research topic in the future. I also thank my collaborators during my internship at Amazon, including Jan Bakus, Heba Elfardy, Kevin Small and Aidan San. They are very supportive, knowledgeable and open-minded researchers who made my internship experience enjoyable.

Over the past few years in Salt Lake City, I have been very fortunate to meet a lot of great people, including Youngjun Kim, Haibo Ding, Xingyuan Pan, Jie Cao, Tao Li, Qingkai Lu, Tianyu Jiang, Yue Yang, Yichu Zhou, Xiaowan Li, Zhenduo Wang, Qingyao Ai, Vivek Gupta, Maitrey Mehta, Mattia Medina Grespan, Ashim Gupta, Nate Stringham, Atreya Ghosal, Fateme Hashemi, Di Wang, Harald Illig, Mian Dai, Haiyan Zang, Paul Hu, Zejian Wu, and Zhongyi Jiang. A Special thanks to Tianyu Jiang for the time we spent together working out. No one else could be as good a gym buddy as he is. Another special thanks to Zejian Wu and Zhongyi Jiang, who have always been open to exploring new experiences with me, such as hiking, bouldering, playing tennis, and trying out different restaurants in Salt Lake City.

I would like to thank Changchen Chen for his support over the years. Our friendship began 13 years ago at the University of Rochester, where we both pursued our Bachelor's degrees, and it has only grown stronger since. Changchen has been a pillar of support throughout my PhD studies and personal life, offering invaluable assistance and guidance. I feel incredibly fortunate to have him as a friend.

I am extremely grateful to my parents, Hua and Xian, and my sister Tian. As a Chinese proverb says, family is a harbor to escape from a storm. My parents and my sister have given me unconditional support during my difficult moments over the past. Without them, it would be impossible for me to complete this dissertation. I am also thankful to my Fiancée Yinlu. We met in a quite romantic and special way four years ago. Since then, her unwavering love, encouragement, and patience have always been my guiding light during challenging times and made me a better person. I am deeply grateful to have her by my side, and I cannot imagine my life without her. Finally, I would like to thank my children

(my cats): Boba, Bomi, Bozi and Bojiu. Despite their talent for chaos, incessant mischief and impressive ability to break things, they have provided immense emotional support for me during this journey. Their unique brand of love is irreplaceable, and I wouldn't have it any other way.

Finally, I would like to thank myself for my persistence and resilience over the years. It was not easy, but we have made it!

CHAPTER 1

INTRODUCTION

Developing computational models to understand affective states of individuals based on their language has long been an intriguing problem in the field of natural language processing (NLP). To date, many NLP tasks have been proposed to identify affective states conveyed in text. One widely studied task is sentiment analysis [208, 115, 179, 198, 124, 162], which recognizes the overall sentiment of a person towards a subject matter. For example, the statement "*I like my research*" expresses a positive sentiment towards the research, and the statement "*The model performance is disappointing*" indicates a negative sentiment towards the model performance. Another well-known task is emotion recognition [3, 127, 1, 125, 30], which identifies the emotions of a person. For example, the statement "*Wow I am going to Disneyland tomorrow*" implies excitement and the statement "*Oh no I failed my exam*" indicates frustration.

While significant advancement has been made in this field, the problem is still far from being well solved. Most current NLP tasks mainly focus on identifying explicit affective expressions (e.g., "*I like this movie*"), so they often struggle with implicit affective expressions that convey affect without using any affect-bearing cues. To illustrate, consider the following examples:

1. *I got an A in this class.*
2. *I got an F in this class.*

In Example 1, the speaker probably possesses a positive affective state such as joy and excitement, because the event of getting an A is an achievement and thus a desirable event for most people. The speaker in Example 2, on the other hand, probably possesses a negative affective state such as disappointment and frustration, since the event of getting an F makes a student fail a class and is thus a negative event for most people. There are no affect-bearing words in Example 1 and 2, so an NLP system must possess the knowledge

about the affective impacts of these events in order to understand the affective states. As another example, consider the following two statements:

3. *Rose's face became red as she walked towards the movie star.*
4. *Rose's face became red as she had walked 10 miles.*

In Example 3, we probably infer that Rose possessed some affective state, because her reddening face is a typical sign of ongoing emotion based on the context, such as embarrassment and shyness. In Example 4, we probably infer that Rose did not have any affective state, since her facial redness resulted from walking over a long distance. For these two examples, an NLP system must be able to interpret the affective states based on the bodily responses. Overall, the four examples can be challenging for most current NLP systems due to the absence of explicit affect-bearing cues.

While there has been little emphasis on learning implicit affective expressions, we argue that it is crucial to focus on their study. This is mainly because implicit affective expressions are commonly used to convey affect in our language [144]. One clear case of this is story-writing. In story-writing, writers often follow the "*Show, don't tell*" principle, which suggests that instead of explicitly stating emotions, character traits, or events, writers should use descriptive language and vivid imagery to allow readers to experience the story through actions, senses, and details. With this technique, the affective state of a character is usually not told explicitly but described implicitly. Consider the two examples in Table 1.1. Example 1, quoted from the book *1984* by George Orwell, describes the tense moment when the two characters, Julia and Winston, are caught by the Thought Police. The intense fear of the characters is suggested by their bodily responses, such as body shaking, teeth chattering and knee buckling. Example 2, quoted from the book *The Adventures of Tom Sawyer* by Mark Twain, describes the joy of a group of boys by enjoyable events such as "*they went whooping and prancing out on the bar*" and "*they ran out and sprawl on the sand*."

Motivated by the need to study implicit affective expressions, this dissertation aims to learn two types of implicit affective expressions that are common and critical for affective text analysis. The first type of implicit affective expressions is *affective events*, which refer to the daily events that impact our affective states when we experience them. While many events in our daily life are mundane and barely impact us, some events could change our affective states positively or negatively when we experience them. For example, a

Table 1.1: Examples of affective states described implicitly in novels.

Example 1
"Now they can see us," said Julia. "Now we can see you," said the voice. "Stand out in the middle of the room. Stand back to back. Clasp your hands behind your heads. Do not touch one another." They were not touching, but it seemed to him that he could feel Julia's body shaking. Or perhaps it was merely the shaking of his own. He could just stop his teeth from chattering, but his knees were beyond his control.
Example 2
After breakfast they went whooping and prancing out on the bar, and chased each other round and round, shedding clothes as they went, until they were naked, and then continued the frolic far away up the shoal water of the bar, against the stiff current ... When they were well exhausted, they would run out and sprawl on the dry, hot sand, and lie there and cover themselves up with it, and by and by break for the water again and go through the original performance once more.

person usually gets upset when he/she loses his/her wallet. On the other hand, a person is usually relaxed and happy when he/she gets some delicious food. The knowledge of affective events is crucial for discovering the affective states of people who experience these events. Suppose someone says "*I got a cold today*." It is very likely that the individual's affective state is negative, since the event of getting a cold usually leads to undesired physical conditions and negatively impacts most people. In contrast, if someone says "*I recovered from a cold today*," the individual's affective state is probably positive, as the event of gaining health is desirable for most people.

Another crucial type of implicit affective expressions that this dissertation focuses on is *embodied emotions*, which refer to the physical responses in our body when emotion arises. In our daily life, experiences of emotion often give rise to physical responses in our body. For example, we may have physiological responses such as heart racing and chills down our spines when we get scared. We may also have visible physical reactions such as clenching our teeth and slamming our fists due to anger. Recognizing these physical responses evoked by emotions benefits recognizing implicit affective states. For example, if we see a person waiting in a line and tapping his/her feet restlessly, we would probably infer that the person is impatient. As another example, if someone throws his/her hands up in the air after hearing some news, we probably infer that the person is excited.

This dissertation thus explores two research problems:

1.1 Improving Affective Event Recognition Systems

There has been prior research [46, 47] on affective event recognition, which classifies the affective impact of an event as positive (e.g., "*I joined a party*"), negative (e.g., "*I dropped my phone in the toilet*") or neutral (e.g., "*I woke up*"). Most of prior research took the approach of building lexical resources of affective events. For example, Ding and Riloff [47] built a knowledge base of affective events called AEKB, which contains about half a million event phrases that are automatically labeled with affective polarities. This dissertation first identifies several limitations of this approach. The first limitation is that the lexical resources of affective events do not generalize well to unseen events. One reason for the insufficient generalization is that the lexical resources do not have sufficient coverage of affective events. This is mainly because an event could be expressed in various forms. For example, the event of getting a cold could be expressed by different phrases such as "*I got a cold*," "*I caught a cold*," "*I become ill with a cold*" and so on. It is also partly because new events always arise in the future and they are not included by the existing lexicons. In addition to the insufficient coverage, the quality of the affective polarities in these lexical resources could be limited as they were developed based on methods that do not capture well the semantics of events. To address this limitation, this work develops a deep learning model, Aff-BERT, to classify affective events. Aff-BERT is a classification model based on fine-tuning the pretrained language model, BERT [44]. Leveraging the powerful pretrained representations produced by BERT, Aff-BERT can better capture the meaning of an event and generalize better to unseen events.

Another limitation of prior work is that the amount of training data is usually small and potentially results in limited model performance. For example, the annotated dataset created by Ding and Riloff [47] contains only 1,490 affective events. To overcome this issue, this dissertation develops two new semi-supervised learning methods to automatically harvest weakly labeled affective events as extra training data. Extensive experiments demonstrate that the semi-supervised learning methods can improve our affective event classifier Aff-BERT.

The first method is *Discourse-Enhanced Self-Training* (DEST). DEST is motivated by the

observation that an event's polarity is often indicated by the sentiment expressions that corefer to the event. Consider the statement "*I just graduated with a PHD degree. This is amazing*." The sentiment expression "*This is amazing*" conveys a positive opinion. As it corefers with the event "*I graduated with a PHD degree*," we can infer that the event is also positive. In this work [225], we refer to these sentiment expressions as coreferent sentiment expressions. Based on this observation, our method first mines a set of unlabeled events and their coreferent sentiment expressions in the local context. To generate new affective events, DEST assigns a polarity label to each unlabeled event based on: 1) the prediction of an affective event classifier that is trained on the training set, and 2) the average polarity of the coreferent sentiment expressions of the unlabeled event. The newly labeled events with high confidence are then added into the training set to improve the affective event classifier. We showed in experiments that DEST can substantially improve the model performance. In addition, we believe that the general idea behind DEST could be useful for many other problems where additional information can be extracted from larger contexts to serve as a secondary signal to help confirm or disconfirm a classifier's predictions.

The second semi-supervised learning method, *Multiple View Co-Prompting*, is motivated by the limitations of mining affective events directly from text, such as the inefficiency and the computational bottleneck of applying text-processing techniques to a large corpus. To avoid these limitations, Multiple View Co-Prompting generates new affective events by prompting language models. Essentially, it is an iterative algorithm where each iteration starts with the *Event Generation* step to generate event phrases by prompting a language model such as GPT2 [157]. Next, it assigns polarity labels to the generated events with the *Polarity Assignment* step. Specifically, for each generated event, this step first collects two independent views of its polarity using two language model prompts. Then the two views are combined to produce an accurate polarity label for the event. Our evaluation [227] demonstrates that the automatically labeled events are of high quality and they can improve the model performance substantially. In addition, we believe that the idea behind this method, which elicits accurate information from language models based on two or more data views, is useful for many prompting methods.

1.2 Recognizing Expressions of Embodied Emotion in Natural Language

The phenomena of embodied emotion has been widely studied in many other research areas such as computer vision and psychology, but it has not been studied by the NLP community before. Learning to identify expressions of embodied emotion can benefit affective text analysis, as embodied emotions are commonly used to convey affective states implicitly in natural language. Motivated by its importance, this dissertation proposes the first study on recognizing expressions of embodied emotion in natural language [226].

This dissertation first formulates the learning problem as a binary classification problem that focuses on body part mentions. Specifically, an input in this task is a text that contains: 1) a sentence containing a body part mention to classify; 2) some preceding sentences as context. The task is to classify the body part mention into one of the following two categories: 1) Embodied Emotion, and 2) Neutral. For example, in the text "*A man walked out from the corner. I saw him and my eyes widened*," the body part "*eyes*" will be labeled as Embodied Emotion. As another example, in the text "*My eyes got watery due to my allergies*," the body part "*eyes*" will be labeled as Neutral. One might wonder if this task could be formulated to identify verbs that indicate embodied emotions (e.g., "*I kicked the wall after I heard the news*"), instead of identifying body parts that are involved in embodied emotion. Focusing on verbs can introduce several issues. First, verbs tend to be highly ambiguous and are often used metaphorically. One such case is "*The film reviewers tore apart Jack's performance in his latest film*," where "*tore*" is metaphorically used to indicate criticism. Furthermore, focusing only on verbs is challenging to operationalize in practice. This is because nearly every sentence contains verbs but only a small fraction of them are related to physical human actions and so embodied emotions. Given these reasons, this study focuses on recognizing expressions of embodied emotion that are associated with body part mentions.

To facilitate the study, this work presents a dataset, CHEER, that contains 7,300 instances with human annotation and is publicly available for the community to conduct future research. To perform classification, a model based on fine-tuning BERT is proposed. Given that the amount of gold training data is relatively small, this work introduces two semi-supervised methods to produce weakly labeled instances. The first method is a pattern-

based method that extracts Embodied Emotion instances from a text corpus by exploiting manner expressions with emotion (e.g., "*frantically*," "*in anger*" and "*with excitement*"). It is motivated by the observation that a body part is usually involved in embodied emotions when it is syntactically connected to a manner expression with emotion (e.g., "*I clenched my fist in anger*"). The second method is a language-model-based method that harvests newly labeled instances by prompting a large language model such as GPT3.5.[1] It is motivated by the findings in our early experiments that some large language models such as GPT3.5 exhibit a strong, though far from perfect, zero-shot learning ability over this task. To generate weakly labeled instances, this method feeds GPT3.5 with a prompt that contains: 1) the task instruction including the definitions, 2) a sentence with a body part to label, and 3) a question to make GPT3.5 answer whether the body part should be labeled as Embodied Emotion or Neutral. Then the input instance is labeled based on the answer generated by GPT3.5. With these two semi-supervised learning methods, we generated a large set of weakly labeled instances, the size of which is almost 10 times the size of the CHEER dataset. Experiments demonstrate that the weakly labeled instances can train an effective learning model without any gold data. Furthermore, they can improve a supervised model when combined with the gold training data, yielding good results for recognizing embodied emotions.

1.3 Dissertation Claims and Research Contributions

The primary contributions of this dissertation are as follows:

> *Claim 1: Accuracy for affective event recognition can be improved with deep learning models that exploit novel semi-supervised algorithms including Discourse-Enhanced Self-Training and Multiple View Co-Prompting.*

This dissertation first identifies the limitations of prior work that focused on creating lexical resources for affective event recognition. To mitigate these limitations, a deep learning model, Aff-BERT, is developed and shown to have better accuracy and coverage for affective event recognition.

To improve Aff-BERT, two novel semi-supervised algorithms are developed to generate new affective events. The first method is Discourse-Enhanced Self-Training (DEST), which

[1]GPT3.5 is available at https://platform.openai.com/docs/models/gpt-3-5-turbo.

generates a polarity label for an unlabeled event based on 1) the prediction of Aff-BERT that is trained on the training set, and 2) the coreferent sentiment expressions of the event. Experiments show that Aff-BERT trained with DEST substantially outperforms strong baselines. The second method is Multiple View Co-Prompting, which generates new affective events by prompting language models. The method first generates new event phrases by prompting a language model such as GPT2. To produce a polarity label for a generated event phrase, the method first extracts two data views of its polarity by two language model prompts. The two data views are then combined to produce the final polarity label. Experiments show that Multiple View Co-Prompting generates weakly labeled affective events with high quality and improves Aff-BERT substantially.

> *Claim 2: Recognizing expressions of embodied emotion in natural language can be improved by training a model specifically for this task and exploiting semi-supervised learning.*

The main contribution of this work is that it proposes the first study on recognizing expressions of embodied emotion in natural language. The learning task is formulated as a binary classification problem focused on body part mentions. In this task, a body part mention is classified as Embodied Emotion or Neutral based on the context. This work also introduces a benchmark dataset that contains 7,300 instances with human annotation. We conduct experiments over the dataset to show that existing systems, such as emotion recognizers and large language models, do not perform well in this task. To improve the task performance, we train a model based on fine-tuning BERT, and show that it substantially outperforms other baseline systems.

Since the amount of training instances is relatively small, this work presents two semi-supervised methods to generate weakly labeled data. The first method extracts Embodied Emotion instances from text by exploiting manner expressions with emotion. The second method produces weakly labeled instances by prompting a large language model. Experiments show that the weakly labeled instances generated by these two methods can train a strong classifier on their own. Furthermore, they can improve a supervised model substantially when combined with gold training data.

1.4 Dissertation Outline

This dissertation is organized as follows:

• **Chapter 2** gives an overview of work related to this dissertation. It first gives a summary of existing work related to affective text analysis, such as sentiment analysis and emotion recognition. Next, it discusses prior work on affective event recognition. It then discusses the background of weakly supervised learning, including self-training, co-training and methods in data augmentation. Finally, it presents prior work in psychology and NLP that is related to embodied emotion recognition.

• **Chapter 3** discusses the limitations of prior work on affective event recognition. To show the limitations of prior work, we introduce a manually annotated dataset that contains affective events extracted from Twitter, and evaluate resources produced by prior work over this dataset. To overcome these limitations, a deep-learning model, Aff-BERT, is presented and shown in experiments to have better accuracy and coverage of affective event recognition.

• **Chapter 4** describes the research on improving affective event recognition systems. It introduces a new semi-supervised learning algorithm, *Discourse-Enhanced Self-Training*, which automatically labels an event based on 1) the prediction of an affective event classifier that is trained on the training data, and 2) the affective polarities of the coreferent sentiment expressions that follow the event. It then presents experiments to show that DEST improves the model performance substantially.

• **Chapter 5** presents follow-up research work on improving affective event recognition systems. To generate weakly labeled affective events, it proposes a simple but effective algorithm, *Multiple View Co-Prompting*, which generates and labels affective events by prompting language models. Experiments demonstrate that the generated affective events are of high-quality and can improve the model performance substantially.

• **Chapter 6** describes the research on recognizing expressions of embodied emotion in natural language. It presents the task formulation, a benchmark dataset with human annotation, and a supervised learning model. To further improve the learning model, it presents two methods to automatically harvest weakly labeled instances. The first method is a pattern-based method that mines Embodied Emotion instances from text by exploiting manner expressions with emotion. The second method is a language-model-based method

that labels instances by prompting a large language model. Experiments show that the weakly labeled instances generated by these two methods can train an effective model without using any gold data. Furthermore, they improve the performance of a supervised model when combined with gold data.

- **Chapter 7** presents the conclusions of this dissertation and also the discussions of future work on affective event recognition and embodied emotion recognition.

CHAPTER 2

BACKGROUND

This dissertation aims to create new NLP models to recognize affective events and embodied emotions. In this chapter, I will discuss prior work in several research areas that are closely related to this dissertation. First, I will give an overview of affective text analysis. Second, related work on weakly supervised learning such as self-training, co-training and data augmentation will be introduced. Since several methods developed in this work are closely related to prompting large language models, I will next present prior work on prompting large language models. Finally, I will provide some background work for the study of embodied emotions.

2.1 Affective Text Analysis

Affective text analysis aims to analyze affect conveyed in text. The term of affect has been defined differently in prior work. Much prior work in psychology used the term of affect as an umbrella term that encompasses multiple concepts such as sentiment, emotion, attitude, moods and etc. Bagozzi et al. [10] considered affect as a general term that refers to emotion, mood and attitude. Scherer et al. [170] later constructed a typology of affective states which includes emotion, mood, interpersonal stance, attitude, and personality traits. There also exists other work that viewed affect differently. For example, prior work [118, 178] proposed that affect is fundamentally different from feelings and emotions as it is a prepersonal, non-conscious experience that exists before personal self-awareness develops. Munezero et al. [133] differentiated affect from feelings and emotions. They proposed that affect is "*a predecessor to feelings and emotions*," while feelings are "*person-centered, conscious phenomena*," and emotions are "*preconscious social expressions of feelings and affect influenced by culture*."

In computer science, much work adopted the idea that affect is an umbrella term that broadly refers to many subjectivity terms, such as emotion, sentiment, attitude and opin-

ion. For example, Picard [149] introduced Affective Computing and used the term "*affective*" to refer to emotion, sentiment, subjectivity, personality, mood and attitudes. This definition was later used in the book *Speech and Language Processing* [86] to explain the term of "affective." Strapparava and Valitutti [183] worked on associating a set of words in the WordNet [122] with affect, which includes emotions, personal traits, attitudes, feelings and etc. Mohammad et al. [126] later worked on identifying affect in tweets and used the term affect interchangeably with emotion.

In NLP, a wide range of tasks have been proposed for affective text analysis. In the rest of this section, I will first give an overview of two popular tasks on affective text analysis: sentiment analysis and emotion recognition. Then I will discuss implicit affect analysis, which is closely related to the topics in my dissertation.

2.1.1 Sentiment Analysis

In NLP, sentiment analysis aims to identify a person's general sentiment, opinion or attitude towards a subject matter. Sentiment analysis has been widely studied for decades. Early NLP research on sentiment analysis [38, 192] mostly worked on extracting sentiment in texts from the finance domain. For example, Das and Chen [38] categorized the sentiment of a message from the stock investors on Yahoo's message board with one of the three categories: bullish (optimistic), bearish (pessimistic) and neutral (either spam messages or messages that are neither bullish nor bearish). Later on, an increasing amount of research work focused on mining opinions from reviews on the web. For example, Turney et al. [197] sampled online reviews of automobiles, banks, movies, and travel destinations and classified a review as recommended (thumbs up) or not recommended (thumbs down). Another example is the work by Pang et al. [139], which identified whether a movie review is a positive review or a negative review. More recent studies on sentiment analysis have been extended to texts in domains other than online reviews, including texts in the political domain [63, 131, 12], news [35, 15], and health-related messages [210, 76].

Sentiment analysis is closely related to subjectivity analysis [206, 208], which aims to detect whether a text is subjective (e.g., opinion expression) or objective (e.g., factual information). A subjective expression was defined by Wilson et al. [208] as "*any word or phrase used to express an opinion, emotion, evaluation, stance, speculation, etc.*," which generally

represents the private states of characters [156] and usually cannot be directly observed or verified by others. An example of an objective statement is the factual statement "*The sun rises in the east*." On the other hand, the statement "*The sunrise is mind-blowing*" is a subjective expression, which conveys awe or amazement towards the sunrise. Prior work [52] has found that objective statements are usually factual statements without any positive or negative sentiment. However, statements with no positive or negative sentiment are not necessarily objective [208]. One example is the statement "*Jerome says the hospital feels no different than a hospital in the states*." It is neutral since Jerome does not have a positive or negative sentiment towards the hospital, and it is subjective because it talks about his personal feelings.

Sentiment analysis could be performed at many different levels of granularity. For example, document-level sentiment analysis focuses on determining the overall sentiment expressed within a complete document, such as a review, a news article and a social media post [139, 197, 212]. This level of analysis provides a broad understanding of the sentiment conveyed by the entire text, without delving into the nuances present in individual sentences or aspects. One of the challenges in document-level sentiment analysis is handling the inherent complexity and variability of natural language. A document can contain mixed sentiments, sarcasm, or ambiguity, making it difficult to accurately capture the overall sentiment. Moving down to a finer granularity, sentence-level sentiment analysis involves analyzing the sentiment expressed within individual sentences [211, 217]. Phrase-level sentiment analysis focuses on identifying and analyzing the sentiment within individual phrases or fragments of text [208, 136, 207]. Some work has focused on aspect-level sentiment analysis, which identifies the sentiment associated with specific aspects of an entity mentioned in the text. For example, the hotel review "*The price is low but the room is not clean*" expresses a positive sentiment towards the aspect of "*price*" but a negative sentiment towards the aspect of "*cleaniness*" regarding the hotel room. This level of granularity is particularly important in domains such as product reviews and feedback analysis, where understanding the sentiment towards different aspects of a product or service is essential for decision-making. Finally, much prior research worked on constructing sentiment lexicons [208, 53, 96], where a word or phrase is usually associated with a positive, negative or neutral sentiment. Sentiment lexicons have been widely used for

sentiment analysis [80, 38, 184, 74, 43, 41], where word/phrase-level sentiments are used as features or aggregated by learning algorithms to infer the sentiment of the input text.

Many learning models have been developed for sentiment analysis. Traditional supervised learning models that have been applied include Naive Bayes [139, 87, 195], Support Vector Machine [139, 130, 110, 219], Logistic Regression [73], Maximum Entropy Classifier [139, 92, 89] and so on. In recent years, deep learning models have been widely developed to improve the task performance. For example, Kim [95] performed text classification with Convolutional Neural Networks [104] and word vectors [121]. These models were found to perform well over multiple text classification tasks including sentiment analysis. Socher et al. [179] proposed the Recursive Neural Tensor Network over the parse tree of a sentence to capture the compositional effects of sentiment in the sentence. Sequential models, such as Recurrent Neural Network (RNN) [163] and Long-short Term Memory networks (LSTM) [75], have also been shown effective for sentiment analysis. For example, Tang et al. [186] presented an LSTM model for target-dependent sentiment classification that takes target information into account. Wang et al. [203] proposed an LSTM with an attention mechanism for aspect-based sentiment analysis, which explores the connection between the content of a sentence and an aspect. More recently, researchers have focused on applying pretrained language models for sentiment analysis, such as BERT [44], RoBERTa [113], and GPTs (e.g., GPT2 [157], GPT3, ChatGPT and GPT4).

2.1.2 Emotion Recognition

Emotion recognition in NLP aims to recognize emotions conveyed in texts, such as anger, fear, joy, sadness, and surprise [81, 50]. This is different from sentiment analysis, which mostly assigns a positive, negative or neutral polarity to a text. Prior work in psychology has defined emotions in many different ways. Kleinginna and Kleinginna [97] defined emotion as "*a complex set of interactions among subjective and objective factors, mediated by neural and hormonal systems, which can a) give rise to affective experiences such as feelings of arousal, pleasure and displeasure; b) generate cognitive processes such as emotionally relevant perceptual affect, appraisals, labeling processes; c) active widespread physiological adjustments to the arousing conditions; and d) lead to behavior that is often, but not always expressive, goal-directed and adaptive*." Scherer [170], on the other hand, believed that the definition cannot

distinguish different affective classes from each other such as attitudes and emotions. He further proposed a new definition, which refers to emotion as "*episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism*." Friedenberg and Silverman [56] proposed that emotion is just "*brief brain and body episode that facilitates a response to a significant event*." Based on definitions in prior work, Munezero et al. [133] proposed that an emotion is determined based on the following factors: "*1) Appraisal (cognition); 2) Physiological reactions of the body, such as increased heartbeat and sweating; 3) Feeling; 4) Expressive display, such as facial expression and bodily expression; 5) Readiness to behave in a particular way*."

Many emotion recognition tasks in NLP are proposed based on different theoretical approaches. The most common approach is the *emotional categories* approach, where emotions are represented by basic categories [150, 58, 51]. One prominent theoretical framework in this approach was proposed by Ekman [51], who introduced the concept of basic emotions, which refer to universal emotional states that are identifiable across cultures. There are six basic emotions in this framework: anger, disgust, fear, happiness, sadness and surprise. The six basic emotions are widely used as foundational emotions in a lot of NLP work. For example, Strapparava and Mihalcea [181] proposed the first task of emotion recognition in NLP to classify news headlines with the six basic emotions. Another work [109] created a dataset of English daily dialogues where each utterance was labeled with one of the six basic emotions. The Plutchik Wheel of Emotions [150] is another famous theoretical framework, where there are eight core emotions: sadness, joy, anger, fear, expectation, surprise, trust and disgust. It is also commonly utilized by prior NLP work on emotion detection [224, 196, 1]. For example, Zhou et al. [223] incorporated the relations between different emotions in the Plutchik Wheel as constraints to improve emotion recognition. Another approach is the *emotional dimensions* approach, such as the circumplex model of affect [164]. This approach represents emotions as points in a multidimensional space. Some common dimensions include the valence dimension that indicates how positive/negative an emotion is, the arousal dimension that indicates the activation/deactivation-level of an emotion, and the dominance dimension that measures how much choice one has over an

emotion. This approach has also been taken by prior NLP work. For example, Mohammad et al. [126] proposed several emotion recognition tasks and one of them is to assign the valence level of a tweet from 0 to 1.

Prior NLP work mostly focused on recognizing emotions in different types of texts, including conversations [109, 62, 39], weblogs [213, 155], tweets [126, 125], and news headlines [181, 22]. Some other work focused on associating words or phrases with emotions. For example, Yang et al. [213] developed a collocation model to associate words from weblogs with emotions and constructed an emotion lexicon. Mohammad and Turney [127] created a high-quality emotion lexicon by crowd-sourcing with Mechanical Turk.

To recognize emotion, diverse learning models have been developed. Popular traditional learning models include Naive Bayes [4, 116, 182], Support Vector Machine [13, 161, 116, 4] and Random Forest [146, 216]. In recent years, researchers have focused on developing deep-learning models to improve the task performance. For example, Abdul-Mageed and Ungar [1] proposed a Gated Recurrent Neural Network to predict Plutchik's 24 fine grained emotions and also 8 primary emotion dimensions. Islam et al. [82] presented a multi-channel convolutional neural network to detect emotions in tweets, which leverages different emotion and sentiment indicators, including hashtags, emojis and emoticons. Ghosal et al. [62] focused on emotion recognition in conversation and proposed a Dialogue Graph Convolutional Network, which leverages self and inter-speaker dependency to better model the conversation context. More recently, many learning models are based on pretrained language models. For example, Demszky et al. [39] fine-tuned BERT over the large-scale GoEmotion dataset to detect emotions in dialogue. Alhuzali and Ananiadou et al. [2] built the SpanEmo model on top of BERT to improve emotion classification by learning associations between emotion labels and words in a sentence.

2.1.3 Implicit Affective Text Analysis

While substantial advancement has been made in affective text analysis in recent years, most prior work aimed at detecting explicit affective expressions where affect is stated explicitly with affect-bearing cues (e.g., "*The movie was amazing!*"). However, prior work [144] has found that many affective expressions are implicit - the affect is not stated explicitly and has to be inferred based on the context or commonsense/world knowledge. For

example, the movie review "*This movie made me feel like walking out within 10 minutes of it beginning*" conveys the author's negative opinion towards the movie. Although there are no negative words, we still can understand the negative opinion based on our world knowledge: people usually leave early if they do not like the movie. Although implicit affective expressions might seem easy for us to understand, many of them are still challenging for current NLP systems, which have not been able to capture world knowledge well or reason based on the context.

In recent years, there has been a growing interest in implicit affective text analysis within the research community. Some work focused on identifying words or phrases that indicate implicit affect. For example, Zhang and Liu [221] found that objective nouns and noun phrases could often indicate opinions. One instance is the word "*valley*" in the statement "*Within a month, a valley formed in the middle of the mattress*." Although the word is an objective noun, it indicates the quality of the mattress to be poor and so conveys a negative opinion. Such objective nouns led to difficulty in recognizing affect since the involved sentences are often objective. Motivated by the observation, they developed a method to automatically identify nouns that indicate opinions, based on the intuition that a noun is probably positive (negative) if a noun occurs in positive (negative) context significantly more than negative (positive) context. There is also a line of work that focused on the connotation of words, senses, and frames and frames [88, 159]. For example, Kang et al. [88] developed a loopy-belief propagation algorithm to create a connotation lexicon ConntationWordNet, which contains word- and sense-level connotative polarities. Rashkin et al. [159] focused on predicting the connotative polarities for the object and the subject of a verb from the entities' and writer's perspective. Feng et al. [54] constructed a large connotation lexicon where words are associated with their connotative polarity, using a connotation induction algorithm guided by multiple selected linguistic insights (e.g., selectional preference and distributional similarity). This line of work is different from our work on affective event recognition, as it focuses on the connotative polarity but ours focuses on the affective polarity.

Another relevant line of work studied affect conveyed in figurative language. Ghosal et al. [61] found that figurative language, such as sarcasm, irony and metaphor, often expresses affect that is significantly different from the polarity of its literal meaning. One

example is the sarcastic statement "*I just love it when my friends throw me under the bus*." The writer expresses a strongly negative opinion towards his/her friends, though the literal meaning is positive as indicated by the word "*love*." Prior work [152, 153] also studied the affect expressed by a simile, which is a comparison between two essentially unlike things (e.g., "*he walked as slowly as a turtle*"). For example, Qadir et al. [152] developed the first NLP task to identify whether the affect expressed by a simile is positive, negative or neutral. For example, "*memory like an elphant*" is a positive sentiment expression and "*memory like a sieve*" is a negative sentiment expression.

2.1.4 Discussion

This dissertation focuses on the tasks of affective event recognition and embodied emotion recognition. They are closely related to but fundamentally different from prior work discussed earlier. Affective event recognition differs from prior work in the following aspects. First, affective event recognition aims to detect the affective polarity of an event, so it is centered on events. Secondly, affective event recognition aims to identify the **stereotypical** affective polarity of an event, which refers to the affective impact of an event on most people without considering the context. On the other hand, most prior work on sentiment analysis and emotion recognition (except research that focused on lexicons) studied the contextual affective polarity. Thirdly, affective event expressions often convey affect implicitly and the polarity has to be inferred based on world knowledge. Most prior work, on the contrary, focused on subjective expressions where the affect is explicitly stated (e.g., using affective words).

The study of embodied emotion recognition is also fundamentally different from prior work in sentiment analysis and emotion recognition. It mainly focuses on the physical manifestation of emotions in our body, such as heart racing and leg shaking, while prior work did not.

2.2 Affective Event Recognition

One major topic in this dissertation is affective event recognition. One line of relevant work is the prior study on events with implicit affective states. For example, Goyal et al. [68, 67] identified negative patient polarity verbs that impart affective polarity on

their patients (e.g., *injured*, *killed*), which were used for generating plot unit representations [105]. They developed two methods to collect patient polarity verbs. The first method collected patient polarity verbs by identifying verbs that co-occur with evil agents (e.g., *monster, villain, terrorist*) or charitable agents (e.g., *hero, angel, rescuer*). The second method leveraged the Basilisk bootstrapping algorithm [188] to iteratively collect both positive and negative patient polarity verbs using a set of seed verbs and a set of conjunction patterns. Vu et al. [200] studied emotion-provoking events, which trigger emotions in people who experience them. They proposed a method to leverage the bootstrapping algorithm Espresso [140], the seed pattern "I am < EMOTION > that <EVENT>," and a set of seed emotional words to iteratively collect more patterns and emotion-provoking events. Li et al. [106] extracted major life events (e.g., *receiving award*) from tweets followed by replies that convey condolences (e.g., "*Sorry to hear that*") or congratulations (e.g., "*Congratulations*," "*Congrats*" and "*Awesome*"). They proposed an iterative method with several key components as follows: 1) extract tweets that are followed by a set of congratulation and condolence responses; 2) apply the LDA algorithm [20] to cluster the collected tweets; 3) have human annotators manually label the clusters with the major life event types (e.g., *getting a job, graduation*); 4) expand the set of congratulation and condolence responses based on the replies of another set of unlabeled tweets. The algorithm eventually harvested 42 major life event types. This study is different from affective event recognition as it does not assign affective polarities to an event nor cover everyday events.

Balahur et al. [14] constructed a commonsense knowledge base, EmotiNet, to store real-life situations with the associated emotions. To construct EmotiNet, they first manually selected examples from an existing corpus, ISEAR [171], which contains self-reports that describe ones' own emotions in certain situations (e.g., "*I felt anger when I had been obviously unjustly treated and had no possibility to prove they were wrong*"). Next, they clustered these examples automatically based on the text similarity and randomly selected cluster representatives from each cluster. Then a semantic role labeling algorithm was applied to these cluster representatives to extract triples of *actor— action—object*. Finally, these extracted triples were paired with the emotion labels in the ISEAR dataset and added into the database. Though EmotiNet was shown to be an appropriate approach for emotion detection, this work did not focus on automatically inferring the polarity labels for real-life

situations.

There has been a line of work that focused on crowd-sourcing daily events and their affective polarities. For example, Asai et al. [9] created a database named "HappyDB" by crowd-sourcing 100,000 happy moments from Mechanical Turk. During the data curation, annotators were asked certain questions (e.g., "*What made you happy in the past 24 hours*") and provided answers for them (e.g., "*My son hugged me in the morning*"). Though this database was shown to be diverse, only positive moments were studied. Furthermore, happy moments were represented by sentences instead of events. Another work by Rashkin et al. [158] built a corpus of 25,000 daily events with their intents and reactions by crowd-sourcing. In their work, they represented an event by a phrase that is extracted using several syntactic patterns. The event phrases were extracted from the ROC Story training set [129], the Google Syntactic N-grams [64], and the Spinn3r corpus [66]. Then the event phrases were further post-processed by replacing predicate subjects and other entities with type variables (e.g., *PersonX, PersonY*) and selectively replacing verb arguments with blanks (_). Afterwards, the intents and reactions of the agent (*PersonX*) of each event phrase were annotated by Mechanical Turkers. In this work, an intent is defined as an "*explanation of why the agent causes a volitional event to occur (or "none" if the event phrase was unintentional)*", and a reaction is defined as "*as an explanation of how the mental states of the agent and other people involved in the event would change as a result*." While the reaction is very similar to the affective polarity, most of it focuses on the fine-grained emotional impact of an event, such as "*feeling alert*" and "*feeling happy*." In addition, the reaction may not be affective (e.g., the event of drinking a cup of coffee makes someone feel awake).

The work on +/- effects [40, 32, 41, 42] is also relevant to affective event recognition. Deng et al. [40] created an annotated dataset that contains benefactive/malefactive events that negatively or positively affect entities. These events include destruction (e.g., the event "*kill the flies*" is bad for the flies) or creation (e.g., the event "*bake a cake*" is good for the cake), gain or loss (e.g., the event "*gain weights*" is good for the weights) and benefit or injury (e.g., the event "*pet the cat*" is good for the cat). Later, these events were renamed to *+/-effect events* [41]. Deng and Wiebe [41] investigated the usefulness of +/- effects of events for sentiment analysis. They developed a Loopy Belief Propagation algorithm to propagate sentiments among entities using the +/- effects of events and a set of implicature rules.

And the proposed model was shown to improve over explicit sentiment classification by 10 points in precision. Deng and Wiebe [42] followed up on the idea and proposed a Probabilistic Soft Logic model (PSL) where explicit sentiments, inference rules and +/- effects of events are combined to make joint predictions for entity-level and event-level sentiment. They showed that the PSL model was able to improve both entity-level and event-level sentiment upon strong baselines. Choi and Wiebe [32] focused on creating the lexicon, +/- EffectWordNet, where WordNet senses are associated with +/- effects. To assign +/- effect to a WordNet sense, they developed a label propagation algorithm to propagate +/- effect in a graph where the nodes are the WordNet senses and the edges between nodes are constructed using WordNet relations such as hypernym and troponym. Overall, this line of work is related to but different from affective event recognition. First, the +/- effects are not necessarily affective. For example. the event "*bake a cake*" has a +effect on the cake, but the effect is not affective. Second, the affected entity is not necessarily an animate object (e.g., a cake), while it has to be a person in affective event recognition.

Another line of related work is Emotion Cause Extraction, which links emotional expressions to the events that cause the emotion [71, 70, 30, 107, 209]. Most existing work uses datasets created from news and microblogs that contain an explicitly mentioned emotion. And this research assigns polarity to events in the context of a specific text passage. As a result, an event can be linked to different emotions in different contexts. In contrast, our work aims to identify the *stereotypical* affective polarity of an event, irrespective of context. Consequently, our classifier can be used to predict the affective polarity of events in contexts that do not contain any explicit emotion or sentiment indicators.

In recent years, there has been work that focused directly on affective event recognition [46, 47, 166]. Ding and Riloff [46] was the first NLP work to specifically study affective event recognition. The task was formalized as classifying the affective polarity of an event with one of the following three categories: positive, negative and neutral. Based on their definition, a positive event is typically desirable or beneficial, a negative event is typically undesirable or detrimental, and a neutral event is: 1) not positive or negative, or 2) so general that it could easily be positive or negative in different contexts. An event is represented by a triple ⟨Agent, Predicate, Theme⟩, which captures a predicate, its agent and its theme (e.g., ⟨*they, have, party*⟩). The event components were extracted from text

using dependency relations such as *dobj* and *nsubj* based on pre-defined heuristic rules. To automatically acquire affective events, they developed a semi-supervised label propagation algorithm to infer polarities for unlabeled events. Specifically, they constructed an event context graph from blog posts where events are connected to each other based on local context, discourse proximity and event-event co-occurrence. Ding and Riloff [47] later revisited this task. Different from the previous work [46], they proposed to represent an event by a quadruple ⟨Agent, Predicate, Theme, Prepositional Phrase⟩. The new event representation additionally includes a prepositional phrase, since it can be important for understanding the event. For example, the event ⟨*I, stay, -, at beach*⟩ is a positive event and the event ⟨*I, stay, -, in prison*⟩ is a negative event. Ding and Riloff [47] also developed a method to automatically create a lexicon of affective events by optimizing the semantic consistency in an event graph. In the graph, event nodes are connected based on three semantic relations: semantic similarity, semantic opposition and shared components. The optimization algorithm then iteratively refines the polarity values of the nodes based on the semantic relations. The algorithm eventually created the Affective Event Knowledge Base (AEKB), which contains over half a million events with affective polarities. Given an event, one can search the event in AEKB and extract the polarity label. Overall, this line of work mostly focused on producing lexicons of affective events. As discussed in Section 3.2, a limitation of lexicons is their limited coverage of events and so insufficient generalization to unseen events. This dissertation aims to produce affective event classifiers based on deep-learning architectures to provide better coverage and generalization of affective events.

More recently, Saito et al. [166] utilized deep learning models for affective event recognition in Japanese, including BiGRU and BERT. In addition, they improved the affective event classifiers by a large amount of weakly labeled affective events. To produce weakly labeled events, they developed a method to propagate the affective polarities of some seed events to other discourse-related events using the discourse relations of CAUSE (e.g., $e1$ causes $e2$) and CONCESSION (e.g., $e1$ in spite of $e2$). Our Discourse-Enhanced Self-Training method is related to but different from this method. Our method uses the coreference relations between events and their co-occurring sentiment expressions to help predict the events' polarity labels, while this method relies on the CAUSE and CONCESSION

relations between events.

2.3 Weakly Supervised Learning

Many NLP tasks rely on supervised learning as the main approach, where a classifier is trained over a set of human-labeled data. However, the amount of gold labeled data is often small due to the difficulty and expense of human annotation. With a small gold training set, classifiers can easily overfit and as a result generalize poorly over unseen data. Different approaches have been developed to get around the problem of the lack of gold labeled data. In this section, I will give an overview of two such research areas that are closely related to the methods in my dissertation. The first one is semi-supervised learning, which improves a classifier by exploiting both labeled and unlabeled data. The second research area is data augmentation, which applies some operations to existing gold labeled data to obtain more training data.

2.3.1 Semi-Supervised Learning

A common form of semi-supervised learning trains a classifier over a set of gold labeled data and a larger set of unlabeled data. As the unlabeled data in many tasks are abundant and easy to obtain, semi-supervised learning could improve the classifier with less human effort and more effectiveness. There have been many semi-supervised learning algorithms, such as EM algorithm [128], self-training [175, 215, 120] and co-training [21, 120]. Next, I will introduce more details of self-training and co-training, which are related to the algorithms designed in this dissertation.

Self-training is a semi-supervised method that iteratively improves a classifier with its own predictions on unlabeled data. Typically, a classifier is first trained over a set of labeled data. Next, the classifier is applied to make predictions for a set of unlabeled data. Then the unlabeled data with the most confident or useful predictions are added into the training set to train a new classifier in the next cycle. Despite its simplicity, self-training has been used widely and shown to be effective in a lot of NLP tasks [215, 160, 119, 48, 165]. However, self-training suffers from several problems. First of all, as the classifier learns from its own predictions, classification errors often reinforce themselves in subsequent iterations, resulting in noisy label predictions and stagnated model improvement. Secondly, using

the most confident predictions may not improve the recall/coverage of class instances, as the new training data are usually very similar to the existing training data. Lowering the confidence threshold could potentially improve the coverage of class instances, but usually it comes with the cost of worse precision in label predictions. Our Discourse-Enhanced Self-Training (DEST) method is a form of self-training but more robust to these problems. In DEST, a prediction for an unlabeled instance is made based on both the classifier's prediction and the information from the local discourse context (e.g., the averaged polarity of the coreferent sentiment expressions). DEST better addresses the error amplification problem, since the contextual information helps confirm the correctness and reject the mistakes in the model predictions. In addition, DEST could also produce a more diverse set of labeled data than using the most confident model predictions alone. This is because instances that are not confident by the model alone could receive a high confidence score and be used as new training data, when the contextual information agrees with the model prediction.

Co-training proposed by Blum and Mitchell [21] is another semi-supervised method that has been widely used in many learning tasks [21, 65, 120, 201]. Essentially, the co-training algorithm produces more training data to improve classifiers based on the following key assumptions: 1) an example is represented by two views/feature sets; 2) the two views are conditionally independent given the class; 3) each view is sufficient for correct classification. Based on the assumptions, co-training trains two classifiers using the two data views. Specifically, in each iteration, it first trains each classifier on the training data represented by the corresponding view. Next, each classifier is applied to make predictions for the unlabeled data, and the unlabeled data with the most confident or useful predictions by each classifier are added into the training set. Then the new training data are used to train two new classifiers in the next cycle. As an example, Blum and Mitchell [21] improved a web page classifier with co-training by using two views of a web page: 1) the words on the web page; 2) the words in hyperlinks pointing to the web page. Since co-training leverages multiple views or representations of the data, it could help make more accurate predictions than self-training when the views contain complementary information. In my dissertation, the DEST algorithm and the Multiple View Co-Prompting algorithm are based on the similar idea that different data views are utilized to make better

predictions. In DEST, the two data views are the view of the event phrase and the view of the associated coreferent sentiment expressions. In Multiple View Co-Prompting, the two data views are the Associated Event View and the Emotion View. Despite the similar idea of exploiting different data views, DEST and Multiple View Co-Prompting differ from co-training in several aspects. First, they train only one classifier, while co-training trains two classifiers. Second, they do not hold the strong assumptions of co-training that each view is sufficient on its own for correct classification and that the two views must be conditionally independent of each other given the class.

2.3.2 Data Augmentation

In recent years, data augmentation has attracted increasing attention in the research community. In general, the goal of data augmentation is to improve a model's performance/robustness with an augmented training set. To produce an augmented training set, most strategies usually apply easy-to-implement transformations to generate slightly different variations of existing training data.

Most existing data augmentation strategies could be categorized into three categories according to the survey by Feng et al. [55]. The first category is the rule-based approaches, which adopt predefined rules for transformation [222, 108, 205]. For example, to improve text classification, Zhang et al. [222] augmented the training set by first selecting a random number of words in a training sentence and then replacing each of these words by a randomly selected synonym from an English thesaurus. Evaluation with different models over six text classification datasets showed that the data augmentation method could substantially reduce testing errors. EDA [205] is another widely used data augmentation method that generates copies of a training sentence by four strategies, including: 1) synonym replacement, which replaces n words from the sentence that are not stop words; 2) random insertion, which inserts a random synonym of a random word that is not a stop word into a random position in the sentence; 3) random swap, which randomly swaps two words in the sentence; 4) random deletion, which randomly deletes each word in the sentence with a probability. Their experiments on five text classification tasks showed that EDA could improve a text classifier when the training data is limited, though the performance gain becomes smaller when there is more training data.

The second category is the example-interpolation-based approaches [220, 199, 29] that produce new labeled instances by interpolating multiple training instances as well as their labels. Essentially, this approach requires that data instances be represented by continuous vectors. For example, Chen et al. [29] developed a method called TMix, which generates a new instance by interpolating the hidden representations, produced by BERT, of two training texts.

The third category is the model-based approaches, which transform existing training data with trained models. For example, Sennrich et al. [176] designed the popular method, back-translation, to generate a new instance by translating a training instance to another language and back to the original language. The new instance is then paired with the label of the original training instance. More recent model-based approaches have tried performing data augmentation using pretrained language models [59, 134]. For example, Ng et al. [134] generated variations of a sentence by arbitrarily replacing some words in a sentence with the [MASK] token and completing the sentence by BERT. Some other work developed methods to first fine-tune language models over the training data and then produce new training data by sampling from the fine-tuned language models [214, 5, 100]. For example, Anaby-Tavor et al. [5] worked on text classification and proposed a data augmentation method, LAMBDA, that first fine-tunes GPT2 [157] over labeled data and then synthesizes weakly-labeled data from it. Specifically, given a set of training sentences with gold labels $\{(x_i, y_i)\}_{i=1}^{n}$, the method organizes the training data into a long sequence: "y_1 SEP x_1 EOS y_2 SEP x_2 EOS...y_n SEP x_n EOS," where SEP is the token to separate a sentence x_* and its label y_*, and EOS is the token to denote the end of the sentence. Then GPT2 is fine-tuned with the next-token prediction task over the long sequence, during which it learns to generate a sentence given a class label. To synthesize weakly labeled data, one can feed GPT2 with the prompt "y SEP" and GPT2 will continue to generate a sentence that belongs to the class y until the EOS token is met.

The Multiple View Co-Prompting method in this dissertation is a model-based data augmentation approach for several reasons. First, it generates event phrases by prompting language models. Second, it extracts data views of event phrases from pretrained language models for label assignment. However, it differs from other model-based methods in the following aspects: 1) it does not require fine-tuning the language models; 2) it seeks dif-

ferent types of data views from language models, while others elicit only one type of data view. In our study of embodied emotion recognition, the weakly supervised method to produce labeled instances using the predictions from language models is a model-based data augmentation method.

2.4 Prompting Large Language Models

Recently, there has been a significant increase in attention towards prompting language model. In general, research on prompting language models aims to query a pretrained language model with a prompt such that the language model generates desired outputs that could be used for downstream NLP tasks [157, 148, 177, 172, 173, 60]. Typically, a prompt could be either a cloze prompt which has an empty slot to be filled in, or a prefix prompt which is an incomplete sequence of tokens to be completed. The cloze prompt is mainly used to elicit information from masked language models which are pretrained to predict a token for a slot (e.g., BERT and RoBERTa [113]), while the prefix prompt is used for language models that are pretrained to generate the next token in an auto-regressive way (e.g., GPT2 and GPT3). Suppose we want to infer the affective polarity of the event phrase, "*I got an A in the final exam*," by prompting pretrained language models. We could query BERT with a cloze prompt: "*I got an A in my final exam! I feel [MASK]*," where [MASK] represents the empty slot. Then BERT could generate emotional terms such as "*happy*" and "*good*" in the slot. We could also query GPT3 with a prefix prompt: "*I got an A in the final exam. I feel*." And GPT3 may complete the prefix prompt as follows: "*I got an A in the final exam. I feel really good*." Based on the completion produced by the language models, we can infer that the event is positive.

Language models have been shown to capture diverse knowledge in the pretraining corpus [148, 84, 185]. For example, Petroni et al. [148] performed an in-depth analysis and showed that BERT contains diverse relational knowledge such as "PlaceOfBirth" (e.g., "*Francesco Bartolomeo Conti was born in Rome.*"), "CapableOf" (e.g., "*Ravens can fly*") and "UsedFor" (e.g., "*A pond is used for swimming*"). They also found that BERT competes effectively with traditional NLP approaches that have some access to oracle knowledge. The observation that language models possess knowledge soon inspired researchers to investigate the capability of language models to perform zero-shot learning (with no ac-

cess to training data) and few-shot learning (with access to a limited amount of training data) [25, 172, 173, 60, 204, 193, 194, 143]. In the zero-shot learning scenario, a language model is typically prompted with a query that contains a task instruction and an input instance to label. In the few-shot learning scenario, a common setup is to use the prompt in the zero-shot learning and insert into it a few training instances and their labels as exemplars. As a few exemplars are introduced, the model could better understand the task and make a better prediction for the input instance. Recent research has found that large language models, such as Llama2 [194], ChatGPT, Falcon [143] and GPT4, can perform impressively well over unseen tasks with zero-shot and few-shot learning.

Although language model prompting has been show to achieve promising results over various learning tasks, prior work [84, 111, 114, 202] has observed that the quality of the feedback provided by language models is usually sensitive to the prompt design, and that prompts that mean the same thing but are worded differently could often guide language models to generate inconsistent feedback. As a result, finding the optimal prompt has become an important task. To avoid the cost of manual prompt engineering, some research has focused on automatically searching for optimal prompts, such as prompt-based fine-tuning, automatic prompt search, and discrete/continuous prompt optimization [177, 154, 172, 173]. For example, Jiang et al. [84] designed an approach to automatically construct prompts based on sentences mined from a large text corpus that contains examples in the training set. Another work [60] used T5 to generate prompts automatically. Besides using only one prompt to elicit desired outputs, some work developed prompt ensembling methods to generate outputs by combining outputs of multiple prompts [84, 172, 173], which could stabilize the model performance and reduce the cost of prompt engineering. For example, to better estimate the log probability of a token at a masked position, Jiang et al. [84] used K prompts to extract K log probabilities for that token and averaged them. To illustrate, to extract the profession of Barack Obama, the set of prompts could be "*Obama worked as a [MASK]*" and "*Obama is a [MASK] by profession*." Qin and Eisner [154] also used K prompts to estimate the probability of a token. Specifically, their work computed the probability of a token by a weighted average of the K probabilities produced by the prompts. The weight of a prompt represents the preference for the prompt, and it could automatically be learned if there is training data.

Prior methods of prompting language models are closely related to the methods developed in this dissertation. Specifically, the Multiple View Co-Prompting method to generate affective events and the LM-based method to generate Embodied Emotion instances are prompt-based methods. However, the Multiple View Co-Prompting method is fundamentally different from previous prompting-based approaches. Essentially, existing methods mostly prompt language models for the same type of label information or data view. For example, Schick et al. [172] developed the pattern-exploiting training (PET) method which trains an ensemble of language models with multiple prompts over weakly-labeled data. To apply PET for affective event recognition, we may use prompts such as "*[EVENT]. I feel [MASK],*" "*[EVENT]. This is [MASK].*" and "*It is [MASK] that [EVENT],*" all of which ask for the same data view - emotional terms that co-occur with the input event. On the other hand, our Multiple View Co-Prompting method utilizes multiple prompts to seek different views (the Associated Event View and the Emotion View) of a data instance. With the independent and complementary information from multiple data views, our method is able to produce more robust labels than methods using only one data view.

2.5 Embodied Emotion

2.5.1 Embodied Emotion in Psychology

In psychology, most emotion theories acknowledge that the body and the mind play important roles in the emotion experience [17]. How the body and the mind participate in the emotional experience, however, is still an open research question. Some emotion theories proposed that the body impacts the mind during an emotion experience [83, 102, 45, 191, 190, 49]. Briefly, this group of theories believed that an emotional experience is a process that starts from external stimuli and undergoes changes in the body before being internally felt by the mind as an emotion. For example, James [83], one of the earliest attempts to explain the cause of the emotion, explained that "*Our natural way of thinking about these standard emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My dissertation on the contrary is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion*." The same theory was later independently proposed by Lange [102] and afterwards renamed as the James-Lange

Theory. To illustrate with another example, suppose we see a mountain lion while we are hiking, and we become terrified. Based on this group of theories, the perception of the mountain lion stimulates some bodily responses such as a racing heart, increased muscle tension, sweating and so on. These bodily responses are then interpreted as fear by our mind, and we feel the emotion of fear.

Another group of theories held the contrary belief that the mind influences the body during an emotion experience [36, 7, 8, 57]. In these models, the mind/mental state corresponds to the entire stimulus situation (e.g., seeing the mountain lion in the previous example). One of the modern pioneering work in this group is the book "The Expression of Emotions in Man and Animals" by Charles Darwin [36]. The book presents three chief principles of expression of emotion, and the first principle, the principle of serviceable associated habits, states that certain states of the mind seek expressions and automatically provoke bodily responses:

> *Certain complex actions are of direct or indirect service under certain states of the mind, in order to relieve or gratify certain sensations, desires, and whenever the same state of mind is induced, however feebly, there is a tendency through the force of habit and association for the same movements to be performed, though they may not then be of the least use. Some actions ordinarily associated through habit with certain states of the mind may be partially repressed through the will, and in such cases the muscles which are least under the separate control of the will are the most liable still to act, causing movements which we recognise as expressive. In certain other cases the checking of one habitual movement requires other slight movements; and these are likewise expressive.*

Another example is the Cannon-Bard Theory [26, 16], which considers that a stimulus evokes emotions and bodily responses simultaneously. In this theory, the thalamus, a part of the brain responsible for relaying sensory information to other parts of the brain, plays a crucial role. Consider the previous example of seeing a mountain lion. Based on the Cannon-Bard theory, the thalamus receives the sensory input of the mountain lion and sends a signal to the amygdala, a structure in the brain responsible for processing strong emotions such as fear and anger. At the same time, the thalamus also sends a signal to the autonomic nervous system, resulting in bodily responses such as a racing heart, increased muscle tension, sweating and so on. In this process, the mental state (the perception of the mountain lion) causes the changes in the body and also the emotional feelings. Another instance in this group is the appraisal theory, proposed by Arnold [7]. The appraisal theory suggests that when a stimulus occurs, a person first evaluates its significance as well as its

value in an automatic way. Then the appraisal evokes the corresponding emotion, which is embodied by various bodily changes.

In addition to the theories of *mind* $\rightarrow$ *body* and *body* $\rightarrow$ *mind*, some theories [101, 18] hold an interactionist view that the mind and the body interact dynamically to shape the emotion experience together. One example along this line is the Schachter-Singer Two-Factor Theory [168], which proposed that "*an emotional state may be considered as a function of a state of physiological arousal and of a cognition appropriate to this state of arousal*." Consider again the example of seeing a mountain lion. According to the Two-Factor Theory, the emotion of fear arises from the following interaction (listed in the temporal order): 1) The stimulus occurs (the perception of the mountain lion); 2) Some physical arousal occurs, such as a racing heart and increased muscle tension; 3) the physical arousal is automatically associated with a cognitive label, which is fear in this case; 4) Fear is felt.

In this dissertation, our study of embodied emotion is closely related to the view that the mind influences the body. Specifically, we define embodied emotion as physical movement or physical arousal that is mainly evoked by emotion. In addition, we apply one more constraint: the physical movement has no other purpose beyond emotion expression. The reason for using this constraint can be found in Chapter 6. With the additional constraint, the embodied emotions we study are a subset of those discussed in the previous literature.

2.5.2 Analysis of Emotion Communication Signals and Emotion Components

Our research on embodied emotion is closely related to prior research on analyzing emotion communication signals. Kim and Klinger [93] studied how emotions are expressed in text using non-verbal communication signals. The study focused on eight emotions: joy, sadness, anger, fear, trust, disgust, surprise, and anticipation, and eight non-verbal communication signals: 1) physical appearance (e.g., "*blushed crimson red*"), 2) facial expressions (e.g., "*rolled his eyes*"), 3) looking behavior (e.g., "*averted her eyes*"), 4) arm and hand gesture (e.g., "*opened her arms*"), 5) movements of body as a whole (e.g., "*slumped his shoulders*"), 6) characteristics of voice (e.g., "*voice getting smaller and smaller*"), 7) spatial relations (e.g., "*leaping into her arms*"), and 8) physical sensations (e.g., "*tingling all over*"). Using an existing dataset [94] where characters in fan fictions had been annotated with emotions, they performed manual annotation and analysis of the non-verbal communication signals

for the characters with emotions in the dataset. Our study of embodied emotion differs from this work in the following two aspects. First, the non-verbal communication signals do not cover all body parts that are covered by our study, including the internal body parts (e.g., heart) and some single body parts (e.g., leg and toes). Another key difference is that they only performed manual analyses over the corpus and did not propose an automated task for recognizing these non-verbal signals.

The study of embodied emotions is also related to the existing work on recognizing emotion components [28, 34]. Casel et al. [28] classified a sentence with one of the five emotion components: 1) Cognitive Appraisal (e.g., "*I wasn't sure what was happening*"), 2) Neurophysiological Symptoms (e.g., "*My heart is racing*"), 3) Motivational Action Tendencies (e.g., "*He wanted to run away*"), 4) Motor Expressions (e.g., "*She smiled*") and 5) Subjective Feelings (e.g., "*I felt so bad*"). For classification, they built multiple classifiers such as a bi-LSTM followed by a convolution layer and a classification layer, and a maximum entropy classifier with TF-IDF bag-of-words features. They further combined an emotion classifier and an emotion component classifier into a joint model, and showed that the additional information of emotion component could improve the performance of the emotion classifier. Cortal et al. [34] also worked on classifying sentences with emotion components and had the similar finding that the information of emotion components helps improve the performance of emotion classifiers. Different from the prior work [28], they studied four different emotion components: 1) Behavior, 2) Feeling, 3) Thinking and 4) Territory. Overall, this line of work differs from ours in two aspects. First, the emotion components are fundamentally different from embodied emotions. For example, Cortal et al. [34] included all behaviors not evoked by emotion during an emotional event (e.g., "*giving a lecture*") and Casel et al. [28] included goal-oriented physical movements (e.g., "*recover the stolen horse*"), while ours does not. In addition, their work focused on teasing apart different emotion components from each other. In essence, it assumes the text to classify is emotional, while ours does not.

2.5.3 Embodied Artificial Intelligence

There has been growing interest in the research on Embodied Artificial Intelligence (AI), which aims to develop AI systems that are integrated with physical or virtual en-

tities and can interact with humans or the environment through sensory inputs and active responses. Unlike traditional AI, which primarily relies on abstract data processing, embodied AI is concerned with the ability to perform tasks in the real world. Sensory inputs to an embodied agent could be various types of data from the environment that enable it to perceive and interact with the world. For example, a robot that manipulates objects receives visual inputs that are captured through imaging devices (e.g., camera) to identify objects and navigate through space. It also receives auditory inputs that are acquired through microphones to detect and interpret sounds. Another type of inputs is tactile input, which is gathered through touch sensors to allow the robot to sense physical contact, texture, pressure, and temperature.

A substantial amount of research on developing embodied AI has been explored in other disciplines. One line of research [6, 72, 123] focused on developing embodied navigation agents that navigate to a goal in a three-dimensional environment with or without external priors or natural language instruction. Various tasks along this line have been proposed, such as motion planning that searches for a collision-free path in a given environment [103], and Simultaneous Localization and Mapping that builds a map of an unknown environment and localizes the agent in the map [69].

Another line of research studied embodied question answering (QA) [37, 218], where an agent physically interacts with its environment to answer questions. For example, we may ask an agent the following question: "*Is there any apple in the fridge?*". To answer the question, an agent needs to navigate to the fridge, open the fridge, and search for apples. The task of Embodied QA is challenging, as it greatly relies on the capabilities for many other tasks. One important capability is to visually ground the spoken language [33, 189, 78], where basic terms in spoken language are associated with visual cues. Consider the scenario where an agent is asked to describe an object (a red apple) on the table. To generate descriptions such as "*There is a red apple on the table*," an agent must be able to associate the visual element (color) with the word "*red*."

The study of social robots is another important topic in embodied AI, which focuses on building robots that interact with people [169, 23]. Social robots can be useful in many domains, such as education (e.g., robotic tutors) [19], healthcare (e.g., robotics assistants to provide healthcare support) [85], and entertainment (e.g., playful robots) [99]. For example,

Cherakara et al. [31] built an embodied conversational agent called FurChat, which can generate open and closed-domain dialogue with emotive facial expressions and interact with people using verbal and non-verbal cues. The workflow in Furchat consists of multiple components: 1) convert the user speech to text using automatic speech recognition, 2) interpret the text using natural language understanding, 3) manage the interaction flow using a dialogue manager, and 4) generate a textual response by natural language generation (e.g., using GPT-3.5), 5) convert the generated text to speech using text-to-speech technology and play the speech through the robot's speaker. To display gestures and facial expressions, the agent first detects the emotion of the user using GPT3.5, and selects an appropriate emoticon (e.g., a sad face for a sad conversation). Given the emoticon, a gesture is selected from a pre-defined set of gestures and displayed along with the generated speech.

The study of embodied emotion recognition in natural language differs from work on embodied AI, because it focuses on identifying textual expressions of bodily responses evoked by emotions and does not involve interactions with the world. However, it is closely related to and could potentially be useful for work on Embodied AI where emotion plays an important role. First, our study can potentially enhance an agent's capability of detecting emotions conveyed in users' speech. Suppose a robot hears a user say "*What you said made my stomach turn*." After converting the speech to a text, a robot should understand that the user got a negative emotion due to the embodied emotion expression "*my stomach turn*" and act accordingly. Our study could also potentially help an embodied agent perform a wider range of gestures to express emotions. Many social robots, such as FurChat, choose an expression from a pre-defined set of gestures to display a specific emotion. The work in this dissertation could be extended in the future to identify diverse bodily manifestations of emotions from text. And researchers in other disciplines could use this harvested set of embodied emotions as a reference to design new gestures to express emotions.

CHAPTER 3

IMPROVING AFFECTIVE EVENT RECOGNITION WITH DEEP-LEARNING MODELS

In recent years, affective event recognition has attracted the attention of researchers in NLP. Prior research has proposed different methods to recognize affective events. While effective, earlier methods suffer several limitations that result in limited model performance and generalization. Motivated by these limitations, this work proposes a deep-learning model, which outperforms the existing methods and also generalizes better to unseen events as shown in Section 3.4.

In this chapter, I will first elaborate on the basic concepts of affective event recognition in Section 3.1. Then I will present details of existing work on affective event recognition and their corresponding limitations in Section 3.2. Finally, a deep-learning model for affective event recognition is introduced in Section 3.3.

3.1 Basic Concepts in Affective Event Recognition

This dissertation mainly adopts the definitions introduced by Ding and Riloff [46, 47] on affective event recognition.

An **event** is defined to be in the form of a tuple ⟨*Agent*, *Predicate*, *Theme*, *Prepositional Phrase* (PP)⟩, where *agent* usually refers to the entity that performs actions, *predicate* refers to the action performed and *theme* refers to the entity affected by the action. For example, in the event phrase "*Jack eats an apple*," *Jack* is the agent, *eat* is the predicate and *apple* is the theme. An event tuple could include a prepositional phrase, which is important for understanding the event. For example, *staying in prison* is an undesired event but *staying at a beach* is usually a desired event.

To extract event tuples, we followed the extraction methods in prior work [46, 47]. Specifically, the extraction methods rely on some heuristic rules based on dependency

relations to extract the event components.[1] For example, the Agent role is extracted using the *nsubj* relation and the Theme role is extracted using the *dobj* relation. Furthermore, each component in the event tuple is represented by lemmatized words. To illustrate, from the sentence *"John is watching the sunset at the beach,"* we extract an event ⟨*John, watch, sunset, at beach*⟩, where the predicate is *"watch,"* the agent is *"John,"* the theme is *"sunset,"* and the prepositional phrase is *"at beach."* In addition, an event could also represent a state, for example, ⟨*I, feel, sad, -*⟩.

Following prior work [46, 47], we define **affective events** to be events that stereotypically impact us in a positive (desirable) or negative (undesirable) way. We use the term **affective polarity** to refer to the affective impact of an event. Affective polarity could be positive (desirable), negative (undesirable), or neutral (neither desirable nor undesirable). For example, the affective polarity is negative for the event ⟨*I, drop, my phone, in toilet*⟩, positive for the event ⟨*I, win, game, -*⟩, and neutral for the event ⟨*I, walk, -, on street*⟩. The impact of an affective event is closely tied to the affective state of the person who experiences it. For example, if we break our legs, we usually experience a negative affective state as the event leads to undesirable injury and pain. On the other hand, if we get a degree from the university, we usually possess a positive affective state since getting a degree is an achievement and opens up more possibilities (e.g., *getting a job*) in the future. Here we emphasize the *"stereotypical"* impact of an event, which is how most people are impacted when experiencing the event. While it is possible that an individual might have an atypical feeling towards an affective event (e.g., someone dislikes getting a degree from the university), we aim to identify its most likely affective polarity in the absence of evidence to the contrary.

Based on the aforementioned definitions, the task of affective event recognition is to classify the stereotypical affective polarity of a given event, which could be positive, negative or neutral.

[1] No semantic role labeling is performed.

3.2 Limitations of Existing Methods for Affective Event Recognition

As introduced in Section 2.2, most prior work on affective event recognition focused on producing lexical resources that contain verbs or event phrases with corresponding affective polarity values [68, 67, 159, 46, 47]. To predict the polarity for an event with a lexical resource, one could look up the event in the lexical resource and obtain its polarity. While lexical resources have been useful for improving affective event recognition, they may not generalize well to unseen events. First, they are likely to have insufficient coverage of affective events. This is mainly because an event could be described in various forms that have different syntactic structures and words. Consider the event "*Jack goes on a trip*." The event could be described with different phrases such as "*Jack travels*," "*Jack starts his trip*" and "*Jack goes on a journey*." As the corpus for building a lexicon is usually limited, a lexicon is not likely to capture all possible forms of all events. In addition, events in our daily life evolve over time, which makes a lexical resource likely to become outdated after some time. Consider the negative event of being infected with COVID-19 (e.g., "*I test positive with COVID-19*"), a disease that emerged after 2019. None of the existing lexical resources of affective events would recognize this event, as they were built from texts that were written before 2019.

A second issue is that the polarity label quality of these lexical resources is far from perfect, as the methods developed to build these resources do not capture well the event semantics. For example, in building the AEKB resource, Ding and Riloff [47] propagated polarities from a set of seed events to events that are similar/dissimilar to them. To measure if two events are similar/dissimilar to each other, they represented each event using the GloVE embeddings [145] and took the cosine similarity between the two events' embeddings. However, GloVE emebddings or static word embeddings in general have been shown to be limited in capturing the textual semantics. As a result, the resulted polarity labels could be noisy. Table 3.1 shows examples of affective events that are correctly and incorrectly labeled in AEKB.

To test if the limited generalization is a real issue, we investigated how well the current largest lexical resource of affective events, Affective Event Knowledge Base (AEKB) [47], generalizes to events in new texts. Briefly, AEKB contains over half a million automatically

Table 3.1: Examples of affective events correctly and incorrectly labeled in AEKB.

Positive Events	
Correct:	
⟨I, win, ribbon, -⟩	⟨I, treasure, our friendship, -⟩
⟨everything, look, okay, -⟩	⟨something, excite, I, -⟩
Incorrect:	
⟨I, not feel, fine, -⟩	⟨I, have to open, it, -⟩
⟨I, say, it, for reason⟩	⟨I, wonder, -, after day⟩
Negative Events	
Correct:	
⟨arm, start to hurt, -, -⟩	⟨my parent, be angry, -, with me⟩
⟨-,distract, I, -⟩	⟨we, not, afford, house⟩
Incorrect:	
⟨rose, laugh, -, -⟩	⟨thought, pass, my mind, -⟩
⟨I, put down, phone, -⟩	⟨I, find, screw, -⟩
Neutral Events	
Correct:	
⟨I, type, blog, -⟩	⟨I, pull out, copy, -⟩
⟨I, go to make, food, -⟩	⟨I, pick up, textbook, -⟩
Incorrect:	
⟨I, watch, movie, -⟩	⟨I, not get, lunch break, -⟩
⟨I, go to watch, sunset, -⟩	⟨I, go, -, to disney⟩

labeled affective event phrases extracted from personal blog posts in the ICWSM 2009 and 2011 Spinn3r datasets. To conduct the experiment, we created a new dataset for affective event recognition in Twitter, where the genre style and the wording are often different from those in blog posts. Specifically, the dataset contains 1,500 events extracted from tweets in Twitter with human annotated polarity labels. The events are represented using the event representation in the AEKB: ⟨Agent, Predicate, Theme, Prepositional Phrase⟩, except that we also allowed adjectival modifiers in the noun phrases (e.g., ⟨*I, have, delicious food*, -⟩) as they often impact the polarity. We will refer to this dataset as **TWITTER**. The full detail of its creation process is elaborated in Section 4.3. To apply AEKB over TWITTER, we matched every event in TWITTER against the AEKB. If the event was found in the AEKB, we assigned the corresponding polarity in the AEKB. Otherwise we assigned the Neutral

polarity.

Table 3.2 reports the performance of AEKB, including the macro-averaged F1, the recall and the precision scores for each of the three polarities: positive (POS), negative (NEG), and neutral (NEU). The first row (Blogs) shows the AEKB results originally reported by Ding and Riloff [47] for events extracted from blog posts, for comparison. Of the 1,500 Twitter events, only 997 events (66%) were found in the AEKB. The second row of Table 3.2 (Twitter-found) shows results for these 997 events. The recall for positive polarity is substantially lower. The low recall for positive polarity is probably because many positive Twitter events are labeled as neutral in the AEKB, as suggested by the lower precision for neutral polarity. For negative polarity, the precision is higher for Twitter data than blog data while the recall is lower. Overall, we see that AEKB is not able to recognize many positive and negative events in Twitter and mistakenly classifies them as neutral events. Another major issue is that about one third (34%) of the Twitter events were not found in the AEKB at all. The third row of Table 3.2 (Twitter-all) shows the results across *all* 1,500 Twitter events, where the missing events are automatically labeled with Neutral. Overall, only 37% of the negative events and 26% of the positive events could be recognized by the AEKB.

In general, these results show that the AEKB, despite its largest size, cannot recognize many affective events for two reasons: (1) many affective events are not present in the knowledge base, and 2) many positive and negative events are incorrectly labeled as neutral. This confirms our concern that existing lexical resources do not generalize well to unseen events.

Table 3.2: Performance of AEKB over the TWITTER dataset.

Method	F1	POS		NEG		NEU	
		Pre	Rec	Pre	Rec	Pre	Rec
Blogs	71.4	75.7	55.1	70.4	63.3	79.3	88.5
Twitter-found	65.2	72.2	40.6	78.7	60.8	65.6	87.9
Twitter-all	50.8	72.2	26.2	78.7	37.1	65.6	61.8

3.3 Aff-BERT: A Deep-Learning Model for Affective Event Recognition

To address the limitations of prior work, we proposed a deep-learning classification model for better coverage and accuracy. We will refer to this model as **Aff-BERT**. Essentially, Aff-BERT is built on top of the pre-trained BERT model [44] and is fine-tuned during the training. Figure 3.1 shows the architecture of Aff-BERT. It takes an event (in the form of tuple) as input and predicts whether the polarity of the event is positive, negative or neutral. First, we concatenate all words in the input event tuple into an event phrase and prepend the special [CLS] token used in BERT to the event phrase. For example, the event tuple ⟨*I, not pass, my history test, -*⟩ is changed to "*[CLS] I not pass my history test*" after this step. We then encode the event phrase with BERT, which outputs an embedding for each token in the input sequence. Then we pass the 768-dimension output embedding of the [CLS] token to a linear layer with softmax to produce a probability distribution over the three polarity classes. Finally, the polarity with the highest probability value is assigned to the event. During training, Aff-BERT is fine-tuned with respect to the cross entropy loss over the gold training data.

The motivation for using BERT as the base encoder is that representations produced by BERT could potentially capture the rich meaning of an event and cluster events that

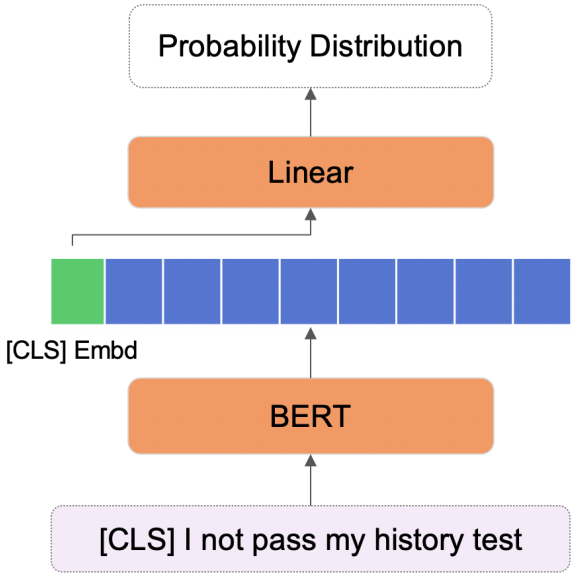

Figure 3.1: The architecture of Aff-BERT.

are semantically similar together, leading to better coverage and generalization. Suppose we want to recognize the affective polarity of the event ⟨*I, develop, COVID-19 symptom, -*⟩ but the training data does not have any event related to COVID-19. Aff-BERT might still successfully predict the event as negative, as BERT could potentially produce an event representation that is similar to the representations of other infection events such as ⟨*I, develop, cold symptom, -*⟩ and ⟨*I, get, allergy symptom, -*⟩.

3.4 Evaluating Aff-BERT

We evaluated Aff-BERT over the dataset created by Ding and Riloff [47], which contains 1,490 manually annotated affective events extracted from personal blog posts. In the dataset, there are 295 positive (20%), 264 negative (18%), and 931 (62%) neutral events. We used the original data split provided by Ding and Riloff [47], where 490 events are used for validation and 1,000 events are used for testing. We will refer to this dataset as **BLOG**.

We compared Aff-BERT to several systems. The first system is the **Affective Event Knowledge Base (AEKB)** produced by Ding and Riloff [47], which contains over a half million events with polarity labels. The second system is **ELMo+LSTM**, an LSTM model with ELMo encoding as inputs [147]. Specifically, ELMo+LSTM first encodes an event phrase with the pretrained language model ELMo [147]. Then it feeds the last layer of ELMo's outputs into a 1-layer LSTM to produce a polarity distribution. We also developed another system, **ELMo + Linear**, where the last layer of ELMo's outputs are averaged and fed into a linear layer to produce a polarity distribution. The last system for comparison is **Aff-BERT(AEKB)**, an Aff-BERT trained with the supervision of AEKB. Specifically, we constructed a training set by collecting from AEKB events with label confidence $\geq 60\%$. Then we trained Aff-BERT over this training set.

We trained all models for 5 epochs with a batch size of 50 and a linear warmup rate of 10 using AdamW optimizer. We also performed a hyperparameter search and selected the best parameter values according to the performance over the validation set. Specifically, the LSTM has a hidden size of 512 and a dropout rate of 0.2. The learning rate is 0.01 for the LSTM, 0.1 for the linear classifier, and 1e-5 for Aff-BERT.

Table 3.3 shows the performance of all systems on the BLOG test set, including the macro-averaged F1 scores as well as the precision and recall scores for each polarity. The

Table 3.3: Performance on the BLOG test set. The F1 score is macro-averaged across polarities.

Method	F1	POS		NEG		NEU	
		Pre	Rec	Pre	Rec	Pre	Rec
AEKB	71.4	**75.7**	55.1	70.4	63.3	79.3	**88.5**
Aff-BERT(AEKB)	73.6	73.2	56.6	75.6	69.5	80.9	**88.5**
ELMo+Linear(Gold)	62.3	56.0	53.7	56.2	51.3	78.2	81.4
ELMo+LSTM(Gold)	70.5	71.4	60.8	70.8	57.3	81.3	**88.5**
Aff-BERT(Gold)	**77.4**	71.7	**66.2**	**78.2**	**77.2**	**85.0**	87.4

first row shows the performance of AEKB reported by Ding and Riloff [47]. The second row shows the performance of Aff-BERT trained with the weakly labeled events in AEKB. We observe that Aff-BERT trained with AEKB data outperforms AEKB. Notably, the improvement is due to the substantial recall and precision gain for negative events. This indicates that a deep learning model (Aff-BERT) trained with AEKB achieves better coverage than AEKB alone, probably due to the semantically rich representations produced by BERT.

We next experimented with learning from gold labeled data by running 10-fold cross validation over the BLOG test data. In each of the 10 runs, we used 8 folds for training, 1 fold for development and 1 fold for testing. The results of ELMo+Linear, ELMo+LSTM and Aff-BERT trained with gold training data are shown in the third, fourth and fifth rows respectively. The linear classifier and the LSTM do not perform as well as the AEKB. But Aff-BERT trained on gold labeled data performs substantially better than both the AEKB and Aff-BERT trained with the AEKB. In particular, the high performance of Aff-BERT(Gold) is due to its substantial gains in the precision scores for Negative and Neutral polarities and the recall scores for Positive and Negative polarities. Overall, the results show that fine-tuning Aff-BERT on a relatively small amount of gold labeled data produces a strong affective event classifier, with respect to both recall and precision.

3.5 Conclusion

In this chapter, we delved into the issue that existing lexical resources for affective event recognition do not generalize well to unseen events. The limitation mainly stems from their insufficient coverage of affective events. In addition, it is also partly because the methods to develop these lexical resources cannot capture the event semantics well. To demonstrate these limitations, we curated a new dataset for affective event recognition on Twitter.

Our experiments over the TWITTER dataset demonstrate that existing lexical resources perform poorly when applied to this new dataset. Motivated by the issue of insufficient generalization, we developed a deep-learning model, Aff-BERT, based on fine-tuning the pretrained BERT model. We demonstrated in experiments that Aff-BERT substantially outperforms other methods and generalizes better to unseen events.

CHAPTER 4

IMPROVING AFFECTIVE EVENT RECOGNITION BY USING DISCOURSE-ENHANCED SELF-TRAINING

In Chapter 3, we demonstrated that existing lexical resources for affective event recognition can not generalize well for affective event recognition. To overcome their limitations, we developed a deep learning model, Aff-BERT, which leverages the power of pretrained language models to better capture the textual semantics of an event. We conducted experiments to fine-tune Aff-BERT over the gold labeled data, and showed that Aff-BERT achieves much better performance than existing lexical resources and strong baselines such as AEKB, and Aff-BERT trained on the AEKB. This indicates that fine-tuning Aff-BERT can produce a strong affective event classifier on a relatively small amount of gold labeled data. The strength of this model led us to wonder whether classification performance could be further improved by semi-supervised methods that introduce more training data. One classical semi-supervised method is self-training, which improves a classifier by using its own predictions on unlabeled instances to generate more training data. However, self-training has limitations. Learning from one's own predictions could simultaneously learn from one's own mistakes, leading to error propagation in the future. Furthermore, using the most confident labels may not improve recall much because the new training instances are similar to the old ones, while using less confident labels often decreases precision because the training data becomes noisier.

To address self-training's limitations, we introduce a novel method, *Discourse-Enhanced Self-Training* (DEST), to further improve Aff-BERT with unlabeled data. DEST is similar to self-training in that it iteratively generates new labeled data to improve the classifier. The key difference is that DEST combines the classifier's predictions with information from local discourse contexts to robustly assign labels to new training instances. The key to this

approach is to exploit unlabeled event phrases that occur near coreferent sentiment expressions. Specifically, we extract event phrases that are followed by a sentiment expression in a syntactic structure that suggests it likely refers to the event. For example, consider the statements below:

a) *I got engaged today. It is exciting.*

b) *I got divorced. This is terrible.*

In Example a), "*it*" co-refers with the act of getting engaged, so the positive sentiment of "*exciting*" applies to that event. In Example b), "*this*" co-refers with the divorce event, so the negative sentiment of "*terrible*" can be propagated to it. Our algorithm then predicts the affective polarity for unlabeled events using both the classifier's prediction for the event phrase as well as the associated sentiment expressions. We show that Discourse-Enhanced Self-Training improves both recall and precision for affective event classification.

4.1 Harvesting Events with Coreferent Sentiment Expressions

The key idea behind our approach is to create a self-training method that uses not only the classifier's own predictions but also a secondary source of information derived from local discourse contexts. Intuitively, the secondary signal confirms the classifier's prediction when they agree, or creates doubt about the classifier's prediction when they disagree. By taking both signals into account, we can assign high-quality labels to a diverse set of new examples in each cycle, which creates a more robust self-training process.

From this point on, we turn our attention to Twitter because it is a vast resource that we can query to acquire a large set of event phrases in specific contexts, and where people share their everyday experiences. We acquire our unlabeled data by searching for event phrases on Twitter that occur with coreferent sentiment expressions. We use a heuristic to identify sentiment expressions that are likely to refer to an event in the preceding sentence. Specifically, we look for sentiment expressions that begin a sentence and match one of the two forms shown in Table 4.1.

Table 4.1: Syntactic patterns of coreferent sentiment expressions.

1. **{this/that/it/I}, {be/feel/seem}, {ADJ+}**
2. **{this/that/it}, {be/feel/seem}, {ADJ* N+}**

In the patterns, the head adjective (ADJ) or head noun (N) is a sentiment term with positive or negative polarity. The sentiment expression cannot be followed by any events in its sentence and must follow a sentence that contains at least one event. Given these restrictions, the pronouns "*this*," "*that*," and "*it*" are likely referring to an event in the previous sentence, although this is not guaranteed. Similarly, the pronoun "*I*" is referring to the speaker who is likely expressing their sentiment toward something that was just mentioned, which is often (though not always) the prior event. We will call the phrases that match these patterns **coreferent sentiment expressions** because they express a sentiment that refers back to something mentioned earlier. Examples of coreferent sentiment expressions include "*this is great*," "*I felt terrible*" and "*It is great news*."

We found that the two syntactic constructions listed in Table 4.1 typically convey a sentiment about an event in the prior sentence, but this heuristic is not perfect. For example, the sentiment sometimes applies to an object in the prior sentence and not an action. One example is the statement: *"I bought a book. It is excellent,"* which describes an excellent book and not an excellent buying experience. Nevertheless, the self-training algorithm will use this data in the aggregate, so some noise can be tolerated. In the following sections, we describe each step of the Twitter data harvesting process.

4.1.1 Creating Sentiment Queries

We create an initial set of sentiment queries for Twitter by instantiating the syntactic patterns shown in Table 4.1 with 3,010 subjective adjectives and 2,023 nouns from the MPQA lexicon [208]. We also use the 1,147 words labeled with "*anypos*" in MPQA as an adjective and a noun to instantiate the patterns. For example, given the adjective *"good,"* we exhaustively generate all phrases that match the regular expression: "{that/this/it/I} {be/feel/seem} *good*," such as *"That is good"* and *"I feel good."*

We then download tweets that contain these phrases. If the context around the sentiment expression satisfies the constraints mentioned earlier, then we extract the events in the previous sentence as affective event candidates. Table 4.2 shows three tweets that were retrieved with queries for the sentiment expressions in *italics* along with the events extracted from each tweet.

Table 4.2: Examples of harvested tweets and extracted events.

Tweet1:	I rode a horse today! *That was fun*.
Events:	⟨I, ride, horse, -⟩
Tweet2:	Someone was abducted on the street right next to mine. *It's terrifying*.
Events:	⟨-, abduct, someone, on street⟩
Tweet3:	Disrupting my daily routine and alienating many people. *I am angry* !
Events:	⟨-, disrupt, my daily routine, -⟩, ⟨-, alienate, people, -, -⟩
Tweet4:	Children are separated from their parents. *It is crime!*
Events:	⟨-, separate, children, -, from their parent⟩
Tweet5:	We did a drink Friday together! *That was a blast!*
Events:	⟨we, do, drink, -⟩

4.1.2 Creating Event Queries

Next we can use the extracted events to harvest more tweets with coreferent sentiment expressions. Searching for phrases that match an event is not trivial. The Twitter API only supports exact phrase matching but an event is represented as a tuple (⟨Agent, Predicate, Theme, PP⟩). Furthermore, the components in an event tuple contain lemmatized head words. We want to construct queries that will retrieve phrases containing morphological variations (e.g., "*drove*" for the lemma "*drive*") as well as modifiers preceding heads (e.g., "*a fancy car*" instead of just "*car*"). To circumvent this problem, we generate text spans for each event tuple from the original tweets that the event was extracted from. The text span contains all words between the leftmost word and the rightmost word of the tuple. Then we apply the PrefixSpan algorithm [167] to compute the frequency of all subsequences of words. For each event tuple, we create queries from the 20 most frequent subsequences that contain all words in the event tuple. For example, ⟨he, drive, car⟩ might yield queries such as *"he drove a fancy car," "he has driven my car,"* etc.

After we retrieve tweets that match an event query, we apply the same constraints as before but in reverse: the sentence that mentions the event must be followed by a coreferent sentiment expression matching our patterns. In this step, we assume that unknown terms in the ADJ or N position of the patterns are sentiment-bearing, allowing us to identify

new sentiment expressions. We found this heuristic to be quite good and produce some interesting affective terms that are not in the MPQA lexicon. For example, the new negative terms include *"cyberbullying," "yucky"* and *"gutless"*, and the new positive terms include *"record-breaking," "reassuring"* and *"heart-warming."* Table 4.3 shows four tweets that were retrieved with queries for the events (searched event) along with the new sentiment terms in *italics*.

4.1.3 Iteratively Harvesting Events

The first step of data harvesting creates sentiment queries from the MPQA lexicon and extracts new event phrases. The second step of data harvesting creates event queries and extracts new sentiment phrases. Given these building blocks, we create a cycle that alternates these steps, iteratively harvesting new events with associated sentiment expressions. In each iteration, we form queries for sentiment or event phrases that have frequency ≥ 5 and have not been used as queries previously. We download 5,000 tweets for each event query and 1,000 tweets for sentiment expression query. Many tweets retrieved by event queries contain no coreferent sentiment expression, so we downloaded more tweets for event queries to increase the number of matched instances. Finally, we discard retweets and duplicated tweets. In this work, we treat a tweet as duplicated if it shares 6 or more consecutive words with another tweet. To be consistent with the criteria used for affective events in the AEKB [47], we also discarded events that did not contain a first-person

Table 4.3: Examples of harvested tweets and new sentiment terms (in italics).

Searched Event:	⟨nothing, be, right, -⟩
Tweet:	It's been one of those days where nothing is right. I feel *yucky*.
Searched Event:	⟨-, cancel, my flight, without notice⟩
Tweet:	My flight's been canceled without any notice. This is *bullshit*.
Searched Event:	⟨-, hit, 100kg, on bench press⟩
Tweet:	Just hit 100kg on the bench press! That is *record-breaking*!
Searched Event:	⟨I, watch, sunrise, from summit⟩
Tweet:	I just watched the sunrise from the summit... It's *breath-taking*!

reference or a family member term. While prior work [47] also discarded events that only mentioned other people, we did not apply this constraint due to the difficulty of recognizing people terms in tweets. We ran the harvesting process over Twitter for 4 iterations, after which few new events were found. The final dataset contains 2,068,600 unique event tuples and 15,494 unique sentiment expressions. Many of these events are potentially affective, but their polarities will be determined in later steps.

4.2 Discourse-Enhanced Self-Training

We designed an enhanced self-training algorithm that learns from unlabeled data by iteratively labeling new instances using both the affective event classifier's prediction as well as polarities associated with the event's discourse contexts. We will refer to this method as *Discourse-Enhanced Self-Training*. Figure 4.1 illustrates how an unlabeled event is scored during Discourse-Enhanced Self-Training.

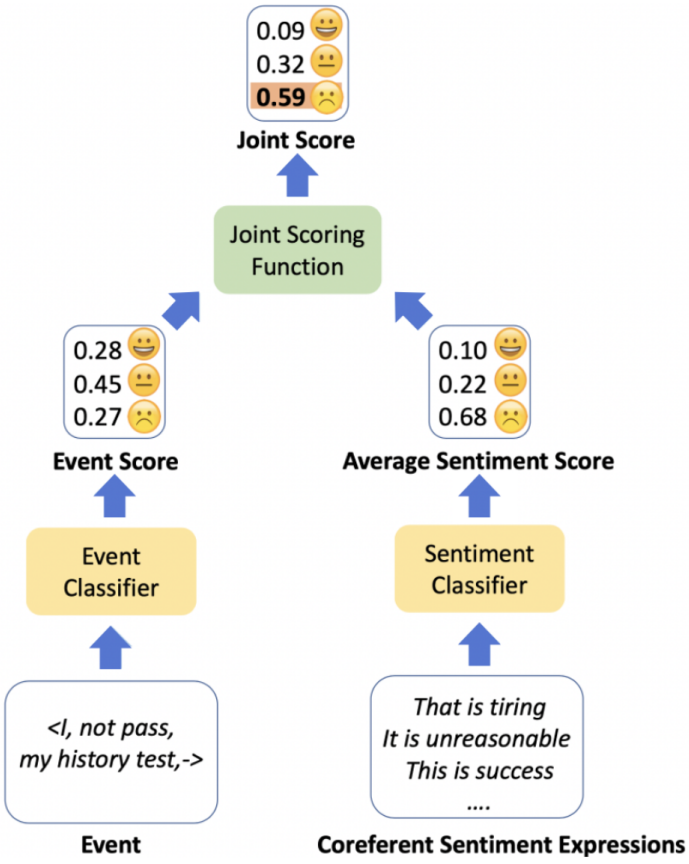

Figure 4.1: Illustration of affective polarity scoring in Discourse-Enhanced Self-Training.

Specifically, each event is paired with the set of coreferent sentiment expressions that occurred with it in our Twitter dataset. For example, the event "$\langle$*I, not pass, my history test,* -$\rangle$" is extracted from tweets such as:

1. *Can't believe I didn't pass my history test AGAIN! That is tiring!*
2. *Ugh, I didn't pass my history test! It is unreasonable! You don't how hard they made it.*
3. *Just found out I didn't pass my history test. This is success in the making, right?*

So the input to the joint scoring function contains the event and also coreferent sentiment expressions such as "*That is tiring*," "*It is unreasonable* " and "*This is success*" and so on.

The affective event classifier is applied to the event and generates a probability distribution over the three polarity values. In parallel, an external sentiment classifier produces a probability distribution over the polarity classes for each of the coreferent sentiment expressions. The probability distributions are then averaged to produce an average probability distribution for the set of sentiment expressions as a whole. Finally, a joint scoring function takes the two probability distributions and produces a joint probability distribution for the event. The polarity with the highest probability is used as the event's label.

Algorithm 1 outlines the procedure in detail. The process begins with a gold labeled set of events E_L, a set of unlabeled events E_U where each event e_i in E_U is paired with a set of coreferent sentiment expressions CSE_i, an external sentiment classifier, and two confidence thresholds θ_{jnt} and θ_{neu}. Each iteration starts by training the event classifier on E_L. The event classifier is then applied to every unlabeled event e_i in E_U to produce an event score vector s_{e_i}. Next, the sentiment classifier is applied to every coreferent sentiment expression *cse* in CSE_i to produce a polarity distribution. Then the polarity distributions of all *cse* in CSE_i are averaged to produce an average polarity distribution $\bar{s}_{\text{CSE}_i}$ for the whole set CSE_i. The joint scoring function then produces a joint score vector s_{jnt_i} for the event e_i by the equation below:

$$s_{jnt_i} = \frac{s_{e_i} \odot \bar{s}_{\text{CSE}_i}}{s_{e_i} \cdot \bar{s}_{\text{CSE}_i}}, \tag{4.1}$$

where $\odot$ denotes element-wise multiplication and $\cdot$ denotes dot product. Conceptually the joint scoring function gives equal weight to the event classifier and the sentiment classifier in the final decision of the label. Finally, each event e_i is assigned the polarity with the highest value in s_{jnt_i}.

Algorithm 1: Discourse-Enhanced Self-Training

Input:

E_L	A set of labeled events
E_U	A set of unlabeled Events, where each event e_i has an associated set of coreferent sentiment expressions CSE_i
θ_{jnt}	Confidence threshold for joint polarity scoring
θ_{neu}	Confidence threshold for neutral polarity labels
SC	An external sentiment classifier
AEC	The affective event classifier being trained

while *E_U is not empty and not maximum iteration* **do**
  Train the affective event classifier AEC over E_L
  For each $e_i \in E_U$, apply AEC to get an event score
  For each $e_i \in E_U$, apply the sentiment classifier SC to each $cse \in CSE_i$ and compute the average cse sentiment score
  Compute the joint score for each $e_i \in E_U$ by Eqn. 4.1
  Label new events (E_{jnt}) based on the joint scores and θ_{jnt}
  Label additional neutral events (E_{neu}) based on the event scores and θ_{neu}
  Update E_L and E_U:
    $E_L = E_L \cup E_{jnt} \cup E_{neu}$
    $E_U = E_U - E_{jnt} - E_{neu}$
end

We generate a set of new labeled events E_{jnt} by assigning labels to unlabeled events that have a polarity probability $\geq \theta_{jnt}$ based on the joint scores. All other events remain unlabeled. However, we found that this process labels relatively few events as neutral. This is because many stereotypically neutral events can be described with positive and negative *contexual* polarities and so followed by positive and negative sentiment expressions. For example, the event "*I read a book*" can be described as a positive event in certain context and co-occur with positive sentiment expressions (e.g., "*I read a book this afternoon. This was relaxing.*"). It can also be described as a negative event in certain context and be followed by negative sentiment expressions (e.g., "*I read a book this morning. I am feeling tired now.*"). As neutral events can co-occur with positive and negative sentiment expressions, the percentage of coreferent neutral sentiment expressions becomes lower, resulting in relatively low neutral scores. To better maintain the distribution of events over all three polarities, we also add a new set of events E_{neu}, which the event classifier predicts as neutral with confidence $\geq \theta_{neu}$.

Overall, DEST is similar to self-training in that a model learns from its own predictions

over the unlabeled data. The key difference is that DEST estimates the polarity label for an event based on two sources of information: 1) the model's prediction for the event, and 2) the average sentiment score of the coreferent sentiment expressions of the event. The use of two sources of information makes DEST have certain advantages over self-training that relies on only one source of information (the model's prediction). First, the label generated by DEST is more precise than the label generated by self-training when the two sources of information agree. Second, DEST could introduce a more diverse set of newly labeled instances than self-training. Typically, the newly labeled instances introduced by self-training are the instances with the highest model confidence. These instances are usually similar to the training data with which the model is trained, and so they provide limited diversity and extra information. On the other hand, DEST could introduce instances that are more diverse and informative for the model. This is because data that the model is not very confident about could be assigned high confidence scores and used as new training data in DEST, due to the effect of the secondary informaton. Consider an event of which the model prediction score is (Positive = 80%, Negative = 10%, Neutral = 10%) and the average sentiment score of the coreferent sentiment expressions is (Positive = 70%, Negative = 10%, Neutral = 20%). Given a confidence threshold of 90%, self-training will not include this event in the new training data. DEST, on the other hand, will include it as a new positive event in the training set, as the joint score is (Positive = 94.9%, Negative = 1.7%, Neutral = 3.4%).

Discourse-Enhanced Self-Training needs an external sentiment classifier, so we fined-tuned a BERT-based model with the gold standard Twitter dataset from SemEval-2017 [162] following the experiment setups in Section 3.3 and Section 3.4. In our experiments, we set θ_{neu} to 0.9 and θ_{jnt} to 0.95 based on the model's performance over the validation set.

4.3 Gold Dataset Creation

We created a gold standard dataset for affective events from Twitter (**Twitter Dataset**) by having two human annotators label 1,500 events that are randomly selected from the harvested events described in Section 4.1.2 and have a frequency ≥ 5. Each event was labeled as positive, negative, or neutral using the same criteria defined by prior work [47] for the AEKB. The pairwise inter-annotator agreement using Cohen's kappa was .75. The

two annotators then adjudicated their disagreements to produce the final set of gold labels. The final dataset contains 435 (29%) positive, 348 (23%) negative and 717 (48%) neutral events. This new evaluation dataset and the collection of the unlabeled harvested events are publicly available at https://github.com/yyzhuang1991/DEST.

4.4 Experimental Results

We performed 10-fold cross validation over the gold Twitter Test set, where each of the 10 runs used 80% of the data (8 folds) for training, 10% of the data (1 fold) for validation/tuning, and 10% of the data (1 fold) for testing. We compare DEST with a strictly supervised learning model and a traditional self-training model. During each iteration of the self-training, the affective event classifier Aff-BERT is applied to each unlabeled event. Events with polarity score ≥ 0.9 are selected as new labeled data. We chose 0.9 as the threshold based on the model's performance on the validation set.

To ensure a rich set of discourse contexts, we only used unlabeled events that (1) had at least 10 distinct coreferent sentiment expressions and (2) did not include "*this*," "*that*" or "*it*" as a subject or object of the event phrase because the event is often vague without knowing what the pronoun refers to. This resulted in 8,532 events in the unlabeled set.

4.4.1 Results

Table 4.4 reports the performance of three classification models after 10 iterations of learning with unlabeled data.[1] The first row shows the results for Aff-BERT trained only with gold labeled data, for comparison. Row 2 shows the results for Aff-BERT with self-training and Row 3 shows the results for Aff-BERT with DEST. From the table we could see that ordinary self-training produced small gains in both precision (76.5% $\rightarrow$ 77.6%) and recall (75.2% $\rightarrow$ 77.2%) as compared to the supervised model with only gold data. On the other hand, our Discourse-Enhanced Self-Training algorithm performed better. It improved precision over the supervised model from 76.5% $\rightarrow$ to 79.6% and improved recall from 75.2% $\rightarrow$ 78.7%, resulting in a gain of 3.3 absolute points in F1 score (75.7% $\rightarrow$ 79.0%). It also outperformed self-training substantially by 2 absolute points in precision and 2 absolute points in F1 score.

[1]For both self-training models, no new examples were labeled after 10 iterations.

Table 4.4: Results for learning from unlabeled data.

Method	Precision	Recall	F1
Supervised	76.5	75.2	75.7
Self-training	77.6	77.2	77.0
DEST	**79.6**	**78.7**	**79.0**

Figure 4.2 shows the learning curves for each method over the 10 iterations based on their F1 score. The flat line is the F1 score for Aff-BERT trained with only gold labeled data. Self-training produced its highest F1 score after the first iteration, then declined and stayed relatively stable without further improvement. In contrast, the learning curve of Discourse-Enhanced Self-Training gradually ascends, reaching its peak in iteration 7 and showing signs that it could potentially exceed that peak with more unlabeled data. Discourse-Enhanced Self-Training produces more robust learning from unlabeled data, and this general approach could be applied to many other problems that have a secondary source of information relevant to the task.

Table 4.5 shows the performance breakdown across the three polarities. Discourse-Enhanced Self-Training improved both precision and recall for all polarities, except the precision was slightly lower for negative polarity. Most notably, it achieved a 6.0% gain in

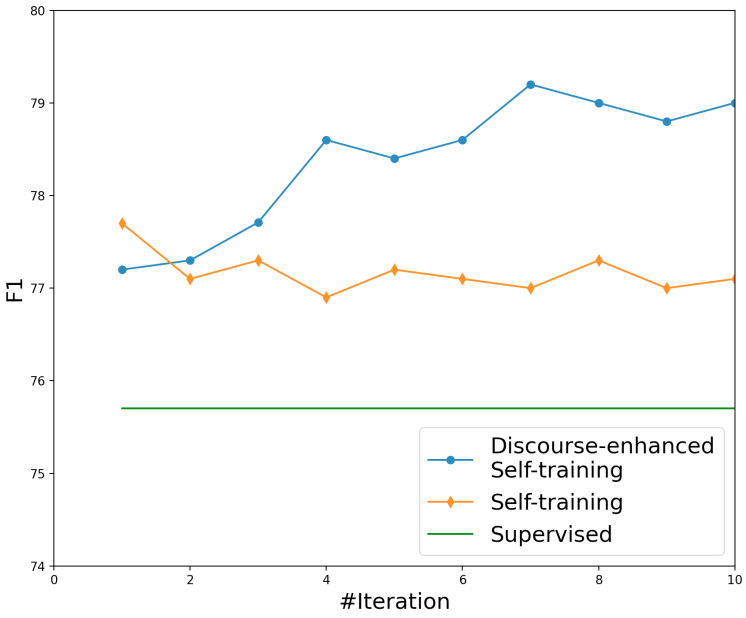

Figure 4.2: Learning curves through 10 iterations.

Table 4.5: Recall and precision across polarities.

Method	POS		NEG		NEU	
	Precision	Recall	Precision	Recall	Precision	Recall
Supervised	74.4	71.5	**79.0**	74.0	76.1	80.1
DEST	**81.8**	**74.8**	78.4	**80.0**	**79.4**	**82.4**

recall for negative polarity, and gained 3.3% of recall for positive polarity alongside a 7.4% gain in precision.

We also generated learning curves for the supervised learner and Discourse-Enhanced Self-Training when trained over different amounts of labeled data. Figure 4.3 shows results when using 50% to 100% of the gold training data in increments of 10%. Discourse-Enhanced Self-Training showed even greater relative improvement over the supervised learner when only 50% of the gold data was used for training. In addition, when using about 60% of the gold data, it achieved performance comparable to the supervised learner trained with 100% of the data.

4.5 Conclusion

This chapter introduces a novel semi-supervised algorithm, Discourse-Enhanced Self-Training (DEST), to improve affective event classification models. DEST is similar to tra-

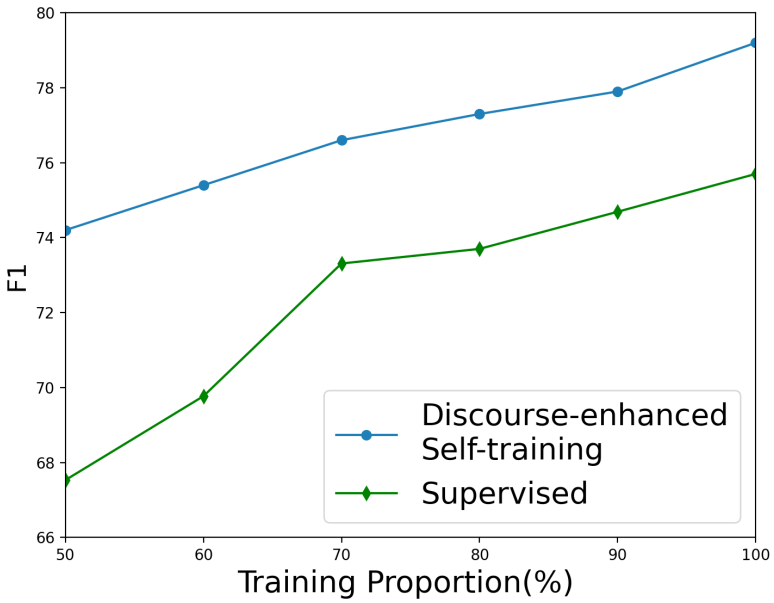

Figure 4.3: Learning curves of models with training sets of different sizes.

ditional self-training as they both leverage unlabeled data and model predictions to iteratively improve the classification model. The key difference is that DEST leverages not only the classification model's predictions but also a secondary source of information to assign labels to unlabeled data. Specifically, DEST combines both the affective event classifier's prediction and polarities of the coreferent sentiment expressions to generate polarity labels for unlabeled events. Our experiments show that DEST can substantially improve upon the supervised learning results. The resulting classification model is substantially more effective for affective event recognition than previous methods. We also believe that the general idea behind our enhanced self-training approach could be useful for many other types of problems where a secondary source of information can be acquired.

CHAPTER 5

IMPROVING AFFECTIVE EVENT RECOGNITION BY MULTIPLE VIEW CO-PROMPTING

In Chapter 4, we showed that the performance of affective event classification models is limited by the amount of gold training data and that it is promising to improve model performance by generating more training data. While successful, prior methods [166, 225] generated training data by mining events from text corpora and assigning polarity labels with weakly supervised methods. However, generating labeled events using text corpora could be challenging in practice. One major challenge is that mining data by going through a large text corpus could be inefficient, as only a small percentage of the text corpus may be relevant. Secondly, extracting data from text corpora could be limited by the computational bottleneck of applying a pipeline of NLP tools to a large text collection and by the brittleness of lexical pattern matching. Given these practical challenges, we explore the following research question in this chapter: *Could we generate automatically labeled affective events without using text corpora?*

To this end, we propose a simpler but more effective method to generate affective events by prompting large language models. As the method relies on only language models, there is no need for text corpora. Specifically, we use one language model prompt to generate affective event candidates, and we introduce a *Co-Prompting* method to automatically label these event candidates with affective polarity. The key idea behind *Co-Prompting* is to design two complementary prompts that capture independent views of an event, reminiscent of co-training [21]. Combining information from two different views of an event produces labels that are more accurate than the labels assigned by either one alone.

Specifically, we acquire affective events in a two-step process: (1) Event Generation, and (2) Polarity Labeling. The first step generates events that are associated with a set of gold

"seed" affective events. For each seed event, we prompt a language model to generate sentences where the seed event co-occurs with some new events. Our hypothesis is that affective events are often preceded or followed by other affective events that are causally or temporally related. For example, if someone breaks his/her leg, a prior event might describe how it happened (e.g., *"fell off a ladder"* or *"hit by a car"*) and a subsequent event might describe the consequences (e.g., *"could not walk"* or *"rushed to the hospital"*).

The second step collects independent views of the polarity for each new event using two complementary language model prompts. One prompt provides an *Associated Event View*, which considers the polarities of the known (labeled) events that co-occur with the new event during Event Generation. The second prompt provides an *Emotion View*, which considers the polarity of the most probable emotion words generated by a language model when prompted with the new event. Finally, we combine information from the two co-prompts to assign an affective polarity label to each new event.

Our experiments show that using these automatically acquired affective events as additional training data for an affective event classifier produces state-of-the-art performance over two benchmark datasets for this task. In addition, the analysis confirms that our co-prompting method utilizing multiple views yields more accurate polarity labels than using either view alone.

5.1 Acquiring Affective Events with Multiple View Co-Prompting

Our research aims to automatically generate labeled affective events to improve classifiers because gold data for affective event classification is only available in limited quantities. Automated methods for data generation offer a cost-effective and practical solution for improving the performance of affective event classifiers, and also could be used to rapidly acquire training data for new domains or text genres. Different from previous methods that perform pattern matching on a large-scale corpus, our method is able to generate high-quality labeled data by only prompting language models. Our method is more practical and also yields better classification performance.

Figure 5.1 shows the flowchart for our approach. The process begins with a modest amount of "seed" data consisting of gold labeled affective events provided as input. The

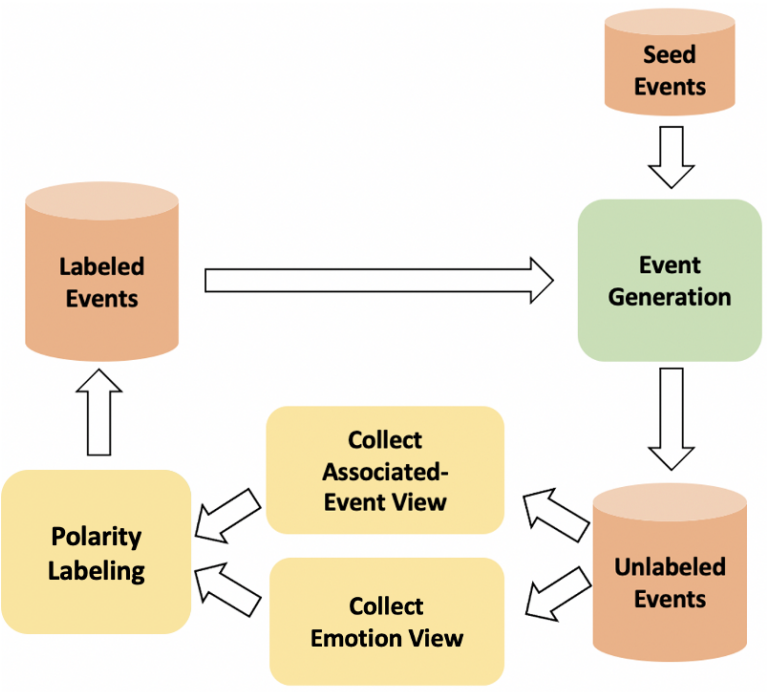

Figure 5.1: Flowchart for acquiring affective events with Multiple View Co-Prompting.

first step (**Event Generation**) uses a language model prompt to elicit events that are associated with each seed event. The second step (**Polarity Labeling**) assigns a polarity label to each new event using *Co-Prompting* to assess polarity from two independent views of the event. Given an event e, the *Associated Event View* considers the affective polarities of labeled events that co-occur with e during Event Generation. The *Emotion View* considers the affective polarities of emotion words that are generated by an Emotion Prompt given the event e. Polarity scores produced from these views are then combined to assign an affective polarity label to the event e.

This process repeats in an iterative fashion, where the newly labeled events are used to discover more affective events in the next cycle. The process ends when no new events are generated or a maximum number of iterations is reached.

5.1.1 Event Generation

The Event Generation process begins with a set of gold affective events and produces a set of new events, many of which we expect to be affective. For each seed event, we create an **Associated Event Prompt** of the following form:

Here are the {*POLARITY*} things that happened to me today: {*EVENT*},

where **{EVENT}** is a placeholder filled by the seed event phrase, and **{POLARITY}** is a placeholder filled by the affective polarity of the seed event. We find that this design in practice can guide a generative language model to complete the sentence by enumerating other events that are likely to co-occur with the given event on the same day. Intuitively, the enumeration behavior is encouraged by the colon ":" and comma. The temporal relation is encouraged by the word "*today*." The polarity placeholder, {POLARITY}, encourages the language model to generate events with the same affective polarity.

For the polarity terms, we used the word *"good"* for events with positive polarity and the word *"bad"* for events with negative polarity. For events with neutral polarity, we simply used an empty string (i.e., *"Here are the things..."*).[1] We expected that this prompt would generate some neutral events, but that it would produce positive and negative events too because people tend to recount events that are interesting or impactful, not boring and mundane. In fact, we do not expect any of these prompts to be perfect. Our goal at this stage is to generate a healthy mix of new events across all three affective polarities (positive, negative, and neutral). The affective polarity for each new event will ultimately be determined later in the Polarity Labeling step. To be consistent with prior work on this topic, we represent each event expression as a 4-tuple of the form: ⟨Agent, Predicate, Theme, Prepositional Phrase (PP)⟩. To create an event phrase for the language model prompt, we concatenate the words in the tuple. Below we show three example prompts initialized with the positive event ⟨*I, get, -, in college*⟩, the negative event ⟨*I, cut, my leg, -*⟩ and the neutral event ⟨*I, walk, -, in class* ⟩:

1. *Here are the good things that happened to me today: I get in college,*
2. *Here are the bad things that happened to me today: I cut my leg,*
3. *Here are the things that happened to me today: I walk in class,*

Note that the resulted event phrase may not be grammatically correct, but our observation is that this did not cause serious problems for the language model.

We used open-source GPT-2_{LARGE} [157] as the generative language model.[2] To obtain

[1] We found that using some neutral words (e.g., "neutral") in the placeholder did not work better.

[2] Code available at https://github.com/openai/gpt-2

diverse outputs, we let GPT-2 generate 200 sentences for each labeled event.[3] For the sampling method, we used nucleus sampling [77] with 0.9 as the top-p threshold, beam search with a beam size of 5 and a temperature of 2.0 to encourage diverse generation. We extracted new events from the sampled sentences to create event tuples, following the same conventions as earlier work [47, 225]. For the sake of robustness, we selected the events that occurred with at least 3 distinct seed events as new events for polarity labeling.

To illustrate, one example sentence generated for the event ⟨*my house, burn down*, -, -⟩ is: "*..., my mom passed away and my family lost everything*." And the events extracted are ⟨*my mom, pass away*, -, -⟩ and ⟨*my family, lose, everything*, -⟩. We show more examples of extracted events in Table 5.1. Overall, the generated events are usually related to the seed event in some way and typically have the same affective polarity (e.g., "*I cut my leg*" → {"*I fall off my bicycle,*" "*I hurt my knee,*" ...}), despite some exceptions (e.g., "*they take my dog*"). For our purposes, it is perfectly fine that some generated events are loosely associated with the seed events, because our goal is simply to harvest new affective events, and their precise relationship to the seed events is irrelevant.

5.1.2 Polarity Labeling with Multiple Views

The next step is to assign affective polarity labels to each new event. We collect affective information from two prompts that provide independent views of an event: (1) we collect affective polarity information from the events generated by the *Associated Event Prompt*, and (2) we use an *Emotion Prompt* to generate emotion terms associated with an event. Finally, we combine the information gathered from these two prompts to assign a polarity label.

5.1.2.1 Emotion Prompting

To acquire another source of information about the affective polarity of an event, we prompt a language model to produce emotion terms with associated probabilities for each event. We design a cloze expression to generate emotion terms associated with an event expression by prompting the masked language model $\text{BERT}_{\text{LARGE}}$. Specifically, we use the following **Emotion Prompt**: **[*EVENT*]. I feel [MASK].**

[3] We discarded samples that did not end with a period, since they are usually incomplete sentences.

Table 5.1: Examples of events generated by the Associated Event Prompt for seed events.

Polarity	Seed Event	Events Generated by Associated Event Prompt
NEG	I cut my leg	I fall off my bicycle, I hurt my knee, I wake up at hospital, I break my rib, I faint, kick me in head, they take my dog, my eye start to water, I break my ankle, I get in car accident
	I not get refund	they take my money, kick me out of game, this happen, freeze my account for hour, I lose money, I get refund, I get angry, ban me, make decision, I get email
	I lose my job	I break up with my girlfriend, I not apply, kick me out of house, arrest me, I go to find out, they try to kill me, I find job, eat my lunch, dump me, I break down
NEU	I walk in class	I start to talk to people, I take seat, I reply, professor tell me, my friend ask me, I take moment, I shake hand, I learn, I sit in front row, I have to explain, I want to tell story
	I close account	I call customer service, message say, I click on link, ban me for day, email tell me, I go, this show me, receive phone call, delete me, I call bank
	I meet someone	I get call from them, I get my drink, I lose weight, I chat for minute, I say something stupid, person tell me, I start to talk, I talk for long time, they invite me, they respect me
POS	I get in college	convince myself, I graduate, I go, I read them, drink coffee, I meet cool people, watch tv, I learn lot about myself, I move, I find good job
	I play match	my team win game, I lose, I go to hotel, I work, I go, I go on stage, I get score, I go to bed, play video game, I get point
	I get house	I pay my tax, I move out of my apartment, I eat my favorite food, I get new job, I start to live, I learn, I pay bill, I care, I afford to eat, I start to look

Specifically, The *[MASK]* token is a special token used by BERT to represent a blank, and will be filled by a predicted token. The word "feel" leads the language model to return words that refer to emotions or other sentiments for the masked token. We expect that positive events will typically be followed by positive emotions, and negative events by negative emotions. For neutral events, we expect to see a mix of both positive and negative emotions because these events can occur in a wide variety of contexts. We used $\text{BERT}_{\text{LARGE}}$ [44] as the masked language model.[4] We store all generated terms and their probabilities produced by BERT for later use.

[4]We also experimented with using GPT-2 to generate emotions, but it was less effective and often produced sentences rather than emotion words, such as "*I break my arm. I feel like this is a real thing.*"

Figure 5.2 illustrates this process for two example events. The top shows the four most probable terms generated from the event tuple <*I, graduate, -, ->*, all of which have positive polarity. The bottom shows the four most probable terms generated from the event tuple <*I, break, my leg, ->*. Three of these terms have negative polarity, but the fourth term has neutral polarity. This example shows that the prompt can produce inconsistent results, but the probability distribution across all of the generated terms typically captures a fairly reliable signal.

5.1.2.2 Multiple View Polarity Scoring

We first define scoring functions to determine the most likely affective polarity for an event from each view independently. Then we present a joint scoring function that combines the scores from the two views to produce a final affective polarity label.

5.1.2.2.1 Associated Event View. This first view captures the degree to which an event co-occurs with labeled events of each polarity during the Event Generation step. Intuitively, we expect that events tend to co-occur with other events of the same polarity. According to this view, we define the *Associated Event Score* (S_A) of an unlabeled event e with respect to a polarity label l as:

$$S_A(l \mid e) = \frac{\sum_{e' \epsilon AEP(e)} I(e', l)}{\mid AEP(e) \mid} \tag{5.1}$$

where $AEP(e)$ is the set of labeled events that co-occur with e in the results produced by

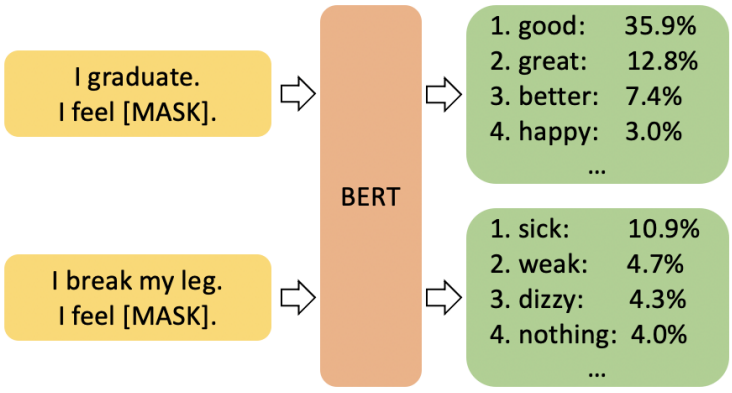

Figure 5.2: Examples of the Emotion Prompt.

the Associated Event Prompt, $I(e', l)$ is an indicator function with a value of 1 if the polarity label of e' is l or zero otherwise, and $|\cdot|$ is the cardinality. Note that a labeled event e' can co-occur with e if either (1) e is generated by the prompt when given e' as input, or (2) e' is generated alongside e by the original prompt and e' was previously labeled (as a seed or during learning).

5.1.2.2.2 Emotion View. This view captures the polarity of the emotion words generated by the Emotion Prompt. Based on this view, we define the *Emotion Score* (S_E) for an unlabeled event e with respect to a polarity label l as:

$$S_E(l \mid e) = \frac{\sum_{w \in D_l} P_{\text{BERT}}(w \mid EP(e))}{\sum_{l' \in L} \sum_{w \in D_{l'}} P_{\text{BERT}}(w \mid EP(e))} \tag{5.2}$$

where D is a gold dictionary of emotion terms, D_l is the subset of words in D that have polarity label l, and $P_{\text{BERT}}(w \mid EP(e))$ is the probability associated with word w produced by the Emotion Prompt (EP) given event e. In short, Eqn. 5.2 computes a polarity score for label l by summing the probabilities of all terms generated by the Emotion Prompt that occur in D with label l. For the gold dictionary D, we collect all of the adjectives and nouns in the MPQA subjectivity lexicon [208] along with their polarity labels.

5.1.2.2.3 Polarity Assignment. We conservatively assign positive and negative polarities to an event only when both S_A and S_E predict the same polarity. Formally, we label an event e with polarity l when both scores for l exceed a confidence threshold θ as follows:

- if $S_A(pos \mid e) \geq \theta$ and $S_E(pos \mid e) \geq \theta$, then e is positive.
- if $S_A(neg \mid e) \geq \theta$ and $S_E(neg \mid e) \geq \theta$, then e is negative.

where θ is a hyperparameter. Note that θ must be greater than 0.5 to avoid multiple label assignments to an event.

For the neutral polarity, we found that the emotion scores $S_E(neu \mid e)$ are low in most cases because the Emotion Prompt tends to generate emotional words even for neutral events. However, we observed that the Emotion Prompt is more likely to generate a mixed set of both positive and negative emotion words for neutral events, presumably because neutral events can occur in both types of contexts. Therefore we assign neutral polarity in a different manner, by looking for a small difference between the positive and negative emotion scores. Specifically, we consider an event e to be **neutral** based on both its neutral

Associated Event Score $S_A(neu \mid e)$ and the absolute difference between its positive and negative Emotion Scores, $S_E(pos \mid e)$ and $S_E(neg \mid e)$:

- if $S_A(neu \mid e) \geq \theta$ and $1 - |S_E(neg \mid e) - S_E(pos \mid e)| \geq \theta$, then e is neutral.

As an example, consider an event with $S_E(neg \mid e) = 0.5$ and $S_E(pos \mid e) = 0.4$, then $1 - |S_E(neg \mid e) - S_E(pos \mid e)| = 0.9$, which indicates that the event is very likely to be neutral. In our experiments, we set the θ value to be 0.9 based on the performance over the development set.

The outline of our approach is shown in Algorithm 2. In summary, our approach generates unlabeled events that are associated with labeled events by prompting GPT2 with the Associated Event Prompt, and then assigns polarities to the unlabeled events based on two different views extracted from language models.

5.2 Evaluation

5.2.1 Datasets

We conducted experiments over two previously used datasets for affective event classification: (1) the **BLOG** dataset constructed by Ding and Riloff [47], which contains 1,490 manually annotated events (20% Positive, 18% Negative and 62% Neutral) extracted from blog posts, and (2) the **TWITTER** dataset developed in our prior work [225] (described in Section 4.3), which contains 1,500 manually annotated events (29% Positive, 23% Negative and 48% Neutral) extracted from Twitter. We performed 10-fold cross-validation on each dataset (8 folds for training, 1 fold for development, and 1 fold for testing).

5.2.2 Generating Newly Labeled Events

To generate newly labeled events for each domain (TWITTER and BLOG), we used the training data as the seed events and ran the process for 15 and 10 iterations, respectively. We chose these stopping points because they produced around 10,000 new events for each domain, and we wanted to keep the number of new events manageable. Between iterations, we added the maximum number of newly labeled events that would maintain the original data distribution of affective polarities. Figure 5.3 shows the number of new events acquired for each iteration. Both curves start at around 1,200 because that is the size of the gold training sets used for seeding. This process ultimately produced (on average,

Algorithm 2: Labeling Data with Multiple View Co-Prompting

Input:

E_L	A set of labeled events
E_{U}	A set of unlabeled events, which is initially empty
E_{used}	A set of events that have been used for data harvesting, which is initially empty
D	A dictionary of emotion terms
θ	Confident threshold
GPT-2_{LARGE}	
$\text{BERT}_{\text{LARGE}}$	

Output: E_L

while *not maximum iteration* **do**

Construct a set of labeled events that have not been used for data harvesting: $E_{seed} = E_L - E_{used}$

For each $e \in E_{seed}$, generate the associated events by prompting GPT-2_{LARGE} with the Associated Event Prompt and store them in E_U.

For each $e \in E_U$, extract the Associated Event View information.

For each $e \in E_U$, extract the Emotion View information by prompting $\text{BERT}_{\text{LARGE}}$ with the Emotion Prompt and using D.

For each $e \in E_U$, assign polarities based on the Associated Event View, the Emotion View and the confidence threshold θ. Create E_{new} to store events that are assigned polarities.

Perform the updates below:

$$E_L = E_L \cup E_{new}$$
$$E_U = E_U - E_{new}$$
$$E_{used} = E_{used} \cup E_{seed}$$

end

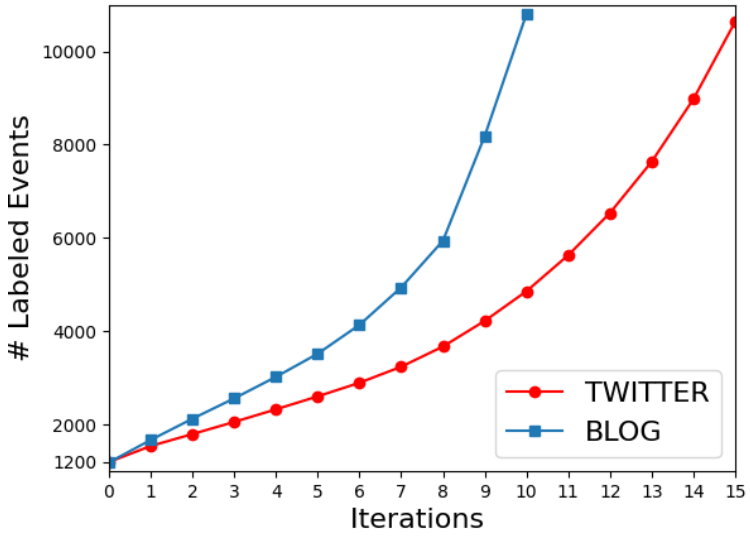

Figure 5.3: Newly labeled events generated across iterations.

across the folds in our cross-validation experiments): 10,636 new events for the TWITTER domain and 10,800 new events for the BLOG domain.

5.2.3 Affective Event Classification Model

We use Aff-BERT (described in Section 3.3) as our classification model, which is an uncased BERT-base model fine-tuned on our data that takes an event tuple as input (we concatenate all of the words into a phrase) and classifies the phrase with respect to three affective polarities (positive, negative, or neutral). We train Aff-BERT with a weighted cross-entropy function, which weights the gold and the new (weakly) labeled data differently: $L = L_G + \lambda L_W$, where L_G is the loss over the gold data, L_W is the loss over the weakly labeled data, and λ is a weight factor. During training, we performed a grid search over all combinations of learning rates (1e-5, 2e-5, 3e-5), epochs (5, 8, 10), batch sizes (32, 64), and λ values (0.1, 0.3, 0.5). We used the values that performed best over the development set.

5.2.4 Comparisons with Prior Work

We compared our method with several other approaches. Three methods were previously proposed by our prior work for affective event classification in Chapter 4: **1)** the Aff-BERT model trained only on gold data; **2)** Aff-BERT with self-training; **3)** Aff-BERT with Discourse-Enhanced Self-Training (DEST). The latter two methods improve Aff-BERT

by providing additional weakly labeled data. For self-training, Aff-BERT is applied to each unlabeled event during each iteration, and events with polarity score ≥ 0.9 are selected as new labeled data.[5] For DEST, we only evaluated it on the TWITTER dataset since it is specific to Twitter. We also evaluated two general-purpose methods for data augmentation: **4)** Back-translation [176], which generates paraphrases of an input phrase via machine translation, and **5)** pattern-exploiting-training (PET) [172], which trains an ensemble of language models with multiple prompts and weakly-labeled data. For Back-translation, we produced one paraphrase for each event phrase in the training set by translating the training event phrase from English to German and then from German back to English, using the wmt19-en-de and wmt19-de-en machine translation models [135]. We then paired the output paraphrase with the original event's polarity label. To train PET, we used $\text{BERT}_{BaseUncased}$ as the language model and used 3 prompts:

1. *"[EVENT]. I feel _."*
2. *"[EVENT]. I felt _."*
3. *"[EVENT]. It was _."*

For hyperparameters, we used 1e-5 as the learning rate, 4 as the batch size, and 5 as the number of training epochs. We selected these values using development data. Since PET requires unlabeled data, we used 20K events randomly collected from the AEKB lexicon produced by Ding and Riloff [47] for experiments with the BLOG data, and we used the 8,532 unlabeled events released in our prior work [225] for experiments with the TWITTER data. The AEKB data can be found at https://github.com/yyzhuang1991/AEKB and the unlabeled data for TWITTER can be found at https://github.com/yyzhuang1991/DEST.

5.2.5 Experimental Results

Tables 5.2 and 5.3 show our experimental results, including the precision and recall for each polarity as well as macro-averaged F1 scores. The *Aff-BERT* row shows the results when trained over only gold labeled data. The other models exploit weakly labeled data for additional training.

On the TWITTER data, Co-Prompting outperforms all other methods. We see a 5.6% absolute F1 score gain compared to Aff-BERT and a 2.3% gain compared to DEST, which

[5] We chose 0.9 as the threshold based on the model's performance on the validation set.

Table 5.2: Experimental results for TWITTER data.

Method	Macro F1	POS		NEG		NEU	
		Precision	Recall	Precision	Recall	Precision	Recall
Aff-BERT	75.7	74.4	71.5	79.0	74.0	76.1	80.1
Back-translation	76.4	80.4	69.2	79.2	75.1	75.3	83.4
Self-training	77.0	78.6	69.5	76.8	**82.3**	77.4	79.8
PET	78.3	78.1	75.6	78.2	81.6	79.2	79.1
DEST	79.0	81.8	74.8	78.4	80.0	79.4	82.4
Co-Prompting	**81.3**	**82.3**	**76.2**	**85.9**	79.7	**79.7**	**86.1**

Table 5.3: Experimental results for BLOG data.

Method	Macro F1	POS		NEG		NEU	
		Precision	Recall	Precision	Recall	Precision	Recall
Aff-BERT	77.4	71.7	66.2	78.2	77.2	85.0	87.4
Back-translation	77.9	79.6	66.1	75.5	74.3	85.3	90.0
PET	78.0	78.5	60.2	81.4	**76.5**	83.8	91.1
Self-training	78.6	76.3	68.3	78.6	76.2	**85.5**	89.0
Co-Prompting	**80.7**	**81.4**	**70.1**	**84.0**	75.3	85.4	**91.8**

is the strongest competitor. Most notably, we see a 3.7% recall gain over DEST for neutral polarity and a 7.5% precision gain for negative polarity.

On the BLOG data, Co-Prompting also consistently outperforms the other methods. It surpasses Aff-BERT by 3.3 absolute points in F1 score, and self-training (the closest competitor) by 2.1 absolute points. In addition, it achieves the highest precision for both positive and negative polarity.

5.2.6 Impact of Multiple Views

We conducted experiments on the TWITTER data to understand the contribution of each view for polarity labeling. First, we assessed the contribution of each prompt and its corresponding view with respect to polarity labeling. Table 5.4 shows the performance of models trained with events labeled by each view alone and by both of them together. The macro-averaged precision, recall and F1 scores across polarities are reported. The first row shows the performance of Aff-BERT with only gold data just for comparison. The next two rows show the results when using only the Associated Event View or the Emotion View for polarity labeling. The last row shows the performance of Co-Prompting. For experiments with only one view, we still used the Associated Event Prompt to generate

Table 5.4: Impact of multiple views on TWITTER data.

Method	Precision	Recall	F1
Aff-BERT	76.5	75.2	75.7
Emotion View	78.9	78.3	78.2
Associated Event View	79.4	78.9	78.8
Both (Co-Prompting)	82.6	80.7	81.3

new events, but assigned the polarity labels based on a single view (Associated Event View or Emotion View). Overall, each view performs well on its own and produces better classification models that outperform Aff-BERT. But Co-Prompting yields a substantially higher F1 score than either view on its own. Our observation is that the labels for neutral events are especially noisy without using both scoring functions.

Next, we investigated how and why the polarity labels change when incorporating both views. Table 5.5 shows the number of labels that are changed correctly or incorrectly when adding the second view. The table on the top shows labels produced by the Associated Event View (AEV) that are changed by Co-Prompting. For example, there are 19 good changes (wrong before, correct now) from neutral to negative (Neu $\rightarrow$ Neg) but 8 bad changes (correct before, wrong now). The Δ column shows the overall net gain in correct labels. Overall, Co-Prompting has the greatest impact by correctly changing neutral labels to be positive or negative. This makes sense because the Associated Event View sometimes had trouble recognizing affective polarity, but the Emotion View specifically tries to identify emotions for each event.

The table at the bottom of Table 5.5 shows labels produced by the Emotion View (EV) that are changed by Co-Prompting. Adding AEV has a big impact in the opposite direction: changing mislabeled negative or positive events to be neutral. Intuitively, this is because EV can be too aggressive about assigning positive and negative polarity and have difficulty recognizing neutral events. These results nicely illustrate the power of Co-Prompting: complementary views have different strengths and weaknesses, and the strengths of one view can compensate for the weaknesses of the other. And more generally, Table 5.5 shows that most of the label changes produced by Co-Prompting were more accurate than the labels produced by one view alone, demonstrating that Co-Prompting with complementary views adds robustness.

Table 5.5: Counts of labels changed by Co-Prompting. Δ: Correct - Incorrect. AEV: Associated Event View. EV: Emotion View.

AEV → Co-Prompting	**Correct**	**Incorrect**	Δ
Neu → Neg	19	8	11
Neu → Pos	24	17	7
Pos → Neu	33	28	5
Neg → Neu	18	13	5
Pos → Neg	3	2	1
Neg → Pos	2	5	-3

EV → Co-Prompting	**Correct**	**Incorrect**	Δ
Neu → Neg	23	7	16
Neg → Neu	29	14	15
Pos → Neu	38	24	14
Neg → Pos	4	3	1
Pos → Neg	0	1	-1
Neu → Pos	22	23	-1

5.2.7 Manual Analysis of Polarity Labels

To directly assess the accuracy of the polarity labels assigned by Co-Prompting for the newly generated events, we asked two people to annotate 200 randomly sampled events from TWITTER. The annotation followed the same annotation guidelines used to create the TWITTER and BLOG datasets as defined by [47]. The pairwise inter-annotator agreement was 89.5% using Cohen's kappa. The annotators then adjudicated their disagreements.

Table 5.6 shows the accuracy of the labels produced by each view alone and by Co-Prompting (Both). The overall accuracy is only 83%-84% for the labels produced by each view alone but 91% for the labels produced by both views. The Associated Event View is most accurate for neutral labels, whereas the Emotion View is most accurate for positive and negative labels. These results again confirm the value of complementary sources of information for labeling data.

Table 5.6: Manual analysis of polarity labels.

Polarity	**AEV**	**EV**	**Both**
POS	50/62 (80.6%)	50/58 (86.2%)	62/68 (**91.2%**)
NEG	33/40 (82.5%)	43/49 (87.8%)	33/35 (**94.3%**)
NEU	85/98 (86.7%)	73/93 (78.5%)	87/97 (**89.7%**)
Overall	168/200 (84.0%)	166/200 (83.0%)	182/200 (**91.0%**)

5.2.8 Learning Curves

We produced learning curves to understand the behavior of training with different amounts of data on the TWITTER domain. Figure 5.4 plots the F1 scores of Co-Prompting when re-training the classification model with the data generated after every 3 iterations.

The dashed line shows the F1 score of Aff-BERT (using only gold data) for comparison. The F1 score of Co-Prompting rises steeply after the first 3 iterations, and continues to improve across later iterations. This graph suggests that running the iterative process even longer could yield further benefits.

We also investigated the effectiveness of our approach with smaller amounts of gold seed data. Figure 5.5 shows the performance of Co-Prompting on the TWITTER data when trained with subsets of the gold data ranging from 50% to 90%. For comparison, we also show the results for the two strongest competitors, DEST and PET, as well as the Aff-BERT baseline. Co-Prompting consistently outperforms the other approaches over all training set sizes. Surprisingly, Co-Prompting trained with only 50% of the gold data achieves the same level of performance as Aff-BERT using 100% of the gold data. This result demonstrates that generating labeled events with our co-prompting method can produce a high-quality classification model even with smaller amounts of gold seed data.

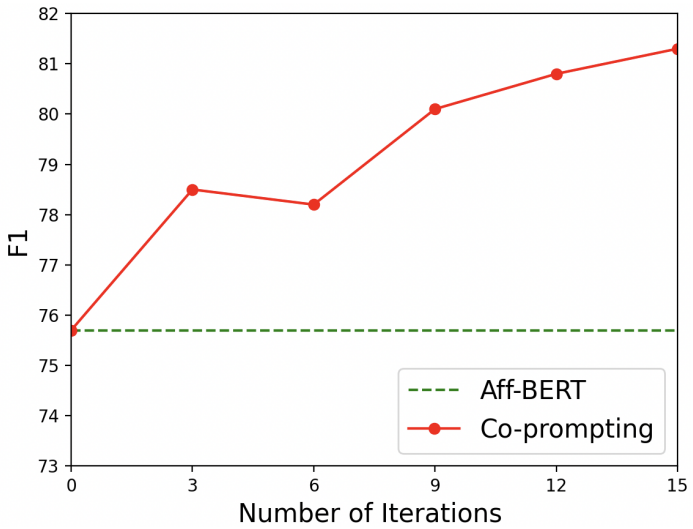

Figure 5.4: Learning curve of Co-Prompting.

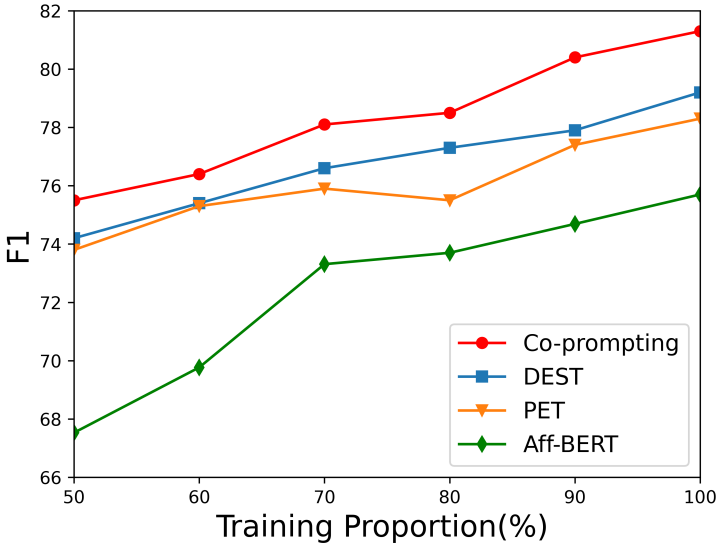

Figure 5.5: Results for different training set sizes.

5.2.9 Behaviors of Language Model Prompts

We also observed some interesting behaviors with the language model prompts. First, language models can exhibit over-generalization behaviors about some events' polarity. For example, GPT-2 repeatedly generated events with the word "throw" (e.g., *"throw a rock," "throw a shoe"*) when the Associated Event Prompt is filled by negative labeled events. This suggests that GPT-2 strongly associates events of throwing anything with negative polarity. Secondly, language models prefer generating events associated with a specific topic. For example, we observed that many generated negative events involved medical issues, such as *"my kidney starts to fail," "my blood sugar drops"* and *"my stomach goes numb."* Using the data generation from language models could potentially cause domain drift. Domain drift can also happen when extracting information directly from a text corpus, but with a neural language model it can be more difficult to understand how or why it is happening.

5.3 Conclusions

Motivated by the challenges of automatically mining affective events from text corpora, this chapter presents a novel approach for eliciting and labeling affective events by Co-Prompting with large language models. The key idea of Multiple View Co-Prompting

is using complementary language model prompts to collect independent views of polarity information, which can then be used jointly as weak supervision to robustly generate new affective events. Our experimental results show that labeling with multiple views is highly effective and that the elicited events substantially improve an affective event classifier. Finally, we believe that Multiple View Co-Prompting is a general idea that should be applicable for other data harvesting tasks as well as tasks that elicit information from language models.

CHAPTER 6

RECOGNIZING EXPRESSIONS OF EMBODIED EMOTION IN NATURAL LANGUAGE

Most people experience emotions every day. When emotions arise, we not only feel them mentally but we also experience them physically via our body. Sometimes an emotion evokes a visible physical reaction. For instance, we may clench our fists or stomp our feet when we feel angry, or raise our hands in the air and dance when we feel happy. We may also have physiological responses when we experience an emotion. For example, we may feel our heart racing or feel a chill down our spine when we get scared. Or we may feel our cheeks flush when we are embarrassed. In general, the physical experience of an emotion via our body is referred to as **embodied emotion** in the psychology literature [101, 151, 138, 17], and it has been recognized as an important component of emotional experiences. Figure 6.1 shows examples of body part references that are and are not associated with embodied emotions.

Recognizing expressions of embodied emotion in natural language is important to identify implicit emotional states, which is a major challenge in emotion recognition [3, 127, 126]. For example, if we read that *"John slammed his fist against the wall,"* we would infer that John is angry. Similarly, if Jane says *"My hands sweated profusely before my presentation,"* we understand that Jane was nervous. In addition, recognizing embodied emotion expressions could help identify behavioral traits and monitor problematic behaviors such as antisocial behaviors [142, 132], which are closely tied to physical responses stimulated by negative emotions.

This chapter introduces the first study on recognizing expressions of embodied emotion in natural language. We formulate the task as a classification problem to determine whether a body part reference describes an embodied emotion. We have created a bench-

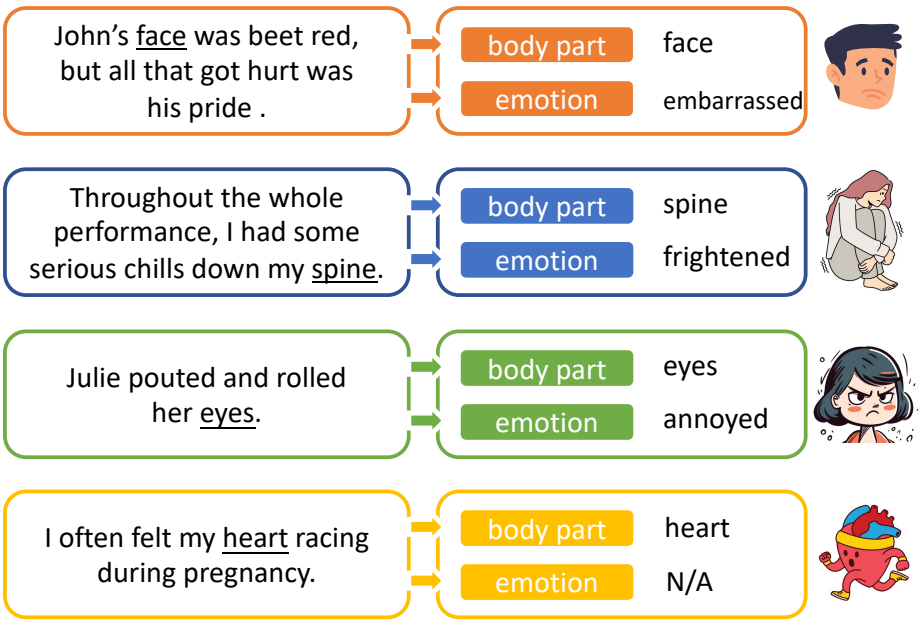

Figure 6.1: Illustration of body part references associated with or not associated with embodied emotions.

mark dataset, **CHEER**, which contains 7,300 body part mentions with human annotations for this task. We conduct extensive experiments to evaluate the effectiveness of multiple existing emotion classifiers on our dataset and show that they do not perform well at recognizing embodied emotion expressions.

We also present two methods to automatically produce weakly labeled data for this task. We develop a pattern-based method that identifies body part words that are syntactically connected to emotion words through manner expressions. For example, *"He slammed his fist in anger"* reveals that *"slammed his fist"* is an embodied reaction to anger. The second method identifies instances of embodied emotion based on prompting a large language model (LLM). Our experiments show that the resulting weakly labeled data can be used to train an effective classifier and also improve classification performance when combined with gold data.

6.1 Task Formulation

We propose a new task to recognize expressions of *embodied emotion* in natural language. While emotion can be embodied in one body part, multiple body parts, or even the whole body, we focus on recognizing expressions of embodied emotion in a *single* body

part, and leave other cases for future work. We formulate the task as a binary classification problem, which classifies a body part word within some context into one of the following two categories: 1) **Embodied Emotion**, where the body part is involved in embodied emotion; 2) **Neutral**, where the body part is not involved in embodied emotion. We define the task as follows:

Definition: *A body part is involved in an embodied emotion if both conditions below are satisfied*:

1) *A physical movement or physiological arousal involving the body part is evoked by emotion.*
2) *The physical movement, if there is any, has no purpose other than emotion expression.*

Condition 1 requires that the physical reaction is caused by emotion. This excludes reactions from other causes, such as weak legs after exercising or watery eyes because of allergies. Condition 2 applies to physical movements (not physiological arousals) and requires that the physical movement has no other purpose. This condition excludes movements that also aim to accomplish a goal. For example, consider the scenario where a house fly is annoying someone, so they slam it with their fist. This action is motivated by emotion, but it is also intended to kill the fly. The set of actions that could be motivated by an emotion is nearly limitless, and the degree to which an emotion causes an action is often ambiguous. Our definition of embodied emotion focuses on movements and physiological arousals that are *solely* emotional and have no additional goal.

One might wonder if the task could be formulated to identify verbs that indicate embodied emotions (regardless of whether a body part is mentioned), instead of identifying body parts that are involved in embodied emotions. One of the advantages of focusing on verbs is that it can cover embodied emotion expressions that do not mention body parts, such as"*I kicked the door after hearing the news*" and "*She jumped when she saw the spider*." However, it can introduce several issues. First, verbs tend to be highly ambiguous and are often used metaphorically. One such case is "*The film reviewers tore apart Jack's performance in his latest film*," where "*tore*" is metaphorically used to indicate criticism. Furthermore, we believe that focusing only on verbs is challenging to operationalize in practice. Nearly every sentence contains verbs, but only a small fraction of them denote physical human actions. Most of this data would not be associated with embodied emotions. We also considered focusing on verbs with lexical semantics that imply a bodily movement (e.g.,

kicked). However, we were not able to find a comprehensive list of such verbs, and also observed many embodied emotion expressions that contained general verbs (e.g., "*raised his eyebrows*"). As a result, we focus our first study on recognizing embodied emotion expressions associated with body parts, and leave these other avenues for future work.

Our task is also contextualized. We identify embodied emotions based on a sentence and its preceding context because physical reactions can be ambiguous without context. For example, the phrase *"my heart is racing"* is likely an expression of embodied emotion in the context of a scary situation, but not in the context of physical exercise.

6.2 Data Collection

Our first goal was to build a dataset of sentences that mention body part words. We began by collecting the terms in two online word lists of body part vocabulary.[1] Then we filtered the list by removing multi-word phrases (e.g., "*index finger*") and plurals. We removed multi-word phrases because most of those phrases in the list referred to internal organs that are rarely discussed and unlikely to be associated with emotions (e.g., "*lumbar vertebrae*"). After the filtering step, the final list contains 162 body part words.

Next, we extracted sentences that mention these body parts in the personal blogs that Ding and Riloff [47] extracted from the ICWSM 2009 and 2011 Spinn3r datasets [90, 91]. This resulted in around 3 million sentences. It is often insufficient to identify embodied emotion based on one sentence in isolation, so we also kept the three preceding sentences. For example, in the sentence "*My hand is shaking*," the shaking could be due to an emotion (e.g., nervousness) or a physical disorder (e.g., tremor) depending on the context.

We next performed several preprocessing steps to clean the collected texts. We used CoreNLP [117] to facilitate this process, such as tokenization and named entity recognition. First, we observed that the data included a lot of sexual descriptions. Sexuality and emotions are often intertwined and determining whether physical responses related to sexual encounters are truly evoked by emotion is challenging, so we decided to exclude texts with sexual descriptions. Specifically, we discarded sentences that contain words in the Sexual category of the LIWC lexicon [187]. We also excluded body part mentions (i.e.,

[1] https://www.collinsdictionary.com/us/word-lists/body-parts-of-the-body and https://www.enchantedlearning.com/wordlist/body.shtml

did not label them) that occur in contexts that mention multiple people because they are also frequently romantic situations. Specifically, we excluded a body part mention if the 5-word window around it contains a plural personal pronoun or at least two different person mentions (personal pronouns or named entities). Note that the 5-word window is only applied within the sentence that contains the body part mentions, and it is not applied across sentences. Finally, we ignored body part mentions that are preceded by a second-person possessive pronoun or a third-person possessive (not pronoun) because these usually refer to another person ("*your eyes*") or a non-human entity (e.g., "*the cat's head*"). We leave for future work the challenge of disentangling emotions and physical actions in multi-person event descriptions.

Finally, we removed infrequent body parts because they usually refer to very specific body parts that are rarely associated with emotions (e.g., "*epiglottis*" and "*ulna*"). We excluded body parts that occurred in less than 0.1% of the sentences. This process produced a final dataset of 868,003 sentences with 56 distinct body parts.

6.2.1 Gold Standard Annotation

We asked two people to produce the gold annotations. An annotation instance is a body part mention in a sentence with the three preceding sentences as context. The annotators produced a binary label (Embodied Emotion versus Neutral) to indicate if the body part is associated with an embodied emotion, following the definition in Section 6.1. The annotators first annotated 2,600 randomly selected sentences that mention a body part. If a sentence mentioned multiple body parts, each mention was presented as a separate instance to annotate. This process produced 2,948 annotated body part mentions. The pairwise inter-annotator agreement measured by Cohen's Kappa was 79%, indicating good agreement. The annotators adjudicated their disagreements to produce the final gold labels. We used this data as the test set. We then asked the annotators to individually label more data and we randomly split these instances into a training set and validation set by the ratio of 7:3. We also made sure that annotation instances that belong to the same sentence went into the same set.

The complete dataset contains 56 distinct body part mentions and 7,300 annotated instances, which consist of 1,350 (18.5%) Embodied Emotion and 5,950 (81.5%) Neutral.

We will refer to this dataset as **CHEER** (a **C**ollection of **H**uman annotations for **E**mbodied **E**motion **R**ecognition). Table 6.1 shows the frequencies of different body parts in CHEER. Table 6.2 shows the statistics of the training, validation and test set. And Table 6.3 shows Embodied Emotion instances in the CHEER data.

6.3 Evaluating Emotion Classifiers

We first conducted experiments to investigate how well existing emotion classifiers can recognize embodied emotion. We evaluated several classifiers that achieved state-of-the-art performance on emotion or affect recognition tasks. The first model is **SpanEmo** [2], which is based on BERT [44] and trained on the affective tweets in SemEval-2018 [126]. To implement this model, we used the code released by the authors at https://github.com/hasanhuz/SpanEmo. The second model, which we will refer to as **GE-BERT**, is a BERT-base model fine-tuned with the GoEmotions data in [39]. As there was no released code, we developed code to train a BERT-base model over the GoEmotion dataset and reported its performance over our dataset. We also evaluated **Seq2Emo** [79], which is a Bi-LSTM model. We used the code released by the authors to train Seq2Emo over the GoEmotions dataset, which can be found at https://github.com/chenyangh/Seq2Emo. Although Seq2Emo was also reported to achieve state-of-the-art performance over the

Table 6.1: Frequencies of different body parts in the CHEER dataset.

head (953), eye (853), hand (691), face (559), heart (384), foot (306), arm (267), leg (255), mouth (251), back (201), shoulder (170), finger (168), ear (143), stomach (138), knee (132), lip (129), chest (129), neck (126), throat (115), nose (111), tooth (104), brain (102), cheek (95), skin (92), tongue (60), ankle (56), lung (55), hip (48), toe (44), thumb (40), forehead (39), spine (31), belly (30), nail (29), jaw (29), eyebrow (28), chin (28), palm (28), wrist (27), waist (25), nerve (25), elbow (22), fist (21), thigh (20), muscle (18), heel (18), rib (15), temple (13), eyelid (13), bone (12), skull (11), vein (11), calf (10), knuckle (8), abdomen (7), forearm (5)

Table 6.2: Statistics of the CHEER dataset in terms of annotated body part mentions.

	Embodied Emotion (%)	**Neutral (%)**	**Total**
Train	578 (19.1%)	2,452 (80.9%)	3,030
Validation	264 (20.0%)	1,058 (80.0%)	1,322
Test	508 (17.2 %)	2,440 (82.8%)	2,948
Total	1,350 (18.5%)	5,950 (81.5%)	7,300

Table 6.3: Embodied Emotion examples in CHEER. The preceding contexts are shortened for brevity.

• It was rather chilly outside due to the rain. I was using the comp late night (LOL) and the comp was downstairs those days back. Kiki came to me and jump onto my lap. I rolled my **eyes** and went "Stupid cat."
• I'm not. You're not. He came home this morning and as I'm sitting at the dining room table checking my e-mail, he sits down and tells me he is going to need my social, and all my names I've had in my life. Immediately my **throat** tightens.
• So, anyway... bad mood yesterday morning. My mom asked if grandma was upset that Mom and I were spending the day together. She said no and stormed off. When we got home, she'd been brooding and pouting and stomping her **feet** as she sulked around the house with nothing to do.
• "I'll never let anyone hurt you again. I promise." She started to shake her head, to deny it all yet again, but something inside her broke, some wall came tumbling down, and she was left standing in the ruins. A loud sob raced up the back of her throat, choking her, and her **knees** buckled.
• "And?" he prompted, the last of his patience vanishing sharply away. "Well, we ate together, and then he took the check before I could get to it." "You let him pay for your meal?" He felt his **eyebrows** fly up in astonishment.
• Looking through these pics today brought and smile to my **face** and tears to my eyes.
• "But he will," Harry asked, nervously wringing his hands, "He will wake up?" "At the moment we can not see any reason why he might remain in this condition for any longer than a week. As long as his condition does not deteriorate then the prognosis is good." Zayn crossed his **arms**, he hated it when people dressed up words.

SemEval2018 dataset, we report the performance of Seq2Emo that was trained over GoEmotions, as it performs better over our dataset. All these models take a text snippet as input and generate multi-label emotions. Finally, we evaluated **Aff-BERT** developed in our prior work [225], an affective event classifier that takes an event phrase as input and identifies its affective polarity. In our experiments, all reproduced models achieved performance that is comparable to the reported performance in the corresponding paper.

These models were trained on different types of input, so we experimented with four strategies for applying each classifier to instances in our CHEER data. Consider the instance below with the underlined "*eyes*" as the targeted body part:

Preceding Context: *Every step he took echoed throughout the room. He stood in front of me, empty eyes locked into mine.*

Sentence: *Then my eyes instantly widened and my mouth dropped open.*

The first two strategies provide full sentences as input to a classifier: a) **Multi-sent**: the input is the preceding context concatenated with the sentence that mentions the body part. b) **Sent**: the input is just the sentence that mentions the body part.

The next two strategies zero in on the context immediately surrounding the body part mention: c) **Window**: the input is the *k*-word window around the body part mention (e.g., the 2-word window is "*Then my eyes instantly widened*"); d) **Event**: the input is the event phrase that mentions the body part (e.g., "*my eyes widened*"). We extract events from dependency parse trees following the same representation used by Aff-BERT. In all cases, the instance is labeled as Embodied Emotion if the classifier recognizes the corresponding input as emotional/affective. If the body part is mentioned in multiple event phrases, we label the instance as Embodied Emotion if any of the phrases is tagged as emotional/affective by the classifier.

6.3.1 Experimental Results

We present the performance of these emotion classifiers using all four input strategies in Table 6.4. For SpanEmo, Seq2Emo and GE-BERT, the *Window* strategy consistently has a higher macro F1 score than *Multi-sent* and *Sent*. For the Embodied Emotion category, we see that the *Window* strategy has higher precision but lower recall than the other two strategies. This is probably because contexts of a smaller scope contain less irrelevant emotion information such as the emotions of other people. The *Window* strategy also outperforms the *Event* strategy except for SpanEmo, mainly because the *Event* strategy has lower recall of Embodied Emotion. This is probably because the emotion classifiers could not recognize emotion in event phrases. Indeed, Aff-BERT achieves a much higher recall of Embodied Emotion than other emotion classifiers with the *Event* strategy, since it is trained to recognize affective polarity for event phrases. However, its recall of Neutral is much lower. This is not surprising, because events that are affective are not necessarily Embodied Emotion. For example, events that describe physical disorder and injury, such as "*My leg feels sore during the exercise*" and "*I hurt my back*," are affective (negative). But they are not Embodied Emotion since these physical conditions are not evoked by emotion.

Overall, GE-BERT produces the best macro F1 score of 58.2%. But it only achieves about 30% recall and precision for recognizing embodied emotions. Aff-BERT$_{\text{Event}}$ achieves the

Table 6.4: Evaluating emotion classification models.

Method	Macro F1	Embodied Emotion			Neutral		
		Precision	Recall	F1	Precision	Recall	F1
SpanEmo							
Multi-sent	26.7	18.1	**92.1**	30.3	**89.0**	13.2	23.0
Sent	32.0	18.1	83.9	29.8	86.3	21.2	34.1
Window	37.6	18.4	74.4	29.5	85.4	31.1	45.6
Event	45.2	18.4	53.7	27.4	83.9	50.4	63.0
Seq2Emo							
Multi-sent	52.3	21.0	33.3	25.7	84.2	74.0	78.8
Sent	53.7	23.3	22.8	23.1	84.0	91.4	87.5
Window	54.4	28.8	16.7	21.2	84.1	91.4	87.6
Event	51.0	24.7	9.4	13.7	83.3	**94.0**	**88.3**
GE-BERT							
Multi-sent	52.6	21.6	36.2	27.1	84.5	72.6	78.1
Sent	54.0	23.2	30.1	26.2	84.5	79.3	81.8
Window	**58.2**	**31.0**	30.3	30.6	85.6	85.9	85.7
Event	53.5	28.4	14.4	19.1	83.8	92.5	87.9
Aff-BERT$_{Event}$	50.3	21.7	56.1	**31.3**	86.4	57.8	69.3

best F1 score of 31.3% for Embodied Emotion, with a higher recall but lower precision as compared to GE-BERT. These results demonstrate that embodied emotions cannot be reliably recognized by existing methods for emotion recognition, which motivates the need for further research on this topic.

6.4 Producing Weakly Labeled Data for Embodied Emotions

Our goal was to create a classifier for recognizing embodied emotion expressions. We produced gold training data, but its amount is relatively small as human annotation is time-consuming. In this section, we introduce two methods to automatically produce a large amount of weakly labeled instances. We will later show that this weakly labeled data can be used to train an effective classifier without any gold data at all, or used in combination with gold data to further improve classification performance.

6.4.1 Labeling Data Using Dependency Patterns

Our first method produces new Embodied Emotion instances by identifying body part words that are syntactically connected to an explicit emotion word through a manner

expression. Specifically, we extract two types of manner expressions using a syntactic dependency parse :

- Prepositional phrases with "in" or "with" and an emotional head noun (e.g., "*My mouth opened **in surprise***" or "*I clapped my hands **with great excitement***").
- Emotional adverb (e.g., "*I **angrily** clenched my fists*" or "*I **impatiently** tapped my finger*").

We observed that emotional manner expressions in the forms above often describe a physical experience when emotion arises (e.g., "*I angrily broke the window*"). As a result, when a body part is syntactically connected to such emotional manner expressions, the sentence tends to describe the physical experience of emotion via the body part.

For emotional nouns in prepositional phrases, we used all positive and negative nouns labeled with strong subjectivity (641 nouns in total) in the MPQA lexicon [208]. For emotional adverbs, we leveraged the WordNet Affect lexicon [183], which associates a subset of words in WordNet [122] with emotions. We extracted the 121 adverbs that are associated with the 6 basic Ekman's emotions [50].

Our pattern-based method first extracts sentences that contain a body part word and one of the emotional manner expressions described earlier. We create an Embodied Emotion instance if a body part word is connected to an emotional manner expression matching one of the dependency relation patterns illustrated in Figure 6.2. In practice, we used CoreNLP [117] to obtain the dependency parse for a sentence. Finally, we remove the emotional manner expression from the sentence so that the classifier cannot use it when learning to recognize embodied emotions.

6.4.2 Labeling Data by LLM Prompting

The pattern-based method is not able to harvest Neutral Instances. In addition, the diversity of the harvested instances may be limited because some body parts rarely co-occur with the emotional manner expressions. To overcome these issues, we also produced new labeled instances by prompting a large language model (GPT3.5). Specifically, we construct a template with an instruction and input placeholders. Given an input instance, we fill the input placeholders with the body part and the sentence that mentions it (see an example in Figure 6.3) and feed it to the language model. We then assign the label based

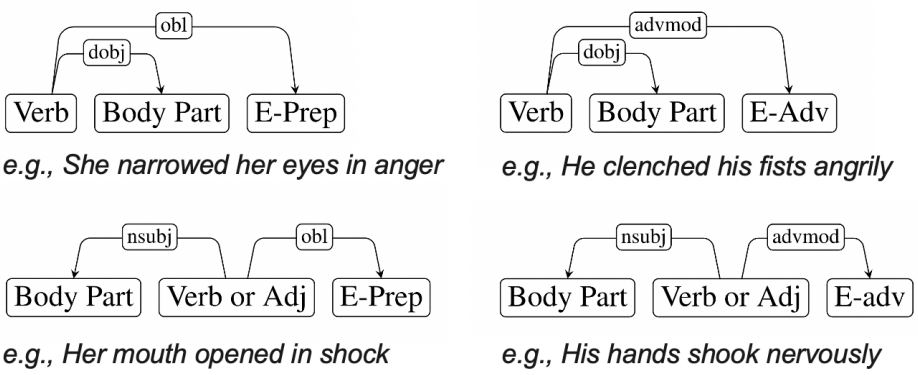

Figure 6.2: Dependency relation patterns. Under each pattern, an example that matches the pattern is shown. E-Prep: a prepositional phrase with emotion head noun. E-Adv: an emotion adverb.

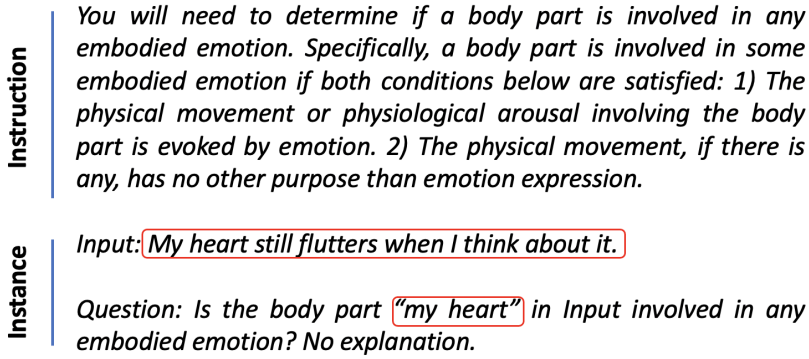

Figure 6.3: Example prompt for GPT3.5. Input placeholders are wrapped by boxes in red.

on the yes-or-no answer. Note that the preceding sentences are not used in the prompt, as we found that using them hurt performance.

6.4.3 Weakly Labeled Dataset

We applied both methods to the subset of the 868,003 sentences in Section 6.2 that were not labeled by the annotators. The pattern-based method produced 7,162 Embodied Emotion instances. For the prompting method, we used GPT3.5 because it achieved the best zero-shot performance (see Section 6.5.1). We first applied the prompting method to collect 7,000 Embodied Emotion instances. We chose the number of 7,000 to make it comparable to the size of the data generated by the pattern-based method. Up to this point, the two methods generated 14,162 Embodied Emotion instances. We then continued to generate Neutral instances using the prompting method (note that the pattern-based

method cannot generate Neutral instances). To maintain a distribution of 20% Embodied Emotion and 80% Neutral, we produced 56,648 Neutral instances with the prompting method.

6.5 Experimental Results

We conducted experiments with classification models trained on weakly labeled data, gold labeled data, or both. We also present results for zero-shot prompting with LLMs as a baseline comparison. For the evaluation metric, we report the macro-averaged F1 score over the test set, as well as Precision, Recall and F1 for each class.

Our classification model is based on fine-tuning the pretrained BERT model [44] with the base-uncased version. Given an input instance, we concatenate the preceding sentences and the sentence that mentions the body part, and insert the CLS token between them. We pass this to BERT and get its last-layer token embeddings. Finally, we produce an embedding for the body part word by averaging the embeddings of its leftmost and rightmost tokens, and then feed it through a linear classification layer to predict the label. For the sake of brevity, we will refer to the classification model as the Embodied Emotion Classifier (**EEC**).

6.5.1 Baselines and Gold Supervision

Large language models (LLMs) have shown impressive zero-shot performance on unseen tasks. So as a point of comparison, we evaluated the performance of several LLMs for zero-shot prompting, including Llama-2-70B [194], Falcon-180B [143] and GPT3.5. Figure 6.3 shows the prompt template that we used. In our preliminary study, we also experimented with few-shot prompting. However, we found that few-shot examples produced worse performance in our task.

The first three rows of Table 6.5 show the zero-shot prompting performance. The best model is GPT3.5, which achieves a macro F1 score of 70.2%. The highest F1 score for Embodied Emotion, however, is only 53.5%. This indicates that all models struggle to reliably recognize embodied emotions. The last row of Table 6.5 (EEC_{gold}) shows the performance of EEC trained with the gold training data (see Section 6.5.3), for comparison. We see that the supervised learning model achieves an F1 score of 83.5%, substantially outperforming

Table 6.5: Zero-shot prompting and gold training results.

Method	Macro	Embodied Emotion			Neutral		
	F1	Pre	Rec	F1	Pre	Rec	F1
Llama-2	43.7	23.1	**95.3**	37.1	**97.2**	33.9	50.2
Falcon	65.8	36.8	79.1	50.2	94.3	71.6	81.4
GPT3.5	**70.2**	**44.0**	68.3	**53.5**	92.5	**81.9**	**86.9**
EEC_{gold}	83.5	73.2	72.1	72.6	94.2	94.5	94.4

the zero-shot prompting results. Compared to the best large language model, GPT3.5, the supervised learning model has a substantially higher precision by 29.2 absolute points and a slightly higher recall by 3.8 absolute points for Embodied Emotion.

6.5.2 Weak Supervision Results

Next, we train EEC using **only** weakly labeled data. We explored different sets of weakly labeled Embodied Emotion instances. Specifically, we trained EEC using:

- $\boldsymbol{E_{PAT}}$: The Embodied Emotion instances (7,162) labeled by the pattern-based method.
- $\boldsymbol{E_{LM}}$: The Embodied Emotion instances (7,000) labeled by the LM-based prompting method.

In experiments, we use the Neutral instances generated by the prompting method, denoted by $\boldsymbol{N_{LM}}$. In each experiment, we randomly selected instances from N_{LM} to enforce a distribution of 20% Embodied Emotion and 80% Neutral (to approximately match the gold distribution). For each set of weakly labeled data, we then randomly selected 2,000 instances for validation and used the rest for training. For the hyperparameters, we set the maximum sequence length in BERT to be 256 and the batch size to 16. We also used the AdamW optimizer with a linear schedule and a warmup rate of 0.1. Before gradient descent, we clipped the gradient norm using the threshold of 1.0. We observed in our early experiments that varying the number of training epochs and the learning rate did not have a significant impact. So we trained the model for 10 epochs with a learning rate of 1e-5 for all experiments.

Table 6.6 presents the results averaged across three runs. The first row shows the performance of zero-shot prompting with GPT3.5 once again, for the sake of comparison. Rows 2 to 6 show the performance of models trained with different sets of weakly labeled data. All of these models outperform zero-shot prompting. The E_{PAT} model achieves a

Table 6.6: Results with weakly labeled data only.

Method	Macro	Embodied Emotion			Neutral		
	F1	Pre	Rec	F1	Pre	Rec	F1
GPT3.5	70.2	44.0	68.3	53.5	92.5	81.9	86.9
EEC with							
E_{PAT}	71.5	**68.0**	40.6	50.8	88.6	**96.0**	92.2
E_{LM}	74.7	52.4	69.2	59.6	93.1	86.8	89.9
$E_{LM} \times 2$	74.5	53.3	66.6	59.2	92.7	87.7	90.1
$E_{PAT} \cup E_{LM}$	**79.3**	62.1	**71.1**	**66.3**	**93.8**	91.0	**92.4**
EEC_{gold}	83.5	73.2	72.1	72.6	94.2	94.5	94.4

macro F1 score of 71.5% , while the E_{LM} model achieves 74.7% F1 score. For the Embodied Emotion class, the E_{PAT} model has higher precision but the E_{LM} model has higher recall. This suggests that the E_{PAT} data is more precise while the E_{LM} data is more diverse.

Next, we tried adding more training data. The $E_{LM} \times 2$ row shows results when using twice as many Embodied Emotion instances (14,000) labeled by the prompting method, and twice as many Neutral instances. This model produces a macro F1 score of 74.5%, which is comparable to the E_{LM} model. This suggests that the value of this weakly labeled data source has maxed out.

Our next experiment trains EEC using both types of weakly labeled data together ($E_{PAT} \cup E_{LM}$). This training set contains 14,162 Embodied Emotion instances, with a corresponding balance of Neutral instances. Table 6.6 shows that training with both sets of data together produces a substantially better classifier, resulting in an F1 score of 79.3%. Importantly, note that training with 14k instances produced by two different methods yields much better results than training with 14k instances produced by the prompting method alone. These results suggest that the instances produced by the two methods are complementary.

The bottom row of Table 6.6 again shows the result of the model trained with gold data, for easy comparison. The model trained with only weakly labeled data ($E_{PAT} \cup E_{LM}$) performs nearly as well as the model trained with gold supervision (just 4.2 points lower in F1 score). We conclude that an embodied emotion classifier can be effectively trained using only weakly labeled data.

6.5.3 Exploiting Both Gold and Weakly Labeled Data

We also investigated whether the weakly labeled data could provide additional benefits when combined with gold labeled data. So we fine-tuned EEC using both the gold training data and the weakly labeled data together. Specifically, we used the best performing weakly labeled data: negative examples from N_{LM} and positive examples from $E_{PAT} \cup E_{LM}$. We used EEC fine-tuned with only gold data for comparison. During training, we optimize the model with respect to the weighted cross entropy loss: $L = L_{gold} + \lambda L_{weak}$, where L_{gold} is the loss over the gold data, L_{weak} is the loss over the weakly labeled data and λ is a hyperparameter. For the number of training epochs, we tried 5 and 10. For the learning rate, we searched through the set of (1e-5, 2e-5, 3e-5). For the weight hyperparameter λ, we searched through the range from 0.1 to 1.0 with an increment of 0.1. We then selected the hyperparameters that performed the best over the gold validation set.

Table 6.7 shows the model performance averaged across three runs. The model trained with only gold data (row 1) yields a macro F1 of 83.5%. When the weakly labeled data is added (row 2), the model improves to achieve an F1 score of 85.4%. This improvement is mainly due to a large increase of 7.4 points in recall of Embodied Emotion (72.1% $\rightarrow$ 79.5%). Overall, the addition of the weakly labeled data helps the model recognize many more instances of embodied emotion with nearly the same precision.

6.6 Analysis

We present several analyses to better understand the behavior of our embodied emotion classifier.

In Section 6.5.3, we showed that combining the gold training data with weakly labeled data improves the performance of our classifier ($\text{EEC}_{gold+weak}$ in Table 6.7). So we further investigated how the different sources of weakly labeled data (E_{LM} and E_{PAT}) impact the model. Table 6.8 shows the performance when we remove one source at a time. Removing

Table 6.7: Using gold and weakly labeled data together.

Method	**Macro**	**Embodied Emotion**			**Neutral**		
	F1	Pre	Rec	F1	Pre	Rec	F1
EEC_{gold}	83.5	**73.2**	72.1	72.6	94.2	**94.5**	94.4
$+weak$	**85.4**	72.9	**79.5**	**76.1**	**95.7**	93.9	**94.7**

Table 6.8: The effects of removing E_{PAT} or E_{LM} from the weakly labeled data, one at a time.

Method	Macro F1	Embodied Emotion			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
All	**85.4**	**72.9**	**79.5**	**76.1**	**95.7**	**93.9**	**94.7**
- E_{PAT}	83.9	72.3	74.9	73.5	94.8	**93.9**	94.3
- E_{LM}	84.1	71.7	76.3	73.9	95.0	93.7	94.3

either source decreases performance, particularly on the recall for embodied emotions which drops from 79.5% down to 74.9% without E_{PAT} or to 76.3% without E_{LM}. These results reinforce the earlier observation that the weakly labeled data produced by the two different methods seem to be complementary and so using them together is beneficial.

Some body parts occur much more frequently than others, as shown in Table 6.1. We expect the classifier to generalize across body parts to some degree, but some body parts are fundamentally different than others (e.g., eyebrows versus spine) so we also expect substantially different language around different body parts. We did an analysis to see how the amount of training data for a specific body part correlates with performance on instances of that body part. We partitioned the 55 body parts into two groups: 27 high-frequency body parts with ≥ 20 training examples and 28 low-frequency body parts with < 20 training examples. Figure 6.4 plots the F1 score for each body part on the y-axis, based on the performance of the $\text{EEC}_{gold+weak}$ model in Table 6.7. Overall, there is a strong correlation between training frequency and performance: most high-frequency body parts show high F1 scores, with a few exceptions. The low-frequency body parts typically have only one or a few instances in the test set so their performance is volatile, but most perform poorly. This analysis strongly suggests that producing additional training data for low-frequency body parts would likely further improve our model.

Finally, we manually analyzed the errors of the best classifier in Table 6.7 ($\text{EEC}_{gold+weak}$) and categorized them into two types. The first error type is **false negative Embodied Emotion**. For most cases of this error, we suspect the classifier failed because it cannot recognize the causal relationship between an emotional experience in the preceding context and the physical reaction. We show two examples in the upper portion of Table 6.9. For instance, "*lip*" in (a) is involved in embodied emotion as the biting results from the negative conversation in the preceding context. The second error type is **false positive Embodied**

Figure 6.4: F1 scores based on body part frequency.

Table 6.9: Error cases.

False Negative Embodied Emotion
(a) He glared up at Ianto. "Thought I told you I didn't want to see your face." Ianto bit his **lip**.
(b) The tragedies that I had brought to my family and friends broke into fragments and stabbed me, as though they were taking revenge ... My **chest** began to tighten ...
False Positive Embodied Emotion
(c) The doctor there told me, "you are having a heart attack even as we speak." My **heart** arrested twice, I was shocked four times.
(d) Eames choked and gasped for air, his **head** already pounding from where he hit the other man.

Emotion. Most cases of this error mention body parts involved in physical disorders. The bottom portion of Table 6.9 shows two examples. Polysemy may explain why the classifier is confused by many of these cases. For example, in (c) the word "shocked" refers to an electrical shock (presumably defibrillation), but it also commonly refers to a reaction of surprise. Similarly, in (d) the word "gasped" simply refers to sharp breathing in this case, but it commonly refers to an emotional response as well.

6.7 Conclusion

This chapter presents the first study on recognizing expressions of embodied emotion in natural language. We first created a dataset that contains 7,300 body part mentions with human annotation for this task, which can be found at https://github.com/

yyzhuang1991/Embodied-Emotions. We then performed extensive experiments to show that this task is challenging for existing emotion recognition methods. Two methods were introduced to automatically produce a large set of weakly labeled instances, including one pattern-based method that extracts manner expressions with explicit emotional words, and one prompting method that exploits a large language model. We showed that the weakly labeled data can be used to train an effective embodied emotion classifier, and that combining it with gold data yields a better classifier than using gold data alone.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This dissertation presents research on learning two types of implicit affective expressions that are common and critical for affective text analysis. The first type of implicit affective expressions is affective events, which refer to events that impact most people in a positive way or negative way. For example, the event "*I graduate with a PHD degree*" is positive and the event "I did not pass my exam" is negative for most people. The second type of implicit affective expressions is embodied emotions, which refer to physical responses in our body when emotion arises. For example, we often clench our fists when we are angry and throw our hands in the air when we are excited. This chapter gives a summary of research contributions presented in this dissertation, and then discusses future research directions based on this dissertation.

7.1 Research Summary and Contributions

In Section 1.3 of Chapter 1, this dissertation presents two research claims. In this section, I will revisit the two research claims and show that they are supported by the results demonstrated in this dissertation.

The first research claim focuses on the task of affective event recognition and is shown below:

> *Claim 1: Accuracy for affective event recognition can be improved with deep learning models that exploit novel semi-supervised algorithms including Discourse-Enhanced Self-Training and Multiple View Co-Prompting.*

In Chapter 3, I first identified several limitations of previous approaches that developed lexicons for affective events, including: 1) lexicons of affective events do not generalize well to unseen events and 2) the quality of polarity labels in these lexicons is not high in some cases. To measure these limitations, we introduced a new dataset for affective events, TWITTER, and evaluated over it the largest existing lexicon of affective events,

AEKB. Experiments demonstrated that only 66% events in TWITTER are found in AEKB, and AEKB only achieves a macro-F1 score of 65.2% over these events. In order to address these limitations, I developed a deep learning model, *Aff-BERT*, based on fine-tuning the pretrained language model BERT. Experiments on TWITTER demonstrated that Aff-BERT substantially outperforms other methods and generalizes better to unseen events.

The performance of Aff-BERT is potentially limited by the small amount of available training data. In Chapter 4, I presented the *Discourse-Enhanced Self-Training* (DEST) method to improve Aff-BERT with weakly labeled data. DEST is motivated by the observation that the polarity of an event is often indicated by the sentiment of coreferent sentiment expressions following the event (e.g., "*I got COVID-19. I feel terrible*."). To generate weakly labeled events, DEST assigns a polarity label to an event based on: 1) the prediction of a classifier (e.g., Aff-BERT) that is trained on the training set, and 2) the average polarity of the coreferent sentiment expressions of the event. Experiments over the TWITTER data demonstrate that Aff-BERT trained with DEST outperforms Aff-BERT with only gold data and Aff-BERT trained with traditional self-training.

Chapter 5 introduced another semi-supervised algorithm, *Multiple View Co-Prompting*, to improve affective event classifiers. Different from DEST, which harvests weakly labeled data from a text corpus, Multiple View Co-Prompting is a simpler but more effective method that generates weakly labeled data by prompting language models. It consists of two key steps: 1) the *Event Generation* step to generate (unlabeled) event phrases by prompting a language model such as GPT2; 2) the *Polarity Assignment* step to assign polarity labels to the generated event phrases. To assign an accurate polarity label to an event, it first extracts two views (*Associated Event View* and *Emotion View*) that contain complementary polarity information for the event. Then the two views are combined to produce a polarity label for the event. Experiments demonstrate that the generated weakly labeled data can substantially improve Aff-BERT when combined with gold training data, achieving state-of-the-art performance for this task. Our analysis also showed that the weakly labeled data has a high quality - the accuracy of the polarity labels of these weakly labeled events is 91%, as assessed by human annotators.

> *Claim 2: Recognizing expressions of embodied emotion in natural language can be improved by training a model specifically for this task and exploiting semi-supervised learning.*

In Chapter 6, I proposed a new task to recognize embodied emotion expressions in natural language. I formalized the learning task as a binary classification problem, which determines whether a body part mention is involved in *Embodied Emotion* or *Neutral*. For example, the body part "*throat*" is categorized as Embodied Emotion in the statement "*I have not seen her for ten years. My throat immediately tightened when I saw her face*," as the tightened throat is evoked by emotions. On the other hand, the "*throat*" is categorized as Neutral in the statement "*I have been sick for a few days and my throat hurts from coughing*," as the throat-hurting symptom is a physical condition. To facilitate the study, I constructed a gold dataset, CHEER, that contains 7,300 instances with human annotations. As shown in experiments, this dataset is challenging for a wide range of existing affect recognition classifiers, including emotion classifiers, affective event classifiers and large language models with zero-shot learning.

To recognize embodied emotion expressions, I proposed an embodied emotion classifier (EEC) based on fine-tuning BERT, which achieves a macro-F1 score of 83.5% over the CHEER dataset. As the training data is small, I also presented two semi-supervised methods to generate weakly labeled data to improve EEC. The first method mines weakly labeled data from a text corpus by exploiting manner expressions with emotion (e.g., "*I crossed my arms with frustration*," "*I clenched my fists in anger*"). The second method produces weakly labeled data by prompting large language models such as GPT3.5. Experiments showed that the weakly labeled data generated by the two methods can train an effective EEC classifier on its own, which achieves a macro-F1 score of 79.3%. It can also improve the EEC classifier when combined with the gold training data, boosting the macro-F1 score from 83.5% to 85.4%.

7.2 Future Research Directions

7.2.1 Improving Affect Recognition in NLP and Other Disciplines with Affective Event Recognition and Embodied Emotion Recognition

The studies of affective event recognition and embodied emotion recognition in this dissertation are motivated by the ultimate goal of improving affect recognition in text. One promising direction for future research is to exploit these two tasks to improve affect recognition in narratives and stories, where the story plot consists of a sequence of experi-

ences of characters. One potential way to improve affect recognition in narratives with the two tasks is to use the knowledge of affective events and embodied emotions as auxiliary features for an affect recognition system. Consider using an affect recognition system to predict the affective state of Jack in the line *''When Jack went past the haunted house, a chill ran down his spine and his heart raced*." We can apply an embodied emotion classifier to the body part mentions ("*spine*" and "*heart*"). At the same time, we can also apply an affective event classifier to the events such as "*Jack went past the haunted house*." Then the predictions of the embodied emotion classifier and the affective event classifier could be used as extra features to help an affect recognition system determine the affective state of Jack. Another potential way is to train an affect recognition system with multi-task learning over the tasks of affect recognition (e.g., sentiment analysis and emotion detection), affective event recognition and embodied emotion recognition. Prior work [27, 112] has found that training a model over multiple tasks that share common knowledge can improve the model performance over each task, as knowledge learned in one task can benefit other tasks. As all three tasks emphasize identifying affect, jointly learning over them could potentially improve an affect recognition system in NLP.

The knowledge of affective events and embodied emotions could also be potentially useful for affect recognition systems in other fields of AI. One such area is the research on social robots, which usually convert the speech of a user to text and then detect emotions by using NLP systems. I believe that it could be valuable to integrate our classifiers of affective events and embodied emotions into the NLP system in a social robot. For example, the knowledge of affective events could help a social robot detect the positive emotion conveyed by the speech: "*I bought a house*." The knowledge of embodied emotion could help a social robot detect that the speech is emotional: "*What you said made my stomach turn*." To do so, one could potentially store an embodied emotion classifier and an affective event classifier in the NLP system of a social robot and apply them when the speech of a user contains an event or a bodily response.

In addition, affect recognition in computer vision (CV) can potentially benefit from our study of affective events and embodied emotions. In recent years, researchers have been interested in developing CV models that detect emotions in images. A line of work focuses on detecting emotions expressed by facial expressions [141, 180] and body gestures [174,

137]. Our study of embodied emotion could be valuable for this line of work. Consider identifying the emotions in an image that shows a man with pursed lips. We could first leverage an image captioning model to generate a caption for the image (e.g., "*the man pursed his lips*"), and apply an embodied emotion classifier to the bodily responses in the caption to determine if there is any emotional state. Then the information provided by the embodied emotion classifier could serve as an extra signal for emotion detection.

Another line of work [98] in CV focuses on detecting emotions based on the scene contexts in images, including the surroundings of a person and the actions occurring around a person. As discussed in the paragraph above, an embodied emotion classifier could potentially benefit models in this research, as it may provide extra signals about the emotional states conveyed by the body gestures and facial expressions in the textual description. In addition, affective event recognition could be valuable for this task. This is because many scene contexts refer to events in our daily life, such as having a birthday party, watching a movie, and resting on the beach. Consider again the method proposed above that generates a caption for an image. The caption may describe the events shown in an image (e.g., "*The child is having a birthday party*"). We could apply an affective event classifier to the events in the caption, and then use the generated information as extra signals to help detect the emotions in the image.

7.2.2 Studying the Intensity of Affective Events

One interesting future direction for affective event recognition is to study the strength or intensity of an affective event. In our daily life, the events we experience do not necessarily affect us to the same degree, even if they have the same polarity. For example, the negative event of losing a pen might be a small matter for most people, but the negative event of losing a car is probably a big deal for most people. I refer to the degree to which an event impacts us positively or negatively as its affective intensity. Studying the affective intensity of an event could benefit a lot of tasks. First, it could help an affect recognition system detect the overall affective state of a person. Consider the online review of a hotel: "*Good thing: I got free breakfast. Bad thing: I could not sleep for 5 nights in a row due to the noise.*" We probably infer that the person has a very negative opinion towards the hotel, as for most people the positive impact of the event "*I got free breakfast*" is much smaller than the

negative impact of the event "*I could not sleep for 5 nights in a row*." Second, it could be beneficial for domains where it is critical to consider the impact of an event. Consider an emergency management system, which often sends more resources and rescues to areas that are more severely impacted by crises, such as hurricanes and earthquakes. To better assess a crisis's impact in the affected areas, information on social media (e.g., tweets posted in Twitter) is usually considered in the decision making. Developing an NLP system that detects the affective intensity of events reported in social media during crisis (e.g., "*some people died here*," "*we are out of water*," "*our houses burned down*" and "*I don't feel any wind here*") could potentially help a management system assess the impact of a crisis in different affective areas, prioritize its responses and make better responses.

7.2.3 Extending the Scope of Embodied Emotion Recognition

In this dissertation, the study of recognizing expressions of embodied emotion in natural language focuses on body part mentions, for the reasons mentioned in Chapter 6. But many expressions of embodied emotion in natural language do not necessarily contain body part mentions. On the other hand, many expressions use verbs that describe body movement, such as "*I kicked the wall because I was angry*," "*I jumped as I saw a spider on the table*," and "*I collapsed after I heard the news*." One valuable future direction for this research is to include these expressions with verbs that describe body part movement. To conduct research in this direction, it could be useful to first collect verbs that are closely related to body part movement, and then focus on expressions with verbs in this lexicon. As a starting point, one might consider collecting verbs under the *body_movement* frame in FrameNet [11], which refers to motions or actions an agent performs with some part of his/her body.

7.2.4 Associating Expressions of Embodied Emotions with Emotional Labels

In this dissertation, our task to recognize expressions of embodied emotion in natural language is formulated as a binary classification problem: categorize a body part mention as Embodied Emotion or Neutral. While it identifies an emotional state based on bodily responses, it does not tell what the emotional state is. A valuable line of future research is to associate expressions of embodied emotion with emotional labels. For example, given the

statement "*My palms were sweaty before I got on the stage*," an embodied emotion classifier should not only categorize "*palm*" as Embodied Emotion, but also associate the sweaty palm with nervousness.

The task of associating expressions of embodied emotions with emotional labels could fundamentally benefit the task of emotion detection in NLP, as it provides more emotional information than the work in this dissertation. Nowadays, most existing tasks of emotion detection are classification problems with at least 6 emotion labels. For example, the GoEmotions dataset [39] contains 27 emotion labels. The study in this dissertation, which performs binary classification, may provide relatively limited information for these emotion detection tasks. If we associate expressions of embodied emotions with emotional labels, this study could not only help these emotion detection systems discover ongoing emotions but also help them identify what these ongoing emotions are.

NLP models that associate expressions of embodied emotions with emotional labels might also enhance the development of more empathetic conversational agents. Consider the potential application of embodied emotion recognition in social robots, as discussed in Section 7.2.1. If an embodied emotion classifier could associate the detected embodied emotions with emotional labels, a social robot could better understand the specific emotional states conveyed by bodily responses in the speech of a user and act accordingly. For example, recognizing that "*sweaty palms*" of a user indicates nervousness, a social robot could respond with calming reassurances or support, which improves user experience and interaction quality.

7.2.5 Addressing the Subjectivity of Affective Norms

The problems studied in this dissertation are highly relevant to knowledge about affective norms. Our study of affective events recognizes the stereotypical affective impact of an event independent of context. For example, it considers the event of going on a vacation a positive event and the event of working overtime a negative event. The study of embodied emotions recognizes the bodily manifestations of emotions, the interpretation of which is heavily based on our commonsense. Consider the text "*my eyes get watery after I heard the news*." Most people would interpret the bodily response "*my eyes get watery*" as a manifestation of emotion. While we consider context in this task, we rely greatly on the

affective norms to recognize embodied emotions.

Knowledge about affective norms is a type of commonsense knowledge that is prevalent and essential in our language, and it has gained a lot of research interests in the research community. However, one major challenge in the current study is that our knowledge about affective norms is subjective and relevant to many factors. For instance, people with different cultural backgrounds may have different knowledge about affective norms. Consider the case where someone receives a watch as a birthday gift. This event is positive in Western culture but very negative in Chinese culture. This is because giving a watch can be interpreted as a reference to the end of life or time running out in Chinese culture. As another example, the hand gesture of joining the thumb and the index finger into a circle (the OK gesture) expresses a positive emotion in the U.S., but it conveys a very negative emotion (e.g., "*you are a loser*") in some other countries such as France, Brazil and Germany. People in different social classes may also have different knowledge about affective norms. For instance, the event of "*getting a phone*" may be very positive for people with very low income but more neutral for people with very high income. In addition, affective norms could be different for people with similar backgrounds, since different people have different personal preferences. For example, the event of having a party could be positive for extroverted persons but negative for some introverted people. Getting pregnant could be positive for some people but negative for others.

I believe that it is crucial to address this subjectivity problem of affective norms in future work. Otherwise, it could be challenging to effectively learn and apply the knowledge about affective norms in practice. One valuable direction is to learn knowledge about affective norms with respect to different languages, cultures and other important factors. Consider the task of embodied emotion recognition. As it is mainly subjective to cultural norms, we could learn embodied emotion expressions in different cultures separately. Suppose we want to recognize expressions of embodied emotions in Chinese culture. To facilitate this study, we could first create benchmark datasets by collecting texts in Chinese. As Chinese texts could be written by people with different cultural backgrounds (e.g., people from Taiwan and people from mainland China), it is also essential to focus the study on data collected from a region with cultural homogeneity (e.g., mainland China).

Regarding the subjectivity issue due to personal preferences, one potential solution is

to capture the likelihoods of affective norms (i.e., how likely is this an affective norm, or how likely is this affective norm true?). Consider the task of affective event recognition. We could capture the likelihoods of different polarities for an event. To build a manually annotated dataset for this task, one could perform large-scale crowdsourcing (e.g., collect human annotations using Mechanical Turk). Suppose we want to collect the likelihoods of different polarities for the event of getting pregnant. We could ask 1,000 mechanical turkers whether this event is positive, negative or neutral. Then based on the statistics, we could estimate the probabilities of polarities for the event (e.g., 80% of the time the event of getting pregnant might be viewed as a positive event and 20% of the time it is viewed as a negative event). I believe that capturing probabilities of affective norms could help address the subjectivity problem due to personal preferences. In addition, it could be useful for comparing and contrasting affective norms across different cultures and demographic sectors.

7.2.6 Addressing the Ethical Concerns of AI Systems with Emotional Intelligence

The ultimate goal of this research is to help build AI systems with emotional intelligence, which can detect emotions and respond with appropriate emotions. I believe that AI systems with emotional intelligence could significantly impact humans in a positive way. For example, they can create more engaging and satisfying interactions for users, leading to better user experiences. As another example, AI with emotional intelligence can provide basic emotional support and companionship, which can be particularly beneficial for individuals who are lonely or dealing with mental health issues.

However, these systems could come with ethical concerns that cause unexpected consequences and harm to our society. In the rest of this section, I will discuss the concerns of AI systems that detect emotion and AI systems that generate emotional responses.

The major concern of AI systems that can recognize users' emotions is that they could be potentially used to manipulate individuals based on their emotional states and reduce their autonomy (i.e., the capability of individuals to make decisions on their own without external control). For example, our shopping behaviors might be manipulated by these AI systems. Companies may create shopping recommendation systems that detect users with negative affective states and suggest retail therapy they do not need. Marketers might

create advertisement systems that exploit emotional data to target consumers at their excited or vulnerable moments and show ads for comfort foods or luxury items. These manipulations can occur in many other domains too. One example is the political realm, where emotionally intelligent AI could be used to analyze the emotions and opinions of users and tailor their messages to change or enhance those users' opinions to impact election outcomes. Another concern with these emotion recognition systems is: *Is it ethically justified to analyze users' emotional states?* The emotional state of an individual is private, so it is worthy of protection. Though people usually express their emotions and opinions in public, many may not want their private states to be analyzed by AI systems.

Our behaviors and decisions may be easily impacted by AI systems that generate emotional responses, as psychology literature has found that our perception of the world and decision-making are influenced by emotion [24]. For example, a conversation system could generate emotional responses to subtly increase one's anger/affection towards a subject matter. Companies could develop shopping AI systems that tell a hesitant person to buy a product by using persuasive tones. Question-answering systems may gain high credibility when they answer in confident tones. Even those AI systems designed for the good could negatively change our behaviors too. For example, AI systems that provide emotional support may make their users overly reliant and diminish their ability to regulate their emotional states on their own. Another concern is that AI systems that generate emotional responses could look very human-like, blurring the lines between human and machine interactions. This ambiguity could potentially make us vulnerable to crimes. For instance, recent reports indicate that scammers have exploited AI-driven chatbots to impersonate real customer service agents or friends, deceiving victims into transferring money or disclosing sensitive personal information.

I believe that addressing these concerns is crucial for the responsible development and deployment of AI systems with emotional intelligence. Without careful consideration and ethical safeguards, the potential benefits of these systems could be outweighed by the risks they pose to individuals and society. For example, we should implement comprehensive ethical guidelines and standards for the development and deployment of emotionally intelligent AI systems. These guidelines should prioritize transparency, ensuring that users are aware when they are interacting with AI and understand the

capabilities and limitations of these systems. In addition, it is important to protect users' privacy. For example, users should be informed about how their data will be used and have the option to decline it. To address the complicated ethical challenges, it is essential to conduct interdisciplinary research that engages experts in psychology, ethics, law, computer science, and other relevant fields. Furthermore, regulatory frameworks must evolve to keep pace with advancements in AI technology. Legislators and policymakers should collaborate with experts across various disciplines to develop regulations that ensure the ethical use of emotionally intelligent AI.

REFERENCES

[1] M. Abdul-Mageed and L. Ungar, *EmoNet: Fine-grained emotion detection with gated recurrent neural networks*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2017, pp. 718–728.

[2] H. Alhuzali and S. Ananiadou, *SpanEmo: Casting multi-label emotion classification as span-prediction*, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 1573–1584.

[3] C. O. Alm, D. Roth, and R. Sproat, *Emotions from text: Machine learning for text-based emotion prediction*, in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 579–586.

[4] S. Aman and S. Szpakowicz, *Identifying expressions of emotion in text*, in International Conference on Text, Speech and Dialogue, Springer, 2007, pp. 196–205.

[5] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, *Do not have enough data? Deep learning to the rescue!*, in Proceedings of the Twentieth AAAI Conference on Artificial Intelligence, vol. 34, Association for the Advancement of Artificial Intelligence, 2020, pp. 7383–7390.

[6] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, *Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers, 2018, pp. 3674–3683.

[7] M. Arnold, *Emotion and Personality*, Columbia University Press, New York City, NY, 1960.

[8] M. B. Arnold, *The Nature of Emotion*, Penguin Books, London, UK, 1968.

[9] A. Asai, S. Evensen, B. Golshan, A. Halevy, V. Li, A. Lopatenko, D. Stepanov, Y. Suhara, W.-C. Tan, and Y. Xu, *HappyDB: A corpus of 100,000 crowdsourced happy moments*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association, 2018, pp. 647–655.

[10] R. P. Bagozzi, M. Gopinath, and P. U. Nyer, *The role of emotions in marketing*, J. Acad. Mark. Sci., 27 (1999), pp. 184–206.

[11] C. F. Baker, C. J. Fillmore, and J. B. Lowe, *The Berkeley Framenet Project*, in COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, Association for Computational Linguistics, 1998, pp. 86–90.

[12] A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes, *Sentiment analysis of political tweets: Towards an accurate classifier*, in Proceedings of the Workshop on Language Analysis in Social Media, Association for Computational Linguistics, 2013, pp. 49–58.

[13] R. C. Balabantaray, M. Mohammad, and N. Sharma, *Multi-class Twitter emotion classification: A new approach*, Int. J. Appl. Inf. Syst., 4 (2012), pp. 48–53.

[14] A. Balahur, J. M. Hermida, and A. Montoyo, *Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model*, IEEE Trans. Affect. Comput., 3 (2012), pp. 88–101.

[15] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, *Sentiment analysis in the news*, in Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association, 2010, pp. 2216–2220.

[16] P. Bard, *A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system*, Am. J. Physiol., 84 (1928), pp. 490–515.

[17] L. Barrett, K. Lindquist, G. Semin, and E. Smith, *The embodiment of emotion*, in Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches, G. R. Semin and E. R. Smith, eds., Cambridge University Press, Cambridge, UK, 2008, pp. 237–262.

[18] L. W. Barsalou and K. Wiemer-Hastings, *Situating abstract concepts*, in Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking, D. Pecher and R. A. Zwaan, eds., Cambridge, UK, 2005, Cambridge University Press, pp. 129–163.

[19] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, *Social robots for education: A review*, Sci. Robot, 3 (2018), eaat5954.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet allocation*, J. Mach. Learn. Res., 3 (2003), pp. 993–1022.

[21] A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, in Proceedings of the 11th Annual Conference on Computational Learning Theory, Association for Computing Machinery, 1998, pp. 92–100.

[22] L. A. M. Bostan, E. Kim, and R. Klinger, *GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception*, in Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 1554–1566.

[23] C. Breazeal, K. Dautenhahn, and T. Kanda, *Social robotics*, in Springer Handbook of Robotics, B. Siciliano and O. Khatib, eds., Springer International Publishing, Cham, Switzerland, 2016, pp. 1935–1972.

[24] T. Brosch, K. Scherer, D. Grandjean, and D. Sander, *The impact of emotion on perception, attention, memory, and decision-making*, Swiss Med. Wkly, 143 (1920), w13786.

[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language models are few-shot learners*, in Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., 2020, pp. 1877–1901.

[26] W. B. Cannon, *The James–Lange theory of emotions: A critical examination and an alternative theory*, Am. J. Psychol., 39 (1927), pp. 106–124.

[27] R. Caruana, *Multitask learning*, Mach. Learn., 28 (1997), pp. 41–75.

[28] F. Casel, A. Heindl, and R. Klinger, *Emotion recognition under consideration of the emotion component process model*, in Proceedings of the 17th Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 49–61.

[29] J. Chen, Z. Yang, and D. Yang, *MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 2147–2157.

[30] Y. Chen, W. Hou, X. Cheng, and S. Li, *Joint learning for emotion classification and emotion cause detection*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 646–651.

[31] N. Cherakara, F. Varghese, S. Shabana, N. Nelson, A. Karukayil, R. Kulothungan, M. Afil Farhan, B. Nesset, M. Moujahid, T. Dinkar, V. Rieser, and O. Lemon, *FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions*, in Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 2023, pp. 588–592.

[32] Y. Choi and J. Wiebe, *+/-EffectWordNet: Sense-level lexicon acquisition for opinion inference*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2014, pp. 1181–1191.

[33] G. Chrupała, *Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques*, J. Artif. Intell. Res., 73 (2022), pp. 673–707.

[34] G. Cortal, A. Finkel, P. Paroubek, and L. Ye, *Emotion recognition based on psychological components in guided narratives for emotion regulation*, in Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, 2023, pp. 72–81.

[35] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis, *SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news*, in Proceedings of the 11th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2017, pp. 519–535.

[36] C. Darwin, *The Expression of Emotions in Man and Animals*, John Murray, London, UK, 1872.

[37] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, *Embodied question answering*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers, 2018, pp. 1–10.

[38] S. R. Das and M. Y. Chen, *Yahoo! for Amazon: Sentiment extraction from small talk on the web*, Manag. Sci., 53 (2007), pp. 1375–1388.

[39] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, *GoEmotions: A dataset of fine-grained emotions*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4040–4054.

[40] L. Deng, Y. Choi, and J. Wiebe, *Benefactive/malefactive event and writer attitude annotation*, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2013, pp. 120–125.

[41] L. Deng and J. Wiebe, *Sentiment propagation via implicature constraints*, in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 377–385.

[42] ——, *Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 179–189.

[43] ——, *MPQA 3.0: An entity/event-level sentiment corpus*, in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2015, pp. 1323–1328.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 4171–4186.

[45] J. Dewey, *The theory of emotion: I: Emotional attitudes*, Psychol. Rev., 1 (1894), pp. 553–569.

[46] H. Ding and E. Riloff, *Acquiring knowledge of affective events from blogs using label propagation*, in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2016, pp. 2935–2942.

[47] ——, *Weakly supervised induction of affective events by optimizing semantic consistency*, in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2018, pp. 5763–5770.

[48] C. Dong and U. Schäfer, *Ensemble-style self-training on citation classification*, in Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, 2011, pp. 623–631.

[49] P. Ekman, *Universals and cultural differences in facial expressions of emotion*, Nebr. Symp. Motiv., 19 (1972), pp. 207–283.

[50] ——, *An argument for basic emotions*, Cogn. Emot., 6 (1992), pp. 169–200.

[51] ——, *Basic emotions*, in Handbook of Cognition and Emotion, T. Dalgleish and M. J. Power, eds., John Wiley & Sons Ltd., New York, NY, 1999, pp. 45–60.

[52] A. Esuli and F. Sebastiani, *Determining term subjectivity and term orientation for opinion mining*, in 11th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2006, pp. 193–200.

[53] ——, *SENTIWORDNET: A publicly available lexical resource for opinion mining*, in Proceedings of the Fifth International Conference on Language Resources and Evaluation, European Language Resources Association, 2006, pp. 417–422.

[54] S. Feng, J. S. Kang, P. Kuznetsova, and Y. Choi, *Connotation lexicon: A dash of sentiment beneath the surface meaning*, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2013, pp. 1774–1784.

[55] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, *A survey of data augmentation approaches for NLP*, in Findings of the Association for Computational Linguistics, Association for Computational Linguistics, 2021, pp. 968–988.

[56] J. Friedenberg and G. Silverman, *Cognitive Science: An Introduction to the Study of Mind*, SAGE, Thousand Oaks, CA, 2005.

[57] N. H. Frijda, *The Emotions*, Cambridge University Press, Cambridge, UK, 1986.

[58] ——, *The laws of emotion*, Am. Psychol., 43 (1988), pp. 349–358.

[59] F. Gao, J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu, *Soft contextual data augmentation for neural machine translation*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 5539–5544.

[60] T. Gao, A. Fisch, and D. Chen, *Making pre-trained language models better few-shot learners*, in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 3816–3830.

[61] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, *Contextual inter-modal attention for multi-modal sentiment analysis*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 3454–3466.

[62] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, *DialogueGCN: A graph convolutional neural network for emotion recognition in conversation*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 154–164.

[63] A. B. Goldberg, X. Zhu, and S. Wright, *Dissimilarity in graph-based semi-supervised classification*, in Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, vol. 2, PMLR, 2007, pp. 155–162.

[64] Y. Goldberg and J. Orwant, *A dataset of syntactic-ngrams over time from a very large corpus of English books*, in 2nd Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, 2013, pp. 241–247.

[65] S. A. Goldman and Y. Zhou, *Enhancing supervised learning with unlabeled data*, in Proceedings of the 7th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 2000, pp. 327–334.

[66] A. S. Gordon and R. Swanson, *StoryUpgrade: Finding stories in internet weblogs*, in Proceedings of the International AAAI Conference on Web and Social Media, Association for the Advancement of Artificial Intelligence, 2008, pp. 188–189.

[67] A. Goyal, E. Riloff, and H. Daumé III, *A computational model for plot units*, Comput. Intell., 29 (2013), pp. 466–488.

[68] A. Goyal, E. Riloff, H. Daume III, and N. Gilbert, *Toward plot units: Automatic affect state analysis*, in Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, 2010, pp. 17–25.

[69] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, *A tutorial on graph-based SLAM*, IEEE Intell. Transp. Syst. Mag., 2 (2010), pp. 31–43.

[70] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du, *A question answering approach for emotion cause extraction*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 1593–1602.

[71] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou, *Event-driven emotion cause extraction with corpus construction*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 1639–1649.

[72] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, *Cognitive mapping and planning for visual navigation*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers, 2017, pp. 2616–2625.

[73] H. Hamdan, P. Bellot, and F. Bechet, *Lsislif: CRF and logistic regression for opinion target extraction and sentiment polarity analysis*, in Proceedings of the 9th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2015, pp. 753–758.

[74] Y. He, C. Lin, and H. Alani, *Automatically extracting polarity-bearing topics for cross-domain sentiment classification*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2011, pp. 123–131.

[75] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput., 9 (1997), pp. 1735–1780.

[76] E. Holderness, P. Cawkwell, K. Bolton, J. Pustejovsky, and M.-H. Hall, *Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records*, in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2019, pp. 117–123.

[77] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, *The curious case of neural text degeneration*, in 8th International Conference on Learning Representations, PMLR, 2020, pp. 1–12.

[78] R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko, *Are you looking? Grounding to multiple modalities in vision-and-language navigation*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 6551–6557.

[79] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou, and O. Zaïane, *Seq2Emo: A sequence to multi-label emotion classification model*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 4717–4724.

[80] A. Huettner and P. Subasic, *Fuzzy typing for document management*, in ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, Association for Computational Linguistics, 2000, pp. 26–27.

[81] C. W. Hughes, *Emotion: Theory, research and experience*, J. Nerv. Ment. Dis., 170 (1982), pp. 315–316.

[82] J. Islam, R. E. Mercer, and L. Xiao, *Multi-channel convolutional neural network for Twitter emotion and sentiment recognition*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 1355–1365.

[83] W. James, *What is an emotion?*, Mind, 9 (1884), pp. 188–205.

[84] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, *How can we know what language models know?*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 423–438.

[85] D. L. Johanson, H. S. Ahn, and E. Broadbent, *Improving interactions with healthcare robots: A review of communication behaviours in social and healthcare contexts*, Int. J. Soc. Robot., 13 (2021), pp. 1835–1850.

[86] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Prentice Hall, Upper Saddle River, NJ, 2009.

[87] H. Kang, S. Yoo, and D. Han, *Senti-lexicon and improved naïve Bayes algorithms for sentiment analysis of restaurant reviews*, Expert Syst. Appl., 39 (2012), pp. 6000–6010.

[88] J. S. Kang, S. Feng, L. Akoglu, and Y. Choi, *ConnotationWordNet: Learning connotation over the Word+Sense network*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 1544–1554.

[89] M. Kaufmann, *JMaxAlign: A maximum entropy parallel sentence alignment tool*, in Proceedings of COLING 2012: Demonstration Papers, Association for Computational Linguistics,, 2012, pp. 277–288.

[90] B. Kevin, J. Akshay, and S. Ian, *The ICWSM 2009 Spinn3r dataset*, in 3rd Annual Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, 2009.

[91] ——, *The ICWSM 2011 Spinn3r dataset*, in Proceedings of the Annual Conference on Weblogs and Social Media, AAAI, 2011.

[92] J. Khairnar and M. Kinikar, *Machine learning algorithms for opinion mining and sentiment classification*, Int. J. Sci. Res., 3 (2013), pp. 1–6.

[93] E. Kim and R. Klinger, *An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling*, in Proceedings of the 2nd Workshop on Storytelling, Association for Computational Linguistics, 2019, pp. 56–64.

[94] ——, *Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 647–653.

[95] Y. Kim, *Convolutional neural networks for sentence classification*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2014, pp. 1746–1751.

[96] S. Kiritchenko, S. Mohammad, and M. Salameh, *SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases*, in Proceedings of the 10th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2016, pp. 42–51.

[97] P. R. Kleinginna and A. M. Kleinginna, *A categorized list of emotion definitions, with suggestions for a consensual definition*, Motiv. Emot., 5 (1981), pp. 345–379.

[98] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, *Context based emotion recognition using emotic dataset*, IEEE Trans. Pattern Anal. Mach. Intell., 42 (2019), pp. 2755–2766.

[99] H. Kozima, M. P. Michalowski, and C. Nakagawa, *Keepon: A playful robot for research, therapy, and entertainment*, Int. J. Soc. Robot., 1 (2009), pp. 3–18.

[100] V. Kumar, A. Choudhary, and E. Cho, *Data augmentation using pre-trained transformer models*, in Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, Association for Computational Linguistics, 2020, pp. 18–26.

[101] G. Lakoff and M. Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, Basic Books, New York City, NY, 1999.

[102] C. Lange, *The Emotions*, Williams & Wilkins, Baltimore, MD, 1885/1922.

[103] S. M. LaValle, *Planning Algorithms*, Cambridge University Press, Cambridge, UK, 2006.

[104] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proc. IEEE, 86 (1998), pp. 2278–2324.

[105] W. G. Lehnert, *Plot units and narrative summarization*, Cogn. Sci., 5 (1981), pp. 293–331.

[106] J. Li, A. Ritter, C. Cardie, and E. Hovy, *Major life event extraction from Twitter based on congratulations/condolences speech acts*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2014, pp. 1997–2007.

[107] X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang, *A co-attention neural network model for emotion cause analysis with emotional context awareness*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 4752–4757.

[108] Y. Li, T. Cohn, and T. Baldwin, *Robust training under linguistic adversity*, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2017, pp. 21–27.

[109] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, *DailyDialog: A manually labelled multi-turn dialogue dataset*, in Proceedings of the Eighth International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, 2017, pp. 986–995.

[110] Y.-M. Li and T.-Y. Li, *Deriving marketing intelligence over microblogs*, in 2011 44th Hawaii International Conference on System Sciences, Institute of Electrical and Electronics Engineers, 2011, pp. 1–10.

[111] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, *What makes good in-context examples for GPT-3?*, in Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, 2022, pp. 100–114.

[112] X. Liu, P. He, W. Chen, and J. Gao, *Multi-task deep neural networks for natural language understanding*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 4487–4496.

[113] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A robustly optimized BERT pretraining approach*, in International Conference on Learning Representations, PMLR, 2020, pp. 1–15.

[114] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, *Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2022, pp. 8086–8098.

[115] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, *Learning word vectors for sentiment analysis*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2011, pp. 142–150.

[116] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, *Using linguistic cues for the automatic recognition of personality in conversation and text*, J. Artif. Int. Res., 30 (2007), pp. 457–500.

[117] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, *The Stanford CoreNLP natural language processing toolkit*, in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2014, pp. 55–60.

[118] B. Massumi, *Notes on the translation and acknowledgements*, in A Thousand Plateaus: Capitalism and Schizophrenia by G. Deleuze, University of Minnesota Press, Minneapolis, MN, 1987, pp. xvi–ix.

[119] D. McClosky, E. Charniak, and M. Johnson, *Effective self-training for parsing*, in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, Association for Computational Linguistics, 2006, pp. 152–159.

[120] R. Mihalcea, *Co-training and self-training for word sense disambiguation*, in Proceedings of the Eighth Conference on Computational Natural Language Learning at HLT-NAACL 2004, Association for Computational Linguistics, 2004, pp. 33–40.

[121] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013, pp. 1–9.

[122] G. A. Miller, *WordNet: A lexical database for English*, Commun. ACM, 38 (1995), pp. 39–41.

[123] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, and D. Kumaran, *Learning to navigate in complex environments*, in 5th International Conference on Learning Representations, PMLR, 2016, pp. 1–16.

[124] S. Mohammad, *A practical guide to sentiment annotation: Challenges and solutions*, in Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2016, pp. 174–179.

[125] S. Mohammad and F. Bravo-Marquez, *Emotion intensities in tweets*, in Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, 2017, pp. 65–77.

[126] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, *SemEval-2018 task 1: Affect in tweets*, in Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2018, pp. 1–17.

[127] S. Mohammad and P. Turney, *Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon*, in Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, 2010, pp. 26–34.

[128] T. K. Moon, *The expectation-maximization algorithm*, IEEE Signal Process. Mag., 13 (1996), pp. 47–60.

[129] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen, *A corpus and cloze evaluation for deeper understanding of commonsense stories*, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 839–849.

[130] T. Mullen and N. Collier, *Sentiment analysis using support vector machines with diverse information sources*, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2004, pp. 412–418.

[131] T. Mullen and R. Malouf, *A preliminary investigation into sentiment analysis of informal political discourse*, in AAAI Symposium on Computational Approaches to Analysing Weblogs, Association for the Advancement of Artificial Intelligence, 2006, pp. 159–162.

[132] M. Munezero, T. Kakkonen, and C. Montero, *Towards automatic detection of antisocial behavior from texts*, in Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, Asian Federation of Natural Language Processing, 2011, pp. 20–27.

[133] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, *Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text*, IEEE Trans. Affect. Comput., 5 (2014), pp. 101–111.

[134] N. Ng, K. Cho, and M. Ghassemi, *SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 1268–1283.

[135] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, *Facebook FAIR's WMT19 news translation task submission*, in Proceedings of the Fourth Conference on Machine Translation, Association for Computational Linguistics, 2019, pp. 314–319.

[136] T. H. Nguyen and K. Shirai, *PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 2509–2514.

[137] M. A. Nicolaou, H. Gunes, and M. Pantic, *Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space*, IEEE Trans. Affect. Comput., 2 (2011), pp. 92–105.

[138] P. M. Niedenthal, *Embodying emotion*, Science, 316 (2007), pp. 1002–1005.

[139] B. Pang, L. Lee, and S. Vaithyanathan, *Thumbs up? Sentiment classification using machine learning techniques*, in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002, pp. 79–86.

[140] P. Pantel and M. Pennacchiotti, *Espresso: Leveraging generic patterns for automatically harvesting semantic relations*, in Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2006, pp. 113–120.

[141] M. Pantic and L. J. M. Rothkrantz, *Expert system for automatic analysis of facial expressions*, Image Vis. Comput., 18 (2000), pp. 881–905.

[142] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*, Psychology Press, London, UK, 2001.

[143] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, *The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only*, in NeurIPS 2023 Datasets and Benchmarks, NeurIPS, 2023, pp. 1–18.

[144] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, *Psychological aspects of natural language use: Our words, our selves*, Annu. Rev. Psychol., 54 (2003), pp. 547–577.

[145] J. Pennington, R. Socher, and C. D. Manning, *GloVe: Global vectors for word representation*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2014, pp. 1532–1543.

[146] I. Perikos and I. Hatzilygeroudis, *Recognizing emotions in text using ensemble of classifiers*, Eng. Appl. Artif. Intell., 51 (2016), pp. 191–201.

[147] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, *Deep contextualized word representations*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2018, pp. 2227–2237.

[148] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, *Language models as knowledge bases?*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 2463–2473.

[149] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[150] R. Plutchik, *A general psychoevolutionary theory of emotion*, in Emotion: Theory, Research, and Experience, R. Plutchik and H. Kellerman, eds., Academic Press, London, UK, 1980, pp. 3–33.

[151] J. Prinz, *Embodied emotions*, in Thinking About Feeling: Contemporary Philosophers on Emotions, R. C. Solomon, ed., Oxford University Press, Oxford, UK, 2004, pp. 44–58.

[152] A. Qadir, E. Riloff, and M. Walker, *Learning to recognize affective polarity in similes*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 190–200.

[153] A. Qadir, E. Riloff, and M. A. Walker, *Automatically inferring implicit properties in similes*, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 1223–1232.

[154] G. Qin and J. Eisner, *Learning how to ask: Querying LMs with mixtures of soft prompts*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 5203–5212.

[155] C. Quan and F. Ren, *Construction of a blog emotion corpus for Chinese emotional expression analysis*, in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2009, pp. 1446–1454.

[156] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman, London, UK, 1985.

[157] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, preprint, Papers with Code, 2019.

[158] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi, *Event2Mind: Commonsense inference on events, intents, and reactions*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2018, pp. 463–473.

[159] H. Rashkin, S. Singh, and Y. Choi, *Connotation frames: A data-driven investigation*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2016, pp. 311–321.

[160] E. Riloff, J. Wiebe, and T. Wilson, *Learning subjective nouns using extraction pattern bootstrapping*, in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Association for Computational Linguistics, 2003, pp. 25–32.

[161] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, *EmpaTweet: Annotating and detecting emotions on Twitter*, in Proceedings of the Eighth International Conference on Language Resources and Evaluation, European Language Resources Association, 2012, pp. 3806–3813.

[162] S. Rosenthal, N. Farra, and P. Nakov, *SemEval-2017 task 4: Sentiment analysis in Twitter*, in Proceedings of the 11th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2017, pp. 502–518.

[163] D. E. Rumelhart, J. L. McClelland, and PDP Research Group, *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*, The MIT Press, Cambridge, MA, 1986.

[164] J. Russell, *A circumplex model of affect*, J. Pers. Soc. Psychol., 39 (1980), pp. 1161–1178.

[165] M. Sachan and E. Xing, *Self-training for jointly learning to ask and answer questions*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2018, pp. 629–640.

[166] J. Saito, Y. Murawaki, and S. Kurohashi, *Minimally supervised learning of affective events using discourse relations*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 5758–5765.

[167] P. Saraf, R. Sedamkar, and S. Rathi, *PrefixSpan algorithm for finding sequential pattern with various constraints*, Int. J. Appl. Inf. Syst., 9 (2015), pp. 37–41.

[168] S. Schachter and J. E. Singer, *Cognitive, social, and physiological determinants of emotional state*, Psychol. Rev., 69 (1962), pp. 379–399.

[169] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, *Experiences with Sparky, a social robot*, in Socially Intelligent Agents: Creating Relationships with Computers and Robots, K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, eds., Springer, New York, NY, 2002, pp. 173–180.

[170] K. R. Scherer, *Psychological models of emotion*, in The Neuropsychology of Emotion, J. C. Borod, ed., Oxford University Press, New York, NY, 2000, pp. 137–162.

[171] K. R. Scherer and H. G. Wallbott, *Evidence for universality and cultural variation of differential emotion response patterning*, J. Pers. Soc. Psychol., 66 (1994), pp. 310–328.

[172] T. Schick and H. Schütze, *Exploiting cloze-questions for few-shot text classification and natural language inference*, in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 255–269.

[173] ——, *It's not just size that matters: Small language models are also few-shot learners*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 2339–2352.

[174] K. Schindler, L. Van Gool, and B. De Gelder, *Recognizing emotions expressed by body pose: A biologically inspired neural model*, Neural Netw., 21 (2008), pp. 1238–1246.

[175] H. Scudder, *Probability of error of some adaptive pattern-recognition machines*, IEEE Trans. Inf. Theory, 11 (1965), pp. 363–371.

[176] R. Sennrich, B. Haddow, and A. Birch, *Improving neural machine translation models with monolingual data*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2016, pp. 86–96.

[177] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, *AutoPrompt: Eliciting knowledge from language models with automatically generated prompts*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 4222–4235.

[178] E. Shouse, *Feeling, emotion, affect*, M/C J., 8 (2005), 2443.

[179] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, *Recursive deep models for semantic compositionality over a sentiment treebank*, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 1631–1642.

[180] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, *Analysis of EEG signals and facial expressions for continuous emotion detection*, IEEE Trans. Affect. Comput., 7 (2015), pp. 17–28.

[181] C. Strapparava and R. Mihalcea, *SemEval-2007 task 14: Affective text*, in Proceedings of the Fourth International Workshop on Semantic Evaluations, Association for Computational Linguistics, 2007, pp. 70–74.

[182] C. Strapparava and R. Mihalcea, *Learning to identify emotions in text*, in Proceedings of the 2008 ACM Symposium on Applied Computing, Association for Computing Machinery, 2008, pp. 1556–1560.

[183] C. Strapparava and A. Valitutti, *WordNet affect: An affective extension of WordNet*, in Proceedings of the 4th International Conference on Language Resources and Evaluation, European Language Resources Association, 2004, pp. 1083–1086.

[184] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, *Lexicon-based methods for sentiment analysis*, Comput. Linguist., 37 (2011), pp. 267–307.

[185] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, *oLMpics-on what language model pre-training captures*, Trans. Assoc. Comput. Linguist., 8 (2020), pp. 743–758.

[186] D. Tang, B. Qin, X. Feng, and T. Liu, *Effective LSTMs for target-dependent sentiment classification*, in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Association for Computational Linguistics, 2016, pp. 3298–3307.

[187] Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: LIWC and computerized text analysis methods*, J. Lang. Soc., 29 (2010), pp. 24–54.

[188] M. Thelen and E. Riloff, *A bootstrapping method for learning semantic lexicons using extraction pattern contexts*, in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002, pp. 214–221.

[189] J. Thomason, M. Shridhar, Y. Bisk, C. Paxton, and L. Zettlemoyer, *Language grounding with 3D objects*, in Conference on Robot Learning, PMLR, 2022, pp. 1691–1701.

[190] S. S. Tomkins, *Affect, Imagery, Consciousness: The Negative Affects*, Springer, New York City, NY, 1962.

[191] ——, *Affect, Imagery, Consciousness: The Positive Affects*, Springer, New York City, NY, 1962.

[192] R. Tong, *An operational system for detecting and tracking opinions in on-line discussions*, in Proceedings of the SIGIR Workshop on Operational Text Classification, Association for Computing Machinery, 2001, pp. 1–6.

[193] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *LLaMA: Open and efficient foundation language models*, 2023, preprint, arXiv:2302.13971 [cs.CL].

[194] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023, preprint, arXiv:2307.09288 [cs.CL].

[195] A. Tripathy, A. Agrawal, and S. Rath, *Classification of sentimental reviews using machine learning techniques*, Procedia Comput. Sci., 57 (2015), pp. 821–829.

[196] E. Tromp and M. Pechenizkiy, *Rule-based emotion detection on social media: Putting tweets on Plutchik's wheel*, 2014, preprint, arXiv:1412.4682 [cs.CL].

[197] P. D. Turney, *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, 2002.

[198] A. Vanzo, D. Croce, and R. Basili, *A context-based model for sentiment analysis in Twitter*, in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, 2014, pp. 2345–2354.

[199] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, *Manifold Mixup: Better representations by interpolating hidden states*, in Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 6438–6447.

[200] H. T. Vu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, *Acquiring a dictionary of emotion-provoking events*, in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 128–132.

[201] X. Wan, *Co-training for cross-lingual sentiment classification*, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, 2009, pp. 235–243.

[202] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, *Self-consistency improves chain of thought reasoning in language models*, in The 11th International Conference on Learning Representations, PMLR, 2023, pp. 1–24.

[203] Y. Wang, M. Huang, X. Zhu, and L. Zhao, *Attention-based LSTM for aspect-level sentiment classification*, in Proceedings of the 2016 Conference on Empirical Methods

in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 606–615.

[204] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, *Finetuned language models are Zero-Shot learners*, in The 10th International Conference on Learning Representations, PMLR, 2021, pp. 1–46.

[205] J. Wei and K. Zou, *EDA: Easy data augmentation techniques for boosting performance on text classification tasks*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 6382–6388.

[206] J. Wiebe, T. Wilson, and M. Bell, *Identifying collocations for recognizing opinions*, in Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, Association for Computational Linguistics, 2001, pp. 24–31.

[207] J. Wiebe, T. Wilson, and C. Cardie, *Annotating expressions of opinions and emotions in language*, Lang. Resour. Eval., 39 (2005), pp. 165–210.

[208] T. Wilson, J. Wiebe, and P. Hoffmann, *Recognizing contextual polarity in phrase-level sentiment analysis*, in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 347–354.

[209] R. Xia and Z. Ding, *Emotion-cause pair extraction: A new task to emotion analysis in texts*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 1003–1012.

[210] S. Yadav, A. Ekbal, S. Saha, and P. Bhattacharyya, *Medical sentiment analysis using social media: Towards building a patient assisted system*, in Proceedings of the 11th International Conference on Language Resources and Evaluation, European Language Resources Association, 2018, pp. 2790–2797.

[211] B. Yang and C. Cardie, *Context-aware learning for sentence-level sentiment analysis with posterior regularization*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 325–335.

[212] C. Yang, K. H.-Y. Lin, and H.-H. Chen, *Emotion classification using web blog corpora*, in IEEE/WIC/ACM International Conference on Web Intelligence, IEEE, 2007, pp. 275–278.

[213] ——, *Emotion classification using web blog corpora*, in IEEE/WIC/ACM International Conference on Web Intelligence, Institute of Electrical and Electronics Engineers, 2007, pp. 275–278.

[214] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, *Generative data augmentation for commonsense reasoning*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1008–1025.

[215] D. Yarowsky, *Unsupervised word sense disambiguation rivaling supervised methods*, in The 33rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 1995, pp. 189–196.

[216] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, *Emotion recognition by textual tweets classification using voting classifier (LR-SGD)*, IEEE Access, 9 (2020), pp. 6286–6295.

[217] H. Yu and V. Hatzivassiloglou, *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*, in Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2003, pp. 129–136.

[218] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, *Multi-target embodied question answering*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers, 2019, pp. 6309–6318.

[219] N. Zainuddin and A. Selamat, *Sentiment analysis using support vector machine*, in 2014 International Conference on Computer, Communications, and Control Technology, Institute of Electrical and Electronics Engineers, 2014, pp. 333–337.

[220] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, *mixup: Beyond empirical risk minimization*, in 6th International Conference on Learning Representations, PMLR, 2017, pp. 1–13.

[221] L. Zhang and B. Liu, *Identifying noun product features that imply opinions*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2011, pp. 575–580.

[222] X. Zhang, J. J. Zhao, and Y. LeCun, *Character-level convolutional networks for text classification*, in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2015, pp. 1–9.

[223] D. Zhou, J. Wang, L. Zhang, and Y. He, *Implicit sentiment analysis with event-centered text representation*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 6884–6893.

[224] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, *Emotion distribution learning from texts*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 638–647.

[225] Y. Zhuang, T. Jiang, and E. Riloff, *Affective event classification with discourse-enhanced self-training*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 5608–5617.

[226] ——, *My heart skipped a beat! recognizing expressions of embodied emotion in natural language*, in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2024, pp. 3525–3537.

[227] Y. Zhuang and E. Riloff, *Eliciting affective events from language models by multiple view co-prompting*, in Findings of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 3189–3201.