Master's thesis

Prompt based zero shot emotion classification

Gita Tabe Jamaat

February 6, 2026

Department of computer science and electrical engineering

Kumamoto University

Abstract

In recent research works, emotion analysis has attracted attention in recent research works and has become a significant area in natural language processing. The objective of emotion classification is assigning emotions to textual units based on existing emotions. Within textual emotion classification, the set of relevant labels should be based on domains and applications.

When doing textual emotion classification, the set of labels might not be known at the time of model development. This conflicts with the method of supervised learning in which the labels need to be predefined. Zero-shot classification can assist us in doing classification without any training data by using pre-trained models.

One approach of zero shot classification can be textual entailment between texts and emotions. Therefore, it can make possible to do classification based on checking the entailment probability between the text and emotion labels by using pre-trained models and selecting the one that has the highest entailment probability as the predicted emotion. Since checking the textual entailment can be context dependent, one idea can be adding descriptions to the emotions which are called prompts. As an example, instead of using only emotion name joy, we can use "The emotion is joy". Moreover, to take advantage of using different terms of emotions, we can use different synonyms of emotions or prompts of synonyms. The synonyms of emotions or prompts of them have been used by different strategies. It can be based on getting average of their entailment probabilities or selecting the one that has the highest entailment probability. As another strategy, the results of two mentioned methods have been combined.

The experiments have been done on two datasets including the labels and texts columns using 3 pre-trained models to check the entailment of texts and emotion labels or prompts of emotions or their synonyms. We can see that using prompt of emotions or synonyms of emotions can sometimes lead to improvement.

Then, the results have been analyzed to check the effects of prompts or using synonyms of emotions. Based on checking the entailment probability for some sample sentences, we can see that the entailment probabilities of true and wrong predictions are close to each other when the sentences are ambiguous and prediction is wrong. The entailment probability is usually high for sentences that their prediction is correct when using prompts of emotion names or synonyms of them.

Acknowledgement

During my master program, I could be provided with an opportunity to enhance my knowledge in Data science and improve my research skills in Natural language processing while doing master thesis and attending courses.

I would like to thank my supervisor, Professor Masayoshi Aritsugi for his invaluable guidance and encouragement in completing my research.

I would also like to thank our lab Assistant Professor Thanda Shew and Associate Professor Israel Mendonca for their guidance and support during my research.

I am grateful to my family for their support to pursue my education.

I am also grateful to my classmates to provide friendly learning environment.

Finally, I am grateful to graduate school of science and technology for their support in providing course instruction and graduate research environment.

Table of contents

Chapter 1

Introduction

The research works on emotion classification have different perspectives. However, they usually need predefined labels for training. Therefore, one solution to solve this problem is zero shot classification [12]. Zero shot classification can make possible to predict classes that have not been seen during training. So, one approach can be checking the entailment of texts and emotions by using Bert based models. To improve the performance, it's possible to take advantage of natural language inference by creating representations of emotions which are called prompts. Also, in addition to prompts of only emotion names, it's possible to use prompts of synonyms of emotions. We can use synonyms of emotions with different strategies such as getting average of their entailment probabilities or selecting the maximum of them or combining the results of two methods.

1.1 Problem statement

There are different approaches for emotion classification. These works have some disadvantages because of their training on datasets with specific labels [1]. So, one idea can be zero shot classification as a sequence pair classifier. The idea of zero-shot classification can solve this problem by considering the emotion as a text and designing classifiers that can do the classification based on textual entailment of texts and emotions [2]. Also, to improve the classification, it's possible to add descriptions at the beginning of the emotion which are called prompts or involve synonyms of emotions by using different strategies.

1.2 Research questions

The goal of this research is to investigate the role of natural language inference in zero shot emotion classification and investigating how it can affect the context of emotions when using pre-trained models to check the entailment of the text and emotions to do classification. Also, it was investigated that whether using synonyms of emotions can give rise to improvement and the results of using different prompts and synonyms have been analyzed. Therefore, the following research questions should be answered.

1. Can adding the templates at the beginning of the emotions affect the context of the emotions and for which cases it can lead to improvement?
2. Which model can lead to more improvement?
3. How would it be better to involve the synonyms of the emotion?
4. Whether using the synonyms of the emotions can lead to improvements?
5. Whether combining the results of using synonyms can lead to any improvements?

Thus finding the answers of these questions can assist us in finding a better perspective and solution regarding the concepts that can be significant to consider in zero shot classification.

1.3 Thesis contribution

In recent studies on emotion classification, the idea of zero shot classification has been investigated based on textual entailment by using pre-trained models [2, 3]. In this research work, we have done more studies on using prompts and more strategies to use synonyms of emotions. We have analyzed the results based on some statistics and some samples of results to find out the effects of prompts and synonyms of emotions deeply when doing textual emotion classification using pre-trained models.

1.4 Thesis orientation

In the rest of thesis, in chapter 2, we have discussed the significant concepts and related works to emotion classification. In chapter 3, we have described the components and ideas of research method for zero shot emotion classification. In chapter 4, we have defined experiment setting and we have shown the obtained results and findings. In chapter 5, we have analyzed the results based on checking prediction and entailment probability for some sample sentences and statistics of results. In addition, we have stated some limitations of this work. In chapter 6, we have conclusion of this research and future work.

Chapter 2

Background and related works

2.1 Background

2.1.1 Emotion classification

Emotion analysis is one of the significant areas in natural language processing. The goal of the emotion classification is at mapping textual units to emotions. The emotions can be determined based on existing emotions in the datasets [14].

2.1.2 Zero shot learning

The goal of zero shot learning is determining classes that have not been seen during training. One approach for zero-shot learning can be sequence pair classification [12]. Therefore, the model takes two sentences as inputs and decide whether they contradict or entail to each other. Therefore, it makes possible to check the textual entailment by using pre-trained models and do classification based on checking entailment probability between text and emotions.

2.1.3 Transformer models

It is a kind of neural network models for understanding sequences and patterns within data based on multi head mechanism. So the models have different components such as self-attention, parallel processing and positional encoding that allow the model to understand the context of data.

The structure of transformers is based on encoder-decoder architecture. The encoder uses positional encoding to calculate the numerical representations of data and decoder to classify data and generate output by identifying words and phrases and identify the most important parts by attention mechanism. The positional encoding is the method of understanding patterns in data. So, the model breaks down the text to the tokens and assigns numerical representations to tokens. As a result, the model can understand the relation of tokens to each other. Another component is self-attention which helps the model to determine which parts are important and which parts don't affect the final output. It can understand how the grammatical structure affects the whole meaning of the sentences [10].

2.1.4 Applications of transformer models

One of the significant applications of transformer models is in natural language processing. Transformer models can understand the patterns in foundation of texts and sequences [10]. Their applications are in tasks such as text summarization, text generation and sentiment analysis.

2.1.5 Natural language inference for zero shot learning

The idea of natural language inference can be adding templates at the beginning of the emotions or using different terms of emotions which are called prompts [6]. So, the task can be defined as sentence pair classification and check whether they are compatible to each other or not.

2.2 Related work

The previous studies have shown the importance of prompts to enhance zero shot emotion classification by using different pre-trained models and different prompt types for checking textual entailment between texts and emotions [1, 2]. However, these studies have limitations in finding the effects of prompts on classification clearly and choosing proper prompts.

In this research, the effects of prompts on zero shot classification have been analyzed based on obtained results of experiments. Also, some more strategies regarding using prompts and synonyms of emotions have been exploited. Then, the results have been analyzed to check the effects of using synonyms of emotions on classification based on checking some statistics of results and textual entailment for some samples.

Chapter 3

Research methods

In this chapter, the components of zero shot classification have been explained. First, the natural language task has been explained to do classification based on textual entailment. Then the idea of prompt generation and ensemble method haves been explained. After that, the idea of using synonyms has been explained and the method for evaluation of results has been stated.

3.1 Natural language inference task

The task can be defined as sentence pair classification and can be a function $f_{(s1,s2)} \rightarrow \{C, N, E\}$ that takes two sentences (premise s1 and hypothesis s2) and the models return 3 scores C,N, E in which C refers to contradiction of the sentences, N refers to neutral output and E refers to entailment of sentences [12]. So, based on considering only the contradiction and entailment scores and converting them into probabilities, the entailment probability can be obtained.

Therefore, if the function takes the text to be classified and each of the emotions, we can obtain entailment probabilities of emotions and select the emotion that has the highest entailment probability as the predicted emotion. Figure 1 illustrates an example regarding this task.

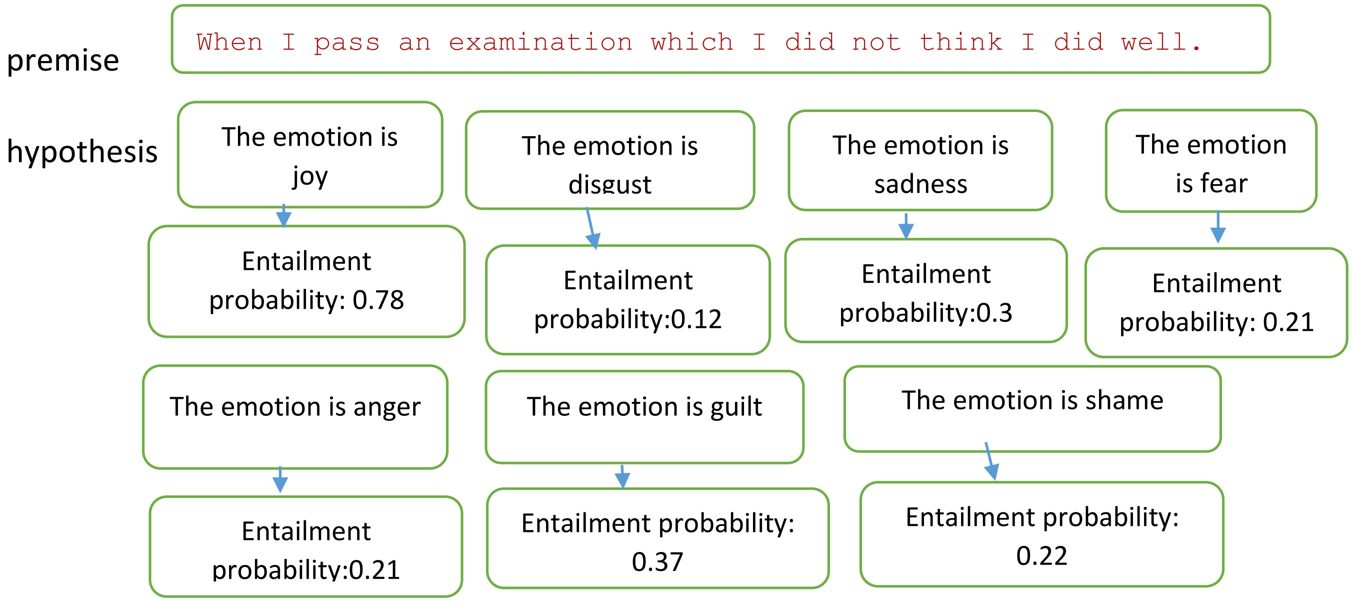
premise
When I pass an examination which I did not think I did well.
hypothesis
The emotion is joy
Entailment probability: 0.78
The emotion is disgust
Entailment probability:0.12
The emotion is sadness
Entailment probability:0.3
The emotion is fear
Entailment probability: 0.21
The emotion is anger
Entailment probability:0.21
The emotion is guilt
Entailment probability: 0.37
The emotion is shame
Entailment probability: 0.22

Figure 1: An example that shows natural language inference task

In figure 1, we can see that the emotion joy has the highest entailment probability and can be selected as the predicted emotion.

3.2 Prompt generation

Since the emotions can be considered as texts, descriptions can be appended at the beginning of emotions which are called prompts [12]. In this work, 4 prompts have been used.

So, the set of prompts are generated using function g(e):

Equation1: $g(e)= c+ r(e)$

Where e represents the emotion, c represents context and r(e) represents the set of emotion representations. In table 1, we can see the types of prompts and examples that have been used in this research work.

Table 1: prompts of emotions

ID	Prompt	example
Emo_name prompt	The emotion name	Joy
Expr_emo prompt	"This text expresses"+ emotion name	This text expresses joy
Feels_emo prompt	"This person feels"+ emotion name	This person feels joyful
Emotion prompt	"The emotion is"+ emotion name	The emotion is joy

3.3 Ensemble method

This method combines the results of using different prompts by getting average of entailment probabilities of emotions in results of different prompts as the entailment probability of the emotion. Since the prompts have different performance, it can generalize the method by combining the results of all of the prompts [9].

3.4 Using synonyms in prompts of emotions

We can replace synonyms of the emotions in the template of the prompts [2]. Table 2 illustrates the prompts of synonyms. (The details of emotion representations for each prompt of synonyms are in appendix A).

Table 2: prompts of synonyms of emotions

ID	Prompt	example
Emo_s prompt	The synonyms	happy
Expr_s prompt	"This text expresses"+ emotion synonym	This text expresses happiness
Feels_s prompt	"This person feels" +emotion synonym	This person feels happy
Emotion-s prompt	"The emotion is" + emotion synonym	The emotion is happiness

So, when considering different synonyms of emotions, it aims to check the entailment of the prompts of different synonyms and text and get average of their entailment probabilities with the text as the entailment probability of the emotion and then consider the emotion that has the highest entailment probability as the predicted emotion.

Another method is setting the maximum of entailment probabilities of the synonyms of emotions with the text as the entailment probability of the emotion and then considering the emotion that has the highest entailment probability as the predicted emotion.

As another idea, we can combine the results of using average of entailment probabilities and using maximum of entailment probabilities and get average of the entailment probabilities of each emotion in two results as the entailment probability of the emotion.

Table 3 illustrates the synonyms that are considered for each emotion.

Table 3: The synonyms of emotions

emotion	synonyms
sadness	sadness, unhappiness, grief, sorrow, loneliness, depression
joy	joy, an achievement, pleasure, the awesome, happiness, the blessing
anger	anger, annoyance, rage, outrage, fury, irritation
disgust	disgust, loathing, bitterness, ugliness, repugnance, revulsion
fear	fear, horror, anxiety, terror, dread, scare
surprise	surprise, astonishment, amazement, impression, perplexity, shock
shame	shame, humiliation, embarrassment, disgrace, dishonor, discredit
guilt	guilt, culpability, responsibility, blameworthy, misconduct, regret

Chapter 4

Experiments and Results

4.1 Datasets

In this research, two public datasets have been used for experiments:

Isear dataset: It includes 7515 data and 7 emotion labels: anger, disgust, fear, guilt, joy, sadness, shame. [5]

Tec dataset: It includes 21051 data and 6 emotion labels: anger, disgust, fear, joy, sadness, surprise. [11]

4.2 Preprocessing of datasets

To improve the quality of datasets for classification, the datasets have been preprocessed before doing classification. So, some signs such as ; , :, *, # , = , (,), & and integers have been removed.

4.3 Models

3 Bert_based models that have been fine-tuned on NLI datasets and are publicly available in hugging face transformers python library [4, 10] have been used for the experiments. The models have different architectures.

Deberta: Decoding enhanced Bert model that uses disentangled attention and enhanced mask decoder. The model microsoft/deberta-v2-xlarge-mnli from hugging face library is used. It contains over 750M parameters [8].

Bart: It is a kind of Bidirectional and autoregressive transformer models. The model facebook/bart-large-mnli from hugging face library is used. It contains over 407M parameters [7].

Roberta: The Roberta model is a version of Bert model that replaces static masking with Dynamic masking and doesn’t include the task of next sentence prediction. The roberta-large-mnli model from hugging face library is used which contains over 355M parameters [13].

4.4 The method of evaluation of results

The performance of predictions has been evaluated using Macro average F1 score and accuracy. The accuracy measures the percentage of correct predictions by the model. The F1 score is based on both precision and recall. So, it is a more reliable metric. The metric macro average F1 is considered for the whole dataset. It is calculated by getting average of F1 scores of predicted classes. These metrics can be calculated based on the predicted class and true class. In the following, there are formulas of these metrics.

Equation 2: accuracy= (TP+TN) /(TP+FP+TN+FN)

Equation 3: precision= TP/(TP+FP)

Equation 4: recall= TP/(TP+FN)

Equation 5: F1 score= 2*precision*recall/ (precision + recall)

Equation 6: Macro average F1= The average of F1 scores of classes

TP: True positive

FP: False positive

TN: True negative

FN: False negative

For evaluation of results, we can check Macro average F1 and accuracy of results when using prompts of emotions or synonyms of emotions. Therefore, we can see in which cases, using prompts can lead to improvements based on comparing Macro average F1. To analyze the results regarding labels, we can get the F1 scores of labels for the cases that there were improvements to them.

To compare the range of entailment probabilities of true predictions and wrong predictions, we can calculate the average of entailment probabilities of true emotion when the prediction is true and when the prediction is wrong.

4.5 The results when using prompts of emotions and ensemble method

In this section, we have shown the results of using prompts of emotions for two datasets. Then, we have obtained some statistics regarding labels.

After that, we have shown the results of ensemble method and compared its performance with the performance of the best individual result of prompts.

4.5.1 The results of using prompts of emotions

Tables 4, 5, 6 and 7 illustrate the results of Isear dataset and Tec dataset. Considering results, we can see that the highest macro average F1 is when using emotion prompt or expr_emo prompt and Deberta model for Isear dataset. The highest macro average F1 is when using emotion name and Deberta model for Tec dataset.

For the first dataset, we can see that expr_emo prompt has the highest macro average F1 when using 3 models and feels_emo prompt has the lowest Macro average F1 when using 3 models. Also for both of the datasets, we can see that Deberta model usually could lead to higher Macro average F1 compared to other models in the result of each model.

Table 4: result for Isear and Tec datasets, using emotion name

	Deberta	Bart	Roberta
Macro average F1 (Isear dataset)	0.54	0.57	0.51
accuracy(Isear dataset)	0.57	0.57	0.51
Macro averaged F1(Tec dataset)	0.41	0.37	0.34
accuracy(Tec dataset)	0.46	0.41	0.43

Table 5: result for Isear and Tec datasets, using emotion prompt

	Deberta	Bart	Roberta
Macro average F1 (Isear dataset)	0.61	0.57	0.5
accuracy (Isear dataset)	0.62	0.57	0.5
Macro average F1 (Tec dataset)	0.4	0.37	0.4
accuracy (Tec dataset)	0.42	0.4	0.42

Table 6: result for Isear and Tec datasets, using expr_emo prompt

	Deberta	Bart	Roerta
Macro average F1 (Isear dataset)	0.61	0.61	0.51
Accuracy (Isear dataset)	0.63	0.61	0.51
Macro average F1 (Tec dataset)	0.37	0.36	0.34
Accuracy (Tec dataset)	0.4	0.4	0.37

Table 7: result for Isear and Tec datasets, using Feels_emo prompt

	Deberta	Bart	Roberta
Macro averaged F1 (Isear dataset)	0.54	0.53	0.48
accuracy (Isear dataset)	0.55	0.53	0.48
Macro averaged F1 (Tec dataset)	0.38	0.36	0.35
accuracy (Tec dataset)	0.4	0.38	0.38

4.5.2 Results regarding labels

In this section, the F1 scores of labels for the cases that using prompt could lead to improvement to them have been obtained. Therefore, we can see that for which labels there are more improvements when using prompt.

The cases that using prompt could lead to improvement compared to using only emotion name for Isear dataset are as the following:

When using Deberta model:

- Emotion prompt: “The emotion is” + emotion
- Expr_emo: “This text expresses” + emotion

When using Bart model:

- Expr_emo: “This text expresses” + emotion

Table 8 illustrates the F1 score for the labels when using only emotion name and Deberta model.

Table 8: results regarding labels (Isear dataset, emotion name, Deberta model)

emotion	anger	disgust	fear	guilt	joy	sadness	Shame
F1 score	0.47	0.46	0.65	0.51	0.9	0.57	0.2

Tables 9 and 10 illustrate F1 score for the labels in the cases that there were improvements for them. We can see that the significant improvements are for labels disgust, shame when using emotion prompt or expr_emo prompt and Deberta model. So, using two different templates could lead to the improvements for the same labels when using Deberta model.

Table 9: results regarding labels (Isear dataset, emotion prompt, Deberta model)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
F1 score	0.48	0.58	0.74	0.55	0.91	0.6	0.43

Table 10: results regarding labels (Isear dataset, Expre_emo prompt, Deberta)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
F1 score	0.51	0.58	0.74	0.54	0.91	0.66	0.34

Table 11 illustrates F1 scores for labels when using Bart model and emotion name.

Table 11: results regarding labels (Isear dataset, emotion name, Bart model)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
F1 score	0.4	0.53	0.69	0.48	0.85	0.63	0.46

Table 12 illustrates the F1 scores for labels when using Bart model and expr_emo prompts

Table 12: results regarding labels (Isear dataset, expr_emo prompt, Bart model)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
F1 score	0.49	0.6	0.75	0.45	0.86	0.64	0.49

Considering tables 11 and 12, the most improvement is for label anger.

The cases that using prompts could lead to improvements for Tec dataset, compared to using only emotion name is as the following:

When using Roberta model:

- Emotion prompt: “The emotion is” + emotion

We can compare the F1 score for the labels in the cases that there have been improvements for them are:

Table 13: Results regarding labels (emotion name, Roberta, Tec dataset)

emotion	anger	disgust	fear	joy	sadness	surprise
F1 score	0.3	0.2	0.34	0.54	0.47	0.38

Table 14: Results regarding labels (emotion prompt, Roberta, Tec dataset)

emotion	anger	disgust	fear	joy	sadness	surprise
F1 score	0.34	0.22	0.39	0.5	0.44	0.38

The more improvement is for label fear compared to other labels.

4.5.3 The results of ensemble method

The ensemble method combines the results of different prompts based on getting average of entailment probabilities of emotions in the results of different prompts as the entailment probability of the emotion.

Table 15 illustrates the results of using ensemble method is as the following:

Table 15: Results, ensemble method

	accuracy	Macro average F1
Deberta (Isear dataset)	0.61	0.59
Roberta (Isear dataset)	0.52	0.52
Bart (Isear dataset)	0.61	0.61
Deberta (Tec dataset)	0.44	0.41
Roberta (Tec dataset)	0.41	0.37
Bart (Tec dataset)	0.41	0.38

Tables 16 illustrates the result regarding the best individual result of prompts. Based on the results, we can see that the ensemble model had the performance nearly equivalent to the best individual method result (the best individual result of prompts).

Table 16: Result, best individual result

	accuracy	Macro average F1
Deberta (Isear dataset)	0.63	0.61
Roberta (Isear dataset)	0.51	0.51
Bart (Isear dataset)	0.61	0.61
Deberta (Tec dataset)	0.46	0.41
Roberta (Tec dataset)	0.42	0.4
Bart (Tec dataset)	0.41	0.37

4.6 The results when using prompts of synonyms of emotions

Since the Deberta model had the better performance compared to other models in last experiments, the experiments regarding using synonyms of emotions have been done using Deberta model.

4.6.1 Using average of entailment probabilities of the synonyms

In this method, the average of entailment probabilities of the synonyms for each emotion is calculated as the entailment probability of the emotion. In table 17, we can see the results for two datasets. Based on the results, we can see that using average of entailment probabilities of synonyms of the emotion could lead to a little improvement for all 4 prompts for Isear dataset. (emo-s: Macro average F1: 0.57 compared to emo-name: Macro average F1: 0.54, expr-s: Macro average F1: 0.62 compared to expr-emo: Macro average F1: 0.61, emotion-s: Macro average F1: 0.63 compared to emotion prompt: Macro average F1: 0.61, feels-s: Macro average F1: 0.56 compared to feels-emo: Macro average F1: 0.54). However, it couldn’t lead to improvements for Tec dataset. It can be because of characteristics of datasets. The texts of Tec dataset are more consistent with emotion names, but using different terms of emotion could improve the entailment of texts and emotions for Isear dataset.

Table 17: Results when using average of entailment probabilities between synonyms of the emotion and text (Isear and Tec datasets, Deberta model)

	Emo_s	Expr_s	Emotion_s	Feels_s
Macro average F1 (Isear dataset)	0.57	0.62	0.63	0.56
Accuracy (Isear dataset)	0.58	0.63	0.63	0.59
Macro average F1 (Tec dataset)	0.32	0.36	0.38	0.36
accuracy(Tec dataset)	0.36	0.41	0.41	0.41

4.6.2 Using maximum of entailment probabilities of synonyms

The second method is setting maximum of entailment probabilities of synonyms of the emotion as the entailment probability of the emotion. Table 18 illustrates the results. Using maximum of entailment probabilities of synonyms could lead to improvement just for feels_emo prompt for Isear dataset compared to using prompts of only emotion names. (Feels_s: Macro average F1: 0.58 to Feels_emo: Macro average F1: 0.54). Thus, considering only stronger synonym cannot give any help.

Table 18: Results when using maximum of entailment probabilities between synonyms of the emotion and text (Isear and Tec datasets, Deberta model)

	Emo_s	Expr_s	Emotion_s	Feels_s
Macro average F1 (Isear dataset)	0.52	0.58	0.6	0.58
Accuracy (Isear dataset)	0.53	0.59	0.6	0.59
Macro average F1 (Tec dataset)	0.37	0.29	0.35	0.34
Accuracy (Tec dataset)	0.42	0.31	0.39	0.38

4.6.3 Combining the results of two methods

As another method we have combined the result of using average of entailment probabilities of synonyms and the result of using maximum of entailment probabilities of synonyms by getting average of entailment probabilities of the emotion in two results. Since the emotion-s prompt had better performance in last experiments, the results of using emotion-s prompt have been combined.

Isear dataset: Macro average F1: 0.64, accuracy:0.65

Tec dataset: Macro average F1: 0.37, accuracy:0.41

Its performance is similar to the performance of using average of entailment probabilities of synonyms, but it could lead to a little improvement for the first dataset.

4.6.4 Results regarding labels

Since emotion-s prompt had better performance compared to other prompts, results regarding labels for emotion-s prompt have been obtained. Therefore, we can see the improvements regarding labels.

Table 19 illustrates the results regarding labels for Isear dataset. The F1 score for some emotions like joy is higher, but for emotions like shame and anger are lower when using 3 mentioned methods.

Table 19: Results regarding labels (Isear dataset, emotion-s prompt, Deberta model)

emotion	anger	Disgust	fear	guilt	joy	sadness	shame
F1 score (when using average of entailment probabilities of synonyms)	0.51	0.52	0.75	0.51	0.91	0.73	0.49
F1 score (when using max of entailment probabilities of synonyms)	0.49	0.53	0.72	0.5	0.86	0.65	0.45
F1 score (when combining two methods)	0.5	0.56	0.72	0.58	0.89	0.72	0.52

Tables 20 illustrates the results regarding labels for Tec dataset. The F1 score for the emotion joy is higher and for emotion disgust is the lowest.

Table 20: Results regarding labels (Tec dataset, emotion-s prompt, Deberta model)

emotion	anger	disgust	fear	joy	sadness	surprise
F1 score (when using average of entailment probabilities)	0.37	0.22	0.39	0.51	0.38	0.38
F1 score (when using maximum of entailment probabilities of synonyms)	0.29	0.26	0.37	0.48	0.36	0.37
F1 score (when combining two methods)	0.35	0.23	0.39	0.5	0.39	0.39

4.6.5 The average of entailment probabilities for each emotion class

To compare the range of entailment probabilities for the cases that prediction is true or wrong, the average of entailment probabilities for each emotion class have been calculated when using average of entailment probabilities of synonyms of emotions and maximum of entailment probabilities of synonyms. Tables 21-24 illustrate the results for two datasets. We can see that the range of average of entailment probabilities is higher for each emotion when the prediction is true.

Table 21: The average of entailment probabilities for each emotion class (Isear dataset, when using average of entailment probabilities of synonyms, emotion-s prompt)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
average (true prediction)	0.87	0.9	0.9	0.86	0.83	0.88	0.91
average(wrong prediction)	0.71	0.62	0.68	0.75	0.35	0.6	0.45

Table 22: The average of entailment probabilities for each emotion class (Tec dataset, when using average of entailment probabilities of synonyms, emotion-s prompt)

Emotion	anger	Disgust	fear	Joy	sadness	surprise
average (true prediction)	0.84	0.86	0.61	0.75	0.79	0.77
average (wrong prediction)	0.47	0.52	0.31	0.23	0.39	0.57

Table 23: The average of entailment probabilities for each emotion class (Isear dataset, when using maximum of entailment probabilities of synonyms of emotions, emotion-s prompt)

emotion	anger	disgust	fear	guilt	joy	sadness	shame
average (true prediction)	0.97	0.97	0.97	0.96	0.95	0.97	0.97
average (wrong prediction)	0.88	0.81	0.85	0.89	0.72	0.8	0.82

Table 24: The average of entailment probabilities for each emotion class (Tec dataset, when using maximum of entailment probabilities of synonyms of emotions, emo-s prompt)

emotion	anger	disgust	fear	joy	sadness	surprise
average (true prediction)	0.86	0.71	0.81	0.79	0.82	0.75
average (wrong prediction)	0.61	0.54	0.4	0.3	0.49	0.6

Chapter 5

Analysis of results and limitations

In this chapter, at first we have checked the prediction by using different prompts for some sample sentences and we have checked how the contents of the sentences are related to the predicted emotions.

After comparing the predictions for some sample sentences when using prompts of emotions, we have checked the effects of prompts on the context of emotions for some sample sentences.

Then, we have analyzed the results based on checking the entailment probabilities for some sample sentences and statistics of results.

5.1 Comparing the prediction for some sample sentences

In this section, we have compared the prediction for some samples to see how the contents of sentences and predicted emotions are related to each other when using prompts.

In tables 25-29, we have seen the comparison of predictions of some sample sentences while using different prompts.

Based on the results in table 25 and 26, we can see that there are some specific tokens in the sentences that are related to the true labels. (like hurt related to anger, lie related to shame, broke related to anger, frustrating related to anger), but prediction by feels_emo prompt can also be reasonable.

Table 25: Some sample sentences that using expr_emo could lead to improvement but feels_emo could not (Isear dataset)

True label	Text	Expr emo	Feels emo
anger	When friends try to put me down or hurt me.	anger	Fear
Shame	Somebody who knows me very well discovered that I had told him a lie.	Shame	Sadness

Table 26: Some sample sentences that using emotion prompt could lead to improvement but feels_emo could not (Tec dataset)

True label	Text	Emotion prompt	Feels emo
anger	I broke my glasses in half	anger	sadness
anger	It's really frustrating when your professors don't email back or show up for their office hours.	anger	fear

In tables 27 and 28, we can see that there are some tokens that are obviously related to the true emotions. (like death related to fear, guilty related to guilt, cheer related to joy, awful related to sadness). Therefore, the prediction by using two different prompts are true.

Table 27: Some sample sentences that both of expr_emo and feels_emo could lead to true prediction to them. (Isear dataset)

True label	Text	Expr emo	Feels emo
fear	Every time I imagine that someone I love or I could contact a serious illness, even death.	fear	fear
guilt	I feel guilty when I realize that I consider material things more important than caring for my relatives. I feel very self-centered.	Guilt	guilt

Table 28: Some sample sentences that both of emotion prompt and feels_emo could lead to true prediction to them. (Tec dataset)

True label	Text	Emotion prompt	Feels emo
joy	the moment when you get another follower and you cheer.	joy	joy
sadness	sounds awful but alot of people are dying recently ((	sadness	sadness

As a result, we can see that for some sentences that their contents are ambiguous, prediction by one of the prompts is true, but prediction by another prompt can also be reasonable. But, for sentences that their contents don't have ambiguity, there are some tokens that are clearly related with their associated emotions.

Table 29: Some sample sentences that using synonyms of emotions could lead to improvement but using prompts of emotion names couldn't

True label	text	Predicted emotion by using prompts of synonyms of emotions	Predicted emotion by using prompts of emotion names
anger	When a car is overtaking another and I am forced to drive off the road.	anger (using emo_s prompt)	fear (using emo_name prompt)
shame	I lied, to be precise I cancelled a meeting with a good friend.	shame (using expr_s prompt)	guilt (using expr_emo prompt)
fear	In a cottage in a large forest, I was alone for a while in the dark.	fear (using emotion_s prompt)	Sadness (using emotion prompt)

Considering table 29, using synonyms of emotions could lead to true predictions for some sample sentences that their contents are ambiguous.

5.2 The effects of prompts on the context of emotions

In the following, there are some sample sentences to compare the entailment probability of emotions with the text before and after using prompts.

Sentence: I heard part of a conversation in which one talked very low about women. (from Isear dataset)

Table 30: Before and after using prompts, Deberta model

emotion	disgust	sadness	anger	fear	guilt	shame	joy
Entailment probability (before using prompts)	0.66	0.76	0.49	0.52	0.44	0.36	0.0028
Entailment probability(after using prompts)	0.83	0.5	0.62	0.32	0.38	0.64	0.0008

Considering table 30, we can see that after using prompt, the entailment probability of the true emotion disgust would be the highest.

Sentence: I realized when I have to wake up from a nap before I want I'm cranky as hell! frustration lol (from Tec dataset, using Roberta model)

Table 31: Before and after using prompt: entailment probability, using Roberta model

emotion	disgust	sadness	anger	fear	surprise	joy
entailment probability(before using prompt)	0.98	0.72	0.97	0.37	0.47	0.017
entailment probability (after using prompt)	0.33	0.07	0.97	0.06	0.1	0.0007

Considering table 31, the entailment probability of the true emotion, anger would be the highest after using prompt.

Therefore, we can see that using prompts can affect the context of the emotion and lead to true predictions.

5.3 Comparing the entailment probability for some sample sentences

In this section, we have compared the entailment probabilities of true and predicted labels while using synonyms of emotions.

In the following, there are sample sentences that the true label and predicted label are different.

When using average of entailment probabilities of synonyms:

- When my partner was attacked and lost three teeth. (from Isear dataset)
  When using emotion-s prompt: True Label: anger, entailment probability: 0.84, predicted label: fear, entailment probability: 0.90

- I hate shopping with my mother (from Tec dataset)
  When using emotion-s prompt: True Label: anger, entailment probability: 0.91, predicted label: disgust, entailment probability: 0.96

When using maximum of entailment probabilities of synonyms:

- In a cottage in a large forest, I was alone for a while in the dark. (from Isear dataset)

  When using emotion-s prompt:

  True label: fear, entailment probability: 0.92,

  predicted label: sadness, entailment probability: 0.98

- I received my parcel trashed. It proves it was teared from backside and thing was removed. cheapservice (from Tec dataset)

  When using emo-s prompt:

  True label: anger, entailment probability: 0.82,

  predicted label: surprise, entailment probability: 0.94

Regarding the above samples, we can see that the sentences are ambiguous and the entailment probabilities of true and predicted labels are close to each other.

In the following, we can see sample sentences that the true label and predicted label are the same.

When using average of entailment probabilities of synonyms:

- Don't believe the lies, look me in the eyes- please don't be scared of me (from Tec dataset, emotion prompt)
  Predicted and true label: fear, entailment probability: 0.6
- When I pass an examination which I did not think I did well. (from Isear dataset, emotion prompt)
  Predicted and true label: joy, entailment probability:0.8

When using maximum of entailment probabilities of synonyms:

- Every time I imagine that someone I love or I could contact a serious illness, even death. (from Isear dataset)

  When using emotion_s prompt:

  Predicted label and true label: fear

  Entailment probability: 0.95

- Be the greatest dancer of your life! practice daily positive habits. fun freedom habits (from Tec dataset)

  When using emotion_s prompt:

  Predicted label and true label: joy, Entailment probability: 0.87

We can see that the entailment probabilities of the true emotions are usually high for these sentences.

5.4 analyzing the results regarding labels

The labels that are mistaken with each other have been checked when using average of entailment probabilities of synonyms and the following findings have been obtained.

For Isear dataset, the labels are mostly mistaken with guilt:

- joy: mostly mistaken with guilt (49 data/7515 data)
- anger: mostly mistaken with guilt (258 data/7515 data)
- disgust: mostly mistaken with guilt (191 data/7515 data)
- fear: mostly mistaken with guilt (103 data/7515 data)
- guilt: mostly mistaken with shame (116 data/7515 data)
- sadness: mostly mistaken with guilt (101 data/7515 data)
- shame: mostly mistaken with guilt (347 data/7515 data)

For Tec dataset, the labels are mostly mistaken with surprise.

- anger: mostly mistaken with surprise (346 data/21051 data)
- disgust: mostly mistaken with surprise (236 data/21051 data)
- fear: mostly mistaken with surprise (798 data/21051 data)
- joy: mostly mistaken with surprise (2830 data/21051 data)
- sadness: mostly mistaken with surprise (1466 data/21051 data)
- surprise: mostly mistaken with joy (947 data/21051 data)

So, for Isear dataset, the emotion guilt lead to many wrong predictions and for Tec dataset the emotion surprise lead to many wrong predictions.

Table 32 indicates the average of entailment probabilities of synonyms of the predicted emotion when the predicted emotion is wrong and it is guilt in Isear dataset.

Table 32: The average of entailment probabilities of synonyms of the emotion guilt when prediction is wrong

synonym	guilt	culpability	responsibility	blameworthy	misconduct	regret
average of entailment probabilities	0.67	0.84	0.81	0.62	0.78	0.64

Therefore, the synonyms culpability and responsibility have the highest average of entailment probabilities and lead to many wrong predictions in Isear dataset.

Table 33 indicates the average of entailment probabilities of synonyms of the predicted emotion when the predicted emotion is wrong and it is surprise in Tec dataset.

Table 33: The average of entailment probabilities of synonyms of the emotion surprise when prediction is wrong

synonym	surprise	astonishment	amazement	impression	perplexity	shock
average of entailment probabilities	0.62	0.54	0.5	0.77	0.57	0.42

So, the synonym impression has the highest average of entailment probabilities and lead to many wrong predictions in Tec dataset.

5.5 The average of entailment probabilities for synonyms of emotions

We have obtained the following findings when getting average of entailment probabilities of synonyms of the true emotion with texts of both of datasets (The details of results are in appendix B).

For Isear dataset, the following findings have been obtained.

- anger: The synonym irritation has the highest average of entailment probabilities (0.91).
  The synonym rage has the lowest average of entailment probabilities (0.67).

- disgust: The synonym repugnance has the highest average of entailment probabilities (0.87).
  The synonym ugliness has the lowest average of entailment probabilities (0.62).

- fear: The synonym scare has the highest average of entailment probabilities (0.92).
  The synonym horror has the lowest average of entailment probabilities (0.76).

- guilt: The synonym culpability has the highest average of entailment probabilities (0.91).
  The synonym responsibility has the lowest average of entailment probabilities (0.73).

- joy: The synonyms happiness and pleasure have the highest average of entailment probabilities (0.87, 0.85).
  The synonym the blessing has the lowest average of entailment probabilities (0.67).

- sadness: The synonyms sadness and unhappiness have the highest average of entailment probabilities (0.89).
  The synonym loneliness and depression have the lowest average of entailment probabilities (0.66, 0.67).

- Shame: The synonym embarrassment has the highest average of entailment probabilities (0.81).
  The synonym disgrace has the lowest average of entailment probabilities (0.73).

For Tec dataset, the following findings have been obtained.

- anger: The synonym irritation has the highest average of entailment probabilities (0.79).
  The synonym rage has the lowest average of entailment probabilities (0.57).

- disgust: The synonym repugnance has the highest average of entailment probabilities (0.78).
  The synonym ugliness has the lowest average of entailment probabilities (0.44).

- fear: The synonym scare has the highest average of entailment probabilities (0.59).
  The synonym horror has the lowest average of entailment probabilities (0.37)

- Joy: The synonyms pleasure and happiness have the highest average of entailment probabilities (0.59).

  The synonyms the blessing and an achievement have the lowest average of entailment probabilities (0.36).

- Sadness: The synonym unhappiness has the highest average of entailment probabilities (0.63).
  The synonyms depression and loneliness have the lowest average of entailment probabilities (0.39, 0.4).

- Surprise: The synonym surprise has the highest average of entailment probabilities (0.83).
  The synonym shock has the lowest average of entailment probabilities (0.56).

So, for both of datasets, for each emotion, the synonyms that have the highest entailment probabilities when checking the entailment of texts with synonyms of their true emotions are:

- sadness: unhappiness
- joy: happiness, pleasure
- anger: irritation

- disgust: repugnance
- fear: scare
- surprise: surprise
- shame: embarrassment
- guilt: culpability

Therefore, these synonyms have more effects on true predictions.

Considering only these synonyms as the representations of the emotions and doing experiments based on checking the entailment probability by using only these synonyms of emotions couldn't lead to improvements compared to last experiments. (Isear dataset: Macro average F1: 0.59, accuracy:0.6, Tec dataset: Macro average F1:0.36, accuracy:0.4). It can be because of the difference in complexity of sentences and considering only the set of synonyms cannot work well for the whole dataset.

The synonyms that have the lower average of entailment probabilities compared to other synonyms when checking the entailment of texts with synonyms of their true emotions in two datasets are:

- anger: rage
- disgust: ugliness
- fear: horror
- guilt: responsibility
- joy: the blessing
- sadness: loneliness, depression
- shame: disgrace
- surprise: shock

Therefore, these synonyms have the less effects on predicting true emotions.

Removing these synonyms from the set of synonyms and getting average of the entailment probabilities of the remaining synonyms had similar performance to the result of using the average of entailment probabilities of all the synonyms. (Isear dataset: Macro average F1: 0.63, accuracy: 0.63, Tec dataset: Macro average F1: 0.38, accuracy: 0.41)

As a result, we can see that considering only synonyms with the highest average of entailment probabilities or removing synonyms with the lowest average of entailment probabilities cannot lead to improvement.

5.6 Limitations of this work

There are some limitations regarding this research. The prompts that have been used in this work are based on some sample prompt types in recent studies [3]. Therefore, we have seen that it can sometimes lead to improvement. Consequently, the more developed prompts may lead to different or improved results.

The texts of datasets that we have used in this research work had so variety in complexity and length of sentences. In addition, we have seen, some of them were also ambiguous. If datasets with more consistent textual characteristics were used, different results might be obtained, and synonym-based methods could lead to more improvements.

Chapter 6

Conclusion and future works

6.1 Conclusion

The goal of this research was doing zero shot classification by using emotion prompts and synonyms of emotions as natural language inference and analyzing the results.
In this research, the experiments have been done using 3 Bert based models. Then, the experiments have been done using different prompts by adding template at the beginning of emotions. We saw that using prompt sometimes could lead to improvements because it can affect the context of emotion word and change the entailment probability between the text and emotions.

Based on analyzing the results, we have seen that there are some tokens related to the emotions for the cases that using prompts could lead to improvements.
Since using different prompts had different performance, to make a general method, we combined the results of different prompts by getting average of entailment probabilities of emotions in different results.

Regarding pre-trained models, we saw that the Deberta model had better performance compared to Roberta and Bart models. As a result, since Deberta model had better performance, the experiments of using synonyms have been done using Deberta model.
We have seen that using average of entailment probabilities of synonyms of emotions could lead to improvements, but selecting maximum of entailment probabilities of synonyms doesn’t usually lead to improvements. As a result, involving all of the synonyms could have better performance.

Based on analyzing some samples of the result, we could see that the entailment probabilities of true emotion and predicted emotion are close to each other for the sentences that have ambiguity in their content. Also, the entailment probability of the predicted emotion is usually high when the prediction is true.

Since the emotions-s prompt had better performance compared to other prompts, the results of two mentioned methods of emotion-s prompts have been combined to increase the effect of stronger synonym while involving all of the synonyms. Combining the results of two mentioned methods could lead to little improvement for one of the datasets, but it couldn’t lead to improvement for another dataset.

We saw that the average of entailment probabilities of some synonyms of true emotions are higher in both of the datasets. Therefore, they have more effects in true predictions. Also, the average of entailment probabilities of some synonyms of true emotions are lower compared to other synonyms in two datasets. However, considering only the synonyms with highest average of entailment probabilities or removing synonyms that have the lowest entailment probability couldn’t lead to improvement.

6.2 Future work

We have seen that using prompts sometimes could lead to improvements. Therefore, a future work regarding this research can be designing a more guided prompt based on the features of the datasets.

We also saw that using synonyms could lead to a little improvement for one dataset. Therefore, as a future work, we can look for a more guided synonym adaptive method that selects the most proper synonyms for each text of dataset based on its characteristics.

References

[1] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 270 280, Beijing, China. Association for Computational Linguistics.

[2] Bareiss Patrick; Klinger Roman; Barnes Jeremy. 2024: English Prompts are Better for NLI-based Zero-Shot Emotion Classification than Target-Language Prompts, in: WWW '24: Companion Proceedings of the ACM on Web Conference 2024, New York,

[3] Flor Miriam Plaza-del-Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistic

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[5] Klaus R. Scherer and Harald G. Wallbott. 1997. The ISEAR Questionnaire and Codebook. Geneva Emotion Research Group

[6] Luana Bulla, Aldo Gangemi, and Misael Mongiovi'. 2023. Towards Distribution-shift Robust Text Classification of Emotional Content. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8256–8268, Toronto, Canada. Association for Computational Linguistics.

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Mar jan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART:Denoising Sequence-to-Sequence Pre training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding enhanced BERT with Disentangled Attention. In International Conference on Learning Representations.

[9] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional Emotion Detection from Categorical Emotion. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[10] Quentin Lhoest, and Alexander Rush. 2020. "Transformers: State-of-the-Art Natural Language Processing". In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics.

[11] Saif Mohammad. 2012. # Emotional tweets. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Work shop on Semantic Evaluation (SemEval2012), pages 246–255

[12] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach

[14] Paul Ekman. 1992. An argument for basic emotions. Cognition and Emotion, 6(3-4):169–200.

Appendix A

List of emotion representations when using prompts of synonyms

	Emo-s	Expr-s	Emotion-s	Feels-s
anger	anger, annoyance, rage, outrage, fury, irritation	anger, annoyance, rage, outrage, fury, irritation	anger, annoyance, rage, outrage, fury, irritation	anger, annoyed, rage, outraged, furious, irritated
fear	fear, horror, anxiety, terror, dread, scare	fear, horror, anxiety, terror, dread, scare	fear, horror, anxiety, terror, dread, scare	fear, horror, anxiety, terrified, dread, scared
joy	joy, achievement, pleasure, awesome, happy, blessed	joy, an achievement, pleasure, the awesome, happiness, the blessing	joy, an achievement, pleasure, the awesome, happiness, the blessing	joyful, accomplished, pleasure, awesome, happy, blessed
sadness	sadness, unhappy, grief, sorrow, loneliness, depression	sadness, unhappiness, grief, sorrow, loneliness, depression	sadness, unhappiness, grief, sorrow, loneliness, depression	sadness, unhappy, grieved, sorrow, lonely, depression
disgust	disgust, loathing, bitter, ugly, repugnance, revulsion	disgust, loathing, bitterness, ugliness, repugnance, revulsion	disgust, loathing, bitterness, ugliness, repugnance, revulsion	disgusted, loathing, bitter, ugly, repugnance, revulsion
guilt	guilt, culpability, blameworthy, responsibility, misconduct, regret	guilt, culpability, responsibility, blameworthy, misconduct, regrets	guilt, culpability, responsibility, blameworthy, misconduct, regrets	guilty, culpable, responsible, blame, misconduct, regretful
shame	shame, humiliate,	shame, humiliation,	shame, humiliation,	shameful, humiliated,

	embarrassment, disgrace, dishonor, discredit	embarrassment, disgrace, dishonor, discredit	embarrassment, disgrace, dishonor, discredit	embarrassed, disgraced, dishonored, discredit
surprise	surprise, astonishment, amazement, impression, perplexity, shock	surprise, astonishment, amazement, impression, perplexity, shock	surprise, astonishment, amazement, impression, perplexity, shock	surprised, astonishment, amazement, impressed, perplexed, shocked

Appendix B

The average of entailment probabilities of synonyms of true emotions with texts of Isear dataset:

anger:

Synonym	anger	annoyance	Rage	Outrage	fury	irritation
The average of entailment probabilities	0.83	0.82	0.67	0.79	0.7	0.91

disgust:

synonym	disgust	loathing	bitterness	ugliness	repugnance	revulsion
The average of entailment probabilities	0.78	0.67	0.68	0.62	0.87	0.75

fear:

synonym	fear	horror	anxiety	terror	dread	scare
The average of entailment probabilities	0.9	0.76	0.87	0.81	0.83	0.92

guilt:

synonym	guilt	culpability	responsibility	blameworthy	misconduct	regrets
The average of	0.85	0.91	0.73	0.85	0.8	0.79

entailment probabilities						

joy:

synonym	joy	an achievement	pleasure	the awesome	happiness	the blessing
The average of entailment probabilities	0.83	0.72	0.85	0.75	0.87	0.67

sadness:

synonym	sadness	unhappiness	grief	sorrow	loneliness	depression
The average of entailment probabilities	0.89	0.89	0.82	0.86	0.66	0.67

shame:

synonym	shame	humiliation	embarrassment	disgrace	dishonor	discredit
The average of entailment probabilities	0.77	0.8	0.81	0.73	0.79	0.76

The average of entailment probabilities of synonyms of the true emotions with texts of Tec dataset:

anger:

synonym	anger	annoyance	Rage	Outrage	fury	irritation
The average of entailment probabilities	0.68	0.71	0.57	0.63	0.58	0.79

disgust:

synonym	disgust	loathing	bitterness	ugliness	repugnance	revulsion
The average of entailment probabilities	0.71	0.61	0.62	0.44	0.78	0.68

fear:

synonym	fear	horror	anxiety	terror	dread	scare
The average of entailment probabilities	0.54	0.37	0.53	0.38	0.45	0.59

joy:

synonym	joy	an achievement	pleasure	the awesome	happiness	the blessing
The average of entailment probabilities	0.44	0.36	0.51	0.48	0.5	0.36

sadness:

synonym	sadness	unhappiness	grief	sorrow	loneliness	depression
The average of entailment probabilities	0.58	0.63	0.5	0.52	0.4	0.39

surprise:

synonym	surprise	astonishment	amazement	impression	perplexity	shock
The average of entailment probabilities	0.82	0.7	0.65	0.74	0.64	0.56