SentiCSE : A Sentiment Representation Learning Framework and a Sentiment-guided Textual Similarity Task for Accurate Sentiment Analysis in a Few-shot Setting

Jaemin Kim[1,2*], Yohan Na[1*], Kangmin Kim[1,3], Sangrak Lee[1] and Dongkyu Chae[1†]

[1]Hanyang University

[2]LG Electronics, [3]BHSN

{jaemink, nayohan, kevin7133, sangrak, dongkyu}@hanyang.ac.kr

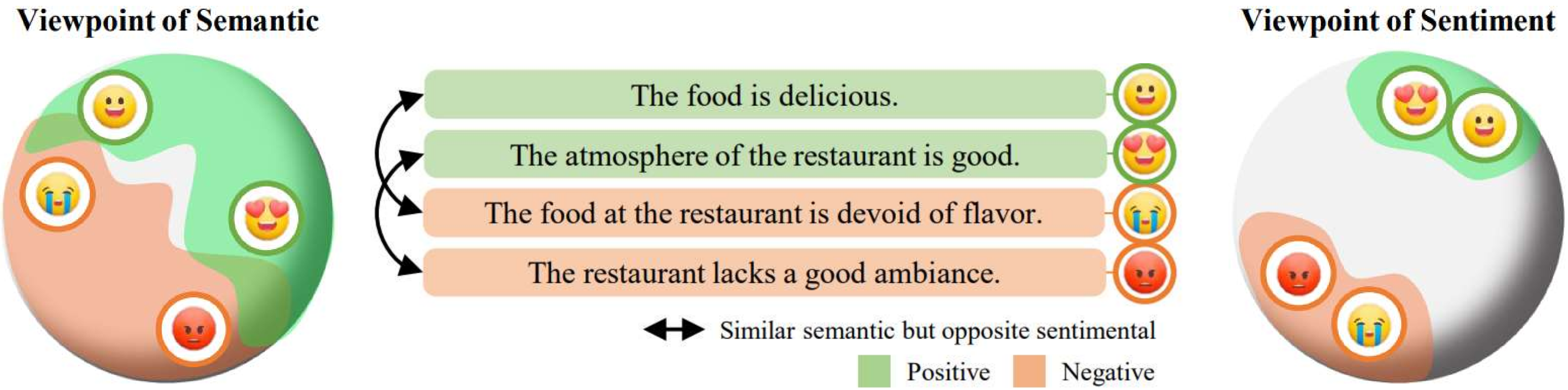
*Co-first authors

†Corresponding author.

Contents

Overview

- The viewpoint of semantic representation and sentiment representation should be differentiated.
  - ✓ Previous studies have neglected the evaluation of sentiment representation quality.
  - ✓ For achieving strong performance in few-shot learning, it is essential to focus on representation learning.
  - ✓ This paper presents **SgTS**, a new task to measure sentiment representation quality
  - ✓ Additionally, we introduce **SentiCSE**, a framework for learning sentiment-focused representations.



Background

❖How is Sentiment Analysis Used in the Real World?

Sentiment Analysis

- ✓ Sentiment analysis is now widely used in various industries.[1, 2]
- ✓ The growth of social media has significantly increased the importance of sentiment analysis.[3]

Brand reputation

Movie reviews

Politics

Employee satisfaction

[1] Zhang et al., AAAI

[2] Yu and Jiang., EMNLP

[3] Yadav and Vishwakarma., AIR, 2020

Background

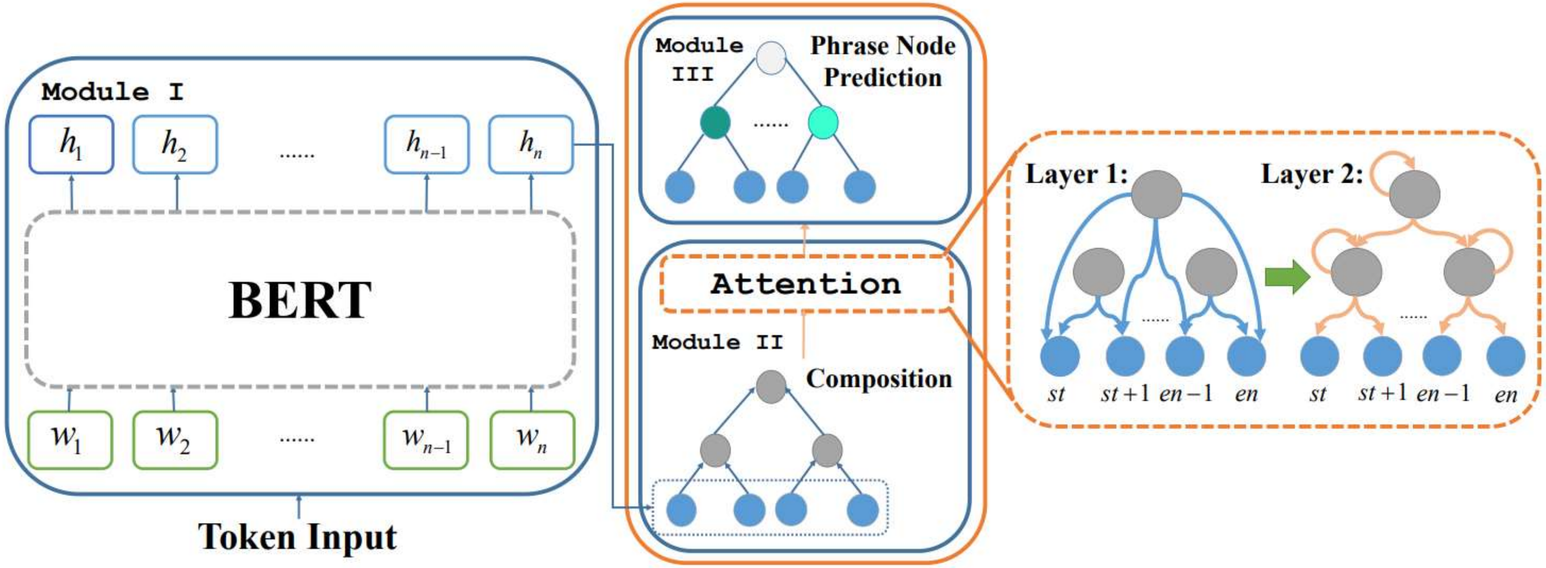
❖ SentiBERT : A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics

Yin, Da et al., ACL

Objectives

✓ Masked Language Modeling : To enable the model to capture contextual information effectively.

✓ Phrase Node Prediction : To capture the compositional sentiment semantics



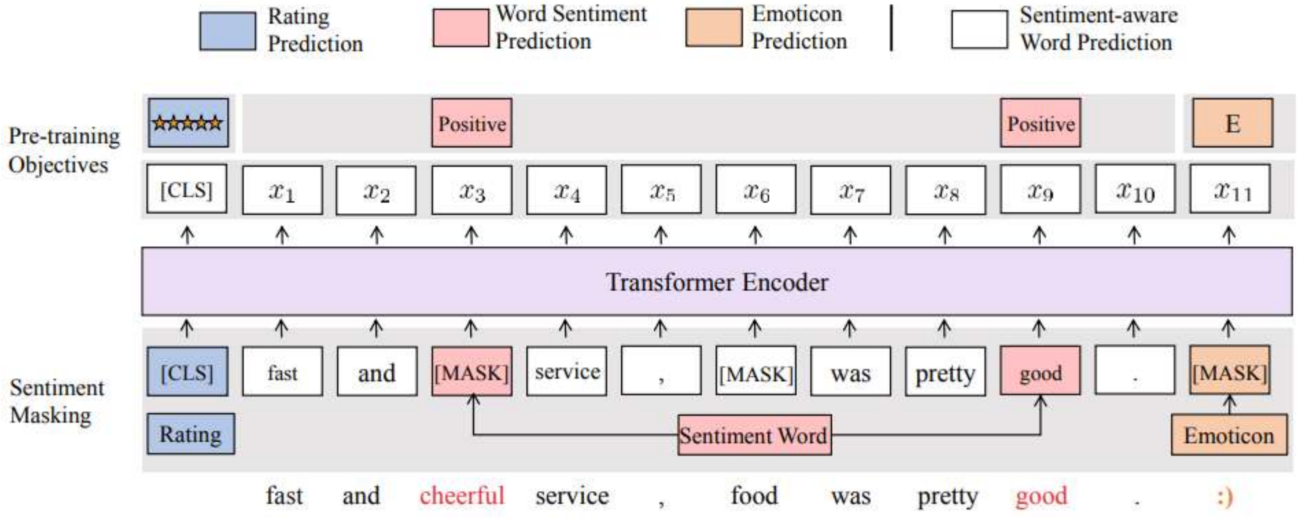
Background

❖ SENTIX: A Sentiment-Aware Pre-Trained Model for Cross-Domain Sentiment Analysis

J Zhou et al., COLING

Objectives

- ✓ Sentiment –aware Word Prediction (SWP) : Similar to Masked Language Modeling (MLM) by masking sentiment words
- ✓ Word Sentiment Prediction (WSP) : Predicting the sentiment polarity of words.
- ✓ Emoticon Prediction (EP) : Similar to Masked Language Modeling (MLM) by masking emoticons
- ✓ Rating Prediction (RP) : Predicting the sentiment polarity rating of sentences.



Background

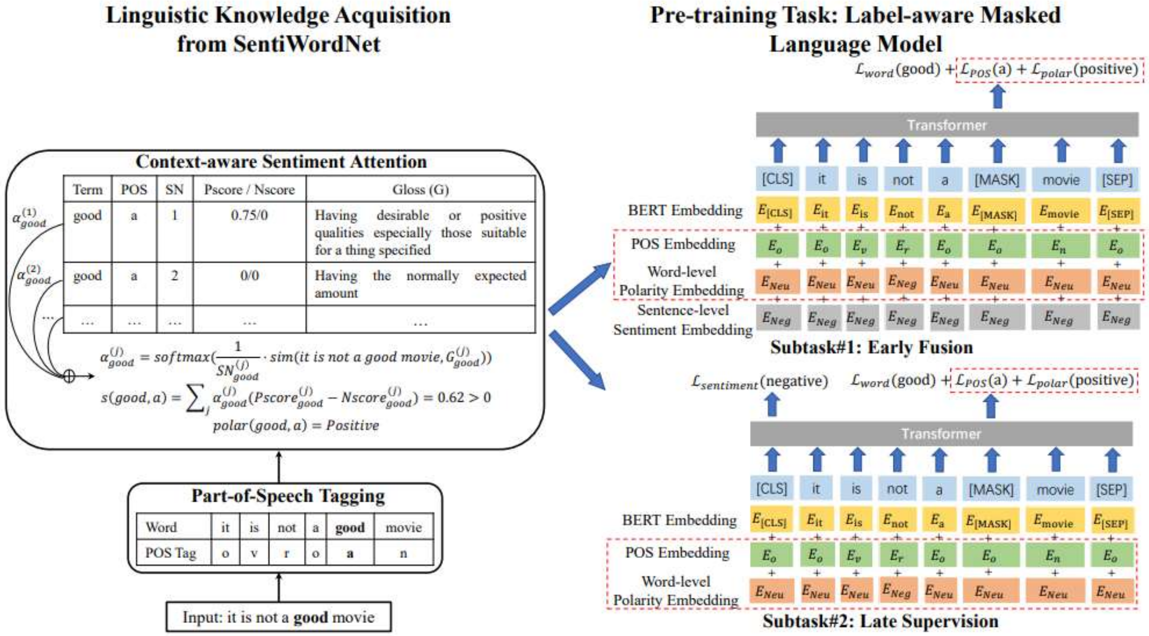
❖ SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge

Ke et al., EMNLP

Objectives

✓ Words-level : Predicting the emotional polarity of masked words by incorporating the part-of-speech (POS) tag information.

✓ Sentences-level : Predicting the sentiment polarity of a sentence.



Background

❖ How is Sentiment Analysis Used in the Real World?

Sentiment Analysis

- ✓ Obtaining labeled training data for each domain requires labor and time costs.[4]
- ✓ A Sentiment-aware Pre-trained Language Model (PLM) capable of robust performance in a few-shot setting is needed.

Brand reputation

Movie reviews

Politics

Employee satisfaction

Train dataset | Train dataset | Train dataset | Train dataset

[4] Socher et al., EMNLP, 2013

Background

Solution

❖ **Sentiment Representation Learning**

✓ If the representation is of sufficiently good quality, it can perform well with only a few samples[5].

[5] DU, Simon S., et al., ICLR

Background

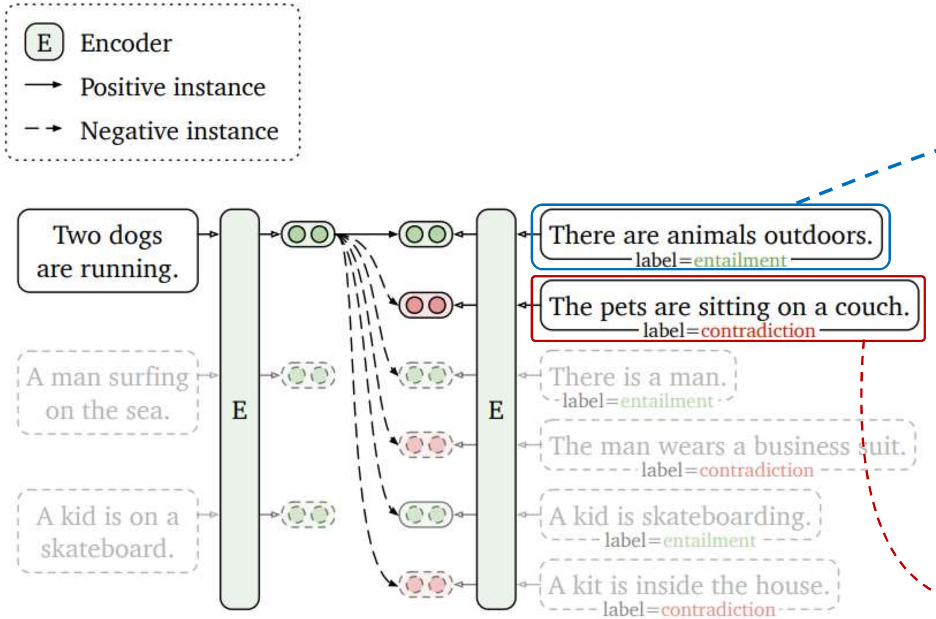
❖ Simple Contrastive Learning of Sentence Embedding (SimCSE)

T Gao et al., EMNLP, 2021 (Cited 1,867 times)

Contrastive Learning

✓ Positive pairs = entailment (premise, hypothesis) pairs

✓ Negative pairs = contradiction (premise, hypothesis) pairs + in-batch negatives



h_i : Premise
h_i^+ : Entailment
h_i^- : Contradiction

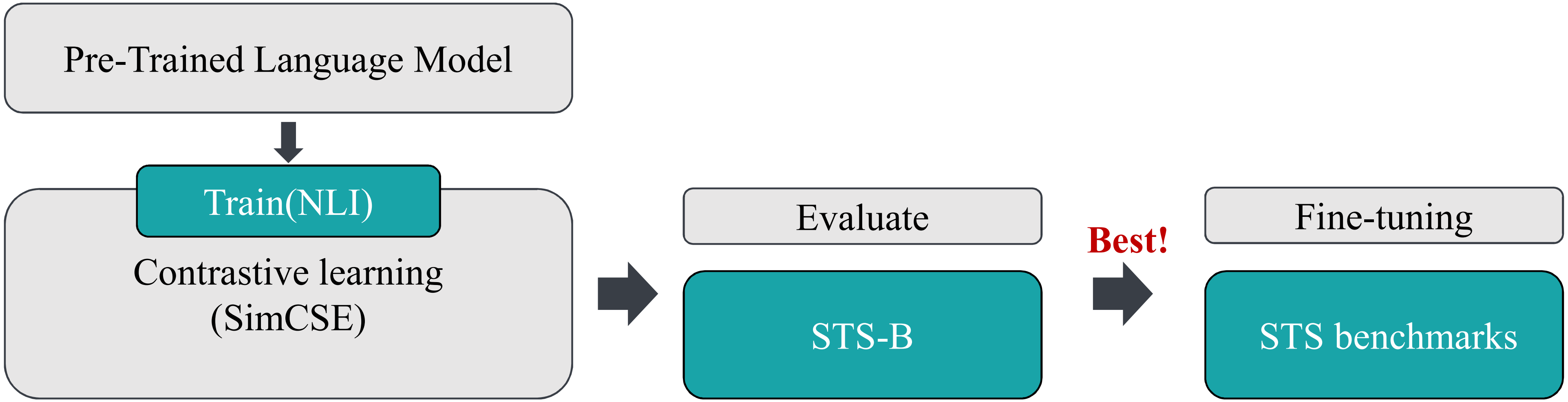
$$Loss_i = -log \frac{\underbrace{\boxed{e^{sim\left(h_i\ ,h_i^+\right)/\tau}}}_{\text{Positive pair}}}{\sum_{j=1}^{N}\Big(\underbrace{e^{sim\left(h_i\ ,h_j^+\right)/\tau}}_{\text{in-batch negative pair}} + \underbrace{\boxed{\alpha^{\mathbb{1}_i^j} \cdot e^{sim\left(h_i\ ,h_j^-\right)/\tau}}}_{\text{Hard negative pair}} \Big)}$$

Background

❖ Simple Contrastive Learning of Sentence Embedding (SimCSE)

T Gao et al., EMNLP, 2021 (Cited 1,867 times)

Contrastive learning



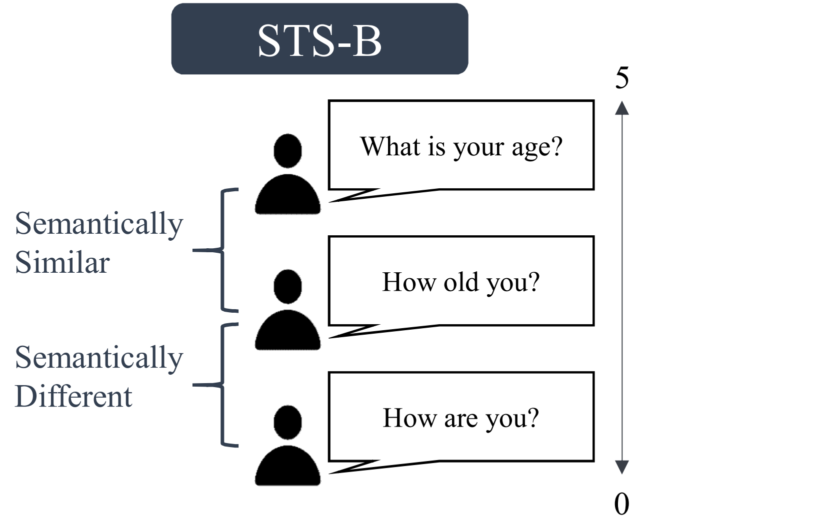
Background

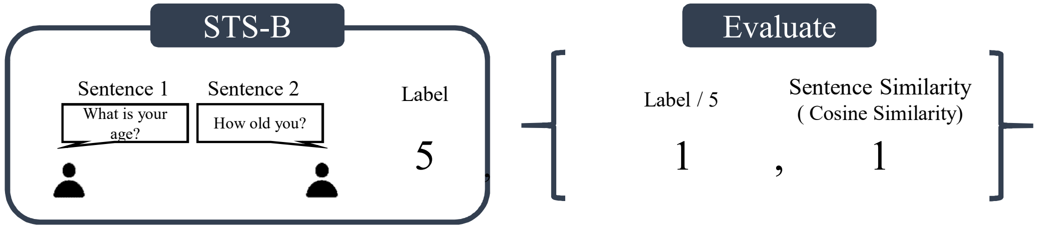
❖ Simple Contrastive Learning of Sentence Embedding (SimCSE)

T Gao et al., EMNLP, 2021 (Cited 1,867 times)

Evaluate on Semantic textual similarity tasks

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				Unsupervised models				
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT$_{base}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT$_{base}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT$_{base}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT$_{base}$♡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT$_{base}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
∗ SimCSE-BERT$_{base}$	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
RoBERTa$_{base}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
RoBERTa$_{base}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-RoBERTa$_{base}$	52.41	75.19	65.52	77.12	78.63	72.41	**68.62**	69.99
∗ SimCSE-RoBERTa$_{base}$	**70.16**	**81.77**	**73.24**	**81.36**	**80.65**	**80.22**	68.56	**76.57**
∗ SimCSE-RoBERTa$_{large}$	**72.86**	**83.99**	**75.62**	**84.77**	**81.80**	**81.98**	**71.26**	**78.90**
				Supervised models				
InferSent-GloVe♣	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder♣	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT$_{base}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT$_{base}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT$_{base}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT$_{base}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
∗ SimCSE-BERT$_{base}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
SRoBERTa$_{base}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa$_{base}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ SimCSE-RoBERTa$_{base}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ SimCSE-RoBERTa$_{large}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**





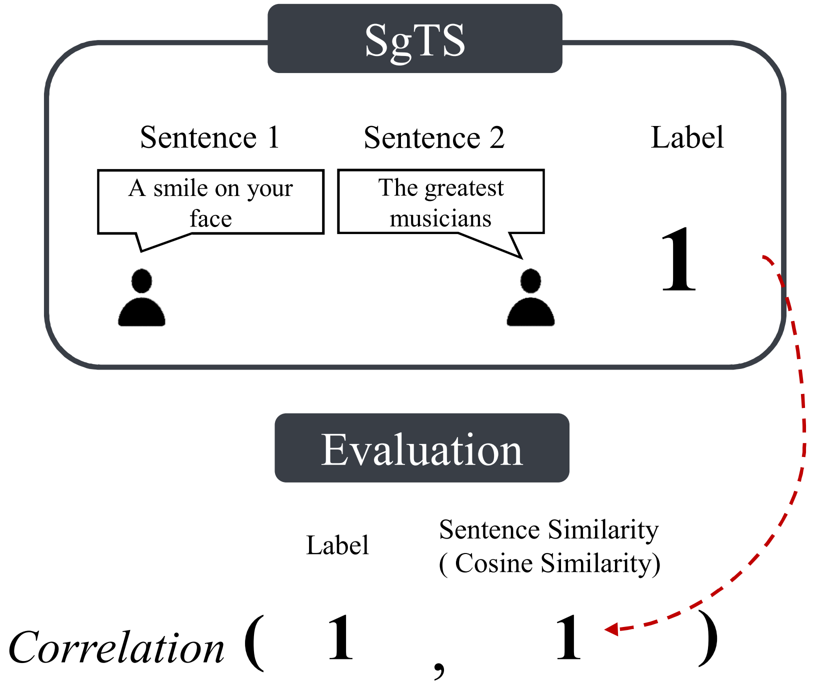
Method

❖ Sentiment-guided Textual Similarity (SgTS)

Sentence	STS
That is not an answer.	4.6
No, it’s not an answer.	
then the captain was gone	1
Then the captain came back	

Sentence	Polarity	**SgTS**
Saw how bad this movie was	NEG	0 (Different)
The greatest musicians	POS	
Is pretty damned funny	POS	1 (Identical)
A smile on your face	POS	

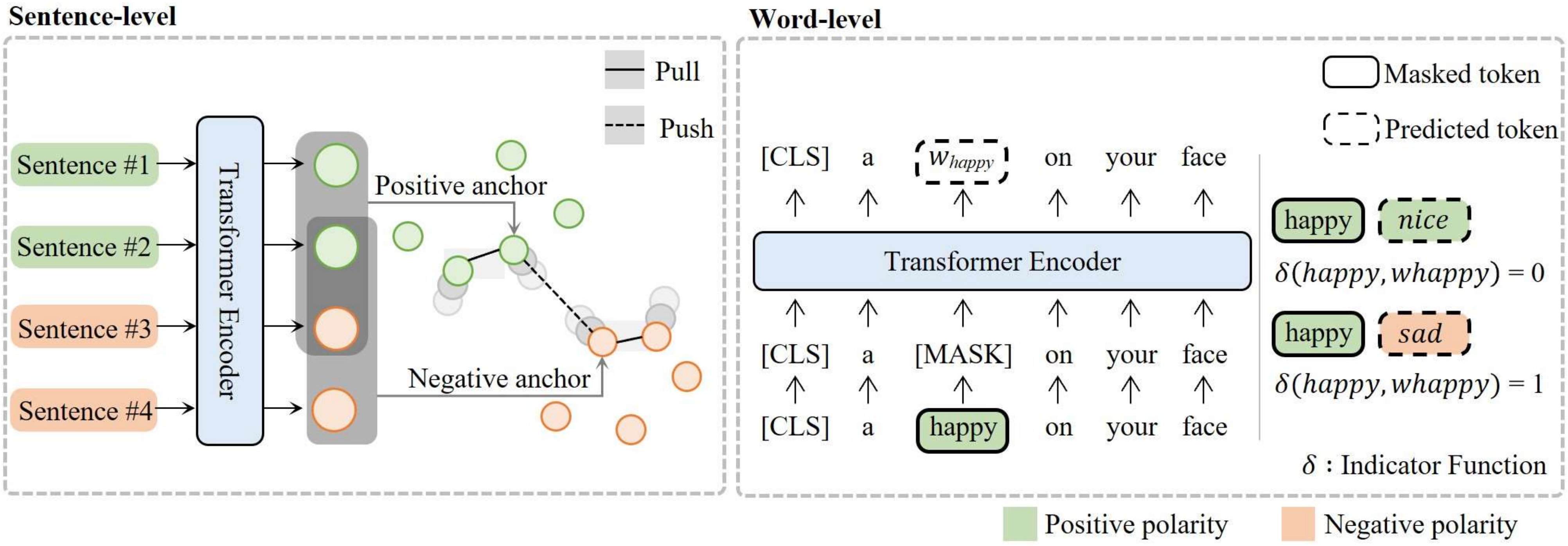
Positive polarity Negative polarity



Method

❖ Sentiment-aware Contrastive Sentence Embedding (SentiCSE)

Objectives



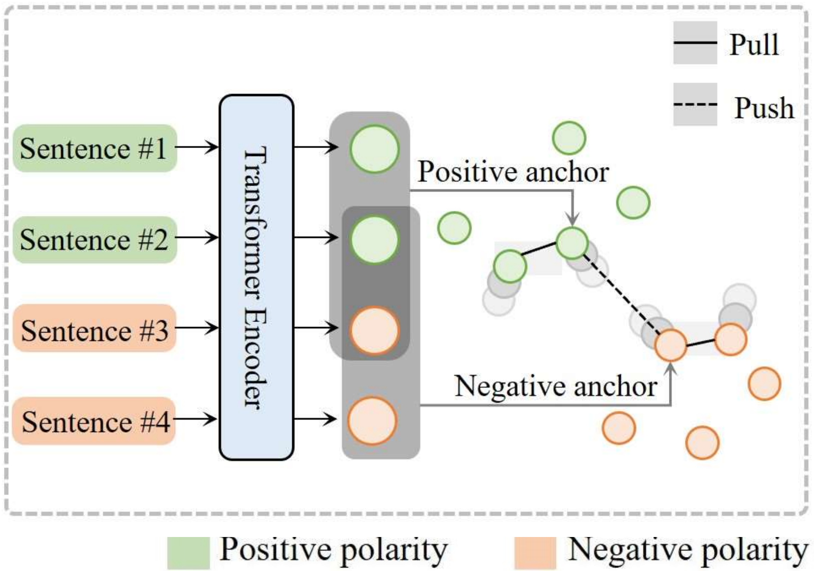
Method

❖ Sentiment-aware Contrastive Sentence Embedding (SentiCSE)

Objectives

- ✓ Due to the nature of sentiment polarity, positive and negative are contrasting labels.
- ✓ quadruple of sentences $q_i : (p_i, {p_i}^+, n_i, {n_i}^+)$

Sentence-level



$$Loss_{pos} = -log \frac{\overset{\text{Positive pair}}{\boxed{e^{sim(p_i\ , p_i^+)/\tau}}}}{\sum_{j=1}^{N}\left(e^{sim(p_i\ , p_i^+)/\tau} + \underset{\text{Hard negative pair}}{\boxed{\alpha \cdot e^{sim(p_i\ , n_j\ )/\tau}}}\right)}$$

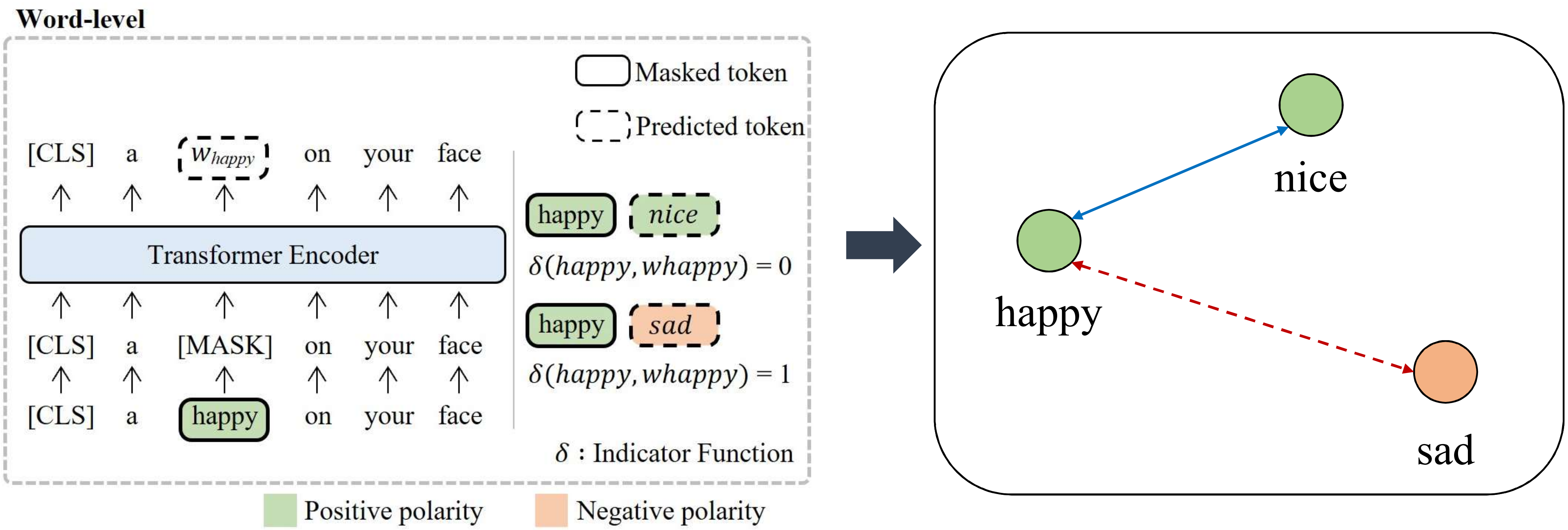
$$Loss_{neg} = -log \frac{\overset{\text{Positive pair}}{\boxed{e^{sim(n_i\ , n_i^+)/\tau}}}}{\sum_{j=1}^{N}\left(e^{sim(n_i\ , n_i^+)/\tau} + \underset{\text{Hard negative pair}}{\boxed{\alpha \cdot e^{sim(n_i\ , p_j^+)/\tau}}}\right)}$$

Method

❖ Sentiment-aware Contrastive Sentence Embedding (SentiCSE)

Objectives

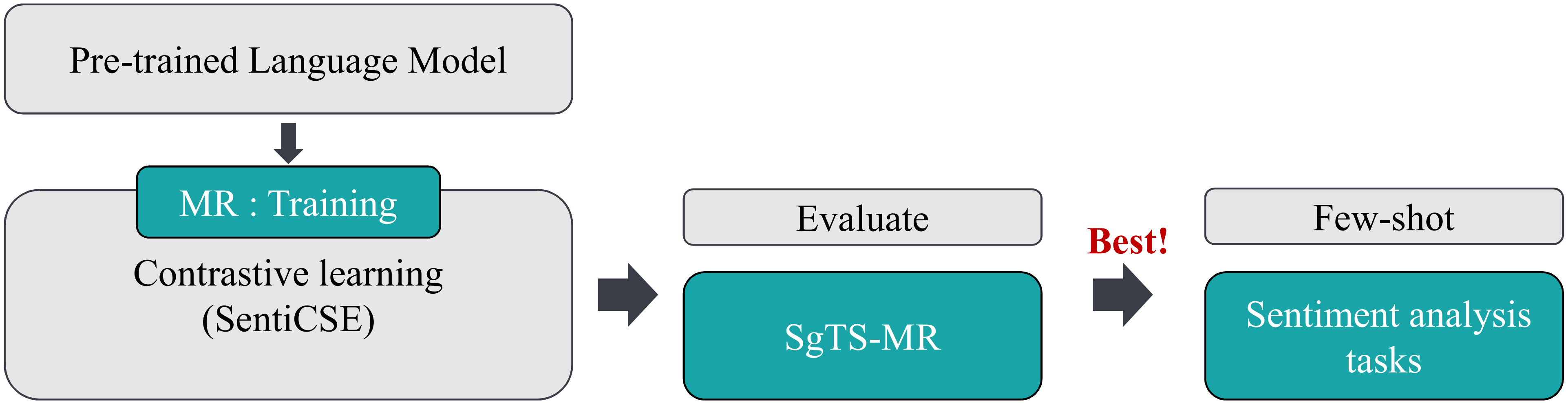
✓ Designed to learn about sentiment semantics from sentiment words.



Method

❖ Sentiment-aware Contrastive Sentence Embedding (SentiCSE)

Contrastive Learning



Experiments

- ❖ Details
  - ✓ Training Dataset

Model	Backbone	Pre-trained dataset					# Sentences
		Wiki	Amazon	Yelp	SST	MR	
SentiBERT	BERT				✓		0.067M
SentiX	BERT		✓	✓			240M
SentiLARE	RoBERTa			✓			6.7M
SentiWSP	ELECTRA	✓					0.5M
SentiCSE	RoBERTa					✓	0.008M

Each model requires between 8 to 48 hours of training time.

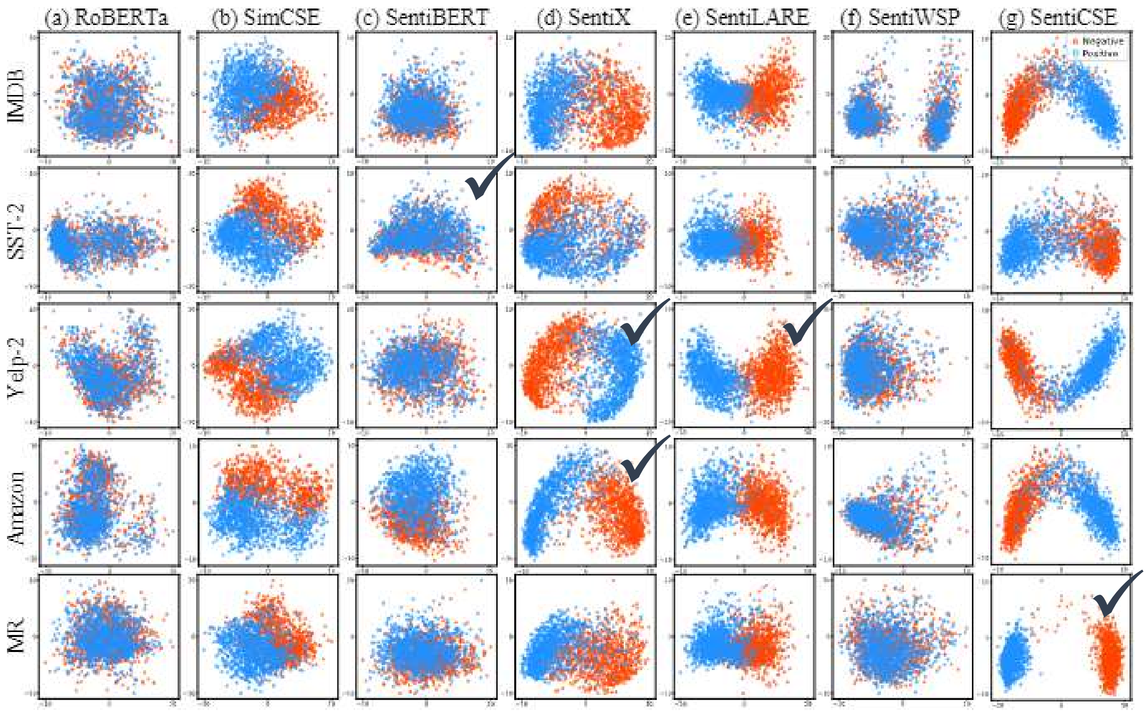
  - ✓ Maximum sentence length : 128, embedding dimension 768, batch size 64
  - ✓ Two NVIDIA A30 GPUs (3 hours)
  - ✓ SentiCSE: Evaluate every 500 steps and utilize the checkpoint at the best performance.

Experiments

❖ Evaluate the quality of sentiment representation (qualitatively).

Visualization of Representation

✓ The representation of SentiCSE reflects sentiment context effectively, as seen by the substantial distance between the positive and negative clusters.

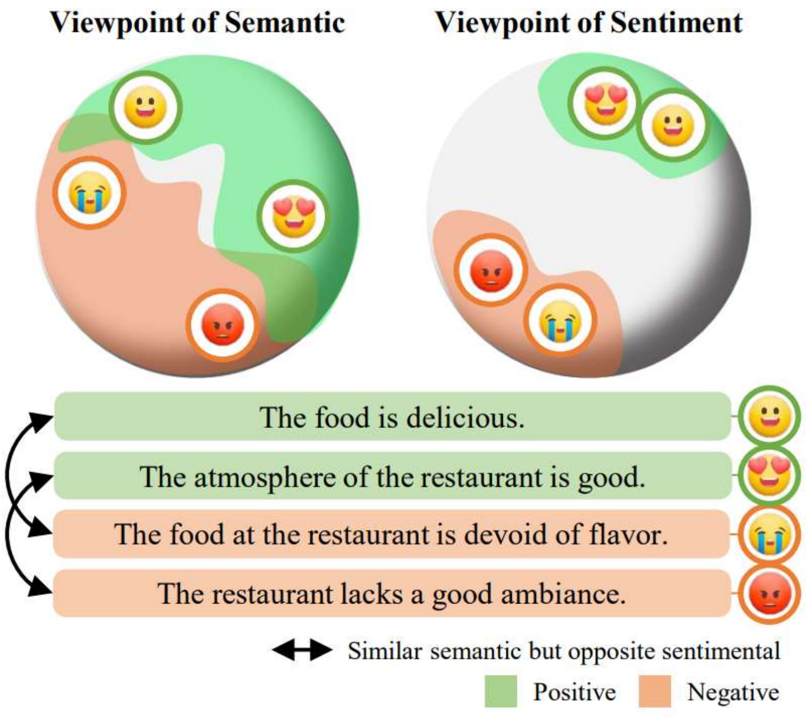


✓ : Training dataset

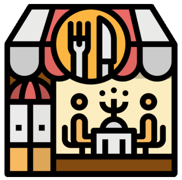
Experiments

❖ Evaluate the quality of sentiment representation (qualitatively).

Sentence Similarity



Atmosphere



The atmosphere of the restaurant is good.

The restaurant lacks a good ambiance.

Taste

The food is delicious.

The food at the restaurant is devoid of flavor.

Service

The restaurant has attentive and friendly staff.

The service at the restaurant is discourteous.

Experiments

❖ Evaluate the quality of sentiment representation (qualitatively).

Sentence: The atmosphere of the restaurant is good.

Top-2 Examples

SentiBERT

Taste	The food is delicious.
Atm	The service at the restaurant is discourteous.

SentiX

Service	The service at the restaurant is discourteous.
Taste	The food is delicious.

SentiLARE

Service	restaurant has attentive and friendly staff.
Atm	The restaurant lacks a good ambiance.

SentiWSP

Taste	The food is delicious.
Service	The service at the restaurant is discourteous.

SentiCSE

Taste	The food is delicious.
Service	The restaurant has attentive and friendly staff.

Experiments

❖ Evaluate the quality of sentiment representation (quantitatively).

Few-shot setting

✓ When evaluating performance on datasets not seen during training, it is apparent that SentiCSE delivers better performance.

Model	1-shot accuracy					5-shot accuracy				
	IMDB	SST2	Yelp-2	Amazon	MR	IMDB	SST2	Yelp-2	Amazon	MR
BERT◇	52.08	50.26	56.76	52.98	52.24	54.02	54.26	62.64	58.10	54.38
SimCSE◇	54.08	61.74	66.20	60.92	61.64	71.26	66.82	81.58	73.58	67.16
SentiBERT◇	51.40	55.60*	59.64	54.90	54.88	57.76	64.84*	70.20	67.02	64.90
SentiX◇	74.64	64.96	87.66*	86.14	**65.06**	**83.68**	72.32	93.40*	92.32*	**76.68**
SentiCSE◇	**76.08**	**87.88**	**81.62**	**82.24**	85.82*	81.84	**93.26**	**87.64**	**84.82**	86.14*
RoBERTa♠	52.00	54.54	56.56	52.84	53.82	60.30	49.80	72.42	64.58	56.78
SimCSE♠	59.04	61.06	68.44	58.40	61.72	74.72	68.08	86.62	75.14	71.56
SentiLARE♠	70.20	74.26	87.00*	84.58	**68.68**	87.18	80.10	93.28*	**91.06**	**82.34**
SentiCSE♠	**82.64**	**92.92**	**89.72**	**89.04**	87.38*	**88.12**	**94.50**	**92.08**	90.40	88.00*

Experiments

❖ Evaluate the quality of sentiment representation (quantitatively).

Few-shot setting

✓ When comparing each model using a standardized training dataset, it is evident that SentiCSE delivers better performance.

Model	Model	1-shot accuracy					5-shot accuracy				
		IMDB	SST2	Yelp-2	Amazon	MR	IMDB	SST2	Yelp-2	Amazon	MR
SST2	SentiBERT◇	51.40	55.60	59.64	54.90	54.88	57.76	64.84	70.20	67.02	64.90
	SentiCSE◇	**74.68**	**91.82**	**82.00**	**81.24**	**86.94**	**81.86**	**92.80**	**88.02**	**86.38**	**90.24**
Yelp2	SentiX◇	74.64	64.96	87.66	86.14*	65.06	83.68	72.32	93.40	92.32*	76.88
	SentiCSE◇	69.22	86.10	91.14	85.48	63.76	84.24	86.48	**95.24**	89.74	80.86
	SentiLARE♠	70.20	74.26	87.00	84.58	68.68	87.18	80.10	93.28	91.06	82.34
	SentiCSE♠	**76.64**	**84.78**	**94.26**	**89.28**	**73.20**	**87.98**	**87.66**	95.12	**92.68**	**86.36**
Amazon	SentiX◇	74.64	64.96	87.66*	86.14	65.06	83.68	72.32	**93.40***	92.32	76.68
	SentiCSE◇	**75.16**	**78.56**	**91.16**	**93.16**	**78.02**	**86.64**	**85.18**	92.98	**93.86**	**85.44**

Experiments

❖ Evaluate the quality of sentiment representation (quantitatively).

SgTS

✓ Comparative performance of SgTS for quantitatively measuring the quality of the proposed sentiment representation.

✓ performs well on the trained data.

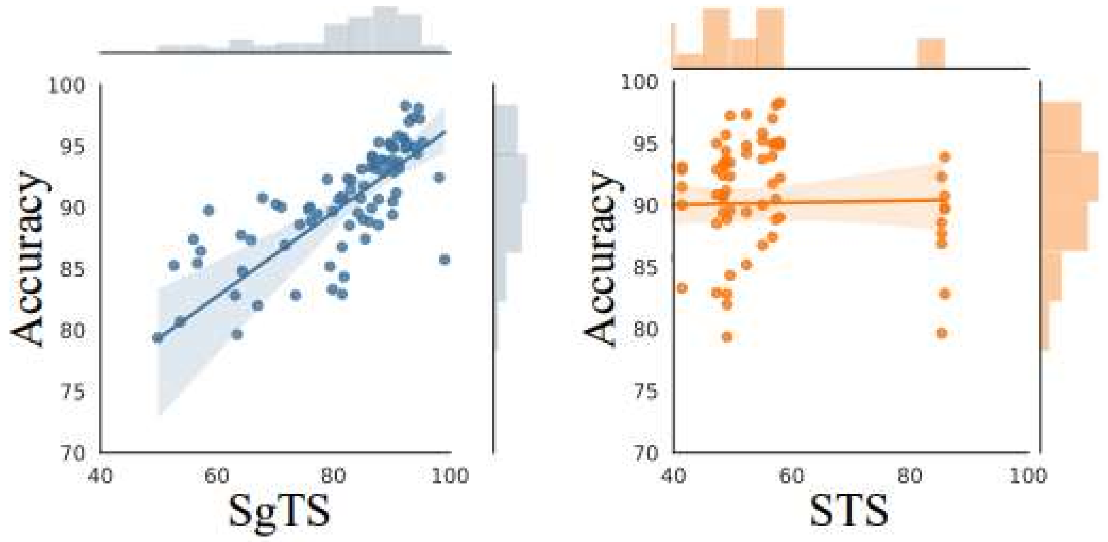
Model	SgTS					
	IMDB	**SST2**	**Yelp-2**	**Amazon**	**MR**	Avg.
BERT◇	0.01	0.08	0.09	0.15	0.07	0.06
SimCSE◇	0.16	0.13	0.24	0.19	0.13	0.18
SentiBERT◇	0.13	0.17*	0.12	0.09	0.18	0.14
SentiX◇	0.62	0.48	0.77*	0.52*	**0.39**	0.56
SentiCSE◇	**0.64**	**0.72**	**0.76**	**0.37**	0.63*	**0.62**
RoBERTa♠	0.06	0.05	0.06	0.02	0.04	0.06
SimCSE♠	0.21	0.11	0.26	0.20	0.19	0.19
SentiLARE♠	0.48	0.38	0.65*	0.36	**0.57**	0.46
SentiCSE♠	**0.77**	**0.72**	**0.82**	**0.56**	0.69*	**0.71**

Experiments

❖ Validity of SgTS

SgTS

- ✓ It is observed that when SgTS shows high performance, the few-shot accuracy is also high.
- ✓ There is a significant correlation above 0.7, significant at the 0.01 level.



(a) SgTS-Accuracy (b) STS-Accuracy

Experiments

❖ Evaluate the quality of sentiment representation (quantitatively).

Linear probing

✓ It is confirmed that each model demonstrates good performance relative to the size of the dataset learned from.

Model	IMDB	SST2	Yelp-2	Amazon	MR
BERT◇	85.25	85.44	89.75	86.44	80.68
SimCSE◇	86.91	87.73	92.29	88.60	79.64
SentiBERT◇	87.40	90.25*	90.76	87.33	84.80
SentiX◇	**94.20**	89.45	97.33*	94.82*	**85.18**
SentiCSE◇	90.63	**95.30**	**93.12**	**89.93**	85.74*
RoBERTa♠	82.82	79.36	88.87	81.98	50.38
SimCSE♠	90.73	89.68	93.89	89.82	82.83
SentiLARE♠	**94.84**	92.20	98.26*	**95.10**	**89.02**
SentiCSE♠	94.03	**95.18**	**95.86**	93.69	89.49*

Conclusion

- ❖ We argue that the representation for the viewpoint of semantic and the viewpoint of sentiment should be distinct.
  - ✓ We propose the first task that can measure the quality of sentiment representation.
  - ✓ Using this, we suggest a framework for learning sentiment representation.
- ❖ We demonstrate superiority in a few-shot setting that can be utilized in the industry.
- ❖ SgTS shows validity in measuring the quality of sentiment representation.