ISSN: 1813-4890

Sentiment Analysis Model based on Contrastive Learning

Muyuan Ouyang
Department of Computer Science, Jinan University, Guangzhou 510632, China
hellomuyuan@163.com

Abstract

As a subtask of natural language processing, sentiment analysis plays a crucial role in various fields. Its task is to help users quickly obtain, organize and analyze relevant features, and predict the unseen data with the identification of the inherent laws of the data, so as to make the best decision. Among existing methods, the research generally focuses on the vectorized representation of text data and how to build high-quality deep learning classifiers while ignoring sentence embedding methods. To generate more discriminative sentence encodings, this paper proposes a novel way of combining contrastive learning with pre-trained language models (such as BERT), and applies a simple contrastive sentence embedding framework SimCSE (Simple Contrastive Learning of Sentence Embbedings) to introduce a self-supervised model BiSE-SimCSE, which applies the input sentence and makes self-predictions in the comparative target, and only uses the standard deviation as noise. Then the model replicates the self-supervised BERT model to form a twin network (BERTs on both sides of the twin network share the same structure and parameters), and input the sentiment text pairs into the two BERT models to code the representation vector of each sentence; and the produced sentence vectors can be used for semantic similarity calculation or for unsupervised clustering tasks. Finally, a single BERT network in the trained twin network is transferred to the supervised SimCSE module for classification tasks, using a supervised approach to incorporate annotation pairs from the natural language inference datasets into our contrastive learning framework. The experimental results on three datasets show that the proposed model outperforms many existing baselines, with the accuracy improved by 1.68%, 1.36% and 1.19% compared with the suboptimal model, respectively.

Keywords

Sentiment Analysis; Contrastive Learning; BERT; BiSE-SimCSE.

1. Introduction

Sentiment analysis is a method to refine the sentimental content of the text and judge its sentimental tendency. For example, given the sentence "the mobile phone is stuck and not easy to use, that is it.", sentiment analysis should recognize that the sentimental content in the sentence is "stuck" and "not easy to use," and judge that the sentimental tendency of the sentence is negative.

As a sub-task[1] of natural language processing, sentiment analysis plays a vital role in various fields. Its task is to help users quickly obtain, sort out and analyze relevant features, and predict the unseen cases based on the internal patterns of the learned data, so as to understand the data better and make the best decision.

The traditional methods to solve sentiment analysis are based on sentiment dictionary and machine learning algorithms, such as Support Vector Machine, Hidden Markov Model[2,3], etc. With the development of deep learning, more and more deep learning models are proposed for

ISSN: 1813-4890

attribute sentiment analysis. Ruder, Tang et al.[4,5] proposed applying the long short-term memory model to analyze sentiment by using the sentimental association between different sentences in long comments. Tang et al.[6] introduced a method with a gated recurrent neural network. Liao et al.[7] proposed a method of coupling local and global information for feature extraction of texts, and a new vector containing local and global information was generated through a combination of models.

Among a large number of existing methods, the research focus is generally on the vectorized representation of text data and how to construct a high-quality deep learning classifier while ignoring the sentence embedding quality. Actually, learning a discriminative embedding is a fundamental problem in natural language processing[8-10].

Therefore, in this paper, we propose a novel way of combining contrastive learning with pre-trained language models (such as BERT), and the contributions are summarized as follows:

(1) We introduce a self-supervised model BiSE-SimCSE to learn from unlabeled or labeled data to produce better sentence embeddings, based on the simple contrastive sentence embedding framework SimCSE[11]. In the self-supervised task for the unlabeled case, each sentence has a consistent number of labels, whether positive or negative. The same sentence goes through BERT twice, resulting in two different but similar vectors as positive pairs.

(2) The single BERT network in the trained Siamese network is further transferred to the supervised SimCSE module for classification tasks. The BERTs on both sides of the twin network have the same structure and parameters, and to input different sentences into the two BERT models can obtain the sentence representation vector of each sentence; and the finally obtained sentence representation vector can be used for semantic similarity calculation.

(3) A new technique for data augmentation is applied in the BERT encoder process, with the method of back-translation, which can better increase the diversity of the text and retain the correct semantic information in the case of changing the grammatical structure.

(4) The proposed model has been implemented and experimental results on three datasets show that the proposed model outperforms many existing baselines.

2. Materials and Methods

2.1. BERT

BERT (bidirectional encoder representations from transformers)[12] is a language model pre-trained on large-scale unlabeled text. Based on its fine-tuning model, it has achieved very good results in natural language processing problems such as sentence-level sentiment classification and part of speech tagging. Bert is composed of multi-layer transformer encoder[13] superposition, and the output of each layer of transformer encoder is used as the input of the next layer of transformer encoder. It is generally believed that the output of the last layer of Bert has rich contextual word-related information, so it is usually used to replace Word2vec[14] and Glove[15] as the input word vector of the model.

2.2. Contrastive Learning

The purpose of contrastive learning is to learn effective representations by pulling together semantically similar neighborhoods and separating non neighborhoods[16]. It learns the general characteristics of data sets by letting the model learn which data points are similar or different without labels. It assumes a set of paired examples $\mathcal{D} = \{(x_i, x_i^+)\}_{i=1}^m$, where x_i and x_i^+ are semantically related. We follow the contrastive framework[17] and take a cross-entropy objective with in-batch negatives[18,19]: let h_i and h_i^+ denote the representations of x_i and x_i^+, the training objective is:

ISSN: 1813-4890

$$\ell_i = \log \frac{e^{\frac{sim(h_i, h_i^+)}{\tau}}}{\sum_{j=1}^{N} e^{\frac{sim(h_i, h_j^+)}{\tau}}} \quad (1)$$

where τ is a temperature hyperparameter, $sim(h_i, h_i^+)$ is the cosine similarity. In this work, we encode input sentences using a pre-trained language model such as BERT, and then fine-tune all the parameters using the contrastive learning objective (Eq. 1).

3. The Proposed BiSE-SimCSE

3.1. Self-supervised Se-SimCSE

The proposed Self-Supervised Se-SimCSE mainly uses auxiliary information to mine its own supervision signals from large-scale unlabeled data, and train the network through the constructed supervision information, so that it can learn valuable representations for downstream tasks. As shown in Fig. 1, In the self-supervised task, each sentence has a consistent number of labels, either positive or negative. Since BERT itself has a random DropOut function, the same sentence passes through BERT twice, and two different but similar vectors are obtained, and the sum of the two vectors obtained by BERT Encoder twice is used as a positive sample pair. In this way, the semantics of the original sample and the generated positive sample are consistent, but the generated embedding is different, so a small enhancement is thus made to the data in the BERT Encoder process. Also, similar approach can be applied for in-batch negatives.

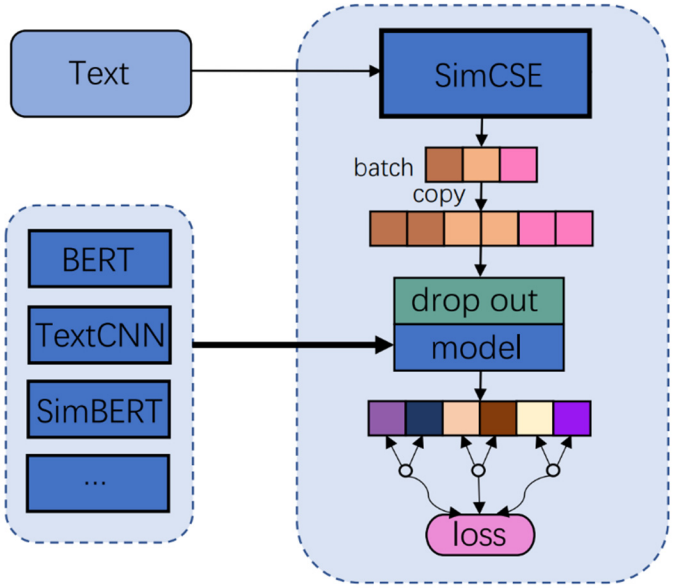

Fig 1. Self-Supervised Se-SimCSE Model

3.2. Siamese Network

Our application of a Siamese Network mitigates the consequences that a single sentence input to BERT and producing fixed-size sentence embeddings can have rather poor sentence encodings. The Siamese network structure enables different fixed-size input sentence vectors to be derived and we can use cosine similarity to find semantically similar sentences.

We use the method of back translation, based on the google translation interface, to translate the Chinese sentiment text of a single sentence into English, and then translate it back to the new Chinese to construct a sentiment analysis text for training.

ISSN: 1813-4890

We then add a pooling operation using the output of the [cls] token to the output of BERT, resulting in a fixed-size sentence embedding.

To optimize BERT, we create conjunctions[20] to update the weights so that the resulting sentence embeddings are semantically meaningful and comparable to cosine similarity. The network structure depends on valid training data. We carry out experiments with the following structure and objective function in Fig. 2:

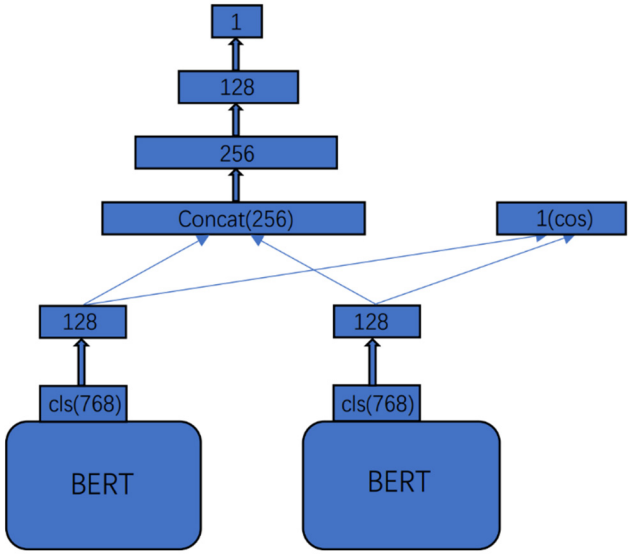

Fig 2. Siamese Network structure: The two BERT networks have tied weights. Compute the cosine similarity of two vectors while regularizing the output of the fully connected layer

While comparing with the cosine similarity, the Concat connection operation is performed on the sentence representations output by the two BERTs. Through the three-layer fully-connected layer, the output of the fully-connected layer is regularized by the softmax function, and the classification of each category is obtained by the model with probability.

The left part of the fully connected layer of the model is used for correlation calculation, and the right part is directly normalized and then used for cosine similarity. Because vector normalization to calculate cosine similarity is equivalent to calculating L_2 distance (metrics of online faiss retrieval distance), the derivation is as follows (Eq. 2):

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{2}$$

where y_i represents the label of sample i with positive class being 1 and negative class 0, p_i represents the probability that sample i is predicted to be a positive class.

3.3. Supervised Se-SimCSE

Studies[10,21] have shown that supervised natural language inference (NLI) datasets[22,23] can predict the difference between two sentences by seeing whether the relationship is entailment, neutrality, or contradiction, which is effective for learning sentence embeddings. Another research[24] leverages labels in the NLI dataset to construct positive and negative examples, demonstrating that adding hard negatives can improve model performance. Based on this reseasrch, the model of SimCSE is inspired and illustrated as follows in Fig. 3:

ISSN: 1813-4890

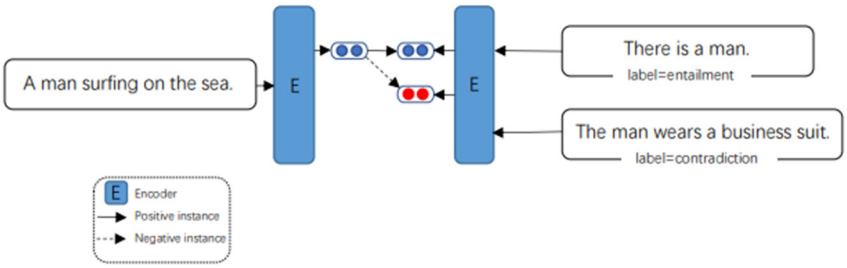

Fig 3. Supervised Se-SimCSE takes the entailment pairs as positives, and contradiction pairs as well as other in-batch instances as negatives

We just transfer the BERT model trained in the Siamese network to perform supervised classification tasks. We directly extract $sim(x_i, x_i^+)$ pairs from the supervised dataset and use them for optimization.

3.4. Sentiment Classification Layer

After passing through the preceding modules, the vector representation corresponding to the text of the last layer is input to the fully connected layer as a classification feature, mapped to the same dimension as the number of emotional categories, and then the softmax function is used to regularize the output of the fully connected layer to get The classification probability of the model for each class as(Eq. 3):

$$p = softmax(h_a W_c + b_c) \tag{3}$$

where W_c and b_c are learnable parameters, and $p \in \mathbb{R}^2$ is the probability vector that classifies the current sample into each class.

4. Experimental Results

4.1. Datasets

Most of the current research is carried out on the English sentiment data set. However, Chinese data is also a huge part of the information, and the emotional expression in Chinese is more complex than that in English. Therefore, we selected three Chinese datasets that are close to life, such as takeaway reviews, hotel reviews and online shopping reviews. Correctly grasping the emotions of consumers and netizens has very high commercial and social value. Automated analysis and use of these massive data can better understand market demands and make optimal decisions.

Experiments are carried out on three Chinese data sets, where the number of positive and negative examples is shown in Table 1.

(1) waimai_ 10K data: more than 12000 user comments collected by a takeout platform, including 4000 positive and 8000 negative;

(2) ChnSentiCorp_ htl_ All data: more than 7000 Hotel Comments, more than 5000 positive comments and more than 2000 negative comments;

(3) online_ shopping_ 10_ Cats data: 10 categories (books, tablets, mobile phones, fruits, shampoo, water heater, Mengniu, clothes, computers, hotels), with a total of more than 60000 comment data, and about 30000 positive and negative comments respectively.

ISSN: 1813-4890

Table 1. Statistics of the datasets.

Datasets	Positive	Negative
waimai_ 10K data	4001	7987
ChnSentiCorp_ htl_ All data	5323	2444
online_ shopping_ 10_ Cats data	31729	31046

All experiments only use "positive" and "negative" samples. We use cross-validation method. 90% of each dataset is randomly selected as training data, and the remaining 10% is used as test data. The experiment was run three times, and the average of the three experimental results was used as the final result data.

4.2. Baseline Models

In order to comprehensively evaluate the model proposed in this paper, this paper compares it with baseline models such as LSTM[25], GRU[26], FastText[27], TextCNN[28], DPCNN[29], BERT[12] and Se-SimCSE.

The baseline models LSTM, GRU, FastText, TextCNN, DPCNN, and BERT are all implemented with open source codes with the original model parameters not adjusted. Se-SimCSE is a variant of SimCSE that combines a simple contrastive learning framework, first using an unsupervised approach to produce better sentence embeddings, with only standard dropout used as noise. Annotated pairs from natural language inference datasets are then incorporated into our contrastive learning framework through a supervised learning method, and finally text classification is performed through softmax.

4.3. Experimental Results

The BERT model is the implementation of the open sourced codes. BERT uses an uncased BERT-base model with a hidden state dimension d^{BERT} of 768 and a total of 12 layers. The batch size is set to 32, and the model learning rate is set to 0.00002. For each training, 90% of the dataset is randomly used as the actual training set, and the remaining 10% is used as the validation set. In the Siamese network, the loss value is output every 500 steps, and the value of the validation set is estimated every 1000 steps of training. In the supervised classification SimCSE model, the waimai_ 10K data dataset and the ChnSentiCorp_ htl_ All data dataset are trained for 10 epochs, and the online_ shopping_ 1_ Cats data dataset is trained for 3 epochs. After each epoch, the performance of the model is verified on the validation set, and finally the model with the highest accuracy rate on the validation set is verified on the test set. We report the average results of the cross-validation experiments as the final results. Using this method to report results can fully avoid random errors and unfair comparison results.

Table 2. Experimental Results

Models	waimai_ 10K data	ChnSentiCorp_ htl_ All data	online_ shopping_ 1_ Cats data
LSTM	84.26	79.33	88.67
GRU	82.63	79.45	88.21
FastText	82.97	79.77	88.73
TextCNN	83.74	79.98	88.74
DPCNN	85.61	80.14	89.01
BERT	89.25	81.88	90.62
Se-SimCSE	91.25	83.81	91.69
BiSE-SimCSE	92.93	85.17	92.88

ISSN: 1813-4890

As can be seen from Table 2 and Fig. 4, the proposed model of BiSE-SimCSE in this paper has achieved best results on the three datasets, and the accuracy is improved by 1.68%, 1.36% and 1.19% compared with the suboptimal model, respectively.

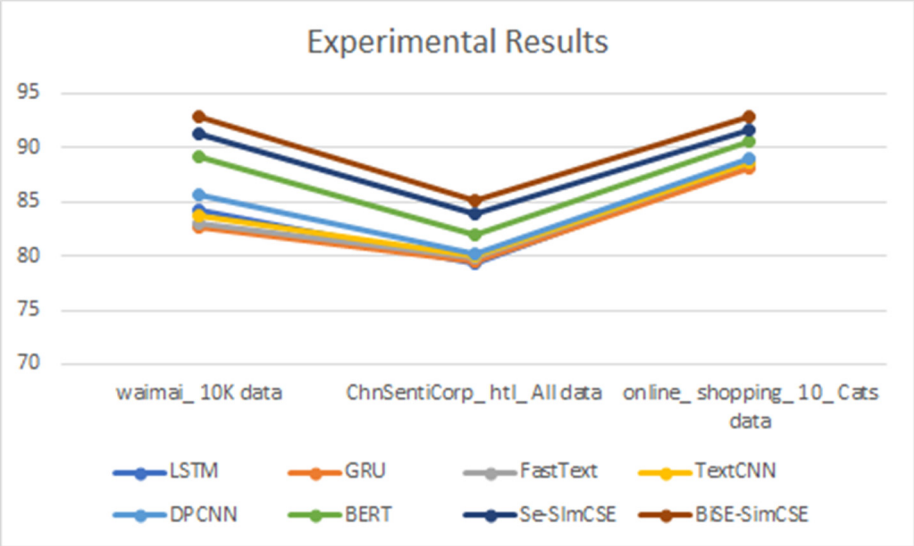

Fig 4. Experimental Results

5. Discussion

Across all the experimental results, we can clearly see that BiSE-SimCSE, effectively integrating contrastive learning, and self-supervised training BERT model with contrastive learning, can better self-predict and obtain higher quality semantic encodings for the downstream tasks, while in Siamese network, a more accurate sentence representation vector for each sentence can be generated, and finally the trained BERT network is applied for the supervised classification task with good results. Therefore, using a supervised approach to incorporate annotated pairs from a natural language inference dataset into our contrastive learning framework enables more accurate complex sentiment classification in Chinese texts. Our experimental results justify that that the model based on contrastive learning has better performance. But it only predicts the sentiment of the entire text, and does not address the fine-grained problem of predicting the sentiment of different aspects of a sentence. Future work will focus on fine-grained sentiment analysis. We believe that contrastive learning can be more widely used in natural language processing.

6. Related Work

Traditional methods for addressing sentiment classification include rule-based methods and statistical learning-based methods. Rule-based methods usually set rules and use sentiment dictionaries based on data and features; statistical learning-based methods usually combine hand-designed features and machine learning algorithms. Based on the Word2Vec word vector obtained by Google News pre-training, Chen et al. [28] constructed four convolutional neural network models by transforming the word vectors. To avoid manually designing features and rules for text classification, Lai et al. [30] proposed a combined model of a bidirectional recurrent neural network and a max pooling layer. Li et al. [31] introduced a text representation model based on optimized TF-IDF and weighted Word2Vec combined with CNN to classify sentiments. Liao et al. [7] put forward an approach for feature extraction of text using coupled local and global information.

ISSN: 1813-4890

In addition, some researchers have improved the effect of text classification models by building better neural network models. For example, Joulin et al. [27] proposed to use the FastText method for text classification, and the results show that it has the advantages of fast training speed and low energy consumption. Felbo et al. [32] designed DeepMoji based on the Embedding layer, Bi-LSTM and attention mechanism, which takes the Embedding layer and Bi-LSTM as input to obtain the vector representation of the document. Yao et al. [33] employed a graph convolutional network to construct a large-scale heterogeneous text graph containing word nodes and document nodes, and applied the co-occurrence information to model global words, and obtained the classification results of sentiment texts based on the classification of nodes.

As we know, implicit emotional expressions also widely exist in texts, especially in some Chinese texts, since Chinese emotional expressions are more implicit. Implicit emotional expressions refer to emotional expressions that do not contain strong polarity markers but still express the emotional polarity of human consciousness clearly in contexts [34]. For example, "the waiter pour water on my hand and walked away", this comment does not express an opinion, but can be clearly read as negative. However, most of the previous methods paid less attention to the modeling of implicit emotion expressions. This motivates us to capture implicit sentiment in a more advanced way to better solve this task.

6.1. Metrics

Our research is motivated by contrastive learning, whose main idea is that the representations of similar samples are close, and the dissimilar ones are far away. Contrastive learning can be applied in self-supervised, unsupervised and supervised scenarios.

Contrastive learning algorithms have two key attributes [35], alignment and uniformity, and many effective contrastive learning algorithms satisfy these two properties. Alignment refers to the degree of approximation between positive samples as (Eq. 4):

$$L_{align}(f;\alpha) \triangleq \mathbb{E}_{(x,y)\sim p_{pos}}[\| f(x)-f(y)\|_2^{\alpha}] \tag{4}$$

uniformity refers to the uniformity of the distribution of eigenvectors on the hypersphere as (Eq. 5):

$$L_{uniform}(f;t) \triangleq \log \mathbb{E}_{x,y \overset{i.i.d.}{\sim} p_{data}}[e^{-t\|f(x)-f(y)\|_2^2}] \tag{5}$$

6.2. SimCLR

Given training data, we need to perform data augmentation to get more positive samples. Correct and effective data augmentation techniques are crucial for learning good representations. Experiments with SimCLR [36] show that, for sentences, deletion or substitution may lead to semantic changes. Generally, in-batch negatives are used in contrastive learning, and the irrelevant data in a batch is regarded as negative samples, while the positive sample pair can be data of two modalities, such as pictures and corresponding descriptions of pictures.

6.3. Data Augmentation

Back translation [37] is a very common data enhancement method in machine translation. Its main idea is to translate a sentence into another language through translation tools, and then translate the translated language back into the original language, and finally get a sentence with similar meaning but different expressions. This method is also a relatively reliable method at present. This technique not only has synonym replacement, word addition and deletion, but also has the effect of adjusting the sentence structure and word order, and can keep the meaning similar to the original sentence, which is proved a very effective data augmentation.

ISSN: 1813-4890

7. Conclusion

In order to overcome the shortcomings of existing sentiment analysis methods, this paper proposes a model BiSE-SimCSE that is integrated with contrastive learning to make the anisotropic space embedded in the pre-training encoding more uniform through contrastive learning. This approach combines the SimCSE model, firstly a self-supervised method is introduced, which applies the input sentence and makes self-predictions in the compared target. Then, input the sentiment text pairs generated by the back-translation method into the Siamese BERT network (the two bert models share parameters, and can also be understood as the same BERT model), and obtain the sentence representation vector of each sentence. Furthermore, utilizing a supervised learning approach, annotated pairs from natural language inference datasets are incorporated into our contrastive learning framework with "entailment" pairs as positives and "contradiction" pairs as hard negatives. Finally, the conducted experiments on three Chinese datasets, compared with multiple state-of-the-art counterparts, justify the effectiveness and superiority of our method.

References

[1] Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. Information 2019, 10. Doi:10.3390/info10040150.

[2] Wagner, J.; Arora, P.; Cortes, S.; Barman, U.; Bogdanova, D.; Foster, J.; Tounsi, L. Dcu: Aspect-based polarity classification for semeval task 4 2014.

[3] Jin, W.; Ho, H.H.; Srihari, R.K. A novel lexicalized HMM-based learning framework for web opinion mining. Proceedings of the 26th annual international conference on machine learning. Citeseer, 2009, Vol. 10.

[4] Ruder, S.; Ghaffari, P.; Breslin, J.G. A hierarchical model of reviews for aspect-based sentiment analysis. arXiv preprint arXiv:1609.02745 2016.

[5] Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for target-dependent sentiment classification. arXiv preprint arXiv:1512.01100 2015.

[6] Tang, D.; Qin, B.; Liu, T. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605.08900 2016.

[7] Liao, M.; Li, J.; Zhang, H.; Wang, L.; Wu, X.; Wong, K.F. Coupling global and local context for unsupervised aspect extraction. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 4579–4589.

[8] Kiros, R.; Zhu, Y.; Salakhutdinov, R.R.; Zemel, R.; Urtasun, R.; Torralba, A.; Fidler, S. Skip-Thought Vectors. Advances in Neural Information Processing Systems; Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; Garnett, R., Eds. Curran Associates, Inc., 2015, Vol. 28.

[9] Hill, F.; Cho, K.; Korhonen, A. Learning Distributed Representations of Sentences from Unlabelled Data. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Association for Computational Linguistics: San Diego, California, 2016; pp. 1367–1377. doi:10.18653/v1/N16-1162.

[10] Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 670–680. doi:10.18653/v1/D17-1070.

[11] Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 2021.

[12] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. doi:10.18653/v1/N19-1423.

[13] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems 2017, 30.

[14] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 2013, 26.

[15] Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[16] Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, Vol. 2, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[17] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning; III, H.D.; Singh, A., Eds. PMLR, 2020, Vol. 119, Proceedings of Machine Learning Research, pp. 1597–1607.

[18] Chen, T.; Sun, Y.; Shi, Y.; Hong, L. On sampling strategies for neural network-based collaborative filtering. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 767–776.

[19] Henderson, M.; Al-Rfou, R.; Strope, B.; Sung, Y.H.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; Kurzweil, R. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652 2017.

[20] Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[21] Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992. doi:10.18653/v1/D19-1410.

[22] Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 632–642. doi:10.18653/v1/D15-1075.

[23] Williams, A.; Nangia, N.; Bowman, S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 1112–1122. doi:10.18653/v1/N18-1101.

[24] Gao, J.; He, D.; Tan, X.; Qin, T.; Wang, L.; Liu, T.Y. Representation degeneration problem in training natural language generation models. arXiv preprint arXiv:1907.12009 2019.

[25] Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural computation 1997, 9, 1735–1780.

[26] Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 2014.

[27] Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 2016.

[28] Chen, Y. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo, 2015.

[29] Johnson, R.; Zhang, T. Deep pyramid convolutional neural networks for text categorization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 562–570.

[30] Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. Twenty-ninth AAAI conference on artificial intelligence, 2015.

[31] Li, L.; Xiao, L.; Jin, W.; Zhu, H.; Yang, G. Text Classification Based on Word2vec and Convolutional Neural Network. International Conference on Neural Information Processing. Springer, 2018, pp. 450–460.

[32] Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524 2017.

[33] Yao, L.; Mao, C.; Luo, Y. Graph convolutional networks for text classification. Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 7370–7377.

[34] Russo, I.; Caselli, T.; Strapparava, C. Semeval-2015 task 9: Clipeval implicit polarity of events. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 443–450.

[35] Wang, T.; Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. International Conference on Machine Learning. PMLR, 2020, pp. 9929–9939.

[36] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[37] Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381 2018.