Self Supervision in Graph Neural Networks for Text Classification

1 Introduction

When analyzing log files or reviews of products, it is essential to map the review to the corresponding product. Methods such as aspect-based sentiment analysis propose ways to determine the sentiment polarity towards a particular aspect, usually the product in online reviews. Recent advances in Large Language Models (LLMs) improved upon aspect-based sentiment analysis by adopting attention-based models to connect aspects with opinion words implicitly. However, this might fail in cases with multiple aspects in a single sentence due to the complexity of the language.

Recent progress from the literature showed the usage of syntax information for effectively encoding a sentence. The most effective syntax encoding method utilizes a reshaped dependency tree of a sentence with an aspect at its root and a Relational Graph Attention Network (RGAT) (Wang et al., 2020) to encode the tree structure. LLMs like BERT (Devlin et al., 2018) that are pre-trained on unlabelled data to learn text representations, so they can be finetuned and deployed with limited resources. Inspired by this, we proposed a self-supervised learning algorithm on tree structures to learn knowledge from syntax information [1]. The success of pre-training schemes in LLMs using unlabelled data showed generalization of the semantic meaning of the sentences.

We hope our pre-training scheme on the RGAT model generalizes the syntax structures and helps solve the problems that may need syntax information by a simple fine-tuning. Our goal is to provide better solutions to cases similar to analyzing reviews with too many mentions of aspects(products, entities etc) in sentences or cases where there is a need to use syntax information for the task and wants a generalized model that can work well with simple fine-tuning on the dataset.

2 Motivation

As shown in the Table 3 below, traditional sentence encoders fail to understand more subtle meaning in the sentences, such as opinions towards a certain aspect or entity. For example, we can see DistilledBERT (Sanh et al., 2019) classifying sentence 7 in Table 3 as having positive sentiment despite the phrase "you won't like roger" causing the ground truth sentiment to be negative. Recent efforts adopt using attention in sentence encoding to measure the contribution of the rest of the words to the current word for downstream task using attention weights.

Techniques such as aspect-based sentiment analysis aims to determine the sentiment towards a particular aspect, usually the aspect is a product or entity in online reviews. However, the research by (Wang et al., 2020) observes that attention fails to capture some empirical details from the sentences that can be learnt structurally, and fails to capture the impact of the opinionated words towards a corresponding aspect in a sentence. Most solutions in the literature solved this problem using LLMs). They have attention based neural network architectures to make connections between aspects and related words that determine the sentiment towards the aspect. However, experiments from the literature showed these models are not capable of handling complex sentences that involve multiple aspects in them, such that making connections of those aspects with the relevant opinion words becomes tougher.

BERT (Devlin et al., 2018), has shown to have performed better than other LLMs for text

[1] Our code can be found here

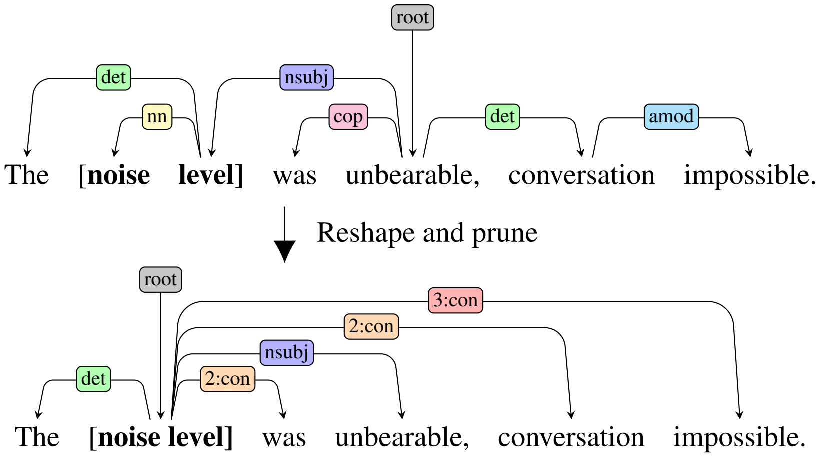
root
det
nsubj
nn
cop
det
amod
The [noise level] was unbearable, conversation impossible.
Reshape and prune
root
3:con
2:con
nsubj
det
2:con
The [noise level] was unbearable, conversation impossible.

Figure 1: An example of dependency tree reshaping.

Category	Example
Unknown token	• The [UNK] is very longer. • And the fact that Apple is driving the 13" [UNK] with the [UNK] graphic chip seems [UNK] to me.
Knowledge based	• This laptop has only 2 usb ports, and they are both on the same side. • The battery life is [UNK] 6-7 hours without charging.
Semantic information	• It's been time for a new laptop, and the only debate was which size of the Mac laptops, and whether to spring for the retina display. • I am still in the process of learning about its features.

Table 1: Some failure cases of our model.

classification tasks. However, it has many limitations. The first limitation of it is that it doesn't work well with neutral reviews as it involves some opinionated words with direct dependency connection towards the aspect. The second limitation is that it's not able to comprehend some texts that need a deeper understanding of the language and context. The third limitation is with sentences that involve advice that only recommend or disrecommend particular aspects but don't include any such opinionated words or clues to indicate the sentiment. Then, recently, there has been research that tries to solve this problem by encoding syntax information and that led to the usage of graph based models in aspect based sentiment analysis.

In particular, recent progress showed Relational Graph Attention Network as the leading algorithm to represent a sentence as a graph, and more sophisticated approaches combined information from LLMs, and graphs to represent embeddings of the sentences. Due to the huge availability of unlabelled data, LLMs like BERT are pre-trained to learn text representations, so they can be finetuned and deployed with limited resources. However, this is not the case for graph based models as little to no research is done to pre-train a graph neural network for text to learn syntax representations.

Apart from such efforts, there have been researchers experimenting with different methods, such as contrastive learning, self-supervision, and new language models or representations to try performing better on such NLP tasks. The research by (Peters et al., 2018) and (Devlin et al., 2018) dive deep into the applications and performances of pre-training schemes like ELMo and Bidirectional Encoder Representations from Transformers (BERT) regarding many such NLP tasks. BERT, specifically, broke the baseline performance on many NLP tasks as it uses masked language modelling during training, which falls under self-supervised learning. Moreover, Constrastive Learning by (Chen et al., 2020) followed by fine-tuning has shown a significant performance gains over base training schemes in many different language tasks. Furthermore, Self-Supervision has been explored in the domain of GNNs.One such paper about Constrastive Learning on graphs, by (You et al., 2020), reported better results on downstream classification tasks after pre-training GNN on citations, cite-seer, and other datasets.

Therefore, we plan to experiment with self-supervised learning, specifically contrastive learning and other data augmentation techniques, on a GNN to see whether the performance of such models improves on text classification related tasks. We believe that by generating contrastive examples from existing sentences, we can improve the robustness of our model more effectively learn sentence representations of sentences with similar structure but different semantic meaning.

3 Literature Suvery

Many recent approaches to text classification use attention-based mechanisms in order to weigh specific words or segments of the sentence that are important to the meaning. Amongst these, BERT is one of the top most models related to such tasks. It was first introduced by (Devlin et al., 2018) as part of the Google Research. The paper proposed BERT(Bidirectional Encoder Representations from Transformers) as the new language representation model that pre-trains deep bidirectional representations using the left and right context together. For sentiment classification, (Wang et al., 2016) uses an attention-based LSTM to classify sentiment surrounding a target aspect. TNET by (Li et al., 2018) employs a similar approach, using an LSTM in addition to Target-Specific Transformations to integrate target aspect information into the word representations. (Zhang et al., 2019) used GCNs on dependency trees for text classification. Relational Graph Attention Networks were used by (Wang et al., 2020), where they combined RGAT and BERT representations for sentiment classification. (Wang et al., 2022) combined three types of negatives for sampling on knowledge graphs to improve training efficiency and that too without much significant computational and memory overhead.

There have also been several approaches to utilizing contrastive learning for NLP. (Wang et al., 2021) generates semantic positive and negative examples by replacing some tokens with their synonyms and others with their antonyms such that it increases the robustness of pre-trained models against adversarial attacks along with its increased sensitivity to small semantic changes. In regards to sentence embeddings, (Gao et al., 2021) proposes two methods - an unsupervised SimCSE as well as a supervised SimCSE. The unsupervised approach uses the same sentence run through an encoder twice with dropout, obtaining two embeddings, as the positive pair, and a different sentence as a negative pair. On the other hand, the supervised approach uses pretrained natural language inference (NLI) along with contrastive pairs to product sentence embeddings. For their contrastive framework, they use entailment pairs as a positive pair along with contradiction pairs as a negative pair.

4 Proposed Method

We implemented a self-supervised framework that generalizes syntax information of the sentence, so the model can be pre-trained using unlabelled data and can be finetuned with limited data. This pre-trained syntax model can be combined with pre-trained LLMs, like BERT, to provide good transferability with fine tuning on limited data. We implemented the framework in the context of having at least an aspect and opinionated words in sentences. We experimented using Twitter tweets, laptop reviews, and restaurant review datasets from SemEval14, as the dataset comes with aspects and corresponding polarity towards it.

Our framework took inspiration from the Contrastive Learning with Semantic Negative Examples for Natural Language Understanding (CLINE) framework proposed by (Wang et al., 2021) to have self-supervision on structural data. We use the given sentence to create a positive sample(sentence with similar semantics, but words are replaced with synonyms) and negative pairs(sentences with different semantics, but words are replaced with antonyms) using the WordNet dataset. Roughly 40% of tokens were replaced with their synonyms for the positive pair and 20% of tokens with their antonyms for the negative pair.

Then, our framework uses a biaffine parser to extract the dependency trees for three sentences (org, pos, neg). As the aspects are unknown during the pre-training, the noun phrases and proper nouns are treated as aspects, and the dependency trees are reshaped to root the aspects as shown in 1. Learnable embedding matrices are used as feature matrices for the dependency tree, with the node features being the word embeddings and edge relationships being the dependency information. A relational graph attention network is used to encode the dependency information into embeddings for all three sentences. These embeddings are contrasted using the CLINE framework to pre-train the RGAT model to generalize the syntax information.

Our hypothesis builds upon the idea that combining a language model that can effectively encode semantic information with a model that effectively encodes syntax information will provide a good sentence representation. Results by (Wang et al., 2020) and several others showed it does improve performance. As we already have pre-trained LLMs that can easily transfer the semantic knowledge after pre-training, there is no model to do the same for syntax knowledge. Our goal is to see if building such a model will actually improve the generalization of syntax information and improve the performance of downstream tasks.

5 Results

Just as how pre-training schemes helped boost the performance of LLMs, we wanted to validate that our pre-training strategy on syntax information works for graph-based models. Therefore, we fine-tuned our pre-trained model for classification tasks and tested it using F1 score, and accuracy as metrics. A successful result would validate our hypothesis and would show significant improvements in downstream classification tasks. We pre-trained our model on the Twitter dataset using the CLINE framework and fine-tuned it on laptop and restaurant datasets. We tested RGAT (no pre-training), RGAT (pre-trained on Twitter dataset), BERT, BERT+RGAT (no pre-training), BERT+RGAT(pre-trained on Twitter dataset) on two SemEval14 benchmark datasets, Restaurant (Rest), and Laptop. Our observations were noted down in the table 5, and 6

Model	**Rest**	**Laptop**
RGAT (no pre-train)	0.74	0.71
RGAT (pre-trained)	0.75	0.71
BERT	0.812	0.76
BERT+RGAT (no pre-train)	0.83	0.765
BERT+RGAT (pre-trained)	0.86	0.76

Table 5: Test accuracy of models with and without pre-training on SemEval14 datasets.

Model	**Rest**	**Laptop**
RGAT (no pre-train)	0.63	0.70
RGAT (pre-trained)	0.68	0.70
BERT	0.77	0.73
BERT+RGAT (no pre-train)	0.79	0.75
BERT+RGAT (pre-trained)	0.792	0.76

Table 6: Test F1 score of models with and without pre-training on SemEval14 datasets.

6 Error Analysis

With respect to the graph-based syntax model, there are three main failure cases observed: failed due to not recognizing aspects, failed due to lack of semantic knowledge base, and failed due to lack of capture of semantic information. As seen with Table 1, our model fails for certain cases when the test dataset includes unknown aspect tokens. This is because our model classifies the text based on the reshaped dependency tree with the aspect as the root. Our approach can't address this as the foundation of our approach is restructuring the sentence with aspect(proper nouns, noun phrase) as the root and determining the polarity wrt the aspect.

Another reason that we identified is the lack of knowledge base related to that context. As shown in Table 1 Row 2(Example 1), the given text includes information about the laptop having two USB ports on the same side. Though the text independently seems to be neutral, the ground truth label is negative as having a laptop that has only two USBs, and that, too, on the same side, isn't great compared to other laptops. Therefore, to determine the actual sentiment of this sentence, the model would have needed knowledge based on the context to see if this quality of a laptop is something great or not.

Lastly, the model can't predict the sentiment to be neutral when it may or may not include positive and negative words for the aspect, but the overall meaning is still neutral, as shown in Table 1 Row 3. This can be resolved by combining the graph-based model with language inference models to understand the overall meaning and context.

Overall, our model succeeded yet failed at certain data points. As shown in Table 2, we can see as to how for certain examples we perform better than other models, especially for sentences that are in the form of advice, and for certain examples either other models perform better than us or none predict right the right sentiment.

7 Conclusion

7.1 Challenges

The first challenge was to find a way to create augmentations of positive and negative sentences to contrast for pre-training. The initial solution was based on subgraph augmentation techniques like node dropping, edge removal, and feature masking in the dependency trees of the sentence. After the feedback from the Professor during the presentations, office hours, and a detailed literature review, we identified CLINE based solution to be a reasonable objective function that solves the problem. We also invested a considerable amount of time and faced challenges in the implementation of RGAT, BERT, RGAT + BERT, CLINE, and subgraph augmentations. Our current state of the project was achieved after several revisions of ideas and iterations of code within a limited time.

7.2 Ethical Considerations

When developing such kinds of sentiment analysis models, one must also consider any ethical impacts the work may cause. For example, WordNet, the lexical database used to generate contrastive pairs for learning, has been found to contain gender biases, associating many neutral words more strongly with one gender than another (Raya and Navarro). As a result, any leveraging model WordNet in its training, including ours, is imbued with the same bias found in WordNet. Gender bias can be socially harmful and expensive, for example, in biomedical research where one gender is overlooked. Some solutions to this issue may be searching for a database without pervasive bias or processing the WordNet data to mitigate gender bias. Raya and Navarro propose an example of the latter solution, which augments the embeddings of biased words to preserve their features without distinction between genders. Another more general ethical issue exists with the potential of abuse caused by sentiment analysis. Like many other NLP applications, powerful groups can misuse sentiment analysis to impact minority groups negatively. One example includes scraping social media data to manage insurance premiums, which may negatively target minorities impacted by bias. In addition, sentiment about a specific topic may be automatically found and censored, which can be dangerous to free speech.

7.3 Limitations

Our model has a few limitations. The first limitation is that the algorithm's performance depends on the parser. The second limitation is caused by the contrastive learning framework we use. As our model creates contrastive pairs by replacing words with their synonyms and antonyms, it can sometimes generate sentences that do not make much sense. An example of this can be seen in the contrastive pair generated from the sentence, "I charge it at night and skip taking the cord with me because of the good battery life." When creating the negative sample from this sentence, our contrastive framework outputs, "I pay cash it at day and skip give the cord with me because of the evil battery life." As we can see, our model has difficulty dealing with words with more than one interpretation and sometimes replaces them with synonyms/antonyms that do not fit the context.

7.4 Future Work

Future work on this project may focus on ways to improve the generation of positive and negative pairs by using a context-aware method of word replacement. This can be done with the help of Natural Language Inference (NLI) models which can answer questions to determine whether the hypothesis logically follows from the premise. Furthermore, it may be worth augmenting our contrastive learning framework with these models as they can help verify whether the positive and negative pairs' meaning has changed appropriately compared to the original sentence.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Claudia Rosas Raya and Ana Marcela Herrera Navarro. Mitigating gender bias in knowledge-based graphs using data augmentation: Wordnet case study.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Dong Wang, Ning Ding, Piji Li, and Hai-Tao Zheng. 2021. Cline: Contrastive learning with semantic negative examples for natural language understanding. *arXiv preprint arXiv:2107.00440*.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. 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 4568–4578, Hong Kong, China. Association for Computational Linguistics.

	Sentence	Our Model	BERT	G.T.
Cases where our model performs worse than BERT	• This laptop has only 2 usb ports, and they are both on the same side.	Neutral	Negative	Negative
	• The mac mini, wireless keyboard/mouse and a hdmi cable is all i need to get some real work done.	Positive	Neutral	Neutral
	• The only solution is to turn the brightness down, etc.	Negative	Neutral	Neutral
	• And as for all the fancy finger swipes – I just gave up and attached a mouse.	Positive	Neutral	Neutral
Cases where our model beats BERT	• Gave a mojito and sit in the back patio	Positive	Neutral	Positive
	• The portions of the food that came out were mediocre.	Neutral	Negative	Neutral
	• Be sure to accompany your food with one of their fresh juice concoctions.	Neutral	Positive	Neutral
	• from the erbazzone emiliana to the mostarda on the cheese plate, the dishes at this restaurant are all handled with delicate care.	Positive	Neutral	Positive
	• finally let into the store 5 at a time, to buy expensive slices from a harried staff.	Neutral	Negative	Neutral
Cases where both models fail	• On start up it asks endless questions just so itune can sell you more of their products.	Positive	Neutral	Negative
	• I bought it to my son who uses it for graphic design.	Positive	Positive	Neutral
	• I wanted a computer that was quite, fast, and that had overall great performance.	Positive	Positive	Neutral
	• price was higher when purchased on mac when compared to price showing on pc when i bought this product.	Negative	Neutral	Positive
	• apple is aware of this issue and this is either old stock or a defective design involving the intel 4000 graphics chipset.	Negative	Negative	Neutral

Table 2: Several examples of when our model performed worse than BERT, beat BERT, and where both models failed.

Sentence	**Prediction**	**G.T**
teen movies have really hit the skids .	Positive	Negative
good film , but very glum .	Negative	Positive
this flick is about as cool and crowd-pleasing as a documentary can get .	Negative	Positive
what better message than ‘ love thyself ’ could young women of any size receive ?	Negative	Positive
no telegraphing is too obvious or simplistic for this movie .	Positive	Negative
as unseemly as its title suggests . .	Negative	Positive
you won’t like roger , but you will quickly recognize him .	Positive	Negative
but it still jingles in the pocket ..	Negative	Positive
sam mendes has become valedictorian at the school for soft landings and easy ways out .	Positive	Negative
manages to show life in all of its banality when the intention is quite the opposite .	Positive	Negative

Table 3: Sentences and wrongly assigned classes by the DistilledBERT model. Here G.T expands as Ground Truth.

Sentences	**Version**
"But the staff was so horrible to us."	Original
"But the stave was needed ugly to us."	Synonym
"But the page was strut rubbish to us."	Antonym

Table 4: Example of Using the CLINE Method of Generative Positive and Negative Pairs by Using Synonyms and Antonyms