EFFECTIVE COMBINATION OF BERT MODEL AND CROSS-SENTENCE CONTEXTS IN ASPECT EXTRACTION

Anh Khoi Le and Truong Son Nguyen

Ho Chi Minh University of Science, National University,
Ho Chi Minh City, Vietnam

ABSTRACT

The Aspect Extraction (AE) field investigates in collecting words which are sentiment aspects in sentences and documents. Despite the pandemic, the number of products purchased online is still growing, which means that the number of product reviews and comments is also increasing rapidly, so the role of the task is gradually crucial. Extract aspects in the text is a difficult task, that requires algorithms capable of deep capturing the semantics of the text. In this work, we combine two models of the two research groups, with the first using the BERT algorithm with multiple concatenated layers and the second using the strategies to enrich the dataset by itself in the training or testing phase.
The source code is available on github.com, researchers can run it through scripts, modify it for further research also. https://github.com/leanhkhoi/AE_BERT_CROSS_SENTENCES

KEYWORDS

Sequence Labeling, Aspect Extraction, BERT, Cross-sentence.

1. INTRODUCTION

In the commercial industry, being able to capture the needs and thoughts of consumers for products is the core factor for businesses. Businesses can seize exactly what aspects of the market product users are interested in, thereby making appropriate improvements to capture customer demands. However, analyzing and extracting emotional aspects in text data such as in a sentence, a paragraph, is not lightweight works because of language complexity as well as the time-consuming dataset labeling process. So far, plenty of machine learning solutions have been proposed to carry out this task such as: POS tagger [1], Dependency parser [2], HMM [3], CRF [4], and have achieved certain results.

In recent years, deep learning algorithms have made a breakthrough and are extremely widely used as a solution to replace or supplement traditional techniques thanks to the development of hardware processing capabilities. One of the areas where deep learning has made a deep impact is NLP with the recent emergence of the Bidirectional Encoder Representations from Transformers (BERT) model [5]. Solutions based on BERT and its variants are now the state-of-the-art of many tasks in NLP. With a multi-layered architecture and trained with a huge data set, BERT is able to encode efficiently language features from syntax to semantics, providing a quality data representation layer for NLP tasks [6, 7] including Aspect Extraction.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.121-129, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111210

Besides, datasets used for Aspect Extraction task are often quite limited because of the time-consuming process of data labeling, so it is possible to use a combination of data in the same dataset come along with BERT mechanism can help to better grasp the patterns of the structure and semantics of text.

Our work is to inherit the architecture based on multiple hidden layers of BERT called parallel aggregation and hierarchical aggregation [8] in preceding research which archived quite a good performance with data enrichment techniques knows as cross-sentence [9] in another research applied for Named Entity Recognition task to create a complete pipeline so that perform Aspect Extraction task gives an even better evaluation. We also keep the configurations in our pipeline the same as those of the previous authors for experiments based comparison objective results.

2. RELATED WORKS

For unsupervised algorithms, the most popular method so far is to use a POS tagger to extract nouns or noun phrases in sentences. Stanford tagger [1] can completely do this well with an accuracy of up to 97 %, but in the problem of extracting aspects, the above accuracy does not bring much benefit because aspects in a sentence are not always a noun or a noun phrase, besides that POS tagger takes all nouns and noun phrases without any restriction according to the context of the sentence. Another approach uses relationship-based graphing, which explores relationships between emotional words and aspects based on parsing sentences into components and their dependencies. Algorithms based on this approach are referred to as Dependency parser [2], Double Propagation [10]. The weakness of these algorithms is to generate many non-aspect components. Frequency-based approaches are also considered as possible solutions. Kelledy [11], proposed a method of using POS tagger to extract all nouns and noun phrases in a sentence, then depend on frequency of occurrence of words and phrases to select the aspects. Endo [12], 2014 presented an improved method using TF-IDF in frequency calculation as well as using a syntactic pattern to remove non-aspect words and phrases.

For supervised algorithms, the two most popular traditional models for aspect extraction are HMM [3] and CRF [4]. Especially, CRF algorithm is dominating in sequential labeling problems like Aspect Extraction, the harmonious combination between BERT output and CRF is the preferred thinking in most of predicting the labels of words studies in the last 1 to 2 years.

Recently, deep learning networks have been applied to give better performance when extracting sentiment aspects, such as LSTM combined with attention mechanism [13], CNN [14]. Xu et al 2018 [15], proposed a model called DD-CNN with the idea of joining 2 embedding layers of text data, the first layer is in-domain embedding, and the second layer is out-domain embedding then feeds them into the CNN deep learning network to perform the classification task. Wang et al. 2020 [16], have introduced a mechanism to automatically concatenate pre-trained different embedding algorithms, whereby for each problem that needs embedding to form the embedding layer. For each combination, the algorithm will calculate error based on results of the training process and then compare it with other combinations to finally find out the most suitable concatenation embedding layer for the problem. This model has achieved state-of-the-art in many problems such as NER, aspect extraction (Aspect extraction), ... but has the disadvantage of expensive training time. With the arrival of BERT, a wide range of tasks in NLP such as text classification [17], summarization [18], question answering [19],… have been greatly improved in performance by using and refining this powerful BERT model. Akbar Karimi et al. 2020 [8], offers an architecture with a mix of pre-trained BERT architecture and CRF, considering Aspect Extraction problem as a sequence labeling problem which is widely applied to implement Named Entity Recognition task. The BERT architecture feature that authors use is that they are not based on only one BERT layer, usually the last layer, but they have used up to the last 4 layers of BERT

because those layers are capable of capturing more language aspects. Our research largely reuses the author's solution with some small improvements to perform Aspect Extraction task, because of the feasibility and good evaluation in experiment.

In addition to picking up the right model, we are also interested in preprocessing data to produce a quality data source. The technique used in our work is based on the ideas of Jouni Luoma et al. 2020 [9], where data will be enhanced by combining the components in itself to increase the size as well as create many interwoven semantic structures. Specifically, each data record, which is considered as a text sentence, will be joined with other data lines in the same training data set to fill the BERT window, then trained for the Aspect Extraction task according to the architecture described in the previous paragraph. Finally, the prediction of tags for each token for each sentence in the test dataset will be calculated based on the model.

3. ASPECT EXTRACTION

Given a dataset consisting of user reviews on a market product. Each line of data will contain information including aspects and respective emotions of the user. The AE's purpose is to find out exactly the aspects. For instance, *"this laptop has a good battery"*, then *"battery"* is an aspect in the sentence above. To accomplish this task, each word in the sentence will be labeled as a character in the set *{B, I, O}*, with *B* representing the starting word of the aspect, *I* representing the word belonging to one of the insides of an aspect and *O* represents the non-aspect word. The job of the algorithm is to predict each word in the sentence corresponding to each of the three characters above. This is called the sequence labeling task.

4. CROSS-SENTENCE IN CONTEXT

For each example in the dataset, it will be filled at both ends by other examples in the same dataset in a given size - BERT window size. The processed data has a fixed size (n, m) where n is dataset size and m is BERT window size. (see figure 1).

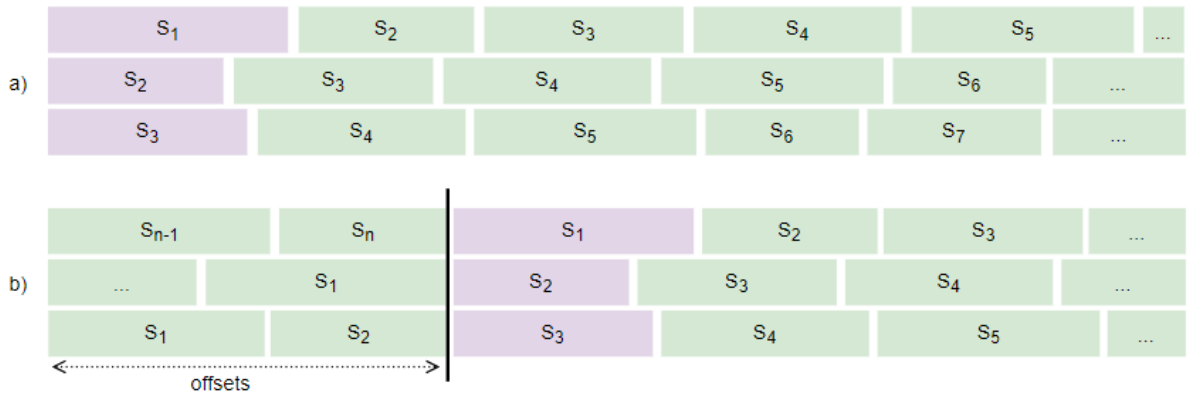

Figure 1. Illustrates two ways to create cross-sentence examples. a) The example of interest is placed at the first position in a BERT window. b) The example of interest will be placed at a position with a certain offsets value, usually 32, 64, 96,...

The data processing according to the above mechanism will create a new dataset with more semantic diversity, enabling the BERT architecture to understand more new patterns that generated from the combiner, thereby increasing the accuracy when performing sequence labeling task.

5. PROPOSAL METHOD

Our work is based entirely on the models of previous authors, bringing them together to form a complete pipeline. So we will keep the components as well as the configuration of the author's model.

Deep network models like BERT have been widely used in many problems because of their ability to understand the semantics of sentences deeply. Usually, most studies will use only last hidden layer of BERT because that is the layer that contains the most insight into the data, but authors realize that adjacent layers also store useful pieces of information [20], so they took advantage of the BERT last layers, combining them in two different ways to create the embedding layer. Besides, to be able to perform well for the sequence labeling problem, the CRF algorithm comes as an optimal solution and is widely used today. CRFs are a type of discriminative undirected probabilistic graphical model, where each data sample is predicted concerning the label of the previous data sample. CRFs are very well suited to sequence labeling tasks and Aspect Extraction can be considered as a labeled sequences problem.

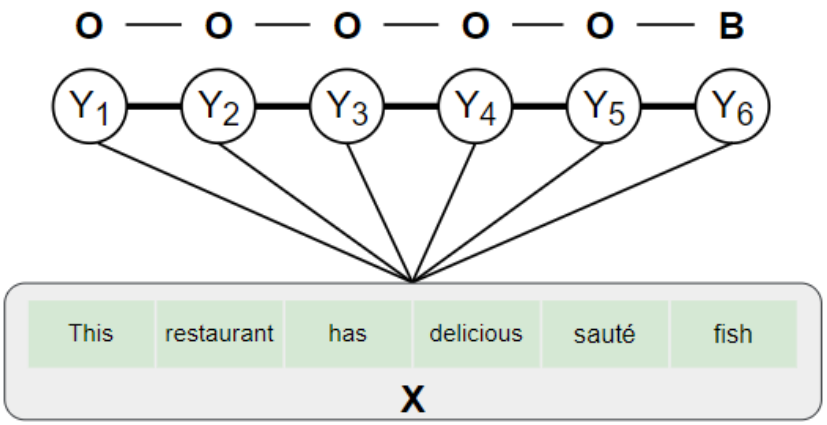

Figure 2. Illustrate the CRF algorithm for a specific example in Aspect Extraction task.

As shown above, prediction for output Y_6 with observation X_6 will be based on the set of observations in X and, importantly, the value of Y_5 that has been predicted before. Also intuitively, we can see Y_4, Y_5 are adjectives, Y_6 is a noun so Y_6 will get a high probability of being an aspect, that's why CRF is the perfect solution for Aspect Extraction.

Pipeline

Based on contributions of the AI research team about cross-sentence, our work will create a pipeline, which consists of 2 phases: training and testing.

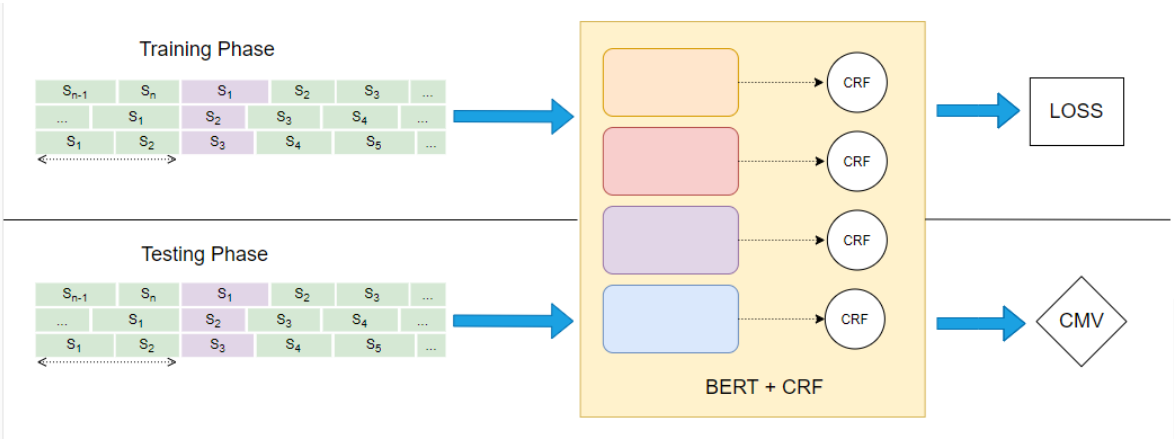

Figure 3. Illustrate the proposed pipeline, including 2 phases of training and testing with the difference in output section.

For each training, we place the sentence of interest at a certain position in the BERT window with a fixed size of 100, then add other sentences in the same data set sequentially at the left and right. Note that the labels of the training data will also be processed according to the above mechanism. (See Figure 3).

At prediction stage, sentences will also be processed in cross-sentence mechanism, except their labels remain the same. The difference is that prediction results will then be passed through a predictive model called the CMV by the research team. Accordingly, prediction results for sentence of interest will be calculated in two ways, the first way is based on the frequency of the label which predicted on each word in sentence of interest and itself but appearing in other predicted sentences that it is not the sentence of interest. And the second way is to calculate the sum of the probability of predicting each word in the sentence of interest and itself but appearing in other predictive sentences and select maximize value representing for the label. (See Figure 3)

6. EXPERIMENTS AND RESULTS

Table 1. SemEval Dataset

	Train		Test	
Dataset	Sentences	Aspects	Sentences	Aspects
Laptop 2014 [21]	3045	2358	800	654
Restaurant 2016 [22]	2000	1743	676	622

To execute the task, we relied on codebase [8] and modified it to accommodate library updates when building the multiple BERT aggregation model. Besides, we also use data process functions of the author group [9], build a cross-sentences dataset with some changes in the naming of variables and functions. For all the remaining model parameters when training, our team keeps the same to ensure the evaluation results show objectively. The model is trained with the last 4 layers of BERT, using Adam optimizer, learning-rate 3e-5, and batch size 16. One slight difference is that we train on GPU (GeForce) GTX 1660S with 6 GB of memory instead of a more powerful GPU because of resource constraints.

Dataset: SemEval is set of a quite famous datasets in Aspect Extraction as well as Aspect -based Sentiment Analysis problem. To perform the Aspect Extraction operation, two datasets are used, SemEval 2014 Laptop and SemEval 2016 Restaurant. Each dataset includes reviews, comments on the subject as well as labeled aspects for each review. (See Table 1)

Analysis: When performing the experiment, we train the model with 10 epochs and observe that the error value varies through each epoch. As shown in the figure below, in both laptop and restaurant sets, at the first epochs, the loss values in both training and validation set are very high, starting from the 4th epoch the model gradually enters convergence point with loss value in training set decreases and at 6th epoch, loss value in validation set is also stable. (See Figure 4.a, 4.b)

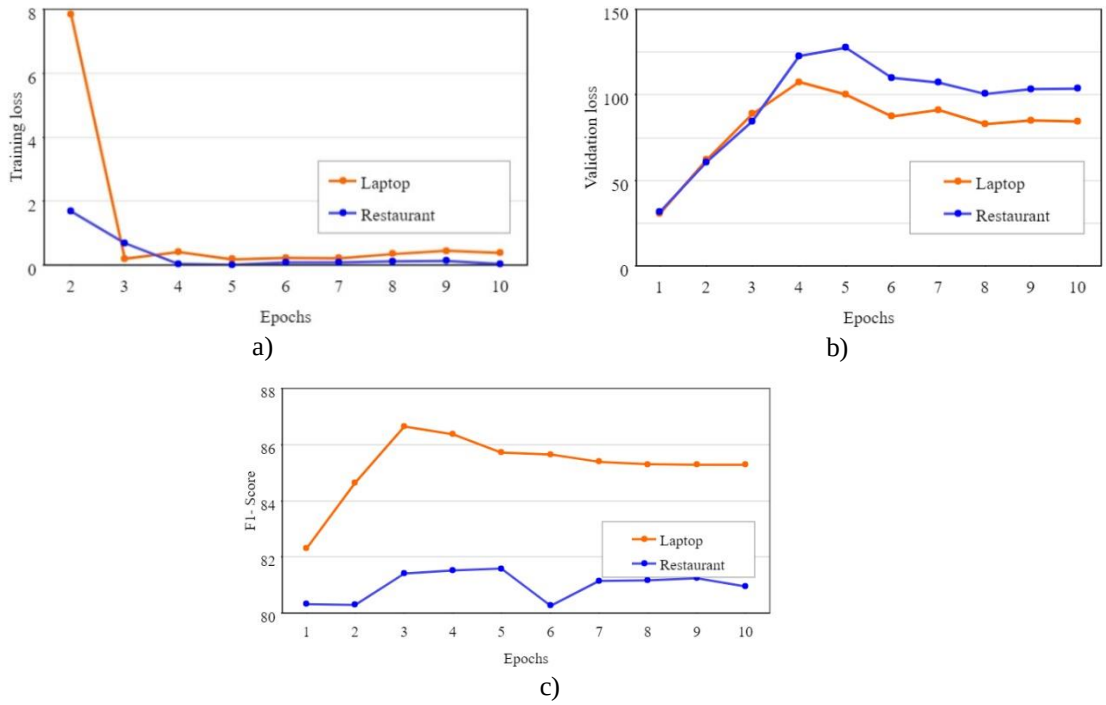

Figure 4. The performance of the pipeline when executing the Aspect Extraction task is measured by the change in the number of epochs during the training period. a) Change on training loss. b) Change on validation loss. c) Change on precision F1

In addition to observing loss value, we also observe the accuracy of prediction on test set when training with different epoch. It can be seen that the accuracy in both the laptop and restaurant will reach the best value when the model is trained from 4 to 6 epochs and will deteriorate if epoch is too high, meaning that it has passed the convergence point. (See Figure 4.c).

Table 2. The table compares evaluation results according to the accuracy F1. The number in bold represents the highest value. Scores not in bold are quoted from [8]. Scores in bold are the superior results of our recommended pipeline. Each result in the table is average of 9 runs.

	Laptop - F1	Restaurant - F1
BERT	79.28	74.10
BERT-PT (30 epochs) [23]	85.93	82.64
P-SUM (4 epochs) [8]	85.94	81.99
Our P-SUM + CMV (4 epochs)	**86.17**	**82.66**
Our P-SUM + CMVP (4 epochs)	**86.28**	**82.71**
Our P-SUM + offset 0 (4 epochs)	**86.17**	**82.21**
Our P-SUM + offset 32 (4 epochs)	**86.29**	**82.80**
Our P-SUM + offset 64 (4 epochs)	**85.96**	**82.67**
H-SUM (4 epochs) [8]	86.09	82.34
Our H-SUM + CMV (4 epochs)	**86.41**	**82.61**
Our H-SUM + CMVP (4 epochs)	**86.43**	**82.83**
Our H-SUM + offset 0 (4 epochs)	**86.24**	82.12
Our H-SUM + offset 32 (4 epochs)	**86.19**	**83.06**
Our H-SUM + offset 64 (4 epochs)	**86.29**	**82.96**

Result: We have tested a lot of measurements to compare with the original model [8]. As be exposed from table 2, with the same number of epochs of 4, most of the results from our pipeline are superior to previously available models. Here, training dataset was processed by cross-sentence mechanism with an offset of 0, while the test dataset is processed and predicted according to CMV (Contextual Majority Voting) and CMVP (Contextual) models. Majority Voting Probability) [9] described in the previous section. Offsets 0, 32, 64 in table 2 are positions of the BERT window where a sentence of interest begins when processing test data. In particular, when performing with P-SUM + offset 32 architecture, accuracy of F1 on our restaurant dataset (82.80) is 0.81 better than P-SUM model (81.99).

7. CONCLUSION

We demonstrated a pipeline with a combination of BERT customization using hidden multilayer integration with a solution that augments the data during training and prediction. We also present a very diverse set of evaluation results based on how the samples in the data are combined. Our idea is very simple, but the obtained results have surpassed the results of previous studies based on the advanced BERT model when executing Aspect Extraction task.

REFERENCES

[1] https://nlp.stanford.edu/software/ tagger.shtml

[2] Wu, Yuanbin, Qi Zhang, Xuan-Jing Huang, and Lide Wu. "Phrase dependency parsing for opinion mining." In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 1533-1541. 2009.

[3] Ghahramani, Zoubin. "An introduction to hidden Markov models and Bayesian networks." In *Hidden Markov models: applications in computer vision*, pp. 9-41. 2001.

[4] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

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

[6] Kantor, Yoav, Yoav Katz, Leshem Choshen, Edo Cohen-Karlik, Naftali Liberman, Assaf Toledo, Amir Menczel, and Noam Slonim. "Learning to combine grammatical error corrections." *arXiv preprint arXiv:1906.03897* (2019).

[7] Kanerva, Jenna, Filip Ginter, and Sampo Pyysalo. "Dependency parsing of biomedical text with BERT." *BMC bioinformatics* 21, no. 23 (2020): 1-12.

[8] Karimi, Akbar, Leonardo Rossi, and Andrea Prati. "Improving BERT Performance for Aspect-Based Sentiment Analysis." *arXiv preprint arXiv:2010.11731* (2020).

[9] Luoma, Jouni, and Sampo Pyysalo. "Exploring cross-sentence contexts for named entity recognition with BERT." *arXiv preprint arXiv:2006.01563* (2020).

[10] Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. "Opinion word expansion and target extraction through double propagation." *Computational linguistics* 37, no. 1 (2011): 9-27.

[11] Smeaton, Alan F., Fergus Kelledy, and Ruairi O'Donnell. "TREC-4 experiments at Dublin City University: Thresholding posting lists, query expansion with WordNet and POS tagging of Spanish." *Harman [6]* (1995): 373-389.

[12] Shimada, Kazutaka, Ryosuke Tadano, and Tsutomu Endo. "Multi-aspects review summarization with objective information." *Procedia-Social and Behavioral Sciences* 27 (2011): 140-149.

[13] Wang, Wenya, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. "Coupled multi-layer attentions for co-extraction of aspect and opinion terms." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1. 2017.

[14] Huang, Binxuan, and Kathleen M. Carley. "Parameterized convolutional neural networks for aspect level sentiment classification." *arXiv preprint arXiv:1909.06276* (2019).

[15] Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. "Double embeddings and cnn-based sequence labeling for aspect extraction." *arXiv preprint arXiv:1805.04601* (2018).

[16] Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. "Automated Concatenation of Embeddings for Structured Prediction." *arXiv preprint arXiv:2010.05006* (2020).

[17] Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. "How to fine-tune bert for text classification?." In *China National Conference on Chinese Computational Linguistics*, pp. 194-206. Springer, Cham, 2019.

[18] Liu, Yang. "Fine-tune BERT for extractive summarization." *arXiv preprint arXiv:1903.10318* (2019).

[19] Yang, Wei, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. "Data augmentation for bert fine-tuning in open-domain question answering." *arXiv preprint arXiv:1904.06652* (2019).

[20] Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. "What does BERT learn about the structure of language?." In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. 2019.

[21] Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 27–35, 01 2014.

[22] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, et al., "Semeval-2016 task 5: Aspect based sentiment analysis*," in Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.

[23] Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. "BERT post-training for review reading comprehension and aspect-based sentiment analysis." *arXiv preprint arXiv:1904.02232* (2019).

AUTHORS

Anh Khôi Lê - currently studying for a master's degree in information systems at the University of Natural Sciences in Ho Chi Minh City. My current research direction is sentiment analysis and related problems. Social behavior analysis is one of my research fields in near future as well.