On Vietnamese Sentiment Analysis: A Transfer Learning Method

Ngoc C. Lê
School of Applied Mathematics and Informatics
Hanoi University of Science and Technology
Institution of Mathematics
Vietnam Academy of Science and Technology
Hanoi, Vietnam
lechingoc@yahoo.com

Nguyen The Lam
School of Applied Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
lamnt.bka@gmail.com

Son Hong Nguyen
School of Applied Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
nguyenhongson.kstn.hust@gmail.com

Duc Thanh Nguyen
School of Applied Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
nguyenthanhduc19975@gmail.com

***Abstract*—Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing (NLP) and text analysis to systematically identify, extract and quantify states and subjective information. Transfer learning (TL) [1] is a study in machine learning focusing on storing knowledge gained while solving one problem and applying it to a different but related problem. In NLP, recent results also demonstrated the effectiveness of models using pre-training on a language modeling task [2], [3]. The transfer learning based models help to rapidly increase understanding of words and sentences arrangement in which semantics and connections are easily grasped. In this paper, we present the results from applying BERT [16], a transfer learning method, in Vietnamese benchmark [17] for one of text classification problems, the Aspect Based Sentiment Analysis problem. The experiments were conducted on two data sets, named Hotel and Restaurant [17], in two task (A) Aspect Detection and (B) Aspect Polarity. The obtained results have outperformed some previous systems [18]–[20] in precision, recall, as well as the F1 measures.**

***Index Terms*—Transfer Learning, Natural Language Processing, Aspect Based Sentiment Analysis, Transformers, Bidirectional Encoder Representations from Transformers, BERT, VLSP**

I. Introduction

Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media for applications that range from marketing to customer service. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. *Aspect based sentiment analysis* refers to determining the opinions or sentiments expressed on different features or aspects of entities, e.g., of a cell phone, a digital camera, or a bank [4]. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, the service for a restaurant, or the picture quality of a camera. The advantage of feature-based sentiment analysis is the possibility to capture nuances about objects of interest. Different features can generate different sentiment responses, for example a hotel can have a convenient location, but mediocre food [5].

The recent growth of machine learning technologies is dramatic, especially natural language processing (NLP) with many different techniques helping to model the complexity information of human language. Even though the lack of labeled data already make the difficulty of training NLP models. One of the most effective solution that can tackle with this problem is transfer learning that is using pre-trained model and fine tuning it in downstream tasks [2], [3]. Transfer learning has demonstrated its efficiencies in many computer vision tasks, by transferring via pre-training on ImageNet [6] or digit recognition [7]. It has also been applied to cancer subtype discovery [8], building utilization [9], [10], general game playing [11].

Transfer learning has been also applied in text classification [12], [13]. However, application of transfer learning in Vietnamese has not been much effective yet, while the performance in NLP tasks have been indicated [2], [14], [15].

Among, transfer learning methods for NLP, BERT (Bidirectional Encoder Representations from Transformers) [16] introduced recently with many promising result in various NLP task of English language. In this research, we aim to verify the performance of BERT in Vietnamese, a low resource language. We focus on sentiment detection problem in VLSP dataset [17] by employing BERT as multi label text classification problem. Our contributions are two-fold:

- We propose a practical model of multi label text classification which employs BERT as an element of transfer

learning. By fine tuning BERT with VLSP dataset, our model achieved significant improvements compared to previous results. To the best of our knowledge, we are the first study of testing performance of BERT in Vietnamese.
- We present a step-by-step processing flow with many techniques that helping BERT to improve its performance. The statistical analysis shows that our model achieves improvements in term of F-score in comparison with strong baselines.

The next section is a very short brief on related work. Section III describes our method. The model is evaluated in Section IV. Some discussion is given in Section V.

II. Related Work

Multi label text classification is an important task classifying text into different categories where a text can be assigned to more than one category. This is more natural but also difficult than normal text single category classification. Among text classification problems, Aspect Based Sentiment Analysis (ABSA) [4] is gained attention by many researchers. It allows us to further analyze the emotions in each user comment by attaching sentiment to specific aspects. This problem involves several sub-problems, e.g., identifying relevant entities, extracting their features/aspects, and determining whether an expressed opinion on each feature/aspect is positive, negative, or neutral. The automatic identification of features can be performed with syntactic methods, with topic modeling [21], [22], or with deep learning [23]. More detailed discussions about this level of sentiment analysis can be found in Liu's work [24].

Vietnamese data for ABSA was proposed in 2018 [19]. Then many researches use Deep Neural Network as the core of models to learn the dependency of words [18], [20], [25]. The method has been shown efficiency for NLP in general, but focused to a specific task and required tons of training data. Based on many different pre-trained word embeddings, deep learning models using the hypotheses that word embedding models can capture the relations between words in the language, which is not always efficient.

Language model is one of the key technologies in NLP used for modeling the characteristic of entire language. Many language modeling methods such as ELMO [2], ULMiT [14], Flair [15],... have already broken through the limitation of labeled dataset and achieved state-of-the-art results. Among them, BERT is introduced as a critical model for NLP based on transformer architecture. It is trained using two different tasks helping to learn the dependency between words in a sentence and between two sentences in a paragraphs. Therefore, the characteristic of the language in which it was trained can be obtained. BERT also provided a pre-trained model in 104 languages [16].

III. Method

A. Datasets

The dataset used in this project is from the VLSP 2018 [17] challenge. It consists of two domains: Hotel and Restaurant. Each domain consists of three parts: train, development, and test sets. Each sample in the dataset is in the form of document-level reviews. The length of these samples are different. This is a matter of concern when handling this dataset. Another matter is imbalanced data. These problems will be addressed in Subsection III-C. As our consideration, somes aspects appear very few in the data. Table I below gives some statistics of the dataset.

Table I
THE STATISTICS OF VLSP DATASET.

Domain	Dataset	#Reviews	#Aspect
Restaurant	Train	2961	9297
	Development	1290	3443
	Test	500	2419
Hotel	Train	3000	13949
	Development	2000	7111
	Test	600	2584

B. Preprocessing

Preprocessing is an essential step in every machine learning and also for our task. In particular, for on-line environment, the document is not often in formal writing. Especially for teenagers, they prefer using a great number of emojis, shortened forms of words, special symbols and characters, misspelling words, grammar mistakes or conjoined words. Before being included in models, the data should follow preprocessing steps:

- **Step 1:** All characters indicating money value such as "100k" or "100,000đ" should be replaced by words "giá tiền" (money value). Also, characters indicating percentage value such as "50%" should be replaced by word "phần trăm" (percentage).
- **Step 2:** All special characters as "=", "?", "#", "+", "-", "$", "@", "&" and so on, and emojis should be removed.
- **Step 3:** Shortened forms of word "không" (not) as "k", "ko", "khong" or "kg" should be corrected to the original words.
- **Step 4:** This is the most important step. The need-to-solved problem is a multi-label classification, so a great number of potentially predicted labels for a review can lead to a confusion for the model. Therefore, we separated long reviews into many sentences, and then relabeled them. These relabels are definitely included in review labels.

C. Model Architecture

a) BERT: The traditional way to obtained word vectors is training the word embedding layer with a large-scale corpora. Then a neural architecture is built next to the embedding layer and trained on a supervised data for the downstream task. There is a problem arising, the shortage of high quality labeled data for model training and data for end tasks containing contextual representations. To solve the problem of data limitations, NLP models can use a training data pre-processing mechanism by transferring from a trained common model from

a large amount of unlabeled data. The subsequent different tasks such as Question Answering and Sentiment Analysis accompanied with small data sets will be refined based on pre-train models. This will result in a significant improvement in accuracy compared to the previously trained models. In this situation, instead of using human knowledge for model designing, we get the knowledge from data. Here, we find the word representation in a context, the process also plays an important role in NLP. A word can be represented variously in real life. For example, The word "car" in the two following sentences - "The car is in the garage." and "The car is on the road.", is different in context. So it must be represented with two different vectors, depending on which sentence it is in. While techniques like Word2vec [26], FastText [27] or Glove [28] only find representations of words through their common context i.e. only one vector for each word, creating a representation of each word based on the other words in the sentence (many vectors representing a word depending on the sentence) will bring more accurate outcomes.

Recently, with the appearance of BERT, a transfer learning model based on contextualized representation has been demonstrated to be superior to traditional word embedding methods by addressing these issues [2], [14], [16]. BERT is released in two sizes:

- **BERTBASE:** 12 layers, 768 hidden dimensions and 12 attention heads (in transformer) with the total number of parameters, 110M;
- **BERTLARGE:** 24 layers, 1024 hidden dimensions and 16 attention heads (in transformer) with the total number of parameters, 340M.

BERT is fine-tuned by using an additional task specific layer, which is a sigmoid layer that converts the problem from multi-label classification to multi-class classification. The two tasks Aspect Detection (AD) and Aspect Polarity (AP) in this research are fine-tuned according to the same architectures as in Fig. 1. The content of the model is explained in detail in the following sections.

b) Aspect Detection: In the VLSP Sentiment 2018 Shared Task [19], Aspect Detection task is Phase A of the ABSA problem. In supervised setting, it is described as a multi-label classification problem because each review can be one or several of the 34 Entity#Attribute (E#A) pairs (the Hotel data set), or 12 pairs (the Restaurant data set). There are two types of methods that can be applied to multi-label classification problems [29], named algorithm adaptation and problem transformation methods. The first one is used in this research.

If an E#A pair appears in the review label, the value of the class corresponding to that E#A pair will be 1. When applied to our task, we use BERT in the single sentence classification tasks. The following explanation is inspired by [30].

The input sentence is a sequence of n words: $x = ([CLS], x_1, x_2, ..., x_n, [SEP])$, where $[CLS]$ token stands for "Classification" and will become apparent later on and $[SEP]$ stands for "Separate" is a dummy token not used for the task. Set $BERT(.)$ as a pre-trained BERT model simulation

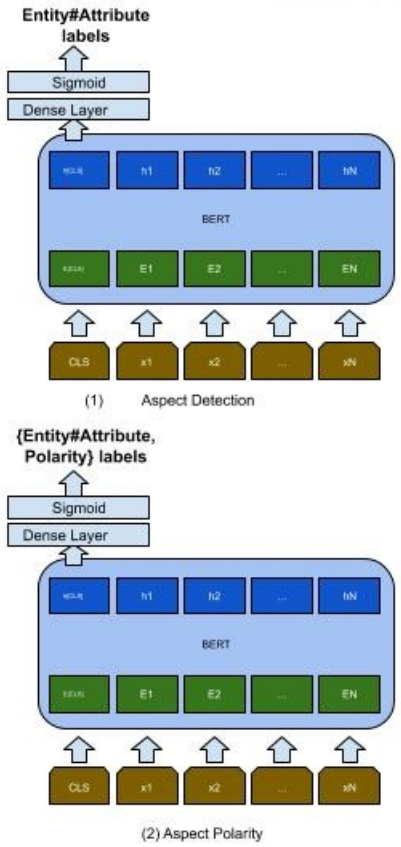

Figure 1. Overview of BERT settings for two tasks.

function. Sentence x after being tokenized and passed to the $BERT()$ function becomes hidden representation $h = BERT(x)$, h has the size $r_h * |x|$ where r_h is the size of the hidden dimension and $|x|$ is the length of the input sentence. We use $h_{[CLS]}$ - the representation of $[CLS]$, which is E#A pairs representation of the whole input sentence. $h_{[CLS]}$ is subsequently passed through a dense layer followed by a sigmoid function $l = sigmoid(W * h_{[CLS]} + b)$, where W belongs to R^{m*r_h}, where m is the number of E#A pairs and b belongs to R^m. Consequently, we have l belongs to $[0, 1]^3$. Training loss is cross entropy on the m E#A pairs. Because the values in vector l (the probability the sentence belongs to a certain class) are independent of each other, we set a threshold θ to decide which classes the review belongs to. Our test shows that the reasonable threshold changes from task to task.

$$f(x) = \begin{cases} 0, x < \theta \\ 1, x \geq \theta \end{cases} \quad (1)$$

Specifically in our experiment, the threshold was set as Table II.

We use the installation based on BERT's Tensorflow 1.12.0. The used BERT version is BERT-base (Multilingual Cased). The Input sentence is padded to 128 tokens. The learning rate is set to 2.5×10^{-5} and the number of training epoch is 50.

As explained in the previous section, we use a preprocessing step, which is to split the review into sentences and then re-

Table II
THE THRESHOLD

Domain	Phase A	Phase B
Restaurant	0.65	0.55
Hotel	0.25	0.45

annotate it. When evaluating models with test sets and dev files, to match the training set, we also split the reviews into sentences, predict the labels for each sentence, and then sum them up for the review.

c) Aspect Polarity: We used a similar model and setting for Phase B - Aspect Polarity. Each input sentence includes n words included in the model will output a set of labels in the form {Entity#Attribute,Polarity} of 102 label sets of Dataset Hotel and 36 label sets of Dataset Restaurant. Fig. 1 - Part (2) show the configuration model for this task. We used exactly the same parameters as the Aspect Detection task. This configuration produces a good result for the Restaurant test dataset but not so good for the Hotel test dataset. The reason is because the number of classes of the Hotel dataset is too large and the phenomenon of imbalanced data occurs. In detail, of the 102 label sets of the Hotel dataset, 54% of the label sets appear in 2.3% of the 3000 reviews in the train set. Therefore, in order to be effective in this task, we removed 54% of the label sets of the Hotel dataset, and only kept 47 label sets during the classifier training. When evaluated in development and test, all removed labels will be presumed not to appear, which greatly improve the accuracy of the model without affecting too much the effectiveness of the problem.

IV. EVALUATION

In this section, we will compare our results with those of the best current models in the VLSP datasets. To evaluate the performance of the system, we used the micro-averaged method, which is specified by the VLSP consortium. Tables III, IV, V, and VI show the results of the comparison between teams participating in the VLSP challenge 2018, Dang [18], and our results. From this comparison, we can see that our model got better results than the other models in all datasets and both Phases A and B. Especially, the results of Phase A processing on the Hotel data set are better than the best previous 10% F1 score.

This can be explained by two reasons. First, the transfer learning method has shown the advantage of contextualized embedding compared to traditional methods. BERT itself is also a large language model with many parameters and a fine-tuned model from BERT has proven to be effective. The second reason is because of the preprocessing step. We found out that the previous models had encountered a problem with the mismatch between train data and test data. Specifically based on observation of Dang [20]: in the Restaurant domain, average length of reviews in test dataset is 158, while in the development dataset, it is 58. This leads to errors when training models with train dataset and evaluate with test dataset. We solved this problem by breaking the review into sentences and then training and making predictions for each sentence and finally combining them to predict the original review.

Table III
THE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER TEAM'S WITH PHASE A OF DEVELOPMENT DATASET.

Domain	Team	Precison	Recall	F1
Restaurant	The 3rd team[1]			
	The 2nd team [25]	0.78	0.65	0.71
	The 1st team [20]	0.75	0.85	0.79
	Dang [18]			
	Our method	**0.7684**	**0.8806**	**0.8207**
Hotel	The 3rd team			
	The 2nd team [25]	0.83	0.51	0.63
	The 1st team [20]	0.75	0.64	0.69
	Dang [18]			
	Our method	**0.7940**	**0.7874**	**0.7907**

Table IV
THE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER TEAM'S WITH PHASE B OF DEV DATASET.

Domain	Team	Precison	Recall	F1
Restaurant	The 3rd team			0.59
	The 2nd team [25]	0.71	0.59	0.64
	The 1st team [20]	0.63	0.71	0.67
	Dang [18]			
	Our method	**0.6616**	**0.7145**	**0.6871**
Hotel	The 3rd team			0.56
	The 2nd team [25]	0.78	0.48	0.6
	The 1st team	0.67	0.58	0.62
	Dang [18]			
	Our method	**0.7943**	**0.5901**	**0.6772**

Table V
THE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER TEAM'S ON PHASE A OF TEST DATASETS

Domain	Team	Precison	Recall	F1
Restaurant	The 3rd team	0.88	0.38	0.54
	The 2nd team [25]	0.62	0.62	0.62
	The 1st team [20]	0.79	0.76	0.77
	Dang [18]	0.8475	0.7648	0.8040
	Our method	**0.7916**	**0.8367**	**0.8135**
Hotel	The 3rd team	0.85	0.42	0.56
	The 2nd team [25]	0.83	0.58	0.68
	The 1st team [20]	0.76	0.66	0.7
	Dang [18]	0.8235	0.5975	0.6925
	Our method	**0.7960**	**0.7972**	**0.7966**

V. CONCLUSION

We presented a transfer learning method that achieved high performance in Vietnamese dataset. In experiment, we treated the sentiment detection as a multi label text classification problem and confirmed the result of our solution based on BERT is significantly improved in comparison with strong baselines. We also proposed a step-by-step processing flow that helped to improve accuracy of BERT. We believe that those processing steps contributed to the final results by making clearer the context behind sentences in the dataset.

[1]The results were obtained from the paper of VLSP 2018 Shared Task

Table VI
THE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER TEAM'S WITH PHASE B OF TEST DATASET.

Domain	Team	Precison	Recall	F1
Restaurant	The 3rd team	0.79	0.35	0.48
	The 2nd team [25]	0.52	0.52	0.52
	The 1st team [20]	0.62	0.6	0.61
	Dang [18]			
	Our method	**0.6327**	**0.6503**	**0.6414**
Hotel	The 3rd team	0.8	0.39	0.53
	The 2nd team [25]	0.71	0.49	0.58
	The 1st team [20]	0.66	0.57	0.61
	Dang [18]			
	Our method	**0.7983**	**0.5882**	**0.6774**

ACKNOWLEDGMENT

We would like to thank VLSP consortium for your great efforts to create this dataset and allow us to use it in this project. The first author also has receive the support from Institute of Mathematics, Vietnam Academy of Science and Technology, Year 2019. The research is also supported by National Foundation for Science and Technology Development (NAFOSTED) of Vietnam, project code: 101.99-2020.16. We also want to express the appreciation to the anonymous reviewers and the editors for their very useful advices, comments, and corrections.

REFERENCES

[1] L. Y. Pratt, "Discriminability-based transfer between neural networks", *NIPS Conference: Advances in Neural Information Processing Systems 5*, Morgan Kaufmann Publishers, pp. 204—211, 1993.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations", arXiv:1802.05365,2018.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", Open AI Technical report, 2018.

[4] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews", *Proceedings of Nineteeth National Conference on Artificial Intellgience(AAAI-2004)*, pp. 755–760, 2004.

[5] M. Cataldi, A. Ballatore, I. Tiddi, and M.-A. Aufaure, "Good location, terrible food: detecting feature sentiment in user-generated reviews", *Social Network Analysis and Mining*, Vol. 3 (4), pp. 1149-–1163, (2013).

[6] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe and L. van der Maaten, "Exploring the limits of weakly supervised pretraining",

[7] D. S. Maitra, U. Bhattacharya, and S. K. Parui, "CNN based common approach to handwritten character recognition of multiple scripts", *in 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1021-–1025, 2015.

[8] E. Hajiramezanali, S. E. Dadaneh, A. Karbalayghareh, Z. Zhou, and X. Qian, "Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data", *in 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018.

[9] I. B. Arief-Ang, F. D. Salim, and M. Hamilton, "DA-HOC: semi-supervised domain adaptation for room occupancy prediction using CO2 sensor data", *4th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys)* Delft, Netherlands. pp. 1-–10, 2017.

[10] I. B. Arief-Ang, M. Hamilton, and F. D. Salim, "A Scalable Room Occupancy Prediction with Transferable Time Series Decomposition of CO2 Sensor Data", *ACM Transactions on Sensor Networks*, Vol. 14:3-–4, pp. 21:1—21:28, 2018.

[11] B. Banerjee and P. Stone, "General Game Learning Using Knowledge Transfer", *IJCAI' 2007*, 2007.

[12] C. B. Do and A. Y. Ng, "Transfer learning for text classification", *in Neural Information Processing Systems Foundation, NIPS*2005*, 2005.

[13] R. Raina, A. Y. Ng, and D. Koller, "Constructing Informative Priors using Transfer Learning", *in Twenty-third International Conference on Machine Learning, ICML*, pp. 713–720, 2006.

[14] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 328-–339, 2018.

[15] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling", *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *NAACL*, 2019.

[17] http://vlsp.org.vn/resources-vlsp2018

[18] D. V. Thin, V. D. Nguyen, K. V. Nguyen, and N. L.-T., Nguyen, "Deep Learning for Aspect Detection on Vietnamese Reviews", *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 2018.

[19] H. T. M. Nguyen, H. V. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, B. X. Ngo, and C. A. Le, "VLSP SHARED TASK: SENTIMENT ANALYS

[20] T. D. Van, K. V. Nguyen, and N. L.-T. Nguyen, "NLP@UIT at VLSP 2018: A Supervised Method for Aspect Based Sentiment Analysis", *Proceedings of the Fifth International workshop on Vietnamese Language and Speech Processing (VLSP)*, 2018.

[21] Z. Zhongwu, B. Liu, H. Xu, and P. Jia, "Constrained LDA for Grouping Product Features in Opinion Mining", *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 448-–459, 2011.

[22] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-grain Topic Models", *Proceedings of the 17th International Conference on World Wide Web WWW '08*, New York, NY, USA: ACM, pp. 111—120, 2008.

[23] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural networking", *Knowledge-Based Systems*, Vol. 108, pp. 42-–49, 2016.

[24] B. Liu, "Sentiment Analysis and Subjectivity", *Handbook of Natural Language Processing*, CRC Press, pp. 627–666, 2010.

[25] T. A. Nguyen and P. Q. N. Minh, "Using Multilayer Perceptron for Aspect-based Sentiment Analysis at VLSP-2018 SA Task", *Proceedings of the Fifth International workshop on Vietnamese Language and Speech Processing (VLSP)*, 2018.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *oceedings of the International Conference on Learning Representations (ICLR 2013)*, 2013.

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information", *Transactions of the Association for Computational Linguistics*, Vol 5., pp. 135-–146, 2017.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation", *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532-–1543, 2014.

[29] G. Tsoumakas and I. Katakis . "Multi-Label Classification: An Overview", *International Journal of Data Warehousing and Mining*, Vol. 3, pp. 1-–13, 2007.

[30] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis", arXiv:1904.02232 (2019).

[31] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web", *Proceedings of WWW*, 2005. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 2227–2237, 2018.